

## Proceedings e report

114



SIS 2017  
Statistics and Data Science:  
new challenges, new generations

28–30 June 2017  
Florence (Italy)

Proceedings of the Conference  
of the Italian Statistical Society

edited by  
Alessandra Petrucci  
Rosanna Verde

FIRENZE UNIVERSITY PRESS  
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.  
(Proceedings e report ; 114)

<http://digital.casalini.it/9788864535210>

ISBN 978-88-6453-521-0 (online)

#### *Peer Review Process*

All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP ([www.fupress.com](http://www.fupress.com)).

#### *Firenze University Press Editorial Board*

A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

This work is licensed under a Creative Commons Attribution 4.0 International License  
(CC BY 4.0: <https://creativecommons.org/licenses/by/4.0/legalcode>)

CC 2017 Firenze University Press  
Università degli Studi di Firenze  
Firenze University Press  
via Cittadella, 7, 50144 Firenze, Italy  
[www.fupress.com](http://www.fupress.com)



## **SOCIETÀ ITALIANA DI STATISTICA**

Sede: Salita de' Crescenzi 26 - 00186 Roma  
Tel +39-06-6869845 - Fax +39-06-68806742  
email: sis@caspur.it web:<http://www.sis-statistica.it>

La Società Italiana di Statistica (SIS), fondata nel 1939, è una società scientifica eretta ad Ente morale ed inclusa tra gli Enti di particolare rilevanza scientifica. La SIS promuove lo sviluppo delle scienze statistiche e la loro applicazione in campo economico, sociale, sanitario, demografico, produttivo ed in molti altri settori di ricerca.

### **Organi della società:**

#### *Presidente:*

- Prof.ssa Monica Pratesi, Università di Pisa

#### *Segretario Generale:*

- Prof.ssa Filomena Racioppi, Sapienza Università di Roma

#### *Tesoriere:*

- Prof.ssa Maria Felice Arezzo, Sapienza Università di Roma

#### *Consiglieri:*

- Prof. Giuseppe Arbia, Università Cattolica del Sacro Cuore
- Prof.ssa Maria Maddalena Barbieri, Università Roma Tre
- Prof.ssa Francesca Bassi, Università di Padova
- Prof. Eugenio Brentari, Università di Brescia
- Dott. Stefano Falorsi, ISTAT
- Prof. Alessio Pollice, Università di Bari
- Prof.ssa Rosanna Verde, Seconda Università di Napoli
- Prof. Daniele Vignoli, Università di Firenze

#### *Collegio dei Revisori dei Conti:*

- Prof. Francesco Campobasso, Prof. Michele Gallo, Prof. Francesco Sanna, Prof. Umberto Salinas (supplente)

## **SIS2017 Committees**

### **Scientific Program Committee:**

Rosanna Verde (chair), Università della Campania “Luigi Vanvitelli”  
Maria Felice Arezzo, Sapienza Università di Roma  
Antonino Mazzeo, Università di Napoli Federico II  
Emanuele Baldacci, Eurostat  
Pierpaolo Brutti, Sapienza Università di Roma  
Marcello Chiodi, Università di Palermo  
Corrado Crocetta, Università di Foggia  
Giovanni De Luca, Università di Napoli Parthenope  
Viviana Egidi, Sapienza Università di Roma  
Giulio Ghellini, Università degli Studi di Siena  
Ippoliti Luigi, Università di Chieti-Pescara “G. D’Annunzio”  
Matteo Mazziotta, ISTAT  
Lucia Paci, Università Cattolica del Sacro Cuore  
Alessandra Petrucci, Università degli Studi di Firenze  
Filomena Racioppi, Sapienza Università di Roma  
Laura M. Sangalli, Politecnico di Milano  
Bruno Scarpa, Università degli Studi di Padova  
Cinzia Viroli, Università di Bologna

### **Local Organizing Committee:**

Alessandra Petrucci (chair), Università degli Studi di Firenze  
Gianni Betti, Università degli Studi di Siena  
Fabrizio Cipollini, Università degli Studi di Firenze  
Emanuela Dreassi, Università degli Studi di Firenze  
Caterina Giusti, Università di Pisa  
Leonardo Grilli, Università degli Studi di Firenze  
Alessandra Mattei, Università degli Studi di Firenze  
Elena Pirani, Università degli Studi di Firenze  
Emilia Rocco, Università degli Studi di Firenze  
Maria Cecilia Verri, Università degli Studi di Firenze

### **Supported by:**

Università degli Studi di Firenze  
Università di Pisa  
Università degli Studi di Siena  
ISTAT  
Regione Toscana  
Comune di Firenze  
BITBANG srl

# Index

Preface	XXV
Alexander Agapitov, Irina Lackman, Zoya Maksimenko <i>Determination of basis risk multiplier of a borrower default using survival analysis</i>	1
Tommaso Agasisti, Alex J. Bowers, Mara Soncin <i>School principals' leadership styles and students achievement: empirical results from a three-step Latent Class Analysis</i>	7
Tommaso Agasisti, Sergio Longobardi, Felice Russo <i>Poverty measures to analyse the educational inequality in the OECD Countries</i>	17
Mohamed-Salem Ahmed, Laurence Broze, Sophie Dabo-Niang, Zied Gharbi <i>Quasi-Maximum Likelihood Estimators For Functional Spatial Autoregressive Models</i>	23
Giacomo Aletti, Alessandra Micheletti <i>A clustering algorithm for multivariate big data with correlated components</i>	31
Emanuele Aliverti <i>A Bayesian semiparametric model for terrorist networks</i>	37

- Giorgio Alleva  
*Emerging challenges in official statistics: new sources, methods and skills* 43
- Rémi André, Xavier Luciani and Eric Moreau  
*A fast algorithm for the canonical polyadic decomposition of large tensors* 45
- Maria Simona Andreano, Roberto Benedetti, Paolo Postiglione, Giovanni Savio  
*On the use of Google Trend data as covariates in nowcasting: Sampling and modeling issues* 53
- Francesco Andreoli, Mauro Mussini  
*A spatial decomposition of the change in urban poverty concentration* 59
- Margaret Antonicelli, Vito Flavio Covella  
*How green advertising can impact on gender different approach towards sustainability* 65
- Rosa Arboretti, Eleonora Carrozzo, Luigi Salmaso  
*Stratified data: a permutation approach for hypotheses testing* 71
- Marika Arena, Anna Calissano, Simone Vantini  
*Crowd and Minorities: Is it possible to listen to both? Monitoring Rare Sentiment and Opinion Categories about Expo Milano 2015* 79
- Maria Felice Arezzo, Giuseppina Guagnano  
*Using administrative data for statistical modeling: an application to tax evasion* 83
- Monica Bailot, Rina Camporese, Silvia Da Valle, Sara Letardi, Susi Osti  
*Are Numbers too Large for Kids? Possible Answers in Probable Stories* 89

Index	IX
Simona Balbi, Michelangelo Misuraca, Germana Scepti <i>A polarity-based strategy for ranking social media reviews</i>	95
A. Balzanella, S.A. Gattone, T. Di Battista, E. Romano, R. Verde <i>Monitoring the spatial correlation among functional data streams through Moran's Index</i>	103
Oumayma Banouar, Saïd Raghay <i>User query enrichment for personalized access to data through ontologies using matrix completion method</i>	109
Giulia Barbati, Francesca Ieva, Francesca Gasperoni, Annamaria Iorio, Gianfranco Sinagra, Andrea Di Lenarda <i>The Trieste Observatory of cardiovascular disease: an experience of administrative and clinical data integration at a regional level</i>	115
Francesco Bartolucci, Stefano Peluso, Antonietta Mira <i>Marginal modeling of multilateral relational events</i>	123
Francesca Bassi, Leonardo Grilli, Omar Paccagnella, Carla Rampichini, Roberta Varriale <i>New Insights on Students Evaluation of Teaching in Italy</i>	129
Mauro Bernardi, Marco Bottone, Lea Petrella <i>Bayesian Quantile Regression using the Skew Exponential Power Distribution</i>	135
Mauro Bernardi <i>Bayesian Factor-Augmented Dynamic Quantile Vector Autoregression</i>	141

- Bruno Bertaccini, Giulia Biagi, Antonio Giusti, Laura Grassini  
*Does data structure reflect monuments structure? Symbolic data analysis on Florence Brunelleschi Dome*  
149
- Gaia Bertarelli and Franca Crippa, Fulvia Mecatti  
*A latent markov model approach for measuring national gender inequality*  
157
- Agne Bikauskaite, Dario Buono  
*Eurostat's methodological network: Skills mapping for a collaborative statistical office*  
161
- Francesco C. Billari, Emilio Zagheni  
*Big Data and Population Processes: A Revolution?*  
167
- Monica Billio, Roberto Casarin, Matteo Iacopini  
*Bayesian Tensor Regression models*  
179
- Monica Billio, Roberto Casarin, Luca Rossini  
*Bayesian nonparametric sparse Vector Autoregressive models*  
187
- Chiara Bocci, Daniele Fadda, Lorenzo Gabrielli, Mirco Nanni, Leonardo Piccini  
*Using GPS Data to Understand Urban Mobility Patterns: An Application to the Florence Metropolitan Area*  
193
- Michele Boreale, Fabio Corradi  
*Relative privacy risks and learning from anonymized data*  
199
- Giacomo Bormetti, Roberto Casarin, Fulvio Corsi, Giulia Livieri  
*A stochastic volatility framework with analytical filtering*  
205

Index	XI
Alessandro Brunetti, Stefania Fatello, Federico Polidoro <i>Estimating Italian inflation using scanner data: results and perspectives</i>	211
Guénael Cabanes, Younès Bennani, Rosanna Verde, Antonio Irpino <i>Clustering of histogram data : a topological learning approach</i>	219
Renza Campagni, Lorenzo Gabrielli, Fosca Giannotti, Riccardo Guidotti, Filomena Maggino, Dino Pedreschi <i>Measuring Wellbeing by extracting Social Indicators from Big Data</i>	227
Maria Gabriella Campolo, Antonino Di Pino <i>Assessing Selectivity in the Estimation of the Causal Effects of Retirement on the Labour Division in the Italian Couples</i>	235
Stefania Capecchi, Rosaria Simone <i>Composite indicators for ordinal data: the impact of uncertainty</i>	241
Stefania Capecchi, Domenico Piccolo <i>The distribution of Net Promoter Score in socio-economic surveys</i>	247
Massimiliano Caporin, Francesco Poli <i>News, Volatility and Price Jumps</i>	253
Carmela Cappelli, Rosaria Simone, Francesca di Iorio <i>Growing happiness: a model-based tree</i>	261
Paolo Emilio Cardone <i>Inequalities in access to job-related learning among workers in Italy: evidence from Adult Education Survey (AES)</i>	267

- Alessandro Casa, Giovanna Menardi  
*Signal detection in high energy physics via a semisupervised nonparametric approach*  
273
- Claudio Ceccarelli, Silvia Montagna, Francesca Petrarca  
*Employment study methodologies of Italian graduates through the data linkage of administrative archives and sample surveys*  
279
- Ikram Chairi, Amina El Gonnouni, Sarah Zouinina, Abdelouahid Lyhyaoui  
*Prediction of Firm's Creditworthiness Risk using Feature Selection and Support Vector Machine*  
285
- Sana Chakri, Said Raghay, Salah El Hadaj  
*Contribution of extracting meaningful patterns from semantic trajectories*  
293
- Chieppa A., Ferrara R., Gallo G., Tomeo V.  
*Towards The Register-Based Statistical System: A New Valuable Source for Population Studies*  
301
- Shirley Coleman  
*Consulting, knowledge transfer and impact case studies of statistics in practice*  
305
- Michele Costa  
*The evaluation of the inequality between population subgroups*  
313
- Michele Costola  
*Bayesian Non-Negative  $l_1$ -Regularised Regression*  
319
- Lisa Crosato, Caterina Liberati, Paolo Mariani, Biancamaria Zavarella  
*Industrial Production Index and the Web: an explorative cointegration analysis*  
327



Index	XIII
Francesca Romana Crucinio, Roberto Fontana <i>Comparison of conditional tests on Poisson data</i>	333
Riccardo D'Alberto, Meri Raggi <i>Non-parametric micro Statistical Matching techniques: some developments</i>	339
Stefano De Cantis, Mauro Ferrante, Anna Maria Parroco <i>Measuring tourism from demand side</i>	345
Lucio De Capitani, Daniele De Martini <i>Optimal Ethical Balance for Phase III Trials Planning</i>	351
Claudia De Vitiis, Alessio Guandalini, Francesca Inglese, Marco D. Terribili <i>Sampling schemes using scanner data for the consumer price index</i>	357
Ermelinda Della Valle, Elena Scardovi, Andrea Iacobucci, Edoardo Tignone <i>Interactive machine learning prediction for budget allocation in digital marketing scenarios</i>	365
Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor <i>Nonparametric classification for directional data</i>	371
Edwin Diday <i>Introduction to Symbolic Data Analysis and application to post clustering for comparing and improving clustering methods by the Symbolic Data Table that they induce</i>	379
Carlo Drago <i>Identifying Meta Communities on Large Networks</i>	387

- Neska El Haouij, Jean-Michel Poggi, Raja Ghozi, Sylvie Sevestre Ghalila, Mériem Jaidane  
*Random Forest-Based Approach for Physiological Functional Variable Selection for Drivers Stress Level Classification*  
393
- Silvia Facchinetti, Silvia A. Osmetti  
*A risk index to evaluate the criticality of a product defectiveness*  
399
- Federico Ferraccioli, Livio Finos  
*Exponential family graphical models and penalizations*  
405
- Mauro Ferrante, Giovanna Fantaci, Anna Maria Parroco, Anna Maria Milito, Salvatore Scondotto  
*Key-indicators for maternity hospitals and newborn readmission in Sicily*  
411
- Ferretti Camilla, Ganugi Piero, Zammori Francesco  
*Change of Variables theorem to fit Bimodal Distributions*  
417
- Francesco Finazzi, Lucia Paci  
*Space-time clustering for identifying population patterns from smartphone data*  
423
- Annunziata Fiore, Antonella Simone, Antonino Virgillito  
*IT Solutions for Analyzing Large-Scale Statistical Datasets: Scanner Data for CPI*  
429
- Michael Fop, Thomas Brendan Murphy, Luca Scrucca  
*Model-based Clustering with Sparse Covariance Matrices*  
437
- Maria Franco-Villoria, Marian Scott  
*Quantile Regression for Functional Data*  
441

Index	XV
Gallo M., Simonacci V., Di Palma M.A. <i>Three-way compositional data: a multi-stage trilinear decomposition algorithm</i>	445
Francesca Gasperoni, Francesca Ieva, Anna Maria Paganoni, Chris Jackson, Linda Sharples <i>Nonparametric shared frailty model for classification of survival data</i>	451
Stefano A. Gattone, Angela De Sanctis <i>Clustering landmark-based shapes using Information Geometry tools</i>	457
Alan E. Gelfand, Shinichiro Shirota <i>Space and circular time log Gaussian Cox processes with application to crime event data</i>	461
Abdelghani Ghazdali <i>Blind source separation</i>	469
Massimiliano Giacalone, Antonio Ruoto, Davide Liga, Maria Pilato, Vito Santarangelo <i>An innovative approach for Opinion Mining : the Plutchick analysis</i>	479
Massimiliano Giacalone, Demetrio Panarello <i>A G.E.D. method for market risk evaluation using a modified Gaussian Copula</i>	485
Chiara Gigliarano, Francesco Maria Chelli <i>Labour market dynamics and recent economic changes: the case of Italy</i>	491
Giuseppe Giordano, Giancarlo Ragozini, Maria Prosperina Vitale <i>On the use of DISTATIS to handle multiplex networks</i>	499

- Michela Gnaldi, Silvia Bacci, Samuel Greiff, Thiemo Kunze  
*Profiles of students on account of complex problem solving (CPS) strategies exploited via log-data*  
505
- Michela Gnaldi, Simone Del Sarto  
*Characterising Italian municipalities according to the annual report of the prevention-of-corruption supervisor: a Latent Class approach*  
513
- Silvia Golia  
*A proposal of a discretization method applicable to Rasch measures*  
519
- Anna Gottard  
*Tree-based Non-linear Graphical Models*  
525
- Sara Hbali, Youssef Hbali, Mohamed Sadgal, Abdelaziz El Fazziki  
*Sentiment Analysis for micro-blogging using LSTM Recurrent Neural Networks*  
531
- Stefano Maria Iacus, Giuseppe Porro, Silvia Salini, Elena Siletti  
*How to Exploit Big Data from Social Networks: a Subjective Well-being Indicator via Twitter*  
537
- Francesca Ieva  
*Network Analysis of Comorbidity Patterns in Heart Failure Patients using Administrative Data*  
543
- Antonio Irpino, Francisco de A.T. De Carvalho, Rosanna Verde  
*Automatic variable and components weighting systems for Fuzzy cmeans of distributional data*  
549
- Michael Jauch, Paolo Giordani, David Dunson  
*A Bayesian oblique factor model with extension to tensor data*  
553

Index	XVII
Johan Koskinen, Chiara Broccatelli, Peng Wang, Garry Robins <i>Statistical analysis for partially observed multilayered networks</i>	561
Francesco Lagona <i>Copula-based segmentation of environmental time series with linear and circular components</i>	569
Alessandro Lanteri, Mauro Maggioni <i>A Multiscale Approach to Manifolds Estimation</i>	575
Tiziana Laureti, Carlo Ferrante, Barbara Dramis <i>Using scanner and CPI data to estimate Italian sub-national PPPs</i>	581
Antonio Lepore <i>Graphical approximation of Best Linear Unbiased Estimators for Extreme Value Distribution Parameters</i>	589
Antonio Lepore, Biagio Palumbo, Christian Capezza <i>Monitoring ship performance via multi-way partial least-squares analysis of functional data</i>	595
Caterina Liberati, Lisa Crosato, Paolo Mariani, Biancamaria Zavanella <i>Dynamic profiling of banking customers: a pseudo-panel study</i>	601
Giovanni L. Lo Magno, Mauro Ferrante, Stefano De Cantis <i>A comparison between seasonality indices deployed in evaluating unimodal and bimodal patterns</i>	607
Rosaria Lombardo, Eric J Beh <i>Three-way Correspondence Analysis for Ordinal-Nominal Variables</i>	613

- Monia Lupparelli, Alessandra Mattei  
*Log-mean linear models for causal inference*  
621
- Badiaa Lyoussi, Zineb Selihi, Mohamed Berraho, Karima El Rhazi, Youness El Achhab, Adiba El Marrakchi, Chakib Nejjari  
*Research on the Risk Factors accountable for the occurrence of degenerative complications of type 2 diabetes in Morocco: a prospective study*  
627
- Valentina Mameli, Debora Slanzi, Irene Poli  
*Bootstrap group penalty for high-dimensional regression models*  
633
- Stefano Marchetti, Monica Pratesi, Caterina Giusti  
*Improving small area estimates of households' share of food consumption expenditure in Italy by means of Twitter data*  
639
- Paolo Mariani, Andrea Marletta, Mariangela Zenga  
*Gross Annual Salary of a new graduate: is it a question of profile?*  
647
- Maria Francesca Marino, Marco Alfò  
*Dynamic random coefficient based drop-out models for longitudinal responses*  
653
- Antonello Maruotti, Jan Bulla  
*Hidden Markov models: dimensionality reduction, atypical observations and algorithms*  
659
- Chiara Masci, Geraint Johnes, Tommaso Agasisti  
*A flexible analysis of PISA 2015 data across countries, by means of multilevel trees and boosting*  
667

Index	XIX
Lucio Masserini, Matilde Bini <i>Impact of the 2008 and 2012 financial crises on the unemployment rate in Italy: an interrupted time series approach</i>	673
Angelo Mazza, Antonio Punzo, Salvatore Ingrassia <i>An R Package for Cluster-Weighted Models</i>	681
Antonino Mazzeo, Flora Amato <i>Methods and applications for the treatment of Big Data in strategic fields</i>	687
Letizia Mencarini, Viviana Patti, Mirko Lai, Emilio Sulis <i>Happy parents' tweets</i>	693
Rodolfo Metulini, Marica Manisera, Paola Zuccolotto <i>Space-Time Analysis of Movements in Basketball using Sensor Data</i>	701
Giorgio E. Montanari, Marco Doretti, Francesco Bartolucci <i>An ordinal Latent Markov model for the evaluation of health care services</i>	707
Isabella Morlini, Maristella Scorza <i>New fuzzy composite indicators for dyslexia</i>	713
Fionn Murtagh <i>Big Textual Data: Lessons and Challenges for Statistics</i>	719
Gaetano Musella, Gennaro Punzo <i>Workers' skills and wage inequality: A time-space comparison across European Mediterranean countries</i>	731

- Marta Nai Ruscone  
*Exploratory factor analysis of ordinal variables: a copula approach* 737
- Fausta Ongaro, Silvana Salvini  
*IPUMS Data for describing family and household structures in the world* 743
- Tullia Padellini, Pierpaolo Brutti  
*Topological Summaries for Time-Varying Data* 747
- Sally Paganin  
*Modeling of Complex Network Data for Targeted Marketing* 753
- Francesco Palumbo, Giancarlo Ragozini  
*Statistical categorization through archetypal analysis* 759
- Michela Eugenia Pasetto, Umberto Noè, Alessandra Luati, Dirk Husmeier  
*Inference with the Unscented Kalman Filter and optimization of sigma points* 767
- Xanthi Pedeli, Cristiano Varin  
*Pairwise Likelihood Inference for Parameter-Driven Models* 773
- Felicia Pelagalli, Francesca Greco, Enrico De Santis  
*Social emotional data analysis. The map of Europe* 779
- Alessia Pini, Lorenzo Spreafico, Simone Vantini, Alessandro Vietti  
*Differential Interval-Wise Testing for the Inferential Analysis of Tongue Profiles* 785
- Alessia Pini, Aymeric Stamm, Simone Vantini  
*Hotelling meets Hilbert: inference on the mean in functional Hilbert spaces* 791



Index	XXI
Silvia Poletini, Serena Arima <i>Accounting for measurement error in small area models: a study on generosity</i>	795
Gennaro Punzo, Mariateresa Ciommi <i>Structural changes in the employment composition and wage inequality: A comparison across European countries</i>	801
Walter J. Radermacher <i>Official Statistics 4.0 – learning from history for the challenges of the future</i>	809
Fabio Rapallo <i>Comparison of contingency tables under quasi-symmetry</i>	821
Valentina Raponi, Cesare Robotti, Paolo Zaffaroni <i>Testing Beta-Pricing Models Using Large Cross-Sections</i>	827
Marco Seabra dos Reis, Biagio Palumbo, Antonio Lepore, Ricardo Rendall, Christian Capezza <i>On the use of predictive methods for ship fuel consumption analysis from massive on-board operational data</i>	833
Alessandra Righi, Mauro Mario Gentile <i>Twitter as a Statistical Data Source: an Attempt of Profiling Italian Users Background Characteristics</i>	841
Paolo Righi, Giulio Barcaroli, Natalia Golini <i>Quality issues when using Big Data in Official Statistics</i>	847
Emilia Rocco <i>Indicators for the representativeness of survey response as well as convenience samples</i>	855

- Emilia Rocco, Bruno Bertaccini, Giulia Biagi, Andrea Giommi  
*A sampling design for the evaluation of earthquakes vulnerability of the residential buildings in Florence*  
861
- Elvira Romano, Jorge Mateu  
*A local regression technique for spatially dependent functional data: an heteroskedastic GWR model*  
867
- Eduardo Rossi, Paolo Santucci de Magistris  
*Models for jumps in trading volume*  
873
- Renata Rotondi, Elisa Varini  
*On a failure process driven by a self-correcting model in seismic hazard assessment*  
879
- M. Ruggieri, F. Di Salvo and A. Plaia  
*Functional principal component analysis of quantile curves*  
887
- Massimiliano Russo  
*Detecting group differences in multivariate categorical data*  
893
- Michele Scagliarini  
*A Sequential Test for the  $C_{pk}$  Index*  
899
- Steven L. Scott  
*Industrial Applications of Bayesian Structural Time Series*  
905
- Catia Scricciolo  
*Asymptotically Efficient Estimation in Measurement Error Models*  
913

Index	XXIII
Angela Serra, Pietro Coretto, Roberto Tagliaferri <i>On the noisy high-dimensional gene expression data analysis</i>	919
Mirko Signorelli <i>Variable selection for (realistic) stochastic blockmodels</i>	927
Marianna Siino, Francisco J. Rodriguez-Cortés, Jorge Mateu, Giada Adelfio <i>Detection of spatio-temporal local structure on seismic data</i>	935
A. Sottosanti, D. Bastieri, A. R. Brazzale <i>Bayesian Mixture Models for the Detection of High-Energy Astronomical Sources</i>	943
Federico Mattia Stefanini <i>Causal analysis of Cell Transformation Assays</i>	949
Paola Stolfi, Mauro Bernardi, Lea Petrella <i>Estimation and Inference of SkewStable distributions using the Multivariate Method of Simulated Quantiles</i>	955
Paola Stolfi, Mauro Bernardi, Lea Petrella <i>Sparse Indirect Inference</i>	961
Peter Struijs, Anke Consten, Piet Daas, Marc Debusschere, Maiki Ilves, Boro Nikic, Anna Nowicka, David Salgado, Monica Scannapieco, Nigel Swier <i>The ESSnet Big Data: Experimental Results</i>	969
Jérémie Sublime <i>Smart view selection in multi-view clustering</i>	977

- Emilio Sulis  
*Social Sensing and Official Statistics: call data records and social media sentiment analysis*  
985
- Matilde Trevisani, Arjuna Tuzzi  
*Knowledge mapping by a functional data analysis of scientific articles databases*  
993
- Amalia Vanacore, Maria Sole Pellegrino  
*Characterizing the extent of rater agreement via a non-parametric benchmarking procedure*  
999
- Maarten Vanhoof, Stephanie Combes, Marie-Pierre de Bellefon  
*Mining Mobile Phone Data to Detect Urban Areas*  
1005
- Viktoriya Voytsekhovska, Olivier Butzbach  
*Statistical methods in assessing the equality of income distribution, case study of Poland*  
1013
- Ernst C. Wit  
*Network inference in Genomics*  
1019
- Dilek Yildiz, Jo Munson, Agnese Vitali, Ramine Tinati, Jennifer Holland  
*Using Twitter data for Population Estimates*  
1025
- Marco Seabra dos Rei  
*Structured Approaches for High-Dimensional Predictive Modeling*  
1033

# Preface

The 2017 SIS Conference aims to highlight the crucial role of the Statistics in Data Science. In this new domain of “meaning” extracted from the data, the increasing amount of produced and available data in databases, nowadays, has brought new challenges. That involves different fields of statistics, machine learning, information and computer science, optimization, pattern recognition. These afford together a considerable contribute in the analysis of “Big data”, open data, relational and complex data, structured and no-structured. The interest is to collect the contributes which provide from the different domains of Statistics, in the high dimensional data quality validation, sampling extraction, dimensional reduction, pattern selection, data modelling, testing hypotheses and confirming conclusions drawn from the data. In the mention that statistics is the “grammar of data science”, statistics has become a basic skill in data science: it gives right meaning to the data. Still, it isn’t replaced by newer techniques from machine learning and other disciplines but it complements them. The Conference is also addressed to the new challenges of the new generations: the native digital generations, who are called to develop professional skills as “data analyst”, one of the more request professionalism of the 21st Century, crossing the rigid disciplinary domains of competence. In this perspective, all the traditional statistical topics are admitted with an extension to the related machine learning and computer science ones. The present volume includes the short papers of the contributions that will be presented in the 4 invited speaker sessions; in the 19 specialized sessions; in the 11 solicited sessions; in the 6 foreign societies sessions and in the 17 contributed sessions as well as, in the panel session.

*Rosanna Verde*  
*President of the Scientific Programme Committee*

*Alessandra Petrucci*  
*President of the Local Organizing Committee*



# Determination of basis risk multiplier of a borrower default using survival analysis

## Determinazione del moltiplicatore di rischio di base di un default mutuario attraverso un'analisi di sopravvivenza

Alexander Agapitov, Irina Lackman, Zoya Maksimenko

**Abstract** The provided research is directed to identification of the predictors affecting at sizes of basis risk multiplier of a loan default for a certain period. Survival models (Cox proportional hazard models) taking into account a grouping sign of rating of reliability of borrowers are put in the basis of calculations. In the conducted research data on loans in the Californian company Lending Club which is engaged in equal crediting were used. The borrower for whom the risk of approach of a default by a certain period was predicted acted as an object of the research.

**Abstract** *Lo scopo dell'analisi è quello di individuare gli elementi che influenzano il valore dei moltiplicatori del rischio base relativamente al mancato pagamento del prestito per un certo periodo. L'analisi è condotta mediante dei modelli di sopravvivenza (modelli a rischi proporzionali; modelli di Cox), tenendo conto del gruppo di reputazione e dell'affidabilità dei debitori. Nella ricerca effettuata sono stati utilizzati i dati sui prestiti di una società californiana Lending Club, che si occupa della concessione del credito. L'oggetto della ricerca era il debitore, per il quale è stato determinato il rischio di insolvenza ad un certo periodo.*

**Key words:** survival analysis, Kaplan-Meier estimator, Cox proportional hazards model

---

Alexander Agapitov

Ufa State Aviation Technical University, Ufa, Russia, e-mail: aleks6321@yandex.ru

Irina Lackman

Ufa State Aviation Technical University, Ufa, Russia, e-mail: lackmania@mail.ru

Zoya Maksimenko

Ufa State Aviation Technical University, Ufa, Russia, e-mail: zubazz@mail.ru

## 1 Introduction

Lending to individuals implies the consideration of all possible risks that could lead to the borrower obligation default. In the context of economic crisis the problem of population debt incurring becomes extremely urgent. In this regard, it is important to be able to identify correctly the multipliers of the basis risk of borrower default for a certain period, taking into account characteristics of both the borrower and features of the credit product. The survival analysis may serve as one of the tools for solving such problems.

The main advantage of the survival analysis compared to other models of credit scoring is the model ability to work with the right censoring data, i.e. when the event (the default) for the object is not observed during the exploration period (for example, the borrower has already paid back the loan in full or he/she is still paying it at the end of the exploration period). Another advantage of survival models is the opportunity not only to assess whether the borrower will pay (will not pay) the loan, but also the possibility of estimating the time of loan obligation fulfillment in good faith. Here, the time of loan obligation fulfillment in good faith will be assumed as a conventional "survival" of the borrower for the Bank, i.e. we consider that for the Bank the borrower has died if he/she ceases to make payment of loan installments on time.

There are many studies using the survival analysis to estimate the default of banks:

1. Survival analysis of private Banks in Brazil in the period of 1994-2007. [1]
2. Bank failure prediction: a two-step survival time approach (the joint research of the University of Vienna and the National Bank of Austria) [3]
3. Start-up banks default and the role of capital (the research of the Bank of Italy according to the data of 1994-2006) [4]

There are also studies where the survival analysis is applied to examine the borrower's default:

1. Survival analysis methods for personal loan data [6]
2. Credit scoring with macroeconomic variables using the survival analysis [2]
3. Survival analysis in credit scoring: A framework for PD estimation [5]

## 2 Data

In our research we used loan data of the Californian company Lending Club, one of the largest peer-to-peer lender in the USA. The summary table on loans was taken from the Kaggle portal a platform holding competitions in the analysis of data. The sample is consisted of 887 379 observations for the period of 2007-2015. For this research we used 36-month loans, the final dataset comprises 602,871 loans.



The risk of event (default) occurrence was predicted for the borrower who was an object of observation. This object was under observation and therefore the borrower was included in the credit risk group: at any period of time there may occur an event when the borrower leaves the risk group. Observation period starts from the moment when the borrower takes a loan and finishes when the borrower default occurs.

The independent variables (predictors) are characteristics of an object, which may influence the risk of event occurrence. We used the following predictors: interest rate on the loan, employee tenure, annual income, the region of the borrower inhabitation, residential property, credit history (the first loan), loan amount, loan purpose and financial reliability of the borrower calculated by Lending Club on the scale from A to D where A is the best possible grade and D the worst.

### 3 Survival analysis

At the first stage of the study the method of Kaplan-Meier was used to identify determinants of the borrower which are predictors of the loan obligation fulfillment in good faith. The graphs of survival functions obtained by using the Kaplan-Meier estimates have shown that the most predictors have significant differences between the group alternatives. Thus, it is possible to make a conclusion about the expediency of survival models application to solve these problems.

At the second stage of the analysis there were tests with a null hypothesis of the survival indistinction groupwise: the log-rank criterion of Mantel-Haenszel and the criterion of Gehan-Wilcoxon. The value of statistics, the number of freedom degrees and significance level for every determinant of the borrower for each test are presented in table 1.

**Table 1** Survival analysis: tests

Variable	Log-rank test			Gehan's Generalized Wilcoxon		
	$\chi^2$ statistic	Degrees of freedom	p-value	$\chi^2$ statistic	Degrees of freedom	p-value
Home ownership	968	2	0	987	2	0
Earliest credit line	742	2	0	749	2	0
Interest rate	10887	3	0	11036	3	0
Annual income	2097	6	0	2121	6	0
Funded Amount	183	4	0	196	4	0
Employment length	499	5	0	512	5	0
Region of the US	130	8	0	131	8	0
Credit purpose	1439	8	0	1395	8	0

Test results showed a statistically significant difference between the groups for each variable. Thus, it was concluded that in order to build the Cox survival model it is necessary to use all predictors of the borrower. The choice preference of Cox

proportional-hazards survival model in comparison with other models was made after the selection procedures based on Akaike and Schwarz information criterion.

At the third stage of the model construction after a preliminary evaluation of the generalized data it was decided to evaluate the proportional hazards model, taking into account the bank's customer groups of "reliable" and "unreliable" clients received by Lending Club. Table 2 shows the results of calculations by multipliers compared to the basis risk calculated using the Cox model for "reliable" and "unreliable" clients respectively.

Table 2: Survival analysis: Cox models

Variable	Level	Good	Bad
Home ownership	Mortgage	0.899	0.905
	Own	0.922	0.946
Earliest credit line	from 1990 to 2000	1.115	1.041
	after 2000	1.150	1.145
Interest rate	>10%	1.864	—
	From 15% to 20%	—	1.416
	>20%	—	2.009
Annual income	From 15 to 30	—	0.998
	From 30 to 50	0.868	0.934
	From 50 to 75	0.706	0.845
	From 75 to 100	0.597	0.741
	From 100 to 150	0.548	0.665
Funded Amount	>150	0.514	0.636
	From 5000 to 10000	1.049	1.228
	From 10000 to 15000	1.084	1.332
	From 15000 to 25000	1.160	1.431
Employment length	>25000	1.204	1.487
	Less than 1 year	0.766	0.943
	1 year	0.690	0.868
	From 2 to 5 years	0.691	0.852
Region of the US	From 6 to 9 years	0.717	0.864
	10 and more	0.699	0.798
	Mountain	1.063	1.000
Region of the US	West North Central	0.992	0.946
	East North Central	0.940	0.924
	West South Central	0.966	0.926
	East South Central	1.128	1.076
	South Atlantic	1.050	0.983
	Mid-Atlantic	1.062	0.957
	New England	0.946	0.870
Purpose	Credit card	0.867	0.843
	Major purchase	0.888	0.874
	Other	1.216	0.955

Variable	Level	Good	Bad
	Car	0.807	0.899
	Medical	1.369	1.136
	Small business	1.980	1.322
	House	1.044	1.134
	Home improvement	1.043	0.978

As a result of the analysis the following conclusions can be made for the "reliable" clients:

1. The risk of debt at higher interest rates remains. The risk for borrowers with high interest rate is in 1.86 times higher than for borrowers with low interest rate.
2. "Reliable" clients risk of debt at the same annual income is lower than in the general model.
3. The size of the loan affects the risk of debt less than other factors. Credit larger than 25 thousand dollars increases the risk in 1.2 times (this indicator increases the risk in 1.5 times in the general model) compared to the baseline.
4. An interesting situation concerning the "credit assignment" variable is observed. The risk of debt of a borrower with a loan to a small business is in 2 times more than customers with basis risk. Also the risk of borrowers with credit for medical services increases; it is in 1.4 times higher compared to the basis risk.

## 4 Conclusions

"Reliable" clients have a lower risk of debt with high socio-economic indicators compared to the conventional model. At the same time, borrowers who took credit for small business have much higher risks.

For "unreliable" borrowers the following risks can be identified:

1. Borrowers who live in owner-occupied dwelling bear the risk by 10% greater compared to borrowers living in rented accommodation.
2. The risk of debt for borrowers with interest rates ranging from 15 to 20 percent is in 1.42 times greater than for borrowers with an interest rate of less than 15%. For customers who have a loan at the interest rate more than 20%, the risk of debt is in 2 times more compared to the basis risk.
3. Clients with annual income of less than \$50 thousand, carry the same high risk of debt.
4. The risk of debt for borrowers whose loan size of more than 15 thousand dollars is in 1.45 times higher than customers with the basis risk.
5. Borrowers living on the East Coast of the USA, on average, carry lower risk of debt compared to the inhabitants of the West Coast and mountain states.
6. The risk of debt for borrowers who took credit for small business is by 32% higher compared to the basis risk. Borrowers with credit for medical services or

real estate purchase have the higher risk by 3%. Like in other models borrowers who took a loan to pay off the credit card debt or to buy a car carry the least risk.

As a conclusion it can be given the following recommendation: in order to reduce the debt risk for "unreliable" customers, Lending Club Company should be more careful selecting customers who want to take a large loan (more than 15 thousand dollars). The high risk of debt is for borrowers with high interest rate.

## References

1. Alves, K., Kalatzis, A., Matias, A.: Analysis of Private Banks in Brazil. No 21500002, EcoMod2009, EcoMod
2. Bellotti, T., Crook, J.: Credit scoring with macroeconomic variables using survival analysis. *J. Oper. Res. Soc.* 60.12, 1699–1707 (2009)
3. Halling M., Hayden E.: Bank failure prediction: a two-step survival time approach. (2006)
4. Libertucci, M., Piersante, F.: Start-up banks default and the role of capital. *Banca D'Italia*, 890 (2012)
5. Man, R.: Survival analysis in credit scoring: A framework for PD estimation. University of Twente (2014)
6. Stepanova, M., Thomas, L.: Survival analysis methods for personal loan data. *Oper. Res.* 50.2, 277-289 (2002)

# School principals' leadership styles and students' achievement: empirical results from a three-step Latent Class Analysis

## *Stili di leadership dei Dirigenti Scolastici e apprendimenti degli studenti: risultati da una three-step Latent Class Analysis*

Tommaso Agasisti, Alex J. Bowers and Mara Soncin

**Abstract** This study exploits the existence of various leadership types in a sample of lower secondary school principals across Italy (N=1,073). Information is derived by a questionnaire provided by INVALSI (National Evaluation Committee for Education) about instructional practices and leadership perceptions. Employing a Latent Class Analysis (LCA), we identify three subgroups of school leaders. We then analyze if some principal's individual characteristics and school context factors are statistically correlated with the probability of having a certain leadership styles' attitude. Finally, we provide evidence that schools where the principal is adopting an "instructional" approach report lower academic test scores.

**Abstract** *La presente ricerca ha lo scopo di indagare l'esistenza di diversi stili di leadership in un campione rappresentativo di scuole secondarie di primo grado italiane (N=1,073). Le informazioni sono tratte dal questionario INVALSI (Istituto Nazionale per la Valutazione del Sistema Educativo) rivolto ai Dirigenti Scolastici, in cui vengono testate le pratiche manageriali e di leadership implementate nelle scuole. Implementando una Latent Class Analysis (LCA), vengono identificati tre approcci alla leadership scolastica. Successivamente, viene analizzata la correlazione tra tali approcci ed una serie di caratteristiche di contesto e del Dirigente Scolastico in capo. Infine, l'analisi mostra come i Dirigenti Scolastici che adottano una leadership "educativa" riportano punteggi inferiori nei test standardizzati.*

**Key words:** Leadership practices, managerial practices, latent class analysis

---

<sup>1</sup> Tommaso Agasisti, Politecnico di Milano; email: [tommaso.agasisti@polimi.it](mailto:tommaso.agasisti@polimi.it)  
Alex J. Bowers, Teachers College, Columbia University; email:  
[Bowers@exchange.tc.columbia.edu](mailto:Bowers@exchange.tc.columbia.edu)  
Mara Soncin, Politecnico di Milano; email: [mara.soncin@polimi.it](mailto:mara.soncin@polimi.it)

## 1. Introduction and existing literature

Research evidence has demonstrated the importance of school leadership in influencing students' success in both cognitive and non-cognitive outcomes (Robinson *et al.*, 2008, Waters *et al.*, 2003). Among school factors, leadership is second only to classroom conditions in influencing achievement (Day *et al.*, 2009, Leithwood *et al.*, 2008). Looking for the existence of different leadership styles, literature has moved from the predominant role of instructional leadership (Smith & Andrews, 1989) to a more comprehensive vision of school leadership, with a growing emphasis on transformational, transactional and distributed leadership (Day *et al.*, 2016, Marks & Printy, 2003, Urlick & Bowers, 2014). Given the fact that the relationship between leadership styles and student achievement/engagement can be both direct and mediated by the role of teachers in the classroom or by school contextual conditions, the search for the most suitable model to measure this association is still fully open (e.g. Grissom *et al.*, 2015). Moreover, school leaders are characterized by different attitudes and approaches in conducting managerial activities. So, part of the literature on the topic is devoted to explore how much of various managerial actions is actually adopted in day-to-day life of school principals (Bloom *et al.*, 2015, Di Liberto *et al.*, 2015). Moreover, the leadership style is not only influenced by the managerial content of principal's activities, but also by a set of contextual conditions and principals' individual characteristics (i.e. mediator factors, Leithwood & Levin, 2005). The current study addresses these issues, aiming at identifying the existence of different leadership types in a sample of Italian school principals and establishing how the different types relate to student achievement. These objectives are pursued through applying a three-step Latent Class Analysis (LCA), a statistical model that both allows for the identification of subgroups of individuals within data and to relate this finding to a distal outcome measured (e.g. Boyce & Bowers, 2016). More specifically, the research questions addressed are:

- i. To what extent is there one or more than one subgroup (latent classes) of leadership types (subgroups) from national-level surveys of principals in Italy around transformational and instructional leadership?
- ii. Which are the main factors associated with the probability that a principal belongs to a specific subgroup of responders?
- iii. To what extent is a typology of school leadership in Italy across transformational and instructional leadership behaviors related to student achievement on standardized tests?

This research is particularly innovative in the Italian context, where studies about leadership styles and managerial practices at school are still in an early stage (Bloom *et al.*, 2015, Di Liberto *et al.*, 2015). Moreover, the topic is particularly interesting in the policy context, given the approval of a law that empowers the role of school principals starting from 2015/16 school year (law 107/2015).

The paper is organised as follows: paragraph 2 refers to data and methodology and paragraph 3 describes the results obtained. Finally, paragraph 4 discusses and concludes.

## 2. Data and Methods

Data used in the study is provided by the National Evaluation Committee for Education (INVALSI) that yearly assesses the competencies of Italian students in reading and mathematics at given grades. Current data refer to grade 8, last year of lower secondary school, and to the school year 2014/15. The test is taken at national level, but every year a set of schools are randomly chosen throughout the country to be part of the National Sample (NS), where assessment is monitored by external evaluators. In 2014/15 wave, the NS is composed by 28,494 students across 1,405 schools. In addition, from 2013/14 school year, NS principals are also asked to fill in a questionnaire about their schools and the way they manage the organisation. Two sections of the questionnaire have primary importance for the current analysis: the one reporting managerial practices used in the school and that containing principals' characteristics. The kind of questions posed to principals are in line with those contained in OECD TALIS 2008 and 2013 (OECD, 2010, 2014). This part is composed by two groups of questions: the first concerns the frequency of application of a set of instructional leadership practices, with a total number of 12 sub-questions posed and a four categories Likert scale as response type. The second question proposes a list of statements (11 items in total) about the leadership role of the principal in the school. The output of the 23 items has been dichotomized to better fit the LCA purpose. Descriptive statistics about the cited items are listed in Table 1. Merging the principal's questionnaire with the students' results, the final sample size is 1,073 schools.

The approach used in the study is a Latent Class Analysis (LCA), a statistical method from mixture modelling that enables to verify the existence of different subgroups within data (Muthén & Muthén, 2000, Muthén, 2004). The current model has been run using Mplus version 7.4.

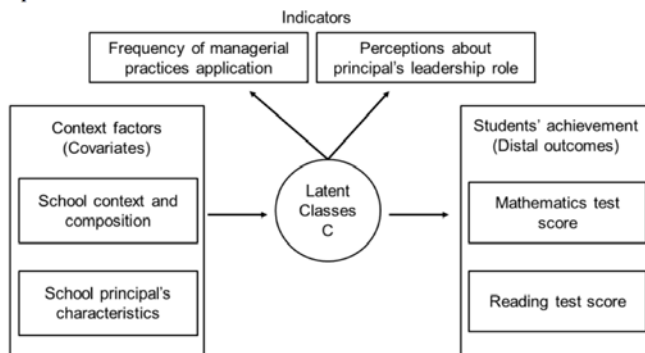
Figure 1 reports the model employed. Indicators are the only factors taken into account when investigating the existence of different subgroups across data (step one of the analysis); in the current analysis, they consist of the two questions posed to principals about the leadership practices. Looking at descriptive statistics in Table 1, it can be noticed that the question concerning the role of the school leader (second half of the table) tends to report a higher level of polarization towards what is considered the *positive* answer (high level of agreement). In order to deal with the trade-off between the number of indicators and the variability across answers, we only delete the three items with a polarization of answers by 99%-1%, obtaining a final sample of 20 indicators. Covariates are then used to characterise the individuals belonging to each group. Being added at step two, none of these factors are able to influence the groups definition (which takes place at step one). On the contrary, it helps to explain group differences, stating how much more/less likely the individuals belonging to a group are to report a specific characteristic (step two of the analysis). Finally, distal outcomes are used as the outputs of the model and defined as factors possibly affected by the belonging of an individual to a class. In other terms, the aim is to identify if across groups there is any statistical difference in the outcome measured (step three of the analysis).

**Table 1.** Descriptive statistics of questions about instructional leadership.

<b>Frequency of use of managerial practices</b>	<b>Seldom</b>	<b>Often</b>
1. I make sure that teachers' professional development activities are in line with the school's educational objectives.	16%	84%
2. I make sure that teachers work in conformity with school educational objectives.	6%	94%
3. I observe educational activities in the classrooms.	46%	54%
4. I provide teachers with suggestions for improving their teaching effectiveness.	51%	49%
5. I supervise students' works	73%	27%
6. When a teacher has a problem in the classroom, I take the initiative to discuss with him/her about it.	7%	93%
7. I inform teachers on opportunities of disciplinary and educational update.	3%	97%
8. I encourage work which is goal oriented and/or based on the Formative Offer Plan	7%	93%
9. I take into account test scores when I make decisions on the school curriculum.	26%	74%
10. I make sure that responsibilities on the coordination of the school curriculum are clearly defined.	14%	86%
11. I deal with bothering behaviors in the classes.	18%	82%
12. I substitute teachers unexpectedly absent.	77%	23%
<b>Opinions about their leadership role</b>	<b>Disagree</b>	<b>Agree</b>
13. In my job, it is important to make sure that educational strategies, approved by the Ministry, are explained to new teachers and applied by more experienced teachers.	6%	94%
14. The use of students' test scores in order to evaluate the teacher's performance reduces the value of his/her professional judgment.	58%	42%
15. Giving teachers a high degree of freedom in choosing the educational techniques can reduce teaching effectiveness.	75%	25%
16. In my job, It is important to make sure that teachers' skills are improving continuously.	2%	98%
17. In my job, It is important to make sure that teachers feel responsible for the achievement of school objectives.	1%	99%
18. In my job, It is important to be convincing when presenting new projects to parents.	13%	87%
19. It is important for the school to verify that rules are respected by everybody.	1%	99%
20. It is important for the school to avoid mistakes in administrative procedures.	5%	96%
21. In my job, It is important to solve timetable problems and/or lesson scheduling problems.	36%	64%
22. It is important that I contribute to maintain a good school climate.	1%	99%
23. I have no possibility to know whether teachers are well performing their teaching tasks or not.	94%	6%

**Figure 1.** Statistical and Conceptual Model of the Latent Class Analysis (LCA) of Principal Leadership Styles.





### 3. Results

#### 3.1. Baseline results: groups of leadership types, individual characteristics and difference in student achievement

Applying the analysis to the INVALSI data about 1,073 school principals across Italy, we identify three different groups of leadership styles. Table 2 reports the main fit statistic tests leading to this finding. The BIC and the LMR test jointly agree about the number of classes to be considered: the BIC starts to increase in correspondence to a number of classes equal to four (from 16,925.6 to 16,988.7), whilst the LMR test is no longer significant with the four classes model ( $p$ -value=0.1387) (Lo *et al.*, 2001, Muthen & Asparouhov, 2006). In addition to that, entropy of the model keeps on level of 0.707 up to 1 and the Akaike Information Criterion (AIC) is equal to 16,616.9. To test for the existence of a local minimum in the best number of classes, we reiterate the model with an additional number of groups. Results from both BIC and LMR test keep on confirming that three is the best number of classes into grouping school principals.

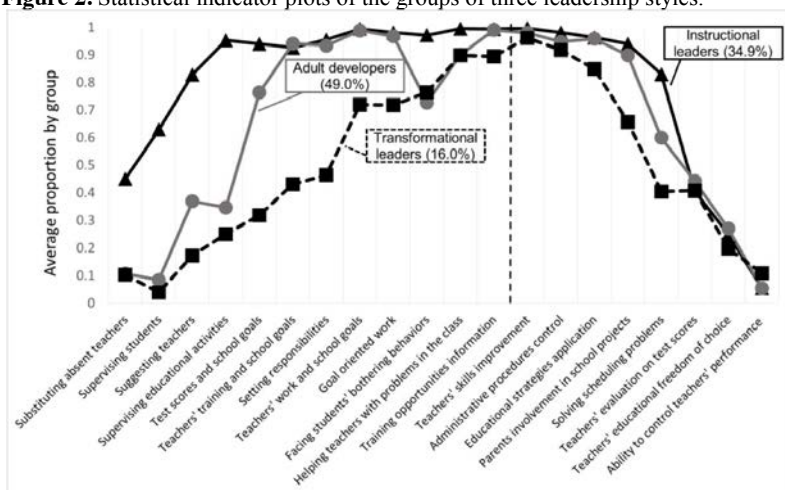
**Table 2.** LCA results and fit statistics.

Classes	AIC	BIC	-Log likelihood	LMR test	p	Entropy
2	16,851.2	17,055.3	- 8,875.0	974.1	0.0000	0.677
3	16,616.9	16,925.6	- 8,246,6	276.3	0.0000	0.707
4	16,575.5	16,988.7	- 8,204.7	82.9	0.1387	0.731
5	16,552.2	17,069.9	- 8,204.7	64.8	0.4366	0.688

Note: AIC=Akaike information criterion; BIC=Bayesian information criterion; LMR=Lo-Mendell-Rubin likelihood ratio test.

Figure 2 reports the average proportion by group of school principals, according to the 20 indicators (reported on the X axis). The vertical dotted line separates the indicators concerning the frequency of application of leadership and managerial practices (on the left) from the indicators about the leadership role (on the right). We named the three groups identified as *Adult developers* (49% of the total), *Instructional leaders* (35% of the total) and *Transformational leaders* (16% of the total). *Adult developers* represents nearly half of the total sample, and show a particularly high focus on supporting teachers' development and training and a lower level of active intervention in classroom activities (Drago-Severson, 2009). In fact, they demonstrate low levels of presence in the classroom, such as substituting teachers, supervising educational activities or facing annoying behaviours among students, so that they can be considered particularly inclined towards adult leadership. *Instructional leaders* represent one third of the total sample (35%) and report an averagely high level of application of the practices reported, which all concern instructional leadership. They are able to cover all the aspects of educational practices happening within the school, with the possible risk of posing their role too close to an operational one. Finally, *transformational leaders* (16%) are so labelled given the high level of orientation towards training opportunity information and the importance of making teachers' skills improve continuously. These are two pillars of transformational leadership in terms of the ability to increase teachers' engagement, skills and ability (Leitwood & Jantzi, 2000, Marks & Printy, 2003, Robinson *et al.*, 2008).

**Figure 2.** Statistical indicator plots of the groups of three leadership styles.



Note: indicators are reported on the X axis. The vertical dotted line divides questions about the frequency of use of leadership practices (on the left), from opinions about the principal's leadership role (on the right). Adult developers, N=526; instructional leaders, N=375; transformational leaders, N=172.

Table 3 reports the results about the covariates used to characterise the groups. Adult developers is kept as the reference group as it is the largest one, so odds ratio are reported for significant measures with reference to it. In details, instructional leaders are 1.92 times more likely to be head of schools located in Central Italy (p-value<0.05), and 4.55 times more likely to be in Southern Italy (p-value<0.01). Moreover, instructional leaders are much more likely (29 times, p-value<0.01) to manage private schools, though it should be noticed the extremely small number of these kind of schools (which represent the 3% of the overall sample, so it can be that estimates are imprecise). With reference to the individual characteristics of school principals, instructional leaders are more likely to be women and older than adult developers (p-value<0.01), who in turn are more likely to have these characteristics than transformational leaders (p-value<0.01). Moreover, instructional leaders are less likely to have a contract of regency (with which principals are in charge of more than one school) than adult developers (p-value<0.05). Somehow, the fact that they do not have to manage different schools can give them higher possibilities to actively intervene in classroom activities. In turn, transformational leaders are more likely to manage more than one school with respect to adult developers (p-value<0.05). Finally, instructional leaders are also less likely to be appointed from less than two years, a span of time that suggests if the principal was already managing the school when the cohort of students analysed (who attends grade 8) entered the lower secondary school, two years before.

Finally, Table 4 reports the school average test score in mathematics and reading per group, the distal outcome employed in the analysis. Results show that students in schools run by instructional leaders report a significantly different and lower average score than students' results in the other two groups and in both the subjects tested. The average school score tends to be higher in reading than in mathematics, with an overall mean respectively of 61.3 and 54.5. Though, instructional leaders report an average school score of 59.6 in reading and 52.8 in mathematics.

**Table 3.** Means and Odd Ratios for Covariates.

	Adult developers		Instructional leaders		Transformational leaders	
	Mean	Odds ratio	Mean	Odds ratio	Mean	Odds ratio
<b>School principals' characteristics</b>						
Average SES index	0.016	-	-0.003		0.013	
Context: school in Central Regions	0.19	-	0.21	1.92**	0.20	
Context: school in Southern Regions	0.33	-	0.54	4.55***	0.25	
Private school	0.01	-	0.07	29.96***	0.01	
Age (years)	55	-	57	1.07***	54	0.96*
Gender (female SP = 1)	0.66	-	0.74	2.27***	0.51	0.44***
Education (PhD = 1)	0.04	-	0.02		0.03	
Experience as SP (years)	9.0	-	10.2		9.2	
Temporary contract	0.03	-	0.04		0.02	
Contract of regency	0.08	-	0.04	0.38**	0.12	1.97*
Appointed in the school from less than 2 years	0.41	-	0.31	0.59**	0.38	

Note: Significance tests are logistic regressions. \*p≤.10. \*\* p≤.05. \*\*\* p≤.01.

**Table 4.** Means and p-values for distal outcomes (grade 8).

School principals' characteristics	Adult developers (1)	Instructional leaders (2)	Transformational leaders (3)	p-value 1 vs 2	p-value 2 vs 3	p-value 1 vs 3
	Mean	Mean	Mean			
School average mathematics test score - grade 8	55.52	<b>52.81</b>	54.86	<b>0.007</b>	<b>0.056</b>	0.508
School average reading test score - grade 8	62.14	<b>59.61</b>	62.60	<b>0.020</b>	<b>0.005</b>	0.612

Note: Significance tests are Pearson chi-square.

#### 4. Discussion and concluding remarks

This study aims at investigating the existence of various leadership types across Italian lower secondary schools. Moreover, each subgroup of school leaders is characterized according to individual and contextual characteristics. Finally, the statistical difference across groups in student achievement is investigated. Applying a Latent Class Analysis (LCA), we define three subgroups of school leaders, namely adult leaders (49%), instructional leaders (35%) and transformational leaders (16%). Groups differ in terms of principals' individual characteristics (age, gender, type of contract) and institutional/contextual factors (public/private ownership and geographical location). Finally, we observe a statistically significant difference across groups in student achievement, with instructional leaders running schools with lower test scores. In interpreting these results, we are cautious about the direction of causality (if any) between school principals leadership styles and school average test scores. In terms of policy implications, results suggest a direction for the evaluation of school principals. Indicators used in this process should focus on stimulating those activities which, in turn, show a higher probability to be associated with better school academic results – for instance, teachers' training and development. On the other hand, principals should be less involved in operational activities that could affect their effectiveness in leading the whole organisation. Future directions of research should aim at finding patterns of leadership types within specific approaches to managerial practices (in terms of areas of management, see Bloom *et al.*, 2015, Di Liberto *et al.*, 2015). This would allow to better investigate the relationship between leadership styles and managerial practices implemented. Moreover, it would be interesting to have additional years of data, in order to analyse whether the effects of leadership are stable over time, in the light of the recent policy changes.

## References

1. Bloom, N., Lemos, R., Sadun, R., & Van Reenen, J. (2015). Does management matter in schools? *The Economic Journal*, 125 (584): 647-674.
2. Boyce, J. & Bowers, A.J. (2016) Principal Turnover: Are there Different Types of Principals Who Move From or Leave Their Schools? A Latent Class Analysis of the 2007-08 Schools and Staffing Survey and the 2008-09 Principal Follow-up Survey. *Leadership and Policy in Schools*, 15(3), 237-272.
3. Day, C., Gu, Q., & Sammons, P. (2016). The Impact of Leadership on Student Outcomes: How Successful School Leaders Use Transformational and Instructional Strategies to Make a Difference. *Educational Administration Quarterly*, 52(2), 221-258.
4. Day, C., Sammons, P., Hopkins, D., Harris, A., Leithwood, K., Gu, Q., Brown, E., Ahtaridou, E., Kington, A. (2009). The impact of school leadership on pupil outcomes - final report. Research Report RR108, Department for Children, Schools and Families.
5. Di Liberto, A., Schivardi, F., & Sulis, G. (2015). Managerial practices and student performance. *Economic Policy*, 30 (84), 683-728.
6. Drago-Severson, E. (2009). Leading adult learning: Supporting adult development in our schools. Corwin Press.
7. Grissom, J., A., Kalogrides, D., & Loeb, S. (2015). Using student test scores to measure principal performance. *Educational evaluation and policy analysis*, 20 (10), 1-26.
8. Leithwood, K., Harris, A., & Hopkins, D. (2008). Seven strong claims about successful school leadership. *School leadership and management*, 28(1), 27-42.
9. Leithwood, K., & Levin, B. (2005). Assessing school leader and Leadership Programme effects on pupil learning: Conceptual and methodological challenges. Research Report RR662. London: Department for Education and Skills.
10. Leithwood, K., & Jantzi, D. (2000). The effects of transformational leadership on organizational conditions and student engagement with school. *Journal of Educational Administration*, 38(2), 112-129.
11. Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3), 767-778.
12. Marks, H. M., & Printy, S. M. (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. *Educational administration quarterly*, 39(3), 370-397.
13. Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 345-370). Thousand Oaks, CA: Sage Publications.
14. Muthen, B., & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive behaviors*, 31(6), 1050-1066.
15. Muthén, B. O., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical & Experimental Research*, 24(6), 882-891.
16. OECD. (2010). TALIS 2008 technical report. OECD Publishing.
17. OECD. (2014). TALIS 2013 Results: An International Perspective on Teaching and Learning. OECD Publishing.
18. Robinson, V., M., J., Lloyd, C., A., & Rowe, K., J. (2008) The impact of leadership on student outcomes: an analysis of the differential effect of leadership types. *Educational administration quarterly*, 44 (5), 635-674.
19. Smith, W. F., & Andrews, R. L. (1989). *Instructional Leadership: How Principals Make a Difference*. Publications, Association for Supervision and Curriculum Development, 125 N. West Street, Alexandria, VA 22314.
20. Urick, A., & Bowers, A. J. (2014). How does Principal Perception of Academic Climate Measure Up? The Impact of Principal Perceptions on Student Academic Climate and Achievement in High School. *Journal of School Leadership*, 24(2), 386-414.
21. Waters, T., Marzano, R., J., & McNulty, B. (2003). *Balanced leadership: what 30 years of research tells us about the effect of leadership on student achievement*. Aurora, CO: Mid-continent Research for Education and Learning.



# Poverty measures to analyse the educational inequality in the OECD Countries

## *Misure di povertà per l'analisi delle diseguaglianze educative nei paesi OECD*

Tommaso Agasisti, Sergio Longobardi and Felice Russo

**Abstract** This paper studies the degree of educational poverty in OECD countries on the basis of last edition (2015) of OECD Programme for International Student Assessment (PISA). The definition of ‘poor in education’, in terms of PISA data, refers to the students below the baseline level of proficiency that is required to participate fully in society. We adopt both one-dimensional and multidimensional approach to measure poverty in education. In this light, the educational poverty is analysed by the poverty metrics developed by Foster, Greer and Thorbecke and those proposed by Alkire and Foster. The main results of our analysis provide a detailed picture of the degree of poverty relative to student learning in OECD countries, and they can be considered an analytical tool to improve the quality of educational systems.

**Sommario** *Il lavoro analizza il grado di povertà educativa nei paesi OECD sulla base dell'ultima edizione (2015) del Programme for International Student Assessment (PISA) dell'OECD. La definizione di povero in ambito educativo fa riferimento, in termini di dati PISA, agli studenti al di sotto della soglia di rendimento richiesta per avere una partecipazione attiva nella società. Per misurare la povertà educativa viene adottato sia un approccio unidimensionale che multidimensionale. In questa ottica, si ricorre sia agli indicatori sviluppati da Foster, Greer e Thorbecke sia quelli proposti da Alkire e Foster. I principali risultati offrono un quadro dettagliato del livello di povertà educativa nei Paesi OECD e costituiscono uno strumento analitico per migliorare il livello qualitativo dei sistemi educativi.*

**Key words:** educational poverty, student’s learning, standardized tests

<sup>1</sup>

Tommaso Agasisti, Politecnico di Milano School of Management; agasisti@polimi.it  
Sergio Longobardi, University of Naples “Parthenope”; longobardi@uniparthenope.it  
Felice Russo, University of Salento (Lecce); felice.russo@unisalento.it

## 1. Introduction

The targeting of educational poverty eradication/alleviation is largely considered a very relevant topic and recently it has captured the attention of governments and international institutions. The main reason of this attention is that poor performance at school has negative impact on the future educational and socio-economic attainment of students (Erickson *et al.*, 2005) and long-term consequences for society as a whole (OECD, 2016).

In this light, the main insight of this paper is to base on the well-developed techniques applied to economics studies about poverty in order to adapt it to some features of the educational data. The main question that arises here is how to quantify the extent of poverty. We guess that a poverty analysis based only on a single education program cannot provide an exhaustive description of the whole learning deprivation matter. As a consequence, our study will not be limited to a single attribute-based approach and we propose both one-dimensional and multidimensional analysis. We use the data on students performance in the three main domains investigated by the OECD PISA (Programme for International Student Assessment): reading, mathematics and science. These scores have statistical properties similar to income data e.g they are both classified as individual and continuous observations. Moreover, these two variables are important predictors of individual and collective well-being. The one-dimensional analysis of poverty in education is performed by estimating three educational deprivation indices developed by Foster, Greer and Thorbecke (1984). The first index (educational deprivation headcount) is the proportion of the student population for whom learning is below the educational poverty line. The second (educational deprivation gap index) considers the student's gap from the educational deprivation line. The third index (educational deprivation severity index) attributes greater weight to the very poor rather than the less poor, taking into account also the inequality among the poor.

Focusing on the multidimensional aspects of educational poverty, we provide an application of additive multidimensional poverty index proposed by Alkire and Foster (2011). The rest of the study is structured as follows. Section 2 presents the OECD data and the methodology. It proposes different tools in order to describe the extent and the changes in poverty. Section 3 discusses the empirical evidence.

## 2. Data and methodology

The analysis of educational poverty and deprivation draws upon the OECD PISA data. The aim of the PISA is to collect highly standardised data that can be used to compare the competencies of representative samples of 15-year-old students in the three main domains of reading, mathematics and science, both within and between countries. Since the first cycle in 2000, PISA has been taking place every 3 years



with a growing number of participating countries and each of these cycles looks in depth at a major domain.

We exploit the last results of PISA 2006 and 2015 editions for all 35 OECD countries and compute some poverty metrics focusing on the test scores in science, as it was the major domain in both PISA rounds.

From a methodological point of view, the analysis of poverty consists of two steps (Sen, 1976): first, the identification of the poor by defining a threshold (poverty line); second, the aggregation of the poor. According to OECD (2010), the threshold is given by a proficiency score corresponding to the lowest limit of level 2 in an ordered scale that goes from 1 (lowest skilled students) to 6 (highly-skilled students) proficiency levels<sup>1</sup>.

In a one-dimensional context, at a given point of time, a poverty statistic  $P$  is a function of the value of learning distribution  $X$  and poverty line  $Z$ . The poverty indexes of Foster-Greer-Thorbecke (FGT) are the most widely used poverty measures (see Foster *et al.*, 1984), it can be defined as:

$$P_{\alpha}(X; Z) = \frac{1}{N} \sum_{1 \leq i \leq q} \left( \frac{Z - x_i}{Z} \right)^{\alpha} \quad (1)$$

where the parameter  $\alpha$  is a non-negative parameter,  $N$  is the total population,  $q$  is the number of units with learning less than  $Z$  and  $x_i$  is the result of standardized test of the unit of observation  $i$ , for  $i=1, 2, \dots, N$ .

According to  $\alpha=0$ , we obtain the educational deprivation headcount. For  $\alpha=1$ , we get the educational deprivation gap index. Finally for  $\alpha=2$ , we obtain the educational severity gap index. Using all three measures gives a fuller view of poverty, reflecting different aspects -incidence, depth, and severity, respectively- of educational poverty. The magnitude and the direction of their changes might differ.

In a multidimensional analysis with, at least, two dimensions a deprived student in both attributes should be considered as poor without ambiguity. However, differently with respect to the one-dimensional case, how should be defined a student deprived in only one learning attribute? The literature suggests two extreme approaches in accordance with the distinction between the «intersection» method and the «union method»: the former approach identifies the observation  $i$  as poor if he/she is poor in both attributes, while in the latter the student is considered poor if his/her learning outcome is below the poverty cut-off in at least one attribute. In this study, we'll present results according to the latter method. In a multidimensional context, Alkire and Foster (2011) advocate a second cut-off,  $k$ , according to the number of dimensions in which the individual has to be deprived in order to be considered globally poor. Indicating with  $c_i$  the number of educational deprivations suffered by observation  $i$ , he/she should be judged educationally globally poor if  $c_i \geq k$ . With this dual cut-off, Alkire and Foster propose the following  $M_{\alpha}$  class of measures:

---

<sup>1</sup> Thus, in what follows we use *absolute* poverty thresholds. The proficiency level 2 is the baseline level of proficiency that is required to participate fully in society. The lowest limit of level 2 corresponds to 407 point for Reading, 420 for Mathematics and 410 for Science.

$$M_\alpha(X; Z) = \frac{1}{Nd} \sum_{i=1}^N \sum_{j=1}^d w_j (g_{ij}(K))^\alpha \quad (2)$$

where  $g_{ij} = \left( \frac{Z_i x_{ij}}{Z_j} \right)$  for the student  $i$  in learning dimension  $j$  and  $g_{ij}=0$  if  $x_{ij} \geq Z_j$ .

The weight assigned to dimension  $j$  is  $w_j$ , such that  $\sum_{j=1}^d w_j = d$ .

The parameter  $\alpha$  is still a non-negative poverty aversion parameter, specific for each dimension. The attributes, which are considered here, are independent - they are neither substitutes nor compliments - and the aggregation procedure is not sensitive to the kind of interrelationship between educational deprivation dimensions. Once the identification step has been accomplished, the AF index can be decomposed into the contribution of each attribute.

### 3. Main results

Firstly, we provide results about FGT educational poverty measures in table 1 by using PISA scores in science. By looking at the proportion of students poorest in education, we notice that there is reduction of the incidence of poverty for 13 countries. That is true in particular for countries in the bottom of the educational poverty outcomes in 2006 (e.g. Mexico, Turkey, Italy, USA, Portugal). For the same countries, this trend is confirmed also for  $\alpha=1$  and  $\alpha=2$ . More interestingly, the extent of poverty declines more rapidly when it is measured by the educational deprivation gap index rather than the educational deprivation headcount, this is a signal that the benefits of poverty reduction accruing to the less poor are lower than those to the very poor. Moreover, the same conclusion is true, with the exception of Poland, when we compare the outcomes of educational deprivation gap index with those of educational deprivation severity index: the value of the latter poverty measure indicates a clearer decline in poverty. Now, we would like to bring the attention of the reader to the fact that, for seven countries, the estimates of the variation have contradictory signs depending on the value of the parameter  $\alpha$ . In contrast with the increase in the incidence and depth of educational poverty, actually the severity of educational poverty decreases in Canada, Chile, Germany, France and Luxembourg. Great Britain and Ireland follow the same path: only with  $\alpha=0$  the educational poverty in science rises between 2006 and 2015. In those cases, the distance between poorly performing students and educational poverty line narrowed over time, while educational deprivation headcount outcomes show an increase from 2006 to 2015. In those countries evidently, the choice of poverty measure does matter. The direction of change is fully reversed when considering the other countries that complete our PISA OECD database: proportionate changes in all our poverty measures are always positive over time.

In particular, we observe a sharp rise in educational poverty in Finland, Hungary, Nederland, Sweden and Slovakia. For those countries, all the values of deprivation

**Table 1:** Index of Foster-Greer-Thorbecke (FGT) for different values of  $\alpha$ , PISA scores in Science.

cnt	Educational deprivation headcount ( $\alpha=0$ )			Educational deprivation gap ( $\alpha=1$ )			Educational deprivation severity Index ( $\alpha=2$ )		
	2006	2015	var.%	2006	2015	var.%	2006	2015	var.%
	AUS	0.127	0.175	38.14%	0.016	0.023	42.00%	0.004	0.005
AUT	0.164	0.209	27.43%	0.020	0.026	27.73%	0.004	0.005	19.21%
BEL	0.171	0.200	16.41%	0.023	0.026	13.17%	0.005	0.005	0.73%
CAN	0.099	0.113	14.90%	0.011	0.012	8.90%	0.002	0.002	-0.45%
CHE	0.159	0.188	18.21%	0.021	0.023	6.55%	0.005	0.005	-7.65%
CHL	0.391	0.346	-11.58%	0.058	0.046	-20.72%	0.014	0.010	-29.07%
CZE	0.155	0.205	32.46%	0.019	0.024	30.63%	0.004	0.004	15.84%
DEU	0.160	0.169	5.94%	0.020	0.021	4.09%	0.004	0.004	-2.11%
DNK	0.180	0.158	-12.07%	0.021	0.019	-12.94%	0.004	0.004	-20.09%
ESP	0.197	0.185	-5.94%	0.024	0.021	-11.86%	0.005	0.004	-22.63%
EST	0.076	0.081	6.22%	0.007	0.008	20.72%	0.001	0.001	31.00%
FIN	0.039	0.116	198.79%	0.004	0.014	267.92%	0.001	0.003	341.67%
FRA	0.216	0.217	0.29%	0.031	0.031	0.10%	0.007	0.007	-4.32%
GBR	0.170	0.176	3.09%	0.024	0.020	-15.57%	0.006	0.004	-33.74%
GRC	0.245	0.333	35.98%	0.035	0.048	35.19%	0.009	0.011	22.89%
HUN	0.151	0.257	69.91%	0.016	0.036	120.73%	0.003	0.008	148.86%
IRL	0.156	0.157	0.80%	0.019	0.017	-10.74%	0.004	0.003	-22.70%
ISL	0.206	0.258	24.86%	0.028	0.033	17.98%	0.006	0.007	6.79%
ISR	0.359	0.310	-13.59%	0.063	0.050	-21.72%	0.017	0.012	-28.81%
ITA	0.254	0.233	-8.28%	0.035	0.030	-13.27%	0.008	0.006	-22.21%
JPN	0.120	0.095	-20.47%	0.016	0.011	-33.52%	0.004	0.002	-46.67%
KOR	0.113	0.141	25.03%	0.013	0.017	29.55%	0.003	0.004	34.46%
LUX	0.220	0.258	17.63%	0.031	0.033	6.52%	0.007	0.006	-8.09%
LVA	0.175	0.171	-2.25%	0.020	0.017	-13.06%	0.004	0.003	-28.53%
MEX	0.511	0.469	-8.28%	0.079	0.062	-21.11%	0.019	0.013	-31.85%
NLD	0.127	0.184	44.78%	0.014	0.022	57.62%	0.002	0.004	73.66%
NOR	0.218	0.186	-14.96%	0.029	0.023	-20.94%	0.007	0.005	-31.26%
NZL	0.134	0.175	30.49%	0.019	0.022	17.64%	0.004	0.004	6.18%
POL	0.173	0.163	-5.75%	0.019	0.017	-8.59%	0.003	0.003	-7.78%
PRT	0.243	0.179	-26.35%	0.030	0.020	-33.66%	0.006	0.004	-39.84%
SVK	0.203	0.305	50.15%	0.026	0.047	80.39%	0.006	0.011	98.43%
SVN	0.141	0.151	7.50%	0.015	0.017	11.02%	0.003	0.003	12.06%
SWE	0.163	0.216	32.42%	0.020	0.031	53.17%	0.004	0.007	64.88%
TUR	0.463	0.447	-3.50%	0.063	0.060	-4.91%	0.013	0.012	-6.86%
USA	0.247	0.198	-20.01%	0.035	0.025	-28.98%	0.008	0.005	-37.74%

Turning to multidimensional analysis, table 2 presents our findings for the AF educational deprivation index when  $\alpha=0$ . This measure both synthesizes the overall learning deprivation and can be decomposed into the contribution of each attribute. We name this statistic *adjusted educational deprivation headcount*, i.e. the total number of dimensions that the multidimensionally poor population experience over  $N \times d$ , the maximum total number of dimensions in which the student population can be deprived. For all learning dimensions (reading, mathematics and science),  $w_j = 1$  for any  $j$ . It emerges that mathematics is largely the learning dimension where educational poverty is higher, while the reading is the attribute that contributes more to learning deprivation only for five countries. The weight of attributes on the overall educational poverty is quite similar for nine countries. Only in four countries, the contribution of the learning attribute is higher than 40%, three times for mathematics, one for reading.

**Table 2:** Multidimensional index of Alkire-Foster, PISA 2015 scores.

Country	Alkire Foster Index		AF Index Decomposition		
	Estimate	Std. err.	Reading	Mathematics	Science
AUS	0.194	0.003	31.57	38.08	30.35
AUT	0.217	0.005	34.05	33.82	32.13
BEL	0.197	0.004	32.83	33.24	33.93
CAN	0.119	0.003	29.12	38.94	31.94
CHE	0.182	0.005	36.97	28.56	34.47
CHL	0.375	0.006	25.05	43.98	30.97
CZE	0.215	0.005	34.3	33.67	32.03
DEU	0.165	0.004	31.47	34.19	34.34
DNK	0.147	0.004	34.68	29.32	36
ESP	0.19	0.004	27.61	39.68	32.71
EST	0.104	0.004	35.99	37.39	26.62
FIN	0.12	0.004	29.2	38.49	32.31
FRA	0.221	0.005	31.99	35.05	32.96
GBR	0.196	0.004	31.21	38.83	29.96
GRC	0.322	0.006	28.35	37.06	34.59
HUN	0.272	0.006	33.97	34.35	31.68
IRL	0.135	0.004	24.16	36.97	38.87
ISL	0.242	0.006	30.87	33.32	35.81
ISR	0.3	0.005	29.42	35.89	34.69
ITA	0.225	0.005	30.03	35.3	34.67
JPN	0.112	0.003	40.01	31.32	28.67
KOR	0.146	0.004	32.41	35.18	32.41
LUX	0.257	0.005	33.48	32.84	33.68
LVA	0.186	0.005	30.56	38.78	30.66
MEX	0.486	0.006	28.79	38.89	32.32
NLD	0.176	0.005	33.79	31.26	34.95
NOR	0.167	0.004	28.58	34.15	37.27
NZL	0.189	0.005	30.65	38.1	31.25
POL	0.162	0.005	30.62	35.41	33.97
PRT	0.198	0.005	28.99	40.56	30.45
SVK	0.304	0.005	35.26	31.15	33.59
SVN	0.154	0.004	32.68	34.55	32.77
SWE	0.2	0.005	30.37	33.45	36.18
TUR	0.453	0.006	29.77	37.21	33.02
USA	0.227	0.005	28.45	42.43	29.12

## References

- Alkire, S. and Foster, J. (2011), Counting and multidimensional poverty measurement, *Journal of Public Economics*, 95(7-8), 476-487.
- Erikson, R., Goldthorpe, J.H., Jackson, M. and Cox, D.R. (2005), On class differentials in educational attainment, *Proceedings of the National Academy of Sciences of the United States of America*, 102(27), 9730-9733.
- Foster, J.E., Greer, J. and Thorbecke, E. (1984), A class of decomposable poverty measures, *Econometrica*, 52(3), 761-766.
- OECD (2016), *Low-Performing Students: Why they fall behind and how to help them succeed*, OECD Publishing.
- OECD (2010), *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science* (Volume 1), OECD Publishing.
- Sen, A.K. (1976), Poverty: An ordinal approach to measurement, *Econometrica*, 44(2), 219-231.

# Quasi-Maximum Likelihood Estimators For Functional Spatial Autoregressive Models

## *Quasi-verosimiglianza stima per funzionale spaziale autoregressivo modello*

Mohamed-Salem Ahmed, Laurence Broze, Sophie Dabo-Niang, Zied Gharbi

**Abstract** We propose a functional linear autoregressive spatial model where the explanatory variable takes values in a function space while the response process is real-valued and spatially autocorrelated. The specificity of the model is the functional nature of the explanatory variable and the structure of a spatial weight matrix which defines the spatial relation and dependency between neighbors. The estimation procedure consists in reducing the infinite dimension of the functional explanatory variable and maximizing a quasi-maximum likelihood. We establish both consistency and asymptotic normality of the regression parameter function estimate. We illustrate the skills of the methods by some numerical results.

**Abstract** *In questo lavoro si propone un modello lineare spaziale autoregressivo in cui la variabile esplicativa prende valori in uno spazio di funzioni e la variabile risposta a valori reali spazialmente correlati. La specificità del modello risiede nella natura funzionale della variabile esplicativa e nella struttura della matrice di prossimità i cui elementi definiscono la relazione spaziale e la dipendenza tra i vicini. La procedura di stima consiste nel ridurre la dimensione infinita della variabile esplicativa funzionale e nel massimizzare una quasi-verosimiglianza. Vengono stabiliti consistenza e normalità asintotica dello stimatore funzionale del parametro di regressione, la cui performance viene illustrata attraverso uno studio di simulazione.*

---

Mohamed-Salem Ahmed

University of Lille, Villeneuve d'ascq, France, LEM-CNRS 9221. e-mail: mohamed-salem.ahmed@univ-lille3.fr

Laurence Broze

University of Lille, Villeneuve d'ascq, France, RimeLab. e-mail: laurence.broze@univ-lille3.fr

Sophie Dabo-Niang

University of Lille, Villeneuve d'ascq, France, LEM-CNRS 9221 and INRIA-MODAL. e-mail: sophie.dabo@univ-lille3.fr

Zied Gharbi

University of Lille, Villeneuve d'ascq, France, LEM-CNRS 9221, and University of Tunis. e-mail: zied.gharbi@etu.univ-lille3.fr

**Key words:** Functional Linear Models, Spatial Autoregressive process, Quasi-Maximum Likelihood Estimators, . . .

## 1 Introduction

This work concerns two different research areas; spatial econometric and functional data analysis. Functional random variables are spreading in statistical analyses due to the availability of high frequency data and of new mathematical strategies to deal with such statistical objects. The field is known as Functional Data Analysis (FDA). Applications of FDA are growing across fields. The functional variables are mainly curves, surfaces or manifolds. For an introduction to this field as well as illustrations and applications, see [21].

In many fields as urban system, agricultural, environmental sciences or economic and many others, one often deals with spatially dependent data. Therefore, modelling spatial dependency in statistical inferences (estimation of spatial distribution, regression, prediction, . . .) is a significant feature of spatial data analysis. Spatial statistic provides tools to solve such problem. Various spatial models and methods have been proposed, particularly within the scope of geostatistics. So far, most of spatial modelling methods are parametric and concern non-functional data. Several types of functional linear models for independent data have been developed over the years, serving different purposes. The most studied is perhaps the functional linear model for scalar response, originally introduced by [14]. Estimation and prediction problems for this model and some of its generalizations have been tackled mainly for independent data (see, e.g., [7], [17], [6], [9]). Some works exist on functional spatial linear prediction using kriging methods (see, e.g., [18], [11], [12], [15] [10]), so highlighting the interest of considering spatial linear functional models. We are interested in a functional spatially autoregressive linear model. One of the well known spatial model is the Spatial Autoregressive Model (SAR) by [5] that extends regression in time series to spatial data. The structure of this model and its estimation has been developed and summarized by many authors as [1], [8], [4] among others. More recently, [16] proposed the Quasi-likelihood estimator for the SAR model for real-valued data and investigated its asymptotic properties under the normal distributional specifications. We extend the previous model to the case where the covariate is a functional random variable. In the following, we provide the functional SAR (FSAR) and its Quasi-likelihood estimator (QML).

## 2 The model

We consider that at  $n$  spatial units, we observe a random real variable  $Y$  considered as *response variable* and a functional covariate  $\{X(t), t \in \mathcal{T}\}$  considered as *explanatory function* corresponding to a square integrable stochastic process on the

interval  $\mathcal{T} \subset \mathbb{R}$ . Assume that the process  $\{X(t), t \in \mathcal{T}\}$  takes values in some space  $\mathcal{X} \subset L^2(\mathcal{T})$ , where  $L^2(\mathcal{T})$  is the space of square integrable functions in  $\mathcal{T}$ . The spatial dependency structure between these  $n$  spatial units is described by a non-stochastic spatial weights  $n \times n$  matrix  $W_n$  that depends on  $n$ . The elements  $W_{ijn}$  of this matrix are usually considered as inversely proportional to distances between spatial units  $i$  and  $j$  with respect to some metric (physical distance, social networks or economic distance, see for instance [20]). Since the weight matrix changes with  $n$ , we consider these observations as triangular arrays observations. This is required to investigate an asymptotic study of the following model that describes the relationship between the response variable  $Y$  and the covariate function  $X(\cdot)$  [22]. We assume that this relationship is modeled by the following Functional Spatial Autoregressive Model (FSAR):

$$Y_i = \lambda_0 \sum_{j=1}^n W_{ijn} Y_j + \int_{\mathcal{T}} X_i(t) \beta^*(t) dt + U_i, \quad i = 1, \dots, n, \quad n = 1, 2, \dots \quad (1)$$

where  $\lambda_0$  (in a compact space  $\Lambda$ ) is the autoregressive parameter,  $\beta^*(\cdot)$  is a parameter function assumed to belong to the space of functions  $L^2(\mathcal{T})$ , and  $(W_{ijn})_{j=1, \dots, n}$  is the  $i$ -th row of  $W_n$ . The disturbances  $\{U_i, i = 1, \dots, n, n = 1, 2, \dots\}$  are assumed to be independent random Gaussian variables such that  $E(U_i) = 0, E(U_i^2) = \sigma_0^2$ . They are also independent to  $\{X_i(t), t \in \mathcal{T}, i = 1, \dots, n, n = 1, 2, \dots\}$ . We are interested in estimating the unknown true parameters  $\lambda_0, \beta^*(\cdot)$  and  $\sigma_0^2$ . Let  $\mathbf{X}_n(\beta^*(\cdot))$  be the  $n \times 1$  vector of  $i$ -th element  $\int_{\mathcal{T}} X_i(t) \beta^*(t) dt$ , then one can rewrite (1) as

$$S_n \mathbf{Y}_n = \mathbf{X}_n(\beta^*(\cdot)) + \mathbf{U}_n, \quad n = 1, 2, \dots \quad (2)$$

with  $S_n = (I_n - \lambda_0 W_n)$ ,  $\mathbf{Y}_n$  and  $\mathbf{U}_n$  are two  $n \times 1$  vectors of elements  $Y_i$  and  $U_i, i = 1, \dots, n$  respectively,  $I_n$  denotes the  $n \times n$  identity matrix.

Let  $S_n(\lambda) = I_n - \lambda W_n$  and  $V_n(\lambda, \beta(\cdot)) = S_n(\lambda) \mathbf{Y}_n - \mathbf{X}_n(\beta(\cdot))$  so the conditional log likelihood function of the vector  $\mathbf{Y}_n$  given  $\{X_i(t), t \in \mathcal{T}, i = 1, \dots, n, n = 1, 2, \dots\}$  is given by :

$$L_n(\lambda, \beta(\cdot), \sigma^2) = -\frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln 2\pi + \ln |S_n(\lambda)| - \frac{1}{2\sigma^2} V_n'(\lambda, \beta(\cdot)) V_n(\lambda, \beta(\cdot)). \quad (3)$$

Maximum likelihood estimates of  $\lambda_0, \beta^*(\cdot)$  and  $\sigma_0^2$  are the  $\lambda, \beta(\cdot)$ , and  $\sigma^2$  that maximise (3). But this likelihood cannot be maximized without addressing the difficulty produced by the infinite dimensionality of the explanatory random function. To deal with this problem, we project as usual, the functional explanatory variable and parameter function in a space of functions generated by a basis of functions with a dimension that increases asymptotically as the sample size tends to infinity. Several truncation techniques exist. [2] proposed to use the estimated eigenbasis of the sample, [3] were limited to a Spline basis adding a penalty that controls the degree of smoothness of the parameter function. [19] proposed to use any basis of functions which verifies some truncation criterion. We shall adapt the alternative proposed by

[19] in order to resolve infinite dimensional problem of the functional space. This method will be denoted *Truncated Conditional Likelihood Method*.

### 3 Truncated Conditional Likelihood Method

Let  $\{\varphi_j, j = 1, 2, \dots\}$  be an orthonormal basis of the functional space  $L^2(\mathcal{T})$ , usually a Fourier or a Spline basis or a basis constructed by the eigenfunctions of the covariance operator  $\Gamma$  where this operator is defined by :

$$\Gamma x(t) = \int_{\mathcal{T}} E(X(t)X(s))x(s)ds, \quad x \in \mathcal{X}, t \in \mathcal{T}. \quad (4)$$

The operator  $\Gamma$  is a linear integral operator whose integral kernel is

$$K(t, v) = E(X(t)X(v)), \quad \text{for all } t, v \in \mathcal{T}. \quad (5)$$

It is a compact self-adjoint Hilbert-Schmidt operator because

$$\int |K(t, v)|^2 dt dv \leq \left( E \left( \int X^2(t) dt \right) \right)^2 < \infty;$$

thus, it can be diagonalized.

We can rewrite  $X(\cdot)$  and  $\beta^*(\cdot)$  in the following way :

$$X(t) = \sum_{j \geq 1} \varepsilon_j \varphi_j(t) \quad \text{and} \quad \beta^*(t) = \sum_{j \geq 1} \beta_j^* \varphi_j(t) \quad \text{for all } t \in \mathcal{T}$$

where the real random variables  $\varepsilon_j$  and the coefficients  $\beta_j^*$  are given by

$$\varepsilon_j = \int_{\mathcal{T}} X(t) \varphi_j(t) dt \quad \text{and} \quad \beta_j^* = \int_{\mathcal{T}} \beta(t)^* \varphi_j(t) dt.$$

Let  $p_n$  be a positive sequence of integers that increases asymptotically as  $n \rightarrow \infty$ , by the orthonormality of the basis  $\{\varphi_j, j = 1, 2, \dots\}$ , we can consider the following decomposition

$$\int_{\mathcal{T}} X(t) \beta^*(t) dt = \sum_{j=1}^{\infty} \beta_j^* \varepsilon_j = \sum_{j=1}^{p_n} \beta_j^* \varepsilon_j + \sum_{j=p_n+1}^{\infty} \beta_j^* \varepsilon_j. \quad (6)$$

The truncation strategy introduced by [19]<sup>1</sup> consists of approximating the left-hand side in (6) by using only the first term of the right-hand side. This is possible when the approximation error vanishes asymptotically, where this error is controlled by

---

<sup>1</sup> Note that our model can be viewed as a particular case of the generalized functional linear models of [19] with the identity link function and  $\lambda_0$  replaced by zero.



a square expectation of the second term in the right-hand side of (6). In particular, the approximation error vanishes asymptotically when one consider the eigenbasis of the variance-covariance operator, by

$$E \left( \sum_{j=p_n+1}^{\infty} \beta_j^* \varepsilon_j \right)^2 = \sum_{j=p_n+1}^{\infty} \beta_j^{*2} E(\varepsilon_j^2) = \sum_{j=p_n+1}^{\infty} \beta_j^{*2} \lambda_j$$

where  $\lambda_j, j = 1, 2, \dots$  are the eigenvalues. Under the truncation strategy,  $\mathbf{X}_n(\beta^*(\cdot))$  will be approximated by  $\xi_{p_n} \beta^*$  where  $\beta^* = (\beta_1^*, \dots, \beta_{p_n}^*)'$  and  $\xi_{p_n}$  is a  $n \times p_n$  matrix of  $(i, j)$ -th element given by

$$\varepsilon_j^{(i)} = \int_{\mathcal{T}} X_i(t) \varphi_j(t) dt, \quad i = 1, \dots, n \quad j = 1, \dots, p_n.$$

Now the truncated Conditional Log Likelihood function can be obtained by replacing in (3)  $\mathbf{X}_n(\beta(\cdot))$  by  $\xi_{p_n} \beta$  for all  $\beta(\cdot) \in L^2(\mathcal{T})$  and  $\beta \in \mathbb{R}^{p_n}$ . The corresponding and feasible Conditional Likelihood is

$$\bar{L}_n(\lambda, \beta, \sigma^2) = -\frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln 2\pi + \ln |S_n(\lambda)| - \frac{1}{2\sigma^2} V_n'(\lambda, \beta) V_n(\lambda, \beta) \quad (7)$$

with  $V_n(\lambda, \beta) = S_n(\lambda) \mathbf{Y}_n - \xi_{p_n} \beta$ . For a fixed  $\lambda$ , (7) is maximized at

$$\hat{\beta}(\lambda) = (\xi_{p_n}' \xi_{p_n})^{-1} \xi_{p_n}' S_n(\lambda) \mathbf{Y}_n \quad (8)$$

and

$$\begin{aligned} \hat{\sigma}^2(\lambda) &= \frac{1}{n} \left( S_n(\lambda) \mathbf{Y}_n - \xi_{p_n} \hat{\beta}(\lambda) \right)' \left( S_n(\lambda) \mathbf{Y}_n - \xi_{p_n} \hat{\beta}(\lambda) \right) \\ &= \frac{1}{n} \mathbf{Y}_n' S_n'(\lambda) M_n S_n(\lambda) \mathbf{Y}_n \end{aligned} \quad (9)$$

where  $M_n = I_n - \xi_{p_n} (\xi_{p_n}' \xi_{p_n})^{-1} \xi_{p_n}'$  and  $A'$  denotes the transpose of a matrix  $A$ . The concentrated truncated Conditional Log likelihood function of  $\lambda$  is:

$$\bar{L}_n(\lambda) = -\frac{n}{2} (\ln(2\pi) + 1) - \frac{n}{2} \ln \hat{\sigma}_n^2(\lambda) + \ln |S_n(\lambda)|. \quad (10)$$

Then the estimator of  $\lambda_0$  is  $\hat{\lambda}$  that maximizes  $\bar{L}_n(\lambda)$ , and  $\hat{\beta}(\hat{\lambda})$ ,  $\hat{\sigma}^2(\hat{\lambda})$  are the estimators of  $\beta^*$  and  $\sigma_0^2$  respectively and denoted by :

$$\hat{\beta}(t) = \sum_{j=1}^{p_n} \hat{\beta}_j(\hat{\lambda}) \varphi_j(t),$$

and

$$\hat{\sigma}^2(\hat{\lambda}) = \frac{1}{n} \mathbf{Y}_n' S_n'(\hat{\lambda}) M_n S_n(\hat{\lambda}) \mathbf{Y}_n.$$

To get identifiability of  $\lambda_0$ ,  $\beta^*$ , and  $\sigma_0^2$  in the truncated model, notice that

$$E(\bar{L}_n(\lambda, \beta, \sigma^2)) = -\frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln 2\pi + \ln |S_n(\lambda)| - \frac{1}{2\sigma^2} E(V_n'(\lambda, \beta) V_n(\lambda, \beta)).$$

We have

$$\begin{aligned} E(V_n'(\lambda, \beta) V_n(\lambda, \beta)) &= E\left((S_n(\lambda) \mathbf{Y}_n - \xi_{p_n} \beta)' (S_n(\lambda) \mathbf{Y}_n - \xi_{p_n} \beta)\right) \\ &= E\left((S_n(\lambda) S_n^{-1} \mathbf{X}_n(\beta^*(\cdot)) - \xi_{p_n} \beta)' (S_n(\lambda) S_n^{-1} \mathbf{X}_n(\beta^*(\cdot)) - \xi_{p_n} \beta)\right) + \\ &\quad \sigma_0^2 \text{tr}(A_n(\lambda)) \\ &= E\left((S_n(\lambda) S_n^{-1} \xi_{p_n} \beta^* - \xi_{p_n} \beta)' (S_n(\lambda) S_n^{-1} \xi_{p_n} \beta^* - \xi_{p_n} \beta)\right) + \\ &\quad E\left(R_n' A_n(\lambda) R_n\right) + \\ &\quad \sigma_0^2 \text{tr}(A_n(\lambda)) + 2E\left((S_n(\lambda) S_n^{-1} \xi_{p_n} \beta^* - \xi_{p_n} \beta)' \mathbf{R}_n(\beta^*(\cdot))\right) \end{aligned}$$

where  $A_n(\lambda) = S_n^{-1} S_n'(\lambda) S_n(\lambda) S_n^{-1}$  and  $\mathbf{R}_n(\beta^*(\cdot)) = S_n(\lambda) S_n^{-1} R_n$ , with elements  $R_i = (\mathbf{X}_n(\beta^*(\cdot)) - \xi_{p_n} \beta^*)_i = \sum_{j>p_n} \beta_j^* \varepsilon_j^{(i)}$ .

The truncation strategy ensures that

$$E\left((S_n(\lambda) S_n^{-1} \xi_{p_n} \beta^* - \xi_{p_n} \beta)' \mathbf{R}_n(\beta^*(\cdot))\right) = o(1) \quad \text{and} \quad E\left(R_n' A_n(\lambda) R_n\right) = o(1).$$

Indeed, in one hand

$$E\left((S_n(\lambda) S_n^{-1} \xi_{p_n} \beta^*)' \mathbf{R}_n(\beta^*(\cdot))\right) = E\left((\xi_{p_n} \beta^*)' A_n(\lambda) R_n\right) = \quad (11)$$

$$\text{tr}(A_n(\lambda)) \sum_{r=1}^{p_n} \sum_{s>p_n} \beta_r^* \beta_s^* E(\varepsilon_r \varepsilon_s),$$

and

$$E\left((\xi_{p_n} \beta)' \mathbf{R}_n(\beta^*(\cdot))\right) = E\left((\xi_{p_n} \beta)' S_n(\lambda) S_n^{-1} R_n\right) = \text{tr}\left(S_n(\lambda) S_n^{-1}\right) \sum_{r=1}^{p_n} \sum_{s>p_n} \beta_r \beta_s^* E(\varepsilon_r \varepsilon_s). \quad (12)$$

The right hand side terms in (11) and (12) are zero when we consider the eigenbasis, otherwise we need to assume that

$$p_n \text{tr}(A_n(\lambda)) \sum_{s>p_n} |\beta_s^*|^2 E(\varepsilon_s^2) = o(1) \quad \text{and} \quad p_n \text{tr}\left(S_n(\lambda) S_n^{-1}\right) \sum_{s>p_n} |\beta_s^*|^2 E(\varepsilon_s^2) = o(1). \quad (13)$$

On the other hand, we have

$$E\left(R_n' A_n(\lambda) R_n\right) = \text{tr}(A_n(\lambda)) \sum_{r>p_n} \sum_{s>p_n} \beta_s^* \beta_r^* E(\varepsilon_r \varepsilon_s). \quad (14)$$

When using the eigenbasis, the term in the right hand side of (14) is of order

$$\text{tr}(A_n(\lambda)) \sum_{s>p_n} \beta_s^{*2} E(\varepsilon_s^2) \quad (15)$$

otherwise it is of order

$$\text{tr}(A_n(\lambda)) \left( \sum_{s>p_n} |\beta_s^*| \sqrt{E(\varepsilon_s^2)} \right)^2. \quad (16)$$

In other words, the term (15) or (16) need to be of order  $o(1)$ .

If this is the case, then

$$\begin{aligned} E(\tilde{L}_n(\lambda, \beta, \sigma^2)) &= \ln|S_n(\lambda)| \\ &- \frac{1}{2\sigma^2} E \left( (S_n(\lambda)S_n^{-1}\xi_{p_n}\beta^* - \xi_{p_n}\beta)' (S_n(\lambda)S_n^{-1}\xi_{p_n}\beta^* - \xi_{p_n}\beta) \right) \\ &\quad - \frac{n}{2} (\ln\sigma^2 + \ln 2\pi) - \frac{\sigma_0^2}{2\sigma^2} \text{tr}(A_n(\lambda)) + o(1). \end{aligned}$$

Let  $I_n + \lambda_0 G_n = S_n^{-1}$  where  $G_n = W_n S_n^{-1}$ , therefore  $S_n(\lambda) S_n^{-1} = I_n + (\lambda_0 - \lambda) G_n$  for all  $\lambda \in \Lambda$ .

Now for a fixed  $\lambda$ ,  $E(\tilde{L}_n(\lambda, \beta, \sigma^2))$  is maximum with respect to  $\beta$  and  $\sigma^2$  at

$$\beta^*(\lambda) = \frac{1}{n} \Gamma_{p_n}^{-1} E \left( \xi_{p_n}' S_n(\lambda) S_n^{-1} \xi_{p_n} \right) \beta^*,$$

where  $\Gamma_{p_n} = \frac{1}{n} E \left( \xi_{p_n}' \xi_{p_n} \right)$  is symmetric and positive definite in case of an eigenbasis. In addition, we have

$$\begin{aligned} \sigma_n^{*2}(\lambda) &= \frac{1}{n} E \left( (S_n(\lambda)S_n^{-1}\xi_{p_n}\beta^* - \xi_{p_n}\beta^*(\lambda))' (S_n(\lambda)S_n^{-1}\xi_{p_n}\beta^* - \xi_{p_n}\beta^*(\lambda)) \right) + \\ &\quad \frac{\sigma_0^2}{n} \text{tr}(A_n(\lambda)) \\ &= \frac{1}{n} (\lambda_0 - \lambda)^2 E \left( (G_n \xi_{p_n} \beta^*)' \left( G_n \xi_{p_n} \beta^* - \frac{1}{n} \xi_{p_n} \Gamma_{p_n}^{-1} E \left( \xi_{p_n}' G_n \xi_{p_n} \beta^* \right) \right) \right) + \\ &\quad \frac{\sigma_0^2}{n} \text{tr}(A_n(\lambda)). \end{aligned}$$

It is clear that  $\beta^*(\lambda_0) = \beta^*$  and  $\sigma_n^{*2}(\lambda_0) = \sigma_0^2$ . However identifiability of  $\beta^*$  and  $\sigma_0^2$  depends on that of  $\lambda_0$ . Let

$$Q_n(\lambda) = E(\tilde{L}_n(\lambda, \beta^*(\lambda), \sigma_n^{*2}(\lambda))) = \ln|S_n(\lambda)| - \frac{n}{2} \ln \sigma_n^{*2}(\lambda) - \frac{n}{2} (1 + \ln 2\pi) + o(1). \quad (17)$$

therefore proving the identifiability of  $\lambda_0$  is equivalent to show that  $\lambda_0$  maximizes  $Q_n(\lambda)$ .

We provide the asymptotic properties of the proposed QMLE estimators of  $\lambda_0$ ,  $\beta^*$ , and  $\sigma_0^2$  assuming some basic conditions; some of them concern the error disturbance and the weighted matrix, others guarantee the equilibrium of the system (1) or deal with the linearity of  $\ln|S_n(\lambda)|$ . Under these assumptions, we prove the convergence in probability and the asymptotic normality of the estimators. We illustrate the skills of the methods by some numerical results.

## References

1. Anselin, L. (1988). Spatial econometrics: Methods and models. *Kluwer Academic Publishers*.
2. Cardot, H., Ferraty, F., & Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45, 11–22.
3. Cardot, H., & Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92, 24–41.
4. Case, A. (1993). Spatial patterns in household demand. *Econometrica*, 52, 285–307.
5. Cliff, A., & Ord, K. (1973). Spatial autocorrelation. *London: Pion Ltd.*
6. Comte, F., Johannes, J. (2012). Adaptive functional linear regression. *Annals of Statistics* 40, 2765–2797.
7. Crambes, C., Kneip, A., Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Annals of Statistics* 37, 35–72.
8. Cressie, A. (1993). Statistics for spatial data. *New York: John Wiley and Sons.*
9. Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *J. Statist. Plan. Inf.* 147, 1-23.
10. Giraldo, R. (2014). Cokriging based on curves, prediction and estimation of the prediction variance, *InterStat*, 2, 1–30.
11. Giraldo, R., Delicado, P., Mateu, J. (2010). Continuous time-varying kriging for spatial prediction of functional data: An environmental application. *Journal of Agricultural, Biological, and Environmental Statistics* 15(1), 66–82.
12. Giraldo, R., Delicado, P., Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics*, 18(3), 411–426.
13. Guyon, X. (1995). *Random fields on a Network: Modeling, Statistics and Applications*. Springer, New York.
14. Hastie, T., Mallows, C. (1993). A statistical view of some chemometrics regression tools: Discussion. *Technometrics*, 35, 140–143.
15. Horvath, L., Kokoszka, P. (2012). *Inference for functional data with applications*, Springer.
16. Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72, 1899–1925.
17. Mas, A., Pumo, B. (2009). Functional linear regression with derivatives. *Journal of Nonparametric Statistics*, 21, 19–40.
18. Nerini, D., Monestiez, P., Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis* 101(2), 409–418.
19. Müller, H.-G., & Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33, 774–805. doi:10.1214/009053604000001156.
20. Pinkse, J., & Slade, M. E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. *Journal of Econometrics*, 85, 125–154.
21. Ramsay, J., & Silverman, B. (1997). Functional data analysis. *New York.*
22. Robinson, P. M. (2011). Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics*, 165, 5–19.

# A clustering algorithm for multivariate big data with correlated components

## *Un algoritmo di clustering per big data multivariati con componenti correlate*

Giacomo Aletti and Alessandra Micheletti

**Abstract** Common clustering algorithms require multiple scans of all the data to achieve convergence, and this is prohibitive when large databases, with millions of data, must be processed. Some algorithms to extend the popular K-means method to the analysis of big data are present in literature since 1998 [1], but they assume that the random vectors which are processed and grouped have uncorrelated components. Unfortunately this is not the case in many practical situations. We here propose an extension of the algorithm of Bradley, Fayyad and Reina to the processing of massive multivariate data, having correlated components.

**Abstract** *I comuni algoritmi di clustering richiedono di esaminare più volte tutti i dati per raggiungere la convergenza, e ciò risulta proibitivo quando devono essere analizzati database enormi, con milioni di dati. In letteratura sono presenti fin dal 1998 [1] alcuni algoritmi che estendono il popolare metodo K-medie all'analisi di big data, ma essi assumono che i vettori aleatori che vengono analizzati e raggruppati abbiano componenti non correlate. Purtroppo tale condizione non è soddisfatta in molti casi pratici. Qui proponiamo un'estensione dell'algoritmo di Bradley, Fayyad e Reina all'analisi di grandi moli di dati multivariati, con componenti correlate fra loro.*

**Key words:** big data, clustering, K-means, Mahalanobis distance

## 1 Introduction

Clustering is the division of a collection of data into groups, or *clusters*, such that points in the same cluster have a small distance from one another, while points in different clusters are at a large distance from one another. When the data are not very high dimensional, but are too many to fit in memory, because they are part of a huge dataset, or because they arrive in streams and must be processed immediately or they are lost, specific algorithms are needed to analyze progressively the data, store in memory only a small number of summary statistics, and then discard the already

---

G. Aletti, A.Micheletti  
ADAMSS Center, Università degli Studi di Milano, Milano, Italy  
e-mail: giacomo.aletti@unimi.it, alessandra.micheletti@unimi.it

processed data and free the memory. Situations like this, in which clustering plays a fundamental role, recur in many applications, like customer segmentation in e-commerce web sites, image analysis of video frames for objects recognition, recognition of human movements from data provided by sensors placed on the body or on a smartphone, etc. The key element in smart algorithms to treat such type of big data is to find methods by which the summary statistics that are retained in memory can be updated when each new observation, or group of observations, is processed. A first and widely recognized method to cluster big data is the Bradley-Fayyad-Reina (BFR) algorithm [1, 7], which is an extension of the classical K-means algorithm. The BFR algorithm responds to the following *data mining desiderata*: 1) Require one scan of the database and thus ability to operate on forward-only cursor; 2) On-line anytime behavior: a "best" answer is always available, with status information on progress, expected remaining time, etc. provided; 3) Suspendable, stoppable, resumable; incremental progress can be saved in memory to resume a stopped job; 4) Ability to incrementally incorporate additional data with existing models efficiently; 5) Work within confines of a limited RAM buffer; 6) Utilize a variety of possible scan modes: sequential, index, and sampling scan, if available. The BFR Algorithm for clustering is based on the definition of three different sets of data: a) the *retained set* (RS): the set of data points which are not recognized to belong to any cluster, and need to be retained in the buffer; b) the *discard set* (DS): the set of data points which can be discarded after updating the sufficient statistics; c) the *compression set* (CS): the set of data points which form smaller clusters among themselves, far from the principal ones and can be represented with other sufficient statistics. Each data point is assigned to one of these sets on the basis of its distance from the center of each cluster. The main weakness of the BFR Algorithm resides in the assumption that the covariance matrix of each cluster is diagonal, which means that the components of the analyzed multivariate data should be uncorrelated. In this way at each step of the algorithm only the means and variances of each component of the cluster centers must be retained. In the following we will describe an extension of the BFR algorithm to the case of clusters having "full" covariance matrix. Since with our method also the covariance terms of the clusters centers must be retained, there is an increase in the computational costs, but such increase can be easily controlled and is affordable if the processed data are not extremely high dimensional.

## 2 An extension of the BFR clustering algorithm

We will use the same three sets of data a)-c) introduced in the BFR algorithm, but using different summary statistics to define the discard set and the compression set.

## 2.1 Data Compression

Like in the BFR algorithm, primary data compression determines items to be discarded (discard set DS), and updates the compression set CS with the sufficient summary statistics of the identified clusters. Secondary data-compression takes place over data points not compressed in primary phase. Data compression refers to representing groups of points by their sufficient statistics and purging these points from RAM. In the following we will always represent vectors as column vectors. Assume that data points  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$  must be compressed in the same cluster. We will retain only the sample mean  $\bar{\mathbf{x}}_n = \sum_{i=1}^n \mathbf{x}_i$ , and the unbiased sample covariance matrix  $S_n = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ . These two sufficient statistics can be easily computed by keeping in memory the following quantities:

$$\begin{aligned} n, \quad \text{sumprod}_{kl}(n) &= \sum_{i=1}^n x_{ik}x_{il}, \quad \text{sumprodcross}_{kl}(n) = \sum_{i=1}^n \sum_{j=1}^n x_{ik}x_{jl}, \\ \text{sumsq}_k(n) &= \sum_{j=1}^n x_{jk}^2, \quad \text{sum}_k(n) = \sum_{j=1}^n x_{jk}, \quad k, l = 1, \dots, p, \quad k < l. \end{aligned}$$

These sufficient statistics can be easily updated when a new data point  $\mathbf{x}_{n+1}$  must be added to the cluster, without processing again the already compressed points. In fact, for  $k, l = 1, \dots, n$ ,  $k < l$ , we have

$$\begin{aligned} \text{sumprod}_{kl}(n+1) &= \sum_{i=1}^{n+1} x_{ik}x_{il} = \text{sumprod}_{kl}(n) + x_{(n+1)k}x_{(n+1)l} \\ \text{sumprodcross}_{kl}(n+1) &= \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} x_{ik}x_{jl} = \text{sumprodcross}_{kl}(n) + x_{(n+1)k}\text{sum}_l(n) \\ &\quad + x_{(n+1)l}\text{sum}_k(n) + x_{(n+1)k}x_{(n+1)l} \\ \text{sumsq}_k(n+1) &= \sum_{j=1}^{n+1} x_{jk}^2 = \text{sumsq}_k(n) + x_{(n+1)k}^2 \\ \text{sum}_k(n+1) &= \sum_{j=1}^{n+1} x_{jk} = \text{sum}_k(n) + x_{(n+1)k} \end{aligned}$$

Thus at each step of the algorithm we have to retain in memory only  $p^2 + p + 1$  sufficient statistics for each cluster, where  $p$  is the dimension of the data points. In addition, note that we should simply sum the corresponding statistics if we want to merge two clusters.

## 2.2 The covariance matrices of the clusters

Note that when a new cluster is formed, it contains too few data points to obtain a positive definite estimate of the covariance matrix, using the sample covariance matrix, at least until  $n \leq p$ . This is a problem since we need to invert this matrix to

compute the Mahalanobis distance, that we will use to assign the observations to the clusters. Recent research methods in estimating covariance matrices include banding, tapering, penalization and shrinkage. We have focused on the Steinian shrinkage method since, as underlined in [8], it leads to covariance matrix estimators that are non-singular, well-conditioned, expressed in closed form and computationally cheap regardless of  $p$ . We use the diagonal matrix  $D_S$  of the sample covariance matrix  $S$  as “target matrix” of the shrinkage method, noting that  $D_S$  was the BRF estimate of the covariance of each cluster used in [1]. In other words, in presence of few data, our method coincides with that of [1], and we allow a progressive influence of correlation as the number of data increases. Summing up, we use a linear shrinkage estimator for the covariance matrix, like that proposed in [3, 4, 6, 8] of the form  $\hat{S} = (1 - \lambda)S + \lambda D_S$ , where  $S$  is the sample covariance matrix,  $D_S$  is its diagonal matrix, and  $\lambda$  is a parameter in  $[0, 1]$ , whose optimal value depends on the number  $n$  of data in the cluster. The parameter  $\lambda$  is initially settled to 1, and then its value is decreasing to 0 when  $n \rightarrow \infty$ . The theoretical optimal value  $\lambda^*$  of  $\lambda$  is found by minimizing the risk function relative to the quadratic loss  $E[\|\hat{S} - \Sigma\|_F^2]$  (see, e.g., [8, 6]) and it is a ratio depending on the unknown  $\Sigma$ . When data are gaussian, the procedure proposed in [3] may be directly implemented to obtain unbiased estimators of numerator and denominator in the formula of  $\lambda^*$ . In non-gaussian setting, a bias due to the fourth moment is present in the numerator and it is corrected [6] with the use of further statistics, as the  $Q$ -statistics introduced in [4] (see also [2]). Unfortunately, it is not possible to compute the  $Q$  statistics on the basis of updatable sufficient statistics, as in our framework. To correct the bias, a new iterative procedure based on three updatable statistics for each cluster has been successfully developed.

### 2.3 Model update

Like in the BFR algorithm, the second step of our algorithm consists of performing K-means iterations over sufficient statistics of compressed, discarded and retained points. In order to assign a point to a cluster we use the Mahalanobis distance from its center (sample mean), i.e. we assign a new data point  $\mathbf{x}$  to cluster  $h$  with center  $\bar{\mathbf{x}}_h$  and estimated covariance matrix  $\hat{S}_h$ , if  $h$  is the index which minimizes  $\Delta(\mathbf{x}, \bar{\mathbf{x}}_h) = (\mathbf{x} - \bar{\mathbf{x}}_h)^T (\hat{S}_h)^{-1} (\mathbf{x} - \bar{\mathbf{x}}_h)$ , and if  $\Delta(\mathbf{x}, \bar{\mathbf{x}}_h)$  is smaller than a fixed threshold  $\delta$ . We also compare  $\mathbf{x}$  with each point  $\mathbf{x}_o$  in the retained set (RS), by computing  $\Delta(\mathbf{x}, \mathbf{x}_o) = (\mathbf{x} - \mathbf{x}_o)^T (\hat{S}_p)^{-1} (\mathbf{x} - \mathbf{x}_o)$ , where  $\hat{S}_p$  matrix is the pooled covariance matrix based on all  $\hat{S}_h$ :

$$\hat{S}_p = \frac{(n_{h_1} - 1)\hat{S}_{h_1} + (n_{h_2} - 1)\hat{S}_{h_2} + \dots + (n_{h_M} - 1)\hat{S}_{h_M}}{n_{h_1} + n_{h_2} + \dots + n_{h_M} - M}, \quad (1)$$

and where  $n_h$  is the number of points in cluster  $h$ . With  $\hat{S}_p$ , we emphasize the weighted importance of directions that are more significant for the clusters when we compute the distance between two “isolated” points. We then approximate locally the distribution of the clusters with a  $p$ -variate Gaussian and we build a confidence



regions around the centers of the clusters (see [5]). We then move  $\bar{\mathbf{x}}_h$  in the farthest position from  $\mathbf{x}$  in its confidence region, while we move the centers of the other clusters in the closest positions with respect to  $\mathbf{x}$  and we check if the cluster center closer to  $\mathbf{x}$  is still  $\bar{\mathbf{x}}_h$ . If yes, we assign  $\mathbf{x}$  to cluster  $h$ , we update the corresponding sufficient statistics and we put  $\mathbf{x}$  in the discard set; if the point is closer to a point  $\mathbf{x}_o$  of the retained set than to any cluster, we form a new new secondary cluster (CS) with the two points and we put  $\mathbf{x}$  and  $\mathbf{x}_o$  in the discard set; otherwise, we put  $\mathbf{x}$  in the retained set (RS).

## 2.4 Secondary data compression

The purpose of secondary data compression is to identify “tight” sub-clusters of points among the data that we can not discard in the primary phase. In [1], this is made in two phases. In the first one, a  $K$ -means algorithm tries to locate subclusters that are merged if they meet a “dense” condition. The candidate merging clusters are chosen sequentially based on a hierarchical agglomerative clustering build on the subclusters. In all this procedure, the euclidean metric was adopted. Finally, the number of clusters is initialized to  $K$ , and it can increase or decrease during the procedure. We adopt the same general idea, but we modify the procedure. First, we change the metric, by taking the pooled covariance  $\hat{S}_P$  given in (1). As for isolated points, we think that this metric is more precise than the euclidean one for this stage. Then, a hierarchical clustering is performed using the Ward’s method: the distance between two clusters  $h_1$  and  $h_2$  with  $n_{h_1}, n_{h_2}$  points and centroids  $\bar{\mathbf{x}}_{h_1}$  and  $\bar{\mathbf{x}}_{h_2}$ , is given by

$$\Delta(A, B) = \frac{n_{h_1}n_{h_2}}{n_{h_1} + n_{h_2}} (\bar{\mathbf{x}}_{h_1} - \bar{\mathbf{x}}_{h_2})^\top \hat{S}_P (\bar{\mathbf{x}}_{h_1} - \bar{\mathbf{x}}_{h_2}).$$

Note that we sequentially merge two clusters only if a suitable dense condition is fulfilled. For example, the total variance (i.e., the trace of the sample covariance matrix) of the union of the two is required to be smaller than a suitable proportion of the sum of the total variances of the single groups.

## 3 Results on simulated data

Synthetic data were created for the cases of 5 and 20 clusters. Data were sampled from 5 or 20 independent  $p$ -variate Gaussians, with elements of their mean vectors (the true means) uniformly distributed on  $[-5, 5]$ . The covariance matrices were generated by computing products of the type  $\Sigma = UHU^\top$ , where  $H$  is a diagonal matrix with elements on the diagonal uniformly distributed on  $[0.7, 1.5]$ , and  $U$  is the orthonormal matrix obtained by the singular value decomposition of a symmetric matrix  $MM^\top$ , where the elements of the  $p \times p$  matrix  $M$  are uniformly distributed on  $[-2, 2]$ . In either cases of 5 or 20 clusters, we generated 10.000 vectors for each cluster, having dimensions  $p = 10, 20, 50$ . This procedure guarantees that these clusters are fairly well-separated Gaussians, an ideal situation for K-Means. We applied

our procedure to these synthetic data, and we computed the secondary data compression after each bucket of 50 or 100 data points. The results are reported in Table 1. We note that the number of clusters is sometimes overestimated, in particular when the dimension  $p$  of the data points is small, which corresponds to the case where the clusters are less separated. In such cases, if the point clouds in different clusters are gathered in particularly "elongated" and rather close ellipsoids, then the correct detection of the clusters may be more difficult. We also note that in case of overestimation of the number of clusters, many of them are composed by 2 or 3 data points, which can then be revisited as small groups of outliers. The method seems to be almost insensitive to the buckets size. We conclude that the method here proposed provides rather good results on synthetic data, even if some improvement could be considered for the secondary data compression. The method is also under testing on real data. An accurate comparison with the BFR algorithm will also be performed.

n. of true clusters	dimension $p$ of data points	n. of data in each bucket	n. of estimated clusters	n. of small clusters	n. of retained points (outliers)
5	10	50	7	1	0
5	20	50	5	0	1
5	50	50	5	0	0
5	10	100	8	1	0
5	20	100	5	0	1
5	50	100	5	0	0
20	10	50	29	6	8
20	10	100	29	6	8

**Table 1** Results of the application of the proposed algorithm to synthetic data. By small clusters we mean clusters containing less than 4 data points

## References

1. Bradley, P.S., Fayyad, U.,Reina, C.: Scaling clustering to large databases, in KDD-98 Proceedings, pp 1-7, American Association for Artificial Intelligence, (1998) .
2. Chen, S. X., Zhang, L.-X., Zhong, P.-S.: Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.*, 105(490):pp. 810-819, (2010).
3. Fisher, T. J., Sun, X.: Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Comput. Statist. Data Anal.*, 55(5):pp. 1909-1918, (2011).
4. Himeno T., Yamada, T.: Estimations for some functions of covariance matrix in high dimension under non-normality and its applications. *J. Multivariate Anal.*, 130:27-44, (2014).
5. Hotelling, H.: The generalization of student's ratio. *Ann. Math. Statist.*, 2(3):360-378, 08 (1931).
6. Ikeda, Y., Kubokawa, T., Srivastava, M. S.: Comparison of linear shrinkage estimators of a large covariance matrix in normal and non-normal distributions. *Comput. Statist. Data Anal.*, 95:95-108, (2016).
7. Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of massive datasets, Cambridge University Press (2014).
8. Touloumis, A.: Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Comput. Statist. Data Anal.*, 83:251-261, (2015).

# A Bayesian semiparametric model for terrorist networks

## *Un modello Bayesiano semiparametrico per reti terroristiche*

Emanuele Aliverti

**Abstract** A recent field of research employs network-analysis' tools to the *dark network* framework, in which pairwise informations about terrorists' activities are available. In this work we focus on the "Noordin Mohamed Top" dataset, developing an asymmetric approach that treats one network as response and the remaining as covariates. The objective is to identify which information may be useful in predicting terrorists' collaboration in a bombing attack, identifying at the same time the most influential subjects involved in these dynamics. Such aim is addressed through an asymmetric Bayesian semi-parametric model for networks that, through a suitable prior specification, integrates a flexible regularization and the detection of leading nodes. Taking advantage of the Pólya-Gamma data augmentation scheme, we develop an efficient Gibbs sampler to make inference on the parameters involved.

**Abstract** *Un recente ambito di ricerca impiega strumenti tipici dell'analisi di reti nei contesti di dark networks, nei quali sono disponibili informazioni riguardanti attività terroristiche sotto forma di legami a coppie. In questo lavoro ci concentriamo sul dataset relativo a "Noordin Mohamed Top", sviluppando un approccio asimmetrico che considera una particolare rete come risposta, e le rimanenti come esplicative. L'obiettivo identificare quale informazione possa essere utile per predire la collaborazione di diversi terroristi in un attentato, identificando contemporaneamente i pi influenti soggetti coinvolti in queste dinamiche. Il problema è affrontato tramite un modello Bayesiano semiparametrico per reti che, attraverso un opportuno specificazione delle distribuzioni a priori, incorpora al suo internouna regolazione flessibile e l'identificazioe dei nodi leader. Sfruttando lo schema Pólya-Gamma per dati aumentati, presentiamo un efficiente Gibbs sampler per fare inferenza sui parametri coinvolti.*

**Key words:** Terrorism, networks, Bayesian semiparametrics, latent space, spike-and slabs prior, matrix factorization

---

Emanuele Aliverti

Università di Padova, Dipartimento di Scienze Statistiche e-mail: [aliverti@stat.unipd.it](mailto:aliverti@stat.unipd.it)

## 1 Introduction

After September 11th, intelligence agencies of different countries employed tools of the network science to serve in the fight against terroristic groups, often named *dark networks*. Great effort has been made to develop tools for identifying key players, that is actors within the network reporting high values in terms of some suitable network statistics. Since aggressive strategies encountered different failures, and the necessity of more sophisticated approaches became evident: [7] for example propose to focus on approaches less aggressive than direct military operations, involving a subtle application of informatics tools in order to gather different informations from various sources. The proper interpretation of retrieved data may provide a deeper description of terrorism, embracing at the same time social, economics and personal aspects, thus useful to develop strategies to defeat the roots of criminals associations.

Our motivating approach rises from the “Noordin Mohamed Top” dataset, drawn from a publication of the International Crisis Group; it consists of different ties among terrorists of the most ruthless group of the southwest Asia.

Data are coded into 10 symmetric relationships between network’s leader, Noordin Mohamed Top, and 78 affiliates, thus naturally coded into a *multilayer simple graphs*, that is a structure  $G = \{V, E_k\}$  where nodes (elements of  $V$ ) represent terrorists and edges (unordered pairs situated in the set  $E_k$ ) the presence of the particular  $k$ -th relationship among two generic subjects.

We expect a certain degree of association among different relationships, since they’re defined over the same set of nodes. Therefore, we would like to propose an approach able to efficiently use the information held inside “simpler” network in order to predict and make inference on the most interesting one, which is the network referred to the co-participation at the same terroristic bombing..

## 2 Proposed approach

Our research objectives can be faced by setting up an asymmetric framework, that threats one network as response and the remaining as covariates. The proposal of [3] is the most appropriate, and hence we will adapt this approach to our purposes by including nodal random effects and a non-parametric matrix factorization that avoids the estimation of different models. Let  $v$  the number of nodes of each network,  $\mathbf{Y}$  the  $v \times v$  adjacency matrix referred to the response network and  $\mathbf{X}$  the  $v \times v \times p$  array containing the  $p$  adjacency matrices referred to the  $p$  explanatory networks. We will consider only undirected and unweighted network (simple graphs), so adjacency matrices associated at are all dichotomous, symmetric and with non-defined elements on the main diagonal. Hence  $y_{ij} = y_{ji} \in \{0, 1\}$  and  $x_{ijk} = x_{jik} \in \{0, 1\} \forall i, j, k$ . Since the response network can assume only two values (presence or absence of edges), it is reasonable to assume a conditional bernoulli distribution for the under-

lying generative mechanism. We parametrize  $\pi_{ij}$ , the probability of observing an edge between node  $i$  and node  $j$ , through its log-odds  $\theta_{ij}$ . Formally:

$$y_{ij} | \pi_{ij} \stackrel{\text{ind}}{\sim} \text{Bin}(1, \pi_{ij}) \quad \theta_{ij} = \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right)$$

Furthermore, we decompose the linear predictor  $\theta_{ij}$  into two components: the first can be regarded as a parametric mixed model component, while the second as a non-parametric matrix factorization.

$$\theta_{ij} = \underbrace{\alpha + \sum_{k=1}^p [\beta_k + b_{ik} + b_{jk}] x_{ijk}}_{\text{Parametric component: fixed and random effects}} + \underbrace{\sum_{h=1}^H \lambda_h z_{ih} z_{jh}}_{\text{Non-parametric component}} \quad (1)$$

$$\alpha \in \mathbb{R}, \quad \beta_k \in \mathbb{R} \quad k = 1, \dots, p \quad z_{ih} \in \mathbb{R}, \lambda_h \in \mathbb{R}^+ \quad i = 1, \dots, v$$

$$\mathbf{b}_i = (b_{i1}, \dots, b_{ip}) \sim \mathbf{G}, \quad i = 1, \dots, v$$

The parametric component describes the relationship between networks and detects potentially influential nodes, which in this application means subjects whose role in some relationships has been particularly different from the average one. The basic interpretation is the following:  $\alpha$  provides an indication of the density of the response network, as an ordinary intercept in the binomial regression. Coefficient  $\beta_k$  are *fixed effects* in a logistic regression, that is the mean variation in the log odds of the outcome attributable to the  $k$ -th explicative network. In order to take advantage of the explicative power of covariates networks, we introduce additive *random effects* referred to the generic nodes  $i$  and  $j$  involved in the  $(i, j)$ -th dyad. In 1  $b_{ki}$  represents the specific deviation of the  $i$ -th node from the main effect  $\beta_k$ , and so can account for his particular propensity in building ties in the response network. For each relationship, the purpose is to identify subjects more (or less) likely to commit an attack with, providing thus a brighter description of those dynamics. Furthermore, additive random effects can account for between-rows heterogeneity contained in the explanatory networks, allowing then a better estimation of the fixed counterpart.

The non-parametric component decompose the residual among response and explanatory networks in a flexible way, that is through a matrix factorization that allows the number of factors to vary adaptively. It can be interpreted as a latent space whose size is at most equal to  $H$ , in which  $z_{ih}$  represents the  $h$ -th latent coordinate of the  $i$ -th node, while  $\lambda_h$  defines the importance of the  $h$ -th dimension of the latent space in defining the final model. This strategy aims to adaptively account for the dependencies in the response not seized by explanatory networks, providing estimates for the parametric component deprived of potentially confounding factors.

### 3 Prior distribution and posterior simulation

For a complete Bayesian definition of the proposed model we need to specify proper prior distributions for the set of parameters involved.

#### 3.1 Parametric component

We specify zero-mean normal distributions over the fixed effects parameter. Formally,

$$\pi(\alpha) \sim N(\mu_{0\alpha}, \sigma_{\alpha_0}^2) \quad \pi(\beta) = \pi(\beta_1, \dots, \beta_p) \sim N_p(\mu_{0\beta}, \Sigma_{\beta_0}) \quad (2)$$

In our application, we expect a certain level of heterogeneity in nodes' behavior, both between different subject and within the same, when involved in different relationships. For example, it's reasonable that dealing directly with leaders may led to a higher propensity in participating at the same terroristic attack. However, certain subjects may have had a central role just in the some specific relationships, such as the school recruitment network, and a marginal position elsewhere; for that, we need a prior distribution able to differentiate particular subjects from standard ones, and hence we specify a *spike and slabs* prior distributions [4] independently for each  $p$ -dimensional vector referred to the generic  $i$ -th subject,  $i = 1, \dots, v$ . Formally:

$$\begin{aligned} \mathbf{G} &\sim N(\mathbf{0}, \mathbf{\Gamma}_i), \quad \mathbf{\Gamma}_i = \text{diag}(\gamma_{i1}, \dots, \gamma_{ip}), \quad \gamma_{ik} = \theta_{ik} \tau_{ik}^2, \quad k = 1, \dots, p \\ \pi(\theta_{ik}) &\stackrel{iid}{\sim} (1 - w_i) \delta_{v_0}(\cdot) + w_i \delta_1(\cdot) \\ \pi(\tau_{ik}^{-2}) &\sim \text{Gamma}(d_1, d_2), \quad \pi(w_i) \sim \text{Uniform}[0, 1] \end{aligned} \quad (3)$$

In 3  $v_0$  is a value close to zero, and the hyper-parameters  $d_1, d_2$  are chosen in order to obtain, for  $\gamma_k = \theta_k \tau_k^2$ , a continuous distribution characterized by a spike in  $v_0$  and a continuous right tail;  $\delta_{v_0}$  and  $\delta_1$  are Formally, a Multiplicative Inverse Gamma (MIG) is specified as prior probability measure over the loading elements  $\lambda_h$  in 1, and standard Gaussian distribution for the latent coordinates. See [2] for a recent discussion regarding the properties of the MIG prior. Formally:

$$\begin{aligned} z_{ih} &\stackrel{iid}{\sim} N(0, 1), \quad i = 1, \dots, v \\ \lambda_h &= \prod_{m=1}^h \frac{1}{\theta_m}, \quad \theta_1 \sim \text{Gamma}(a_1, 1), \quad \theta_{h \geq 2} \stackrel{iid}{\sim} \text{Gamma}(a_2, 1) \end{aligned} \quad (4)$$

with  $a_1 > 0$  and  $a_2 > 1$  fixed hyper parameters.

### 3.2 Posterior Simulation

Adapting the Pólya-Gamma data augmentation strategy proposed by [6] in the logistic regression framework, we can obtain the full-conditional distributions for the parameters involved in our model, and hence implement a Gibbs sampling strategy.

## 4 Results

The effects of different network is heterogeneous: for example, a tie in the communication network increments, in mean, the log odds of collaborating in the same bombing operations of around 2 times; furthermore, if two terrorist had been in the same terroristic organization the log odds is lowered of an amount around 1.33 times, that is not so trivial. As for influential nodes, the spike and slabs strategy identify several terrorists, confirmed to be such in the Indonesian reports. The predictive performance recorded an average area under the ROC curve equal to 0.864, a false positive rate equal to 0.225 and a total negative rate of 0.220, using as estimates for the missing edges the mean of the posterior predictive density and, where needed, the overall density of the response network as cutoff value.

## References

1. Bhattacharya, A and D B Dunson (2011). Sparse Bayesian infinite factor models. In: *Biometrika* 98.2, pp. 291-306.
2. Durante, Daniele (2017). A note on the multiplicative gamma process. In: *Statistics & Probability Letters* 122, pp. 198-204.
3. Hoff, Peter (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In: *Advances in Neural Information Processing Systems*, pp. 657-664.
4. Ishwaran, H., and Rao, J. S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies". *Annals of Statistics*, 730-773.
5. Polson, Nicholas G., James G. Scott, and Jesse Windle (2013). Bayesian Inference for Logistic Models Using Pólya-Gamma Latent Variables. In: *Journal of the American Statistical Association* 108.504, pp. 1339-1349.
6. Roberts, Nancy and Sean F Everton (2011). Strategies for Combating Dark Networks. In: *Journal of Social Structure* 12,





# Emerging challenges in official statistics: new sources, methods and skills

Giorgio Alleva

Official statistics is challenged to provide an increasingly complete picture of the complexity of our societies with a compelling demand for data. At the same time, it is facing human resources and budget constraints. The development and dissemination of new digital technologies have removed many obstacles, first of all the cost for production, storage and analysis of information. A leap forward in efficiency is therefore in order, if we are to meet our responsibilities and to guarantee ever increasing quality standards. As sampling surveys are expensive, response rates are decreasing and response burden must be reduced, data collection need to be optimised. The emergence of new data sources and availability of Big data and the opportunity of a massive exploitation of those already at hand (like administrative data) require new tools and methodologies. The response to these thematic, methodological and organizational challenges lies in “integration”: of sources, of methods, and of skills. Multiple use of data sources should be based on a re-engineering of the production process of official statistics. At Istat, the core of the new organisation aims at moving away from the ‘silo’ approach, typical of traditional statistical agencies, towards the enhancement of horizontal services: management, methodology and IT innovations drive the integration process, linking sources to boost coherence, tailoring new products to the different users’ needs, reducing the response burden through the reuse of available data and information, increasing the use of technology, and resulting in significant efficiency and time saving. This new organisational model supports the Integrated System of Statistical Registers, a single logical data asset resulting from the integration of survey data, administrative data as well as data coming from new sources. Pillars of this system are the Population Register, the Business Register and the Territorial Register which are interconnected with one another through the Activity Register. The Integrated System allows achieving units and variables identification and estimation consistency as single cohesive units, which will make several new analyses possible (including a longitudinal approach). Hence, the system will not only improve efficiency by means of economies of scale,

---

Giorgio Alleva

Presidente Istituto Nazionale di Statistica (ISTAT), e-mail: [giorgio.alleva@uniroma1.it](mailto:giorgio.alleva@uniroma1.it)

but also high quality and richer statistical outputs. The path towards integration cannot leave the research activity of the Institute aside: research and development of new techniques and methodologies are indeed at the center of Istat's modernisation project. Istat has set up a three-year plan for methodological and thematic research. Methodological research will move along four strategic research areas: integrated system of registers; censuses obtained by data integration; big data; and the unique process. Innovation will emerge from the so-called "Innovation lab", a place where researchers will share new ideas, test new solutions, new processes and new products. The frontier of data integration for official statistics is represented by the increasing opportunities to use big data to produce timely high-quality statistics with greater detail, and competing with growing numbers of new, non-official, players. NSIs are compelled to speed their production and make it more effective and less burdensome for respondents. In Istat, as in most NSIs in Europe, several projects using big data sources for the production of statistics are currently ongoing. Some projects are in the early stages of implementation, some other are still in the experimental phase. We expect to reap the first results in the near future. All those projects need to tackle three key issues: quality, privacy and security issues, partnerships. The use of big data in the production of official statistics is part of a wider strategy on "Experimental statistics" including new indicators from integration of sources, new tools for new phenomena, and unconventional classifications. Outputs from these innovative work will need to be treated accordingly. Data dissemination is another key issue in official statistics' innovation with the progressive opening of our data at its core. Open data are a key enabler of data driven innovation. When official statistics meets open data, several benefits are generated: from the possibility to reach users more easily to the enrichment of the published information with metadata that allow a proper interpretation. Much has already been done in the last years to disseminate them. Through the Linked Open Data portal users can now access interconnected and structured information through graphical interfaces that can be directly queried by external applications, independently of the technologies adopted. Finally, on the way towards innovation, high level skills and a change-driven culture are essential. Statistical institutions need data specialist able to produce, integrate and interpret data and to work with big and open data, but such skills are also strongly requested by the market, everywhere in Europe. Italy needs to engage further to urgently foster these new professions. Statistical organisations can also benefit greatly from each other and from mutual exchange and support and networking.

# A fast algorithm for the canonical polyadic decomposition of large tensors

## *Un algoritmo veloce per la decomposizione di grandi tensori*

R. André, X. Luciani and E. Moreau

**Abstract** The canonical polyadic decomposition is one of the most used tensor decomposition. However classical decomposition algorithms such as alternating least squares suffer from convergence problems and thus the decomposition of large tensors can be very time consuming. Recently it has been shown that the decomposition can be rewritten as a joint eigenvalue decomposition problem. In this paper we propose a fast joint eigenvalue decomposition algorithm then we show how it can benefit the canonical polyadic decomposition of large tensors.

**Abstract** *La decomposizione canonica di tensori è usata in diversi campi tra cui quello del data science. Tuttavia, nei classici algoritmi di decomposizione, come l'alternating least squares, si possono riscontrare problemi di convergenza. Proprio per questo motivo, la decomposizione di grandi tensori può essere molto dispendiosa in termini di tempo di calcolo. Recentemente, sono stati sviluppati algoritmi di decomposizione canonica veloci, basati sulla diagonalizzazione di un insieme di matrici su una base comune di autovettori. In questo articolo proponiamo un algoritmo originale per risolvere quest'ultimo problema. In seguito mettiamo in evidenza l'aspetto più interessante di questo approccio al fine di effettuare la decomposizione canonica di grandi tensori.*

**Key words:** Tensor, Canonical polyadic decomposition, PARAFAC, algorithms, Joint eigenvalues decomposition

---

Rémi André, Xavier Luciani and Eric Moreau  
Aix Marseille Université, Université de Toulon, CNRS, ENSAM, LSIS, Marseille, France, e-mail:  
luciani@univ-tln.fr

## 1 Introduction

In many data sciences applications, collected data have a multidimensional structure and can thus be stored in multiway arrays (tensors). In this context, multi-way analysis provides efficient tools to analyze such data sets. In particular, the Canonical Polyadic Decomposition (CPD) also known as PARAllel FACTor analysis (PARAFAC) has been successfully applied in various domains such as chemometrics, telecommunications, psychometrics and data mining, just to mention a few [7].

The CPD models the data thanks to multilinear combinations as described below. Let us consider a data tensor  $\mathcal{F}$  of order  $Q$  (*i.e.* a  $Q$ -dimensions array) and size  $I_1 \times \dots \times I_Q$ . Its CPD of rank  $N$  is then defined by:

$$\mathcal{F}_{i_1 \dots i_Q} = \sum_{n=1}^N F_{i_1 n}^{(1)} \dots F_{i_Q n}^{(Q)} + \mathcal{E}_{i_1 \dots i_Q} \quad (1)$$

where  $\mathbf{F}^{(q)}$  is the  $q$ -th factor matrix of size  $I_q \times N$  and  $\mathcal{E}$  is the error tensor. One crucial point here is that this decomposition has usually an unique solution up to trivial scaling and permutation indeterminacy. The idea is then that the meaningful information lies in the factor matrices. Thus we want to estimate these matrices from the data. Several algorithms were proposed in this purpose. The most popular is the Alternating Least Squares algorithm (ALS) [6]. This iterative algorithm is very simple to implement and usually provides accurate results. However it suffers from well known convergence problems. In particular the convergence is very sensitive to the initialization and the algorithm can be easily stuck in a local minimum of the cost function. A smart initialization is always possible but in practice one had better to perform several runs of the algorithm with random initialization. A second consequence is that it is difficult to set efficiently the threshold of the stopping criterion. Indeed it frequently occurs that the algorithm escapes from a local minimum after a very large number of iterations and during these iterations the variations of the cost function can be very small. This issue becomes significant when the computational cost per iteration is high *i.e.* for high rank CPD of large tensors. Thereby the decomposition performed with ALS can have a high effective computational cost and thus can be time consuming when dealing with high rank CPD of large tensors. Several other iterative algorithms were proposed to solve those convergence problems but in practice the computational cost of these solutions remains high. More details about ALS convergence problems and other iterative CPD algorithms can be found in [7], [3] and [1].

Recently several authors showed how to rewrite the CPD as a Joint EigenValues Decomposition (JEVD) of a matrix set [5, 8, 10]. The JEVD consists in finding the eigenvector matrix  $\mathbf{A}$  that jointly diagonalizes a given set of  $K$  non-defective matrices  $\mathbf{M}^{(k)}$  in the following way:

$$\mathbf{M}^{(k)} = \mathbf{A} \mathbf{D}^{(k)} \mathbf{A}^{-1}, \quad \forall k = 1, \dots, K. \quad (2)$$

This approach allows to reduce the computational cost of the CPD because JEVD algorithms converge in few iterations with an excellent convergence rate. Furthermore, it is less sensitive to the overestimation of the CPD rank than ALS.

Several JEVD algorithms have been proposed in the last decade [4, 8, 9]. In a recent paper we have introduced an algorithm called JDTE [2]. This algorithm offers a good trade-off between speed and precision but its performances decrease with the matrix size. In the CPD context, it means that it is not suitable for high rank CPD. As a consequence, we propose in this paper an improved version of this algorithm.

The paper is organized as follow. In the next section we recall a simple and economic way to rewrite the CPD as a JEVD problem. Then in section 3 we describe the proposed JEVD algorithm. Finally, in section 4 we evaluate our approach for the decomposition of large tensors by means of numerical simulations.

In the following, the operator  $\text{Diag}\{\cdot\}$  represents the diagonal matrix built from the diagonal of the matrix argument, the operator  $\text{ZDiag}\{\cdot\}$  sets to zero the diagonal of the matrix argument and  $\|\cdot\|$  is the Frobenius norm of the argument matrix or tensor.

## 2 From CPD to JEVD

There are several ways to rewrite the CPD as a JEVD problem. Here we use the method described in [8] because the associated algorithm, called DIAG, has the lowest numerical complexity.

We consider the tensor  $\mathcal{T}$  and its CPD of rank  $N$  defined in introduction. The first step consists in rearranging entries of  $\mathcal{T}$  into an unfolding matrix  $\mathbf{T}$  of size  $\prod_{p=1}^P I_p \times \prod_{q=p+1}^Q I_q$  by merging the first  $P$  modes on the rows of  $\mathbf{T}$  and the  $Q - P$  other modes on its columns. Defining for all couple of integers  $(a, b)$  with  $a \leq b$ :

$$\mathbf{Y}^{(b,a)} = \mathbf{F}^{(b)} \odot \mathbf{F}^{(b-1)} \odot \dots \odot \mathbf{F}^{(a)}, \quad (3)$$

where  $\odot$  is the Khatri-Rao product, we can thus rewrite (1) in a matrix form:

$$\mathbf{T} = \mathbf{Y}^{(P,1)} (\mathbf{Y}^{(Q,P+1)})^\top. \quad (4)$$

Of course, other merging of the tensor modes could have been chosen, leading to other unfolding matrices. The choice of the unfolding matrix can have a huge impact on the numerical complexity of the DIAG algorithm [8]. As a rule of thumb, when all tensor dimensions are large we recommend to chose  $P = Q - 2$  and to place the smallest dimension at the end ( $I_Q \leq I_q, \forall q$ ). In the following we assume that the rank of  $\mathbf{T}$  is not greater than  $N$ .

The second step is the Singular Value Decomposition (SVD) of  $\mathbf{T}$ , truncated at order  $N$ . We denote  $\mathbf{U}$ ,  $\mathbf{S}$  and  $\mathbf{V}^\top$  the matrices of this truncated SVD.

At this stage, there exists a unique non singular square matrix  $\mathbf{A}$  of size  $N \times N$  such that:

$$\mathbf{Y}^{(P,1)} = \mathbf{U}\mathbf{A} \quad \text{and} \quad (\mathbf{Y}^{(Q,P+1)})^\top = \mathbf{A}^{-1}\mathbf{S}\mathbf{V}^\top. \quad (5)$$

$(\mathbf{Y}^{(Q,P+1)})^\top$  can be seen as an horizontal block matrix:

$$(\mathbf{Y}^{(Q,P+1)})^\top = \left[ \phi^{(1)} (\mathbf{Y}^{(Q-1,P+1)})^\top, \dots, \phi^{(I_Q)} (\mathbf{Y}^{(Q-1,P+1)})^\top \right], \quad (6)$$

where  $\phi^{(1)}, \dots, \phi^{(I_Q)}$  are the  $I_Q$  diagonal matrices built from the  $I_Q$  rows of matrix  $\mathbf{F}^{(Q)}$ . Then, (5) and (6) yield:

$$\mathbf{S}\mathbf{V}^\top = \left[ \Gamma^{(1)\top}, \dots, \Gamma^{(I_Q)\top} \right], \quad (7)$$

where  $\Gamma^{(k_1)} = \mathbf{Y}^{(Q-1, P+1)} \phi^{(k_1)} \mathbf{A}^\top$  for  $k_1 = 1, \dots, I_Q$ . Assuming that matrices  $\Gamma^{(k_1)}$  and matrix  $\mathbf{Y}^{(Q-1, P+1)}$  are full column rank, then they all admit a Moore-Penrose matrix inverse denoted by  $\sharp$ . Thereby, we can define for any couple  $(k_1, k_2)$  with  $k_1 = 1, \dots, I_Q - 1$  and  $k_2 = k_1 + 1, \dots, I_Q$ :

$$\mathbf{M}^{(k_1, k_2)} \stackrel{\text{def}}{=} \left( \Gamma^{(k_1)\sharp} \Gamma^{(k_2)} \right)^\top, \quad (8)$$

$$= \mathbf{A}\mathbf{D}^{(k_1, k_2)} \mathbf{A}^{-1}, \quad (9)$$

where  $\mathbf{D}^{(k_1, k_2)} = \phi^{(k_2)} \phi^{(k_1)\sharp}$  are diagonal matrices. As a result,  $\mathbf{A}$  performs the JEVD of the set of matrices  $\mathbf{M}^{(k_1, k_2)}$  and can be estimated using a JEVD algorithm. An important observation have to be made here. In the previous step we have built  $I_Q(I_Q - 1)/2$  matrices  $\mathbf{M}^{(k_1, k_2)}$ . When dealing with large tensors this value can be very high with respect to the matrix size. In practice this does not help to improve the estimation of  $\mathbf{A}$  significantly and dramatically increases the numerical complexity of the JEVD step. Thereby, we propose as an alternative to build only a subset of  $I_Q - 1$  matrices, for instance by taking  $k_2 = k_1 + 1$  in (8). This can be seen as an economic version of the DIAG algorithm.

After the JEVD, matrices  $\mathbf{Y}^{(P, 1)}$  and  $\mathbf{Y}^{(Q, P+1)}$  are immediately deduced from  $\mathbf{A}$  using (5). Finally, we can easily deduce  $\mathbf{F}^{(a)}, \dots, \mathbf{F}^{(b)}$  from  $\mathbf{Y}^{(b, a)}$  as explained in [8].

In the next section, we propose an algorithm to solve the JEVD step. In order to simplify the notations, subscripts  $k_1$  and  $k_2$  are replaced by unique subscript  $k$  so that equation (9) becomes:

$$\mathbf{M}^{(k)} = \mathbf{A}\mathbf{D}^{(k)} \mathbf{A}^{-1}, \quad \forall k = 1, \dots, K, \quad (10)$$

where  $K = I_Q(I_Q - 1)/2$  or  $K = I_Q - 1$  depending on whether we choose the original DIAG algorithm or the economic version.

### 3 A fast JEVD algorithm

We propose here a fast algorithm to compute an estimate of  $\mathbf{A}$ , denoted  $\mathbf{B}$ , up to a permutation and scaling indeterminacy of the columns. This indeterminacy is inherent to the JEVD problem.

We want that  $\mathbf{B}$  jointly diagonalizes the set of matrices  $\mathbf{M}^{(k)}$ . It means that matrices  $\widehat{\mathbf{D}}^{(k)}$  defined by:

$$\widehat{\mathbf{D}}^{(k)} = \mathbf{B}^{-1} \mathbf{M}^{(k)} \mathbf{B}, \quad \forall k = 1, \dots, K \quad (11)$$

must be as diagonal as possible.  $\mathbf{B}$  is called the diagonalizing matrix. This kind of problem can be efficiently solved by an iterative procedure based on multiplicative updates. Before the first iteration, we set  $\widehat{\mathbf{D}}^{(k)} = \mathbf{M}^{(k)}$ , then at each iteration, matrices  $\mathbf{B}$  and  $\widehat{\mathbf{D}}^{(k)}$  are updated by a new matrix  $\mathbf{X}$  as follow:

$$\mathbf{B} \leftarrow \mathbf{B}\mathbf{X} \text{ and } \widehat{\mathbf{D}}^{(k)} \leftarrow \mathbf{X}^{-1}\widehat{\mathbf{D}}^{(k)}\mathbf{X}, \quad \forall k = 1, \dots, K. \quad (12)$$

The strategy that we now propose to compute the updating matrix  $\mathbf{X}$  can be seen as a modified version of the one we proposed in [2]. The main difference is that here we resort to a sweeping procedure. It means that  $\mathbf{X}$  is built from a set of  $N(N-1)/2$  matrices, denoted  $\mathbf{X}^{(i,j)}$  ( $i = 1, \dots, N-1$  and  $j = i+1, \dots, N$ ) as follow:

$$\mathbf{X} = \prod_{i=1}^{N-1} \prod_{j=i+1}^N \mathbf{X}^{(i,j)}. \quad (13)$$

As a consequence, at each iteration, the updates in (12) consist now in  $N(N-1)/2$  successive  $(i,j)$ -updates of  $\mathbf{B}$  and  $\widehat{\mathbf{D}}^{(k)}$ , defined as:

$$\mathbf{B} \leftarrow \mathbf{B}\mathbf{X}^{(i,j)} \text{ and } \widehat{\mathbf{D}}^{(k)} \leftarrow \left(\mathbf{X}^{(i,j)}\right)^{-1}\widehat{\mathbf{D}}^{(k)}\mathbf{X}^{(i,j)}, \quad \forall k = 1, \dots, K. \quad (14)$$

Furthermore, because of the scaling indeterminacy of the JEVd problem we can impose the following structure to matrices  $\mathbf{X}^{(i,j)}$ :  $\mathbf{X}^{(i,j)}$  is equal to the identity matrix at the exception of entries  $X_{ij}^{(i,j)}$  and  $X_{ji}^{(i,j)}$  that are equal to two unknown parameters:

$$X_{ij}^{(i,j)} = x_1^{(i,j)}, \quad (15)$$

$$X_{ji}^{(i,j)} = x_2^{(i,j)}. \quad (16)$$

We now explain how these parameters are computed for a given couple  $(i,j)$ . First of all, let us define the function  $C$  as

$$C(\mathbf{X}^{(i,j)}) = \sum_{k=1}^K \|\text{ZDiag}\{(\mathbf{X}^{(i,j)})^{-1}\widehat{\mathbf{D}}^{(k)}\mathbf{X}^{(i,j)}\}\|^2. \quad (17)$$

$C$  is a classical diagonalization criterion that is equal to zero if the  $K$  updated matrices are diagonal. Therefore we look for  $\mathbf{X}^{(i,j)}$  that minimizes  $C$ .

Matrix  $\mathbf{X}^{(i,j)}$  can be decomposed as  $\mathbf{X}^{(i,j)} = (\mathbf{I} + \mathbf{Z}^{(i,j)})$ , where  $\mathbf{Z}^{(i,j)} = \text{ZDiag}\{\mathbf{X}^{(i,j)}\}$ . The criterion can then be written as:

$$C(\mathbf{X}^{(i,j)}) = \tilde{C}(\mathbf{Z}^{(i,j)}) = \sum_{k=1}^K \|\text{ZDiag}\{(\mathbf{I} + \mathbf{Z}^{(i,j)})^{-1}\widehat{\mathbf{D}}^{(k)}(\mathbf{I} + \mathbf{Z}^{(i,j)})\}\|^2. \quad (18)$$

We consider in fact an approximation of  $C(\mathbf{X}^{(i,j)})$  assuming that we are close to the diagonalizing solution *i.e.*  $\mathbf{X}^{(i,j)}$  is close to the identity matrix. This implies that  $\|\mathbf{Z}^{(i,j)}\| \ll 1$  and thus the first order Taylor expansion of  $(\mathbf{I} + \mathbf{Z}^{(i,j)})^{-1}$  yields:

$$(\mathbf{I} + \mathbf{Z}^{(i,j)})^{-1}\widehat{\mathbf{D}}^{(k)}(\mathbf{I} + \mathbf{Z}^{(i,j)}) \simeq (\mathbf{I} - \mathbf{Z}^{(i,j)})\widehat{\mathbf{D}}^{(k)}(\mathbf{I} + \mathbf{Z}^{(i,j)}) \quad (19)$$

$$\simeq \widehat{\mathbf{D}}^{(k)} - \mathbf{Z}^{(i,j)}\widehat{\mathbf{D}}^{(k)} + \widehat{\mathbf{D}}^{(k)}\mathbf{Z}^{(i,j)} - \mathbf{Z}^{(i,j)}\widehat{\mathbf{D}}^{(k)}\mathbf{Z}^{(i,j)} \quad (20)$$

$$\simeq \widehat{\mathbf{D}}^{(k)} - \mathbf{Z}^{(i,j)}\widehat{\mathbf{D}}^{(k)} + \widehat{\mathbf{D}}^{(k)}\mathbf{Z}^{(i,j)}. \quad (21)$$

In the same way, matrices  $\widehat{\mathbf{D}}^{(k)}$  can be decomposed as  $\widehat{\mathbf{D}}^{(k)} = \Lambda^{(k)} + \mathbf{O}^{(k)}$ , where  $\Lambda^{(k)} = \text{Diag}\{\widehat{\mathbf{D}}^{(k)}\}$  and  $\mathbf{O}^{(k)} = \text{ZDiag}\{\widehat{\mathbf{D}}^{(k)}\}$ . Here our assumption means that matrices  $\widehat{\mathbf{D}}^{(k)}$  are almost diagonal and thus  $\|\mathbf{O}^{(k)}\| \ll 1$ . It yields:

$$\widehat{\mathbf{D}}^{(k)} - \mathbf{Z}^{(i,j)} \widehat{\mathbf{D}}^{(k)} + \widehat{\mathbf{D}}^{(k)} \mathbf{Z}^{(i,j)} \simeq \Lambda^{(k)} + \mathbf{O}^{(k)} - \mathbf{Z} \Lambda^{(k)} + \Lambda^{(k)} \mathbf{Z} \quad (22)$$

and finally we can approximate  $\tilde{\mathcal{C}}(\mathbf{Z}^{(i,j)})$  by  $C_a(\mathbf{Z}^{(i,j)})$ :

$$C_a(\mathbf{Z}^{(i,j)}) = \sum_{k=1}^K \|\mathbf{Z} \text{Diag}\{\mathbf{O}^{(k)} - \mathbf{Z}^{(i,j)} \Lambda^{(k)} + \Lambda^{(k)} \mathbf{Z}^{(i,j)}\}\|^2 \quad (23)$$

$$= \sum_{k=1}^K \sum_{\substack{m,n=1 \\ m \neq n}}^N (O_{mn}^{(k)} + Z_{mn} \Lambda_{mm}^{(k)} - Z_{mn} \Lambda_{nn}^{(k)})^2 \quad (24)$$

$$= \sum_{k=1}^K \left( (O_{ij}^{(k)} - x_1^{(i,j)} (\Lambda_{jj}^{(k)} - \Lambda_{ii}^{(k)}))^2 + (O_{ji}^{(k)} - x_2^{(i,j)} (\Lambda_{ii}^{(k)} - \Lambda_{jj}^{(k)}))^2 + \sum_{\substack{m,n=1 \\ m \neq n}}^N (O_{mn}^{(k)})^2 \right). \quad (25)$$

We can then easily show that  $C_a$  is minimum for :

$$(x_1^{(i,j)}, x_2^{(i,j)}) = \left( \frac{\sum_{k=1}^K O_{ij}^{(k)} (\Lambda_{jj}^{(k)} - \Lambda_{ii}^{(k)})}{\sum_{k=1}^K (\Lambda_{ii}^{(k)} - \Lambda_{jj}^{(k)})^2}, \frac{\sum_{k=1}^K O_{ji}^{(k)} (\Lambda_{ii}^{(k)} - \Lambda_{jj}^{(k)})}{\sum_{k=1}^K (\Lambda_{jj}^{(k)} - \Lambda_{ii}^{(k)})^2} \right). \quad (26)$$

We call this algorithm SJDTE for Sweeping Joint eigenvalue Decomposition based on a Taylor Expansion.

## 4 Numerical Simulations

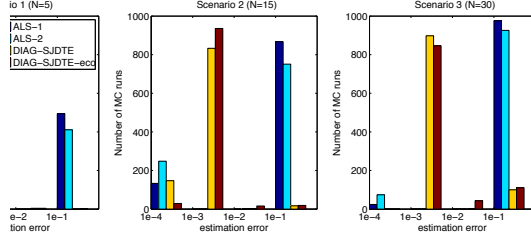
We have included SJDTE in the DIAG procedure described in section 2 for the JEVD step. Two versions of this CPD algorithm were implemented corresponding to original and economic versions of DIAG. In the following, these are referred as DIAG-SJDTE and DIAG-SJDTE-eco respectively and are compared with the ALS for the CPD of large tensors of order 3. In this purpose we define the tensor reconstruction error:  $r_T = \|\mathcal{T} - \widehat{\mathcal{T}}\| / \|\mathcal{T}\|$  and the factor matrices estimation error:  $r_F = \sum_{q=1}^3 \|\mathbf{F}^{(q)} - \widehat{\mathbf{F}}^{(q)}\| / \|\mathbf{F}^{(q)}\|$  where  $\widehat{\mathbf{F}}^{(1)}$ ,  $\widehat{\mathbf{F}}^{(2)}$  and  $\widehat{\mathbf{F}}^{(3)}$  are the factor matrices estimated by an algorithm and  $\widehat{\mathcal{T}}$  is the tensor reconstructed from these matrices. Other comparison criteria are the number of computed iteration,  $n_{it}$  and the cputime of Matlab (elapsed time during the algorithm run),  $t_{cpu}$ . Of course the cputime strongly depends on the implementation of the algorithms and for this reason the computational cost might be preferred. However in the present case, numerical complexities of compared algorithms involve subroutines such as truncated SVDs whose numerical complexity are hard to evaluate with precision. Furthermore, all the algorithms compared in this section were carefully implemented in house in Matlab language and optimized for it.

We have implemented two versions of ALS. In the first version (ALS-1), the ALS procedure is stopped when the relative difference between two successive values of  $r_T$  is lower than  $10^{-4}$  or when  $n_{it}$  reaches 50. In the second version (ALS-2), we set these two values to  $10^{-8}$  and 200. SJDTE is stopped when the relative dif-



**Table 1** Average reconstruction error, estimation error, number of iterations and cputime.

Algorithm	Scenario 1 : $N = 5$				Scenario 2 : $N = 15$				Scenario 3 : $N = 30$			
	$r_T$	$r_F$	$n_{it}$	$t_{cpu}$	$r_T$	$r_F$	$n_{it}$	$t_{cpu}(s)$	$r_T$	$r_F$	$n_{it}$	$t_{cpu}$
ALS-1	0.2	0.35	7	3.41	0.22	0.43	10	7.02	0.2	0.42	13	17.5
ALS-2	0.16	0.29	10	14.5	0.17	0.34	16	52.8	0.16	0.34	21	167
DIAG-SJDTE	0.01	0.0006	4	5.78	0.012	0.005	5	38.9	0.014	0.0162	5	275
DIAG-SJDTE-eco	0.01	0.0016	4	4.36	0.013	0.006	5	7.36	0.017	0.019	5	14

**Fig. 1** Distributions of the estimation error for the four algorithms according to the value of  $N$ .

ference between two successive values of  $C$  is lower than  $10^{-3}$  or when  $n_{it}$  reach 10. Comparisons are made by means of Monte-Carlo (MC) simulations. For each MC run, three new factor matrices of size  $200 \times N$  are randomly drawn from a normal distribution and a new tensor is built from the CPD model. A white Gaussian noise is then added to its entries in order to obtain a signal to noise ratio of 40 dB. Then the four algorithms are run to compute the CPD of rank  $N$  of the noisy tensor. We distinguish three scenarios according to the chosen value of  $N$ :  $N = 5$  (scenario 1),  $N = 15$  (scenario 2) and  $N = 30$  (scenario 3). For each scenario, average values of  $r_T$ ,  $r_F$ ,  $n_{it}$  and  $t_{cpu}$  are computed from 1000 MC runs. Results are reported in table 1 for each algorithm. In order to have a more precise idea of the convergence rate of the algorithms, we show in figure 1 the distribution of  $r_F$  in the ranges  $[10^{-4}; 10^{-3}]$ ,  $[10^{-3}; 10^{-2}]$ ,  $[10^{-2}; 10^{-1}]$  and  $[10^{-1}; 1]$ . Convergence problems of ALS clearly appear from these results. Whatever the considered scenario, the average value of  $r_T$  and of  $r_F$  remains high for both ALS-1 and ALS-2. Figure 1 shows that ALS behavior is binary. For instance in the first scenario ( $N = 5$ ), less than 60% of the values of  $r_F$  fall in the range  $[10^{-4}; 10^{-3}]$  and all the other values are greater than  $10^{-1}$ . Moreover, the proportion of  $r_F$  values below  $10^{-1}$  dramatically decreases with  $N$ : 25% for  $N = 15$  and 7% for  $N = 30$  with ALS-2. In these conditions, cputimes of ALS-1 are very low and compete with those of DIAG but considering the previous observation about the convergence rates, these values are misleading. Indeed, in practice ALS should be run from different starting point in order to obtain satisfying convergence hence increasing the total cputime. Furthermore, comparing ALS-1 and ALS-2 results, it appears that decreasing the threshold of the stopping criterion had little impact on the convergence. Conversely, DIAG-SJDTE and DIAG-SJDTE-eco offer good results in term of average reconstruction and estimation errors. This is mainly due

to good convergence rates as it can be seen on figure 1, with a little advantage for DIAG-SJDTE. In addition, these performances are quite stable with respect to the value of  $N$ . For instance for  $N = 30$  more than 80% of  $r_F$  values are still lower than  $10^{-2}$ . Now, regarding the average cputime, DIAG-SJDTE-eco is very less time consuming than DIAG-SJDTE for  $N = 15$  (7s against 39s) and  $N = 30$  (14s against 275s) whereas the average iteration numbers of both algorithms is the same. Considering the small difference between both algorithms regarding  $r_F$  criterion, we can thus clearly recommend the use of DIAG-SJDTE-eco when  $N$  is large.

## 5 Conclusion

We have proposed an original JEVD algorithm and showed how it can help for computing the canonical polyadic decomposition of large tensors. Preliminary results showed in this work point out that this approach provides very good convergence rates comparing to a reference CPD algorithm. Moreover it converges in very few iterations and the computing times are very low, including for high rank CPD. Further studies will be conducted to refine this conclusion. In particular, we want now to evaluate the impact of the choice of the subset of matrices  $\mathbf{M}^{(k)}$  and of the JEVD algorithm inside the DIAG procedure.

## References

1. Acar, E., Dunlavy, D.M., Kolda, T.G.: A scalable optimization approach for fitting canonical tensor decompositions. *Journal of Chemometrics* **25**(2), 67–86 (2011)
2. Andre, R., Trainini, T., Luciani, X., Moreau, E.: A fast algorithm for joint eigenvalue decomposition of real matrices. In: *European Signal Processing Conference (EUSIPCO 2015)*
3. Comon, P., Luciani, X., de Almeida, A.L.F.: Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics* **23** (2009)
4. Fu, T., Gao, X.: Simultaneous diagonalization with similarity transformation for non-defective matrices. In: *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2006)*, vol. 4, pp. 1137–1140
5. Hajipour, S., Albera, L., Shamsollahi, M., Merlet, I.: Canonical polyadic decomposition of complex-valued multi-way arrays based on simultaneous schur decomposition. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pp. 4178–4182
6. Harshman, R.A.: Foundation of PARAFAC procedure: Models and conditions for an 'explanatory' multi-mode factor analysis. *UCLA working papers in Phonetics* (16), 1–84 (1970)
7. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM Review* **51**(3), 455–500 (2009)
8. Luciani, X., Albera, L.: Canonical polyadic decomposition based on joint eigenvalue decomposition. *Chemometrics and Intelligent Laboratory Systems* **132**(0), 152 – 167 (2014)
9. Luciani, X., Albera, L.: Joint eigenvalue decomposition of non-defective matrices based on the lu factorization with application to ica. *IEEE Transactions on Signal Processing* **63**(17), 4594–4608 (2015)
10. Roemer, F., Haardt, M.: A semi-algebraic framework for approximate {CP} decompositions via simultaneous matrix diagonalizations (secsi). *Signal Processing* **93**(9), 2722 – 2738 (2013)

# On the use of Google Trend data as covariates in nowcasting: Sampling and modeling issues

## *L'utilizzo dei dati Google Trend come covariate per il nowcasting: problemi di campionamento e modellizzazione*

M.Simona Andreano, Roberto Benedetti, Paolo Postiglione, Giovanni Savio

**Abstract** The use of Big-data, and more specifically of Google Trend data, in now-and forecasting, has become common practice, even by Institutes and Organizations in charge of producing official statistics around the world. However, such data will have many implications in the model estimation, which can roughly impact final results. In this paper, starting from a MIDAS-AR model with Google Trend covariate, we are focussing on the main issues concerning the sampling error and the time domain context.

**Abstract** *L'uso di Big-data, e più nello specifico di Google Trend data, è divenuto prassi comune nell'ambito delle previsioni e del nowcasting, anche per gli istituti nazionali ufficiali. In realtà, il ricorso a tali dati pone diverse problematiche nell'ambito della stima del modello, che possono avere rilevanti ripercussioni sul risultato finale. Nel presente lavoro, partendo dalla stima di un modello MIDAS-AR con covariata Google Trend, si sono affrontate le principali problematiche riguardanti l'errore campionario e le specificità nel dominio delle serie storiche.*

**Key words:** Google trend, MIDAS model, repeated survey.

## 1 Introduction

The emergence of Big Data, and their capacity to help in now-casting, forecasting, disaggregating, and filling in gaps of conventional statistics data sources, is now history. However, the use of Big Data in economic now-casting and forecasting is still full of open questions, which concerns, among others:

- (a) representativeness of Internet data sources;
- (b) the synthesis of the information contained in the data;
- (c) the presence of non stationarity and seasonality
- (d) the estimation methods and modelling for disaggregating purposes;
- (e) the now- or forecasting model evaluation.

In the present paper we focus on the Mixed Data Sampling Models (MIDAS) with Google Trend as covariate to now- and forecasting a target variable  $y_t$   $h$ -step ahead, where the lowest frequency series  $y_t$  is regressed on the higher frequency one, through a distributed lag operator:

$$y_{t+mh} = y_{t_m+h_m} = \beta_0 + \beta_1 B(L^{1/m}; \theta) x_{t_m-\omega}^{(m)} + \varepsilon_{t_m+h_m}^{(m)}$$

and  $B(L^{1/m}; \theta) = \sum_{k=0}^K B(k; \theta) L_m^k$  denotes a weighting function,  $t$  indexes the basic time unit,  $m$  is the frequency mixture and  $\omega$  is the number of values of the indicators that are available earlier than the lower-frequency variable to be estimated.

In the next Section we will summarize the main issues arising from the estimation of such a model, with the purpose to highlight the open questions coming from the use of Big Data in empirical applications. Some final remarks will conclude the paper.

## 2 Modeling and sampling issues

Google Trend provides an index of the relative volume of search queries conducted through Google, and provides aggregated indices of search queries, which are classified into a total of 605 categories and sub-categories using an automated classification engine. Choi and Varian (2006, 2012) first showed the relevance of such data in predicting consumer behavior and initial unemployment claims for the

US. In our MIDAS-AR model the covariate  $x_t$  is the weekly query index of Google Trend, and it is used to nowcasting the (monthly) target variable  $y_t$ . For further properties on MIDAS models we refer to Ghysels et al. (2006a, 2006b).

First of all, we have to highlight, that Internet data sources, like Google Trend data, are not a probabilistic sample, but a self-selected sample created by the Internet users. Therefore, it can be a systematic bias in the sample of Internet people and if we ignore these biases and assume they will be resolved through sheer sample size, we compromise the utility of the findings of our research. For example, in the Google search of job, the Internet users will probably be young people, leaving in big cities. These biases could be considerable, when the forecasting model is applied for getting more granular information of the phenomena, and all aspects are disaggregated with respect to the geographical location, sex, group, sectoral activity etc.. This typical problem in the new era of big data arise because investigators are more likely to be separated from the data collection process and have less intimate knowledge about the texture and the quality of the many elements in their datasets.

The use of Google trend data in a time series domain causes other more specific shortcomings. The target variable  $y_t$  should be observed through a panel survey, where the same individual provides responses on repeated occasions. This is the case for Unemployment from Labor Force survey or for Consumption from Household survey, variables often forecasted through Google trend data (Askitas and Zimmermann, 2009; D'Amuri, 2009; Fondeur and Karamé, 2012; Schmidt and Vosen, 2010). In these cases, the sample overlap induces a correlation structure in the sampling errors of the time series of estimates, which affects the analysis of them. Estimators that ignore these correlations are generally inefficient relative to the minimum variance linear unbiased estimator (MVLUE). Bell and Wilcox (1993) assessed the sensitivity of parameter estimates for time series models of retail sales data to the treatment of sampling error, through the application of ARMA models on the sample error, while Binder and Dick (1989) used state-space models. More recently, Steel and McLaren (2009) examined the interaction between the design of a repeated survey and the methods used for estimation and reviewed the different forms of estimators.

The Google trend covariate  $x_t$  can also be seen as a time series drawn from a repeated survey, where the design is unknown. The same Internet-users will supposedly make their search repeated in the short term, therefore we have an overlap with adjacent days (weeks) observations that induces correlation.

Both these sampling error structure, coming from  $y_t$  and  $x_t$  should be properly being considered in our model estimation.

Always looking at the issues on the estimation of time series with MIDAS model, we need to spend some words on the presence of trend and seasonality in our variables  $y_t$  and  $x_t$ . To this regard, we outline two different questions: one arising from the aforementioned problem of sampling error in repeated survey observations, and the other from structural characteristics in the trend of Internet data.

In the first case, we note that the sampling errors can have important effects on the seasonal autocorrelation properties of the observed time series, and if we don't remove the survey error component, we may end up with spurious correlations as

part of our time series estimates, affecting the trend and seasonal components (Bell and Wilcox, 1993). Ignoring (seasonally correlated) sampling error can make seasonality appear much more variable than it appears, when the sampling error is accommodated in the model. The application of different seasonal adjustment procedures (X11-ARIMA or TRAMO-SEATS) will affect in a different way the final estimates of the model, with a lower bias when the model based procedure is applied. However, the identification of seasonality in the weekly Google trend data is not immediate and different seasonal frequencies may overlap on each other. Methods for producing variance estimates for seasonally adjusted and trend estimates have been considered by Wolter and Monsour (1981) and Pfeffermann (1994) and reviewed by Scott et al. (2005).

Secondly, we note that Google Trend data are proposed as a weekly query index that indicates the percentage deviation from the date to which the data are normalised and the series go back at most to 2004. However, the Internet users from 2004 until nowadays are significantly increased and the trend observed over this period should be affected by this tendency. Therefore, spurious trend relationships could occur when dealing with Google Trend data. One solution may be to remove the trend from the series and applying the MIDAS model on stationary time series  $y_t$  and  $x_t$ .

Until now we have not dealt with the problem concerning the synthesis of the information contained in the Internet data, because, in our case, we have only one Google trend covariate  $x_t$  and the topic is too wide to be exhaustively discussed here. We only note that, also in our case, before to choose the appropriate Google trend covariate to insert in the model, we needed to explore many different query options and a large-scale data analysis should be made. As pointed by Fisher et al. (2013), most data analytics developed for standard data reduction process may not be able to be applied directly to big data and there exist different efficient methods to solve the dimensionality problem: sampling, data condensation, density-based approaches, incremental learning, machine learning techniques, boosting, bagging, etc. In addition to the issues of data size, Laney (2001) presented a well-known definition (also called 3Vs) to explain what is the “big” data: volume, velocity, and variety. The 3Vs imply that the data size is large, will be created rapidly, and will be existed in multiple types and captured from different sources. These three characteristics strongly influence the choice of the appropriate reduction technique to apply on the data.

The majority of traditional forecasting techniques that perform relatively well in the case of standard data sets, are more likely to distort the accuracy of forecast when applied on Big data, because of the presence of high noise in Big data series. This suggests that there is a need for employing and evaluating the use of forecasting techniques, which can filter the noise in Big Data and forecast the signal alone. With Big data, there is an increased complexity in differentiating between randomness and statistically significant outcomes, as there is an increased chance of reporting a chance occurrence as a statistically significant outcome and misleading the stakeholders interested in the forecast. Nonlinear and non standard model are more appropriate when forecasting with Big data (and Google Trend data).

### 3 Conclusions

In the present paper we made an overview of the several problems arising from the estimation of a time series model with Google trend covariate, focussing on the main issues concerning the sampling and the time domain context. Thereafter we note a set of key challenges that at present hinder and restrict the accuracy and effectiveness of Big Data forecasts. Many questions are still open and a more depth analysis of the estimations troubles should be faced up, to avoid misleading forecast outcomes.

### References

1. Askatas, N., Zimmermann, K.: Google Econometrics and Unemployment Forecasting. IZA Discussion Paper, 4201 (2002)
2. Bell, W, Wilcox, D.: The effect of sampling error on the time series behavior of consumption data. *Journal of Econometrics*, 55, 235-265, (1993)
3. Binder, D.A., Dick, J.P.: Modelling and estimation for repeated surveys, *Survey Methodology* 15, 29-45, (1989).
4. Choi, H., Varian, H.: Predicting Initial Claims for Unemployment Insurance Using Google Trends', Technical report, Google. Available from: [http:// research.google.com/ archive/ papers/ initialclaimsUS.pdf](http://research.google.com/archive/papers/initialclaimsUS.pdf), (2006)
5. Choi, H., Varian, H.: Predicting the Present with Google Trends. *The Economic Records*, 88, 2-9 (2012)
6. D'Amuri, F.: Predicting unemployment in short samples with internet job search query data. MPRA Working Paper 18403, (2009)
7. Fischer U., Schildt, C., Hartmann, C., Lehner, W.: Forecasting the data cube: A model configuration advisor for multi-dimensional data sets. In: *IEEE 29th International conference on data engineering, (ICDE)*, Brisbane, 8–12, (2013)
8. Fondeur, Y., Karamè, F.: Can Google data help predict French youth unemployment? *Economic Modelling*, 30, 117-125 (2013)
9. Ghysels, E., Santa-Clara, P., Valkanov, R.: Predicting volatility: getting the most out of return data sampled at different frequencies. *J. of Econometrics*, 131, 59-95 (2006a)
10. Ghysels, E., Santa-Clara, P., Valkanov, R.: MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26, 53-90 (2006b)
11. Laney D (2001) 3D Data management: controlling data volume, velocity and variety. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Accessed 13 Dec 2013
12. Pfeiffermann, D.: A general method for estimating the variances of X-11 seasonally adjusted estimators. *Journal of Time Series Analysis*, 15, 85-116, (1994)
13. Schmidt, T., Vosen, S.: A monthly consumption indicator for Germany based on internet search query data, *Ruhr Economic Papers* 208 (2010)
14. Scott, S., Sverchkov, M.I, Pfeiffermann, D.: Variance measures for X-11 seasonal adjustment: A summing up of empirical work. *ASA Proceedings of the Joint Statistical Meetings*, 3534-3545. American Statistical Association (Alexandria, VA), (2005)
15. Steel, D., McLaren, C.: Design and Analysis of Surveys Repeated over Time, *Handbook of Statistics*, 29, 289-313, (2009)
16. Wolter, K.M., Monsour, N.J.: On the Problem of Variance Estimation for a Deseasonalised Series. In *Current Topics in Survey Sampling*, (eds. D. Krewski, R. Platek and J.N.K. Rao), New York, Academic Press, 367-403, (1981)





# A spatial decomposition of the change in urban poverty concentration

## *Una scomposizione spaziale della variazione nella concentrazione della povertà urbana*

Francesco Andreoli and Mauro Mussini

**Abstract** This paper explores the change in the concentration of poor individuals in the neighborhoods of a city, taking into account neighborhood locations on urban map. Urban poverty concentration is measured and the change over time in urban poverty concentration is broken down into different components. Each of these components is further split into spatial components explaining the extent to which spatial dependence affects the change in urban poverty concentration.

**Abstract** *L'articolo indaga la variazione nella concentrazione dei poveri nei quartieri di una città, considerando le posizioni dei quartieri sulla mappa urbana. Si misura la concentrazione della povertà urbana e si scompone la sua variazione nel tempo in diverse componenti. Ciascuna di queste componenti è divisa nelle sue componenti spaziali, che spiegano quanto la dipendenza spaziale influisca sulla variazione nella concentrazione della povertà urbana.*

**Key words:** Administrative data, Decomposition, Poverty, Spatial Inequality

## 1 Introduction

The growing availability of administrative data enables to develop research on poverty at a finer level of territorial disaggregation. When information on poverty status (poor or non-poor) for residents of neighborhoods in a city is complemented by spatial neighborhood information, the analysis of poverty distribution across neighborhoods can be linked with the analysis of spatial dependence in the dis-

---

Francesco Andreoli  
Luxembourg Institute of Socio-Economic Research, LISER. MSH, 11 Porte des Sciences, L-4366  
Esch-sur-Alzette/Belval Campus, Luxembourg, e-mail: francesco.andreoli@liser.lu

Mauro Mussini  
Department of Economics, University of Verona, Via Cantarane 24 - 37129 Verona, e-mail:  
mauro.mussini@univr.it

tribution of poverty in the city. This paper focuses on urban poverty concentration, which is measured by means of the Gini index. The change over time in the Gini index of urban poverty is broken down into three components, explaining the roles of changes in population proportions of neighborhoods, re-ranking of neighborhoods and changes in disparities between neighborhood poverty rates. Each component of the change in urban poverty concentration is further split into spatial components, separating the contribution of changes occurred between neighboring neighborhoods and that of changes occurred between non-neighboring neighborhoods. This decomposition over time and space is used to analyze the change in urban poverty concentration across the census tracts in the City of Los Angeles.

## 2 Concentrated poverty

The spatial distribution of poor people within urban space is considered, with urban space partitioned into  $n$  administrative units. The  $n$  administrative units detected by a space partition are referred to as neighborhoods. A city is hence partitioned into  $n$  neighborhoods. For instance, neighborhoods may be defined at the census tract level, however the setting can be extended to other geographic levels. Every individual living in a neighborhood is assigned with a poverty status (poor or non-poor) according to the fact that his income is below or above a poverty line. In this framework, a urban poverty configuration is a collection of counts of residents and poor residents across the city neighborhoods.

The concept of urban poverty concentration is here linked with the fact that poor residents tend to be distributed disproportionately across neighborhoods. It is a relative concept that can be expressed by comparing the distribution of poor population shares across neighborhoods with the distribution of population proportions across the same neighborhoods. Hence, one does not value the fact that in one urban poverty configuration there are more poor individuals than in another, but rather that the proportion of poor individuals in the population is larger and less evenly spread out across neighborhoods in one configuration compared to another. To measure the degree of concentration of poor individuals across neighborhoods, the Gini index is used. The Gini index of urban poverty is expressed by applying the matrix formulation of the Gini index suggested by Mussini and Grossi [2] and further developed by Mussini [1], as this matrix expression is useful to decompose the change over time in the index and to measure the spatial components of this change.

Let  $\mathbf{p} = (p_1, \dots, p_n)^T$  be the  $n \times 1$  vector of neighborhood poverty rates sorted in decreasing order and  $\mathbf{s} = (s_1, \dots, s_n)^T$  be the  $n \times 1$  vector of the corresponding population shares.  $\mathbf{1}_n$  being the  $n \times 1$  vector with each element equal to 1,  $\mathbf{P}$  is the  $n \times n$  skew-symmetric matrix:

$$\mathbf{P} = \frac{1}{\bar{p}} (\mathbf{1}_n \mathbf{p}^T - \mathbf{p} \mathbf{1}_n^T) = \begin{bmatrix} \frac{p_1 - p_1}{\bar{p}} & \dots & \frac{p_n - p_1}{\bar{p}} \\ \vdots & \ddots & \vdots \\ \frac{p_1 - p_n}{\bar{p}} & \dots & \frac{p_n - p_n}{\bar{p}} \end{bmatrix}, \quad (1)$$

where  $\bar{p}$  is the overall poverty rate in the city. The elements of  $\mathbf{P}$  are the  $n^2$  relative pairwise differences between the neighborhood poverty rates as ordered in  $\mathbf{p}$ . Let  $\mathbf{S} = \text{diag}\{\mathbf{s}\}$  be the  $n \times n$  diagonal matrix with diagonal elements equal to the population shares in  $\mathbf{s}$ , and  $\mathbf{G}$  be a  $n \times n$   $G$ -matrix (a skew-symmetric matrix whose diagonal elements are equal to 0, with upper diagonal elements equal to  $-1$  and lower diagonal elements equal to  $1$ ) [4]. The Gini index of urban poverty is expressed in matrix form:

$$G(\mathbf{s}, \mathbf{p}) = \frac{1}{2} \text{tr}(\tilde{\mathbf{G}} \mathbf{P}^T), \quad (2)$$

where the matrix  $\tilde{\mathbf{G}} = \mathbf{S} \mathbf{G} \mathbf{S}$  is the weighting  $G$ -matrix, a generalization of the  $G$ -matrix introduced by Mussini and Grossi [2] to add weights in the calculation of the Gini index.

### 3 Decomposing changes in urban poverty concentration

Suppose that poverty rates and population shares of  $n$  neighborhoods are observed in times  $t$  and  $t + 1$ . Let  $\mathbf{p}_t$  be the  $n \times 1$  vector of the  $t$  poverty rates sorted in decreasing order and  $\mathbf{s}_t$  be the  $n \times 1$  vector of the corresponding population shares. Let  $\mathbf{p}_{t+1}$  be the  $n \times 1$  vector of the  $t + 1$  poverty rates sorted in decreasing order and  $\mathbf{s}_{t+1}$  be the  $n \times 1$  vector of the corresponding population shares. The change in urban poverty concentration from  $t$  to  $t + 1$  is measured by the difference between the Gini index in  $t + 1$  and the Gini index in  $t$ :

$$\Delta G = G(\mathbf{s}_{t+1}, \mathbf{p}_{t+1}) - G(\mathbf{s}_t, \mathbf{p}_t) = \frac{1}{2} \text{tr}(\tilde{\mathbf{G}}_{t+1} \mathbf{P}_{t+1}^T) - \frac{1}{2} \text{tr}(\tilde{\mathbf{G}}_t \mathbf{P}_t^T). \quad (3)$$

Equation 3 can be broken down into three components explaining the roles of changes in population shares, ranking of neighborhoods and disparity of poverty rates. Let  $\mathbf{p}_{t+1|t}$  be the  $n \times 1$  vector of  $t + 1$  neighborhood poverty rates sorted in decreasing order of the respective  $t$  neighborhood poverty rates, and  $\mathbf{B}$  be the  $n \times n$  permutation matrix re-arranging the elements of  $\mathbf{p}_{t+1}$  to obtain  $\mathbf{p}_{t+1|t}$ , that is  $\mathbf{p}_{t+1|t} = \mathbf{B} \mathbf{p}_{t+1}$ . Let  $\lambda = \bar{p}_{t+1} / \bar{p}_{t+1|t}$  be the ratio of the actual  $t + 1$  overall poverty rate to the fictitious  $t + 1$  overall poverty rate which is the weighted average of  $t + 1$  poverty rates where the weights are the corresponding population shares in  $t$ . Matrix  $\mathbf{P}_{t+1|t} = (1/\bar{p}_{t+1|t}) (\mathbf{1}_n \mathbf{p}_{t+1|t}^T - \mathbf{p}_{t+1|t} \mathbf{1}_n^T)$  contains the  $n^2$  relative pairwise differences between the neighborhood poverty rates as arranged in  $\mathbf{p}_{t+1|t}$ . Applying the Mussini and Grossi decomposition [2], the change in urban poverty concentration between  $t$  and  $t + 1$  is split into three components:

$$\Delta G = \frac{1}{2}tr(\mathbf{W}\mathbf{P}_{t+1}^T) + \frac{1}{2}tr(\mathbf{R}\lambda\mathbf{P}_{t+1}^T) - \frac{1}{2}tr(\tilde{\mathbf{G}}_t\mathbf{D}^T) = W + R - D, \quad (4)$$

where  $\mathbf{W} = \tilde{\mathbf{G}}_{t+1} - \lambda\tilde{\mathbf{G}}_{t|t+1}$ ,  $\mathbf{R} = \tilde{\mathbf{G}}_{t|t+1} - \mathbf{B}^T\tilde{\mathbf{G}}_t\mathbf{B}$  and  $\mathbf{D} = \mathbf{P}_t - \mathbf{P}_{t+1|t}$ . Component  $W$  measures the effect of changes in the population shares of neighborhoods. A positive value of  $W$  indicates that the weights assigned to more unequal pairs of neighborhoods are larger in  $t+1$  than in  $t$ , increasing urban poverty concentration from  $t$  to  $t+1$ . A negative value of  $W$  indicates that the weights assigned to more unequal pairs of neighborhoods are smaller in  $t+1$  than in  $t$ , reducing urban poverty concentration from  $t$  to  $t+1$ . Component  $R$  measures the effect of re-ranking of neighborhoods from  $t$  to  $t+1$  and its contribution to the change in urban poverty concentration is always non-negative. The nonzero elements of  $\mathbf{R}$  detect the pairs of neighborhoods which have re-ranked from  $t$  to  $t+1$ . Component  $D$  measures the effect of disproportionate change between neighborhood poverty rates. The generic  $(i, j)$ -th element of  $\mathbf{D}$  compares the relative difference between the  $t$  poverty rates of the neighborhoods in positions  $j$  and  $i$  in  $\mathbf{p}_t$  with the relative difference between the  $t+1$  poverty rates of the same two neighborhoods in  $\mathbf{p}_{t+1|t}$ . A positive value of  $D$  means that relative disparities in poverty rates have overall decreased from  $t$  to  $t+1$ , reducing urban poverty concentration. A negative value of  $D$  indicates that relative disparities in poverty rates have overall increased from  $t$  to  $t+1$ , increasing urban poverty concentration. If all neighborhood poverty rates have changed by the same proportion from  $t$  to  $t+1$ , then  $D = 0$ .

### 3.1 The spatial components of urban poverty concentration

The components of the change in urban poverty concentration quantify the impacts of different distributional changes on urban poverty, however they do not explain the extent to which these changes have occurred between neighborhoods which are geographically close or not. The spatial location of neighborhoods is neglected by  $\Delta G$ ,  $W$ ,  $R$  and  $D$  as they would remain the same if neighborhoods exchanged their positions on urban map. The spatial components of  $\Delta G$ ,  $W$ ,  $R$  and  $D$  can be separated by using the approach suggested by Rey and Smith [3] to decompose the Gini index into a neighbor component of inequality and a non-neighbor component of inequality.

Let  $\mathbf{N}_t$  be the  $n \times n$  binary spatial weights matrix having its  $(i, j)$ -th entry equal to 1 if and only if the  $(i, j)$ -th element of  $\mathbf{P}_t$  is the relative difference between the poverty rates of two neighboring neighborhoods, otherwise the  $(i, j)$ -th element of  $\mathbf{N}_t$  is 0. Using the Hadamard product,<sup>1</sup> the relative pairwise differences between the poverty rates of neighboring neighborhoods can be selected from  $\mathbf{P}_t$ :

$$\mathbf{P}_{N,t} = \mathbf{N}_t \odot \mathbf{P}_t. \quad (5)$$

<sup>1</sup> Let  $\mathbf{X}$  and  $\mathbf{Y}$  be  $k \times k$  matrices. The Hadamard product  $\mathbf{X} \odot \mathbf{Y}$  is defined as the  $k \times k$  matrix with the  $(i, j)$ -th element equal to  $x_{ij}y_{ij}$ .

For each pair of neighborhoods, the relative difference between the  $t + 1$  poverty rates of the two neighborhoods in  $\mathbf{P}_{t+1|t}$  has the same position as the relative difference between their  $t$  poverty rates in  $\mathbf{P}_t$ . Thus,  $\mathbf{N}_t$  also selects the relative pairwise differences between neighboring neighborhoods from  $\mathbf{P}_{t+1|t}$ :

$$\mathbf{P}_{N,t+1|t} = \mathbf{N}_t \odot \mathbf{P}_{t+1|t}. \quad (6)$$

Since  $\mathbf{D} = \mathbf{P}_t - \mathbf{P}_{t+1|t}$ , the Hadamard product between  $\mathbf{N}_t$  and  $\mathbf{D}$  produces the matrix with nonzero elements equal to the elements of  $\mathbf{D}$  pertaining to neighboring neighborhoods:

$$\mathbf{D}_N = \mathbf{P}_{N,t} - \mathbf{P}_{N,t+1|t} = \mathbf{N}_t \odot (\mathbf{P}_t - \mathbf{P}_{t+1|t}) = \mathbf{N}_t \odot \mathbf{D}. \quad (7)$$

$\mathbf{P}_{N,t+1}$  being the  $n \times n$  matrix whose nonzero elements are the relative pairwise differences between the  $t + 1$  poverty rates of neighboring neighborhoods, the decomposition of the change in the neighbor component of urban poverty concentration is obtained by replacing  $\mathbf{P}_{t+1}$  and  $\mathbf{D}$  in equation 4 with  $\mathbf{P}_{N,t+1}$  and  $\mathbf{D}_N$  respectively:

$$\Delta G_N = \frac{1}{2} \text{tr}(\mathbf{W}\mathbf{P}_{N,t+1}^T) + \frac{1}{2} \text{tr}(\mathbf{R}\lambda\mathbf{P}_{N,t+1}^T) - \frac{1}{2} \text{tr}(\tilde{\mathbf{G}}_t\mathbf{D}_N^T) = W_N + R_N - D_N. \quad (8)$$

$\mathbf{P}_{nN,t+1}$  and  $\mathbf{D}_{nN}$  being the matrices with the relative pairwise differences between non-neighboring neighborhoods, the decomposition of the change in the non-neighbor component of urban poverty concentration is obtained by replacing  $\mathbf{P}_{t+1}$  and  $\mathbf{D}$  in equation 4 with  $\mathbf{P}_{nN,t+1}$  and  $\mathbf{D}_{nN}$  respectively:

$$\Delta G_{nN} = \frac{1}{2} \text{tr}(\mathbf{W}\mathbf{P}_{nN,t+1}^T) + \frac{1}{2} \text{tr}(\mathbf{R}\lambda\mathbf{P}_{nN,t+1}^T) - \frac{1}{2} \text{tr}(\tilde{\mathbf{G}}_t\mathbf{D}_{nN}^T) = W_{nN} + R_{nN} - D_{nN}. \quad (9)$$

Given equations 8 and 9, the decomposition over time and space is

$$\Delta G = \Delta G_N + \Delta G_{nN} = W_N + W_{nN} + R_N + R_{nN} - (D_N + D_{nN}) = W + R - D. \quad (10)$$

## 4 Application

The decomposition is used to analyze the change in urban poverty concentration in the City of Los Angeles from 1980 to 2014. The administrative units are the census tracts [5]. For each census tract, poverty line is known in both 1980 and 2014. To check for spatial autocorrelation in poverty distribution across census tracts, the Rey and Smith test based on random permutations is applied [3]. The hypothesis of randomness in poverty distribution is rejected in both 1980 and 2014.<sup>2</sup> Table 1 shows the spatial decomposition of each component of the change over time in the Gini index of urban poverty. Most of urban poverty concentration is explained by the

<sup>2</sup> The pseudo p-value obtained from 99 permutations is equal to 0.01 in both 1980 and 2014.

Table 1: Decomposition over time and space, Los Angeles, 1980-2014.

component	$G_{2014}$	$G_{1980}$	$\Delta G$	$W$	$R$	$D$
$N$	0.10111	0.10991	-0.00880	-0.00145	0.02405	0.03140
$nN$	0.27417	0.30090	-0.02674	-0.00314	0.05852	0.08212
total	0.37527	0.41082	-0.03554	-0.00459	0.08257	0.11352

disparities between poverty rates of non-neighboring census tracts in both 1980 and 2014, as the non-neighbor component of the Gini index of urban poverty overcomes the neighbor component. Urban poverty concentration decreases from 1980 to 2014. The decrease of disparities between poverty rates (0.11352) plays a major role in reducing urban poverty concentration, however its equalizing effect is partially offset by the impact of re-ranking (0.08257). The change in the relative frequency distribution of population across census tracts reduces urban poverty concentration, but it plays a minor role in the reduction of urban poverty concentration ( $-0.00459$ ).

## 5 Conclusion

A decomposition of the change over time in urban poverty concentration is shown. The decomposition links inequality in poverty distribution across city neighborhoods with the spatial dependence in poverty distribution. The decomposition explains the roles of changes in population distribution across neighborhoods, re-ranking of neighborhoods and changes in disparities between neighborhood poverty rates. Each component of the change in urban poverty concentration is broken down into spatial components.

## References

- [1] Mussini M (2017) Decomposing Changes in Inequality and Welfare Between EU Regions: The Roles of Population Change, Re-Ranking and Income Growth. *Social Indicators Research* 130:455–478
- [2] Mussini M, Grossi L (2015) Decomposing changes in CO<sub>2</sub> emission inequality over time: the roles of re-ranking and changes in per capita CO<sub>2</sub> emission disparities. *Energy Economics* 49:274–281
- [3] Rey SJ, Smith RJ (2013) A spatial decomposition of the Gini coefficient. *Letters in Spatial and Resource Sciences* 6:55–70
- [4] Silber J (1989) Factor components, population subgroups and the computation of the Gini index of inequality. *The Review of Economics and Statistics* 71:107–115
- [5] United States Census Bureau (2015) Metropolitan and micropolitan statistical areas data. <https://www.census.gov/population/metro/>, accessed 12/30/2016

# How green advertising can impact on gender different approach towards sustainability

## *L'impatto della "pubblicità verde" sul diverso approccio di genere alla sostenibilità.*

Margaret Antonicelli, Vito Flavio Covella

### **Abstract**

In the last decades, concerns on protection of the environment have really increased among consumers. Initially, people were interested in discovering main environmental problems but, actually, consumers have started to exercise their decision making process in the purchase of products.

Performing a first pre-test and subsequently the final analysis, the probit model study analyses both the statistical/econometric and the substantive significance of gender differences in customer expectations, considering the "effect" of a green advertising. This model is estimated jointly with an ordered probit model analyzing the magnitude of this different gender approach. Results of the joint estimation and the conventional single equation ordered probit model were presented for comparison.

### **Abstract**

*Negli ultimi decenni, le preoccupazioni in materia di tutela dell'ambiente sono notevolmente aumentate tra i consumatori. Inizialmente, le persone erano interessate a scoprire i principali problemi ambientali, ma, in realtà, i consumatori hanno iniziato ad esercitare il loro processo decisionale per l'acquisto di prodotti.*

*Effettuando prima un pre test e poi l'analisi finale, il modello probit utilizzato in questo studio analizza sia da un punto di vista statistico/econometrico che sostanziale le differenze di genere nelle aspettative dei clienti, considerando l'effetto della "una pubblicità verde". Il modello probit ordinato sottolinea la grandezza di questo diverso approccio di genere. I risultati della stima congiunta e della singola equazione del modello probit convenzionale sono risultati fondamentali per effettuare il confronto.*

**Key words:** Green advertising, gender difference, probit model, econometric approach

## 1 Introduction

Although much has been written about sustainable durable goods consumption in the last few decades, obtaining reliable information on consumer preferences for new social/ethical and eco-labeled products can be an arduous task.

The literature on business ethics, corporate social responsibility and sustainability includes many studies on gender differences, however the results are often contrasting. In particular, there has not yet been full agreement on the role and significance of gender differences in customer expectations and perceptions of responsible corporate conduct. The current study analyses both the statistical and the substantive significance of gender differences in customer expectations and perceptions of corporate responsibility, also examining the influence of age and education<sup>1</sup>. In particular, the purpose of this study was to understand how male and female consumers differently evaluate sustainability claims from brands and how brands' sustainability efforts and the presence/absence of information transparency in the claims affect their brand schemas differently.

## 2 Literature review and hypothesis development

Recently, societies have become more concerned about environment protection<sup>2</sup>. As a result, many consumers are modifying their consumption practices, choosing products with reduced environmental impacts<sup>3</sup>. Sustainable consumption, in this study, refers to the purchase and use of products with lower environmental impacts<sup>4</sup> and that result in pro-social behaviours<sup>5</sup>. The first big observed phenomenon is the change concerning conscious consumer attention: if once they were more interested in the sustainability of products, today, the green community looks first at sustainable approach of companies and only later, to the products.

These observations underline the importance of communication in sustainable reputation creation process: in fact, in order to consider a sustainable company, it is important that "the company communicates in a transparent way to the consumer"<sup>6</sup>. Reputation that is fundamental when choosing the product to buy.

In this way, communication and reputation have a primary role in sustainability. According to past research, sustainable consumers are mainly female, aged between 30 and 44 years old, well educated, in a household with a high annual income.

---

<sup>1</sup> Calabrese A., Costa R., Rosati F., Gender differences in customer expectations and perceptions of corporate social responsibility

<sup>2</sup> Corraliza and Berenguer, 2000

<sup>3</sup> Schaefer and Crane, 2005

<sup>4</sup> Follows and Jobber, 2000; Pedersen, 2000; Gordon et al., 2011

<sup>5</sup> Diego Costa Pinto, Marcia M. Herter, Patricia Rossi, Adilson Borges, 2014

<sup>6</sup>Roveda, 2014



Female participants are more likely to engage in sustainable consumption because they hold stronger attitudes towards the environment than male participants<sup>7</sup>. In addition, women tend to be more socially responsible, environmentally concerned and ecologically conscious than men and tend also to consider the impacts that their consumption may cause on others more carefully than men<sup>8</sup>. Female participants are also more willing to change their lifestyle in order to reduce the negative environmental impacts of consumption than male participants<sup>9</sup> and are willing to buy and to pay more for an environmentally friendly product than male participants. Furthermore, studies showed that the adoption of sustainable practices may depend upon reasons beyond conservation of the environment<sup>10</sup>. In this work, we thus developed the following hypothesis: high propensity and high knowledge for men to buy sustainable durable goods and extensibility of the results obtained in the pre-test to the entire sample.

### **3. Methodology**

#### ***3.1 Data***

The present study analyses both the econometric and the substantial significance of gender differences in customer expectations, as well as the perception of corporate responsibility, by additionally examining the influence of age and education. The work, based on the Italian macro-context, explains sustainable discourses in advertising. The aim is to define sustainability broadly and explain the issue of inequality, particularly gender inequality, as originating in various forms of ascendancy over nature. More specifically, was drawn up a quantitative survey on the attitude of Italian citizens towards production systems. A cross-sectional survey has been performed to a sample of no less than 1200 units on the entire Italian territory. Delving into more detail, the questionnaire employed has been pre-tested to reduce error through possible misinterpretation. Regarding the models used, this work is based on a comparison of three different methods: Joint estimation of Probit and Ordered Probit and Single Equation Model.

---

<sup>7</sup> Diamantopoulos et al., 2003; Jain and Kaur, 2006

<sup>8</sup> Roberts, 1996b; Mainieri et al., 1997; Straughan and Roberts, 1999; Noble et al., 2006

<sup>9</sup> Abeliotis et al., 2010

<sup>10</sup> Diego Costa Pinto, Marcia M. Herter, Patricia Rossi, Adilson Borges, 2014

### ***3.2 Empirical results***

Previous studies have identified a variety demographic and attitudinal characteristic that may affect consumer propensity to buy sustainable durable goods. For empirical implementation, the explanatory variables of equation are gender, age and education. In addition, the importance of level of knowledge and satisfaction about sustainability durable goods, frequency of purchase of sustainable durable goods, willingness to pay an additional fee to sustainable durable goods and level of information available on the sustainability. Specifically, it is expected that female would be more attentive to green advertising and would be more willing to buy a sustainable durable goods if sustainability were considered as an important attribute to making produce purchases. The maximum likelihood estimates of the generalized binary- ordinal probit model are presented in Table 1 about entire analysis . For comparison, results of the conventional single-equation ordered probit estimation are also presented. It is evident from Table 1 that the single-equation model performs poorly as compared to the binary- ordinal probit model judging from pseudo-R2s that were computed as a measure of goodness-of-fit for to estimated models. In general, regarding to the hypothesis, after finding extensibility of the pre-test, results showed the main effects of gender and identity on sustainable consumption. In particular, this research suggests that female participants will have higher levels of sustainable consumption than male participants. This evidence it is verified in all three models, observing “gender”. The results provide further evidence for past research<sup>11</sup>, suggesting that female participants are likely to engage more in sustainable consumption than male participants. It is also important to emphasize that, not only is the second hypothesis is verified but also in the second model the great majority of the independent variables result to be highly significant.

---

<sup>11</sup> Roberts, 1996a

**Table 1:** Result of Joint estimation of Probit and Ordered Probit and Single Equation Model

Variable	Joint estimation				Single equation estimation
	Probit		Ordered		ordered probit
Constant	1.71845	***	2.27266	***	2.060921 **
	(-3.801)		-1.6015		(-1.349)
Gender	0.01009	**	0.14198	***	0.19281 **
	(-1.099)		-0.0465		-0.2014
Age	-0.00593		-0.01102	***	-0.01382 **
	(-6.012)		(-0.4491)		(-0.1408)
Education	-0.07423	*	0.07281	***	0.01545
	(-14.719)		-0.2493		(0.3556)
Level of knowledge about sustainability durable goods	0.03162	**	-0.31418	***	-0.40419 *
	(-0.170)		(-0.0728)		(-1.3193)
Willingness to pay an additional fee to sustainable durable goods	-0.05559	*	-0.06562	**	-0.07148 **
	(-1.536)		(-1.8640)		(-0.5389)
Frequency of purchase of sustainable durable goods	0.41152		0.17971	*	0.08281
	-12.126		-0.9323		-0.1347
Level of satisfaction in knowing that the property purchased is sustainable	-0.19884	**	0.1857	**	0.11809 *
	(-3.586)		-1.3396		-0.1984
Level of information available on the sustainability	-0.53425	*	0.51809	**	0.44372 **
	(-4.412)		-1.1324		-1.0803
$\mu_1$			1.303	**	0.812 **
			-0.5483		(9.821)
$\mu_2$					1.91 **
					-17.787
$\beta$			0.147		
			-3.522		
Log likelihood			-407.065		-421.681
Pseudo R2			0.339		-0.245
Sample	1200		1136		1200

Numbers in parentheses are t-ratios

\*Indicates statistical significance at the level 0,10, \*\* 0,005 and \*\*\* 0,001

## Discussion

Despite the large literature regarding to sustainability, it is clear that a higher level of information is a key factor for a positive acceptance. It means that sustainability, with different communication policies, are able to create a good image, modifying consumer behaviour<sup>12</sup>. This study has one important limitation that can guide future studies: the sample consisted only Italian participants with access to Internet. Although such a sample may have biased the results, it is important to note that, in

<sup>12</sup> Antonicelli M., Calace D., Morrone D., Russo A., Vastola V., 2015

Italy, Internet penetration rate is 87,1% (Istat, 2016). Moreover, this could explain the predominance of young participants in the sample. Future research could focus on another method of data collection in order to consider a different, wider age range. This could make possible to investigate whether gender and identity effects change in different age groups. In addition, convenience sampling is a limitation of the study. Future studies could use representative samples to investigate the effects of gender and identities on sustainable consumption.

## References

1. ANTONICELLI M., CALACE D., MORRONE D., RUSSO A., VASTOLA V., 2015, Information or confusion? The role of ecolabels in agrifood sector, *Analele Universității din Oradea, Fascicula Ecotoxicologie, Zootehnie și Tehnologii de Industrie Alimentară*, Vol. XIV/A, pp. 187-195.
2. BERKOWITZ L., LUTTERMAN K.G., 1968, The traditional socially responsible personality, *Public Opinion Quarterly*, 32, 169–185.
3. CALABRESE A., COSTA R., ROSATI F., 2016. Gender differences in customer expectations and perceptions of corporate social responsibility, *Journal of Cleaner Production*, Vol. 116, pp. 135-149.
4. COSTA PINTO D., HERTER M. M., ROSSI P., BORGES A., 2014. Going green for self or for others? Gender and identify salience effects on sustainable consumption, *International Journal of Consumer Studies*, Vol. 38, pp. 540-549.
5. CHERRIER H., 2006, Consumer identity and moral obligations in nonplastic bag consumption: a dialectical perspective, *International Journal of Consumer Studies*, 30, 515–523.
6. CORRALIZA J.A., BERENGUER, J., 2000, Environmental values, beliefs, and actions: a situational approach, *Environment and Behavior*, 32, 832–848.
7. HORNE R.E., 2009, Limits to labels: the role of eco-labels in the assessment of product sustainability and routes to sustainable consumption, *International Journal of Consumer Studies*, 33, 175–182.
8. JAIN S.K., KAUR G., 2006, Role of socio-demographics in segmenting and profiling green consumers: an exploratory study of consumers in India, *Journal of International Consumer Marketing*, 18, 107–146.
9. LUCHS M., MOORADIAN T., 2012, Sex, personality, and sustainable consumer behaviour: elucidating the gender effect. *Journal of Consumer Policy*, 35, 127–144.
10. ROBERTS J., 1993, Sex differences in socially responsible consumers' behaviour. *Psychological Reports*, 73, 139–148.
11. ROBERTS J.A., 1996b, Green consumers in the 1990s: profile and implications for advertising, *Journal of Business Research*, 36, 217–231.
12. SALAZAR H.A., OERLEMANS L., VAN STROE-BIEZEN S., 2013, Social influence on sustainable consumption: evidence from a behavioural experiment, *International Journal of Consumer Studies*, 37, 172–180.
13. SCHAEFER A., CRANE A., 2005, Addressing sustainability and consumption, *Journal of Macromarketing*, 25, 76–92.
14. SCHULTZ P.W., 2001, The structure of environmental concern: concern for self, other people, and the biosphere, *Journal of Environmental Psychology*, 21, 327–339.
15. STOCK J. H., WATSON M. W., 2011, *Introduction to Econometrics*, 3/E, Pearson Higher Education

# Stratified data: a permutation approach for hypotheses testing

## *Dati stratificati: un approccio di permutazione per test d'ipotesi*

Rosa Arboretti, Eleonora Carrozzo, Luigi Salmaso

**Abstract** The present work aims at presenting a general nonparametric alternative to the well known van Elteren test for two-sample stratified analysis. We developed a solution based on permutation tests that considers the Nonparametric Combination (NPC) methodology for reducing the dimensionality of the problem. A simulation study to compare performances of proposed test with those of the usual van Elteren test and of aligned rank test has been performed considering both continuous and ordinal data. Results shows the respect of nominal  $\alpha$ -level under  $H_0$  even for small sample sizes. A real application example is also presented.

**Abstract** *Il presente lavoro ha l'obiettivo di proporre un'alternativa non parametrica generale al test di van Elteren per analisi a due campioni stratificati. La soluzione sviluppata è basata sui test di permutazione e considera la metodologia della Combinazione Non Parametrica (NPC) per ridurre la dimensionalità del problema. È stato eseguito uno studio di simulazione per confrontare le prestazioni del test proposto con quelle del test di van Elteren e del test Aligned Rank, considerando sia dati continui che ordinali. I risultati mostrano il rispetto dell' $\alpha$  nominale sotto  $H_0$  anche per basse numerosità campionarie. È inoltre presentata un'applicazione ad un caso reale.*

**Key words:** Stratified test, Nonparametric Combination methodology, Permutation tests

---

Carrozzo E.

Department of Management and Engineering, Univeristy of Padova, Stradella S. Nicola 3, Vicenza (Italy) e-mail: carrozzo@gest.unipd.it

Salmaso L.

Department of Management and Engineering, Univeristy of Padova, Stradella S. Nicola 3, Vicenza (Italy) e-mail: luigi.salmaso@unipd.it

Arboretti R.

Department of Civil Environmental and Architectural Engineering, Univeristy of Padova, Via Marzolo 9, Padova (Italy) e-mail: rosa.arboretti@unipd.it

## 1 Introduction

Let us suppose to have two treatments and we are interested in detecting differences among their effects. Suppose also that treatments may be influenced by a confounding factor which is taken into consideration by stratification. In this situation when performing the analysis we must take into consideration the presence of these strata.

Literature on stratified experiments is vast and in particular in the context of the so called multicenter clinical trials it revolves around the van Elteren test [10], which is the optimal test when there is no interaction among treatment effect and strata. However if treatment effect is not constant across strata, Van Elteren test can become inefficient to detect differences between treatments. Therefore in literature we found alternatives to Van Elteren test which present good operating characteristics [2, 4, 3, 5, 6].

Our interest on stratified tests arose dealing with a real industrial problem which was also affected by a very small sample size. Thus we wondered if existing tests are suitable for our purpose, and we decided to provide a general solution for the problem at hand, and make a comparison with existing ones.

In the present paper we want to describe our proposed solution for stratified problems, where variables can be of different nature (continuous, discrete, ordinal etc.). The proposed approach is nonparametric based on permutation tests and considers the NonParametric Combination (NPC) methodology [9] as tool to reduce the dimensionality of the problem. This implies that factors of stratification can be also more than one.

Section 2 is aimed at presenting and formalizing the problem. The idea at basis of the proposed procedure is described and after defining main notations and assumptions a detailed algorithm for achieving the NPC-based procedure is provided.

In Section 3 we report the results of a simulation study aimed at comparing the performance of our proposed method with that of van Elteren test and of the Aligned Rank test proposed in [6]. We consider small sample sizes commonly of interest in practice. We investigate the case in which effect were constant across strata and case where effect are varying across strata, both under the null hypothesis than under the alternative.

Finally, in order to illustrate usefulness of the proposed method in a practical context, in Section 4 we analyze data from an industrial problem.

## 2 NPC-based permutation test for stratified analysis

In the present section we describe the permutation approach proposed to deal with stratified problems. Such nonparametric solution is based on NonParametric Combination (NPC) methodology which allows to overcome methodological difficulties related to the presence of one or more stratification factors.

Let  $X_{ish} \sim F(x + \gamma_s + \delta_h)$  be a response variable for the  $i$ -th observation in the  $s$ -th stratum for the treatment  $h$ ,  $\gamma_s$  represents the location effect of stratum  $s$  and  $\delta_h$

is the effect of the treatment  $h$ ,  $i = 1, \dots, n_{hs}$ ,  $s = 1, \dots, S$ ,  $h \in \{A, B\}$ , and  $S$  is the number of strata.

We are interested in testing for:

$$\begin{cases} H_0^G : \delta_A = \delta_B \\ H_1^G : \delta_A \begin{matrix} (<) \\ (>) \end{matrix} \delta_B \end{cases} \quad (1)$$

Taking into account the possible effect of stratification factors, we break down the problem into sub-problems for each stratum, i.e.:

$$\begin{cases} H_{0(s)} : \delta_{A(s)} = \delta_{B(s)} \\ H_{1(s)} : \delta_{A(s)} \begin{matrix} (<) \\ (>) \end{matrix} \delta_{B(s)} \end{cases} \quad (2)$$

The idea of the NPC-based procedure is to suitably combine the p-values from each stratum. It is important to note that the effect of treatments may be multivariate. In order to clarify the steps of the procedure, in Section 2.1 we report the related algorithm.

## 2.1 An algorithm for NPC-based stratified test

In this section we describe the algorithm to achieve the NPC-based stratified procedure. For an overview on NPC-based testing and its properties see for example [1], [7], [8], [9]. Before listing the steps of the algorithm let us define some important notations. Given:

$$\mathbf{x}_h = [x_{1(h,1)} \cdots x_{n_{h_1}(h,1)} \cdots x_{1(h,S)} \cdots x_{n_{h_S}(h,S)}]$$

a sample of size  $n_h = \sum_{s=1}^S n_{hs}$  for treatment  $h \in \{A, B\}$  from an unknown distribution  $F$ , let us define the whole sub-sample  $x_{(s)}$  for stratum  $s$  of size  $n_{A_s} + n_{B_s}$ , and the related permuted sub-sample  $x_{(s)}^*$ , being  $u^*$  any random permutation of labels  $u = 1, \dots, (n_{A_s} + n_{B_s})$ :

$$x_{(s)} = [x_{1(A,s)} \cdots x_{n_{A_s}(A,s)} x_{n_{A_s}+1(B,s)} \cdots x_{n_{A_s}+n_{B_s}(B,s)}] \text{ and} \\ x_{(s)}^* = [x_{u_{[1]}^*(A,s)} \cdots x_{u_{[n_{A_s}]^*}^*(A,s)} x_{u_{[n_{A_s}+1]}^*(B,s)} \cdots x_{u_{[n_{A_s}+n_{B_s}]^*}^*(B,s)}].$$

These are the steps of the procedure:

1.  $\forall s = 1, \dots, S$ :
  - 1.1. In order to testing (2) on  $x_s$  compute a suitable test statistic  $T_{(s)}$ , for example the difference of means for continuous data:
$$T_{DM(s)} = \frac{1}{n_{A_s}} \sum_{i=1}^{n_{A_s}} x_{i(A,s)} - \frac{1}{n_{B_s}} \sum_{j=1}^{n_{B_s}} x_{j(B,s)}$$
or the Anderson Darling test statistic in case of ordinal data:
$$T_{AD(s)} = \sum_{i=1}^{v-1} N_{i(B_s)} [N_{i(\bullet s)} ((n_{A_s} + n_{B_s}) - N_{i(\bullet s)})]$$

where  $v$  is the number of categories,  $N_{i(\bullet s)} = N_{i(A_s)} + M_{i(B_s)}$  in which  $N_{i(A_s)}$  and  $N_{i(B_s)}$  are cumulative frequencies of the category  $i$  for stratum  $s$  in the treatment group  $A$  and  $B$  respectively (see [9, sec. 2.8.3]);

- 1.2. perform a random permutation  $u^*$  obtaining  $x_{(s)}^*$ ;
- 1.3. compute the permuted value of test statistics  $T^*$ ;
- 1.4. independently repeat  $B$  times step (1.2)-(1.3) to obtain the permutation distribution of test statistic  $T_{(s)}$ .
- 1.5. Estimate p-value:  $\lambda_s = \sum_{b=1}^B I(T_{(s)}^{*b} \geq T_{(s)})/B$  and the related empirical significance function  $\lambda_s^{*b} = \frac{1}{2} + \sum_{j=1}^B I(T_{(s)}^{*j} \geq T_{(s)}^{*b})/(B+1)$ ,  $b = 1, \dots, B$ ;
2. Through a suitable combination function  $\Phi(\cdot)$  combine the p-values related to different strata obtaining:  $T_{(\cdot)} = T_{(\cdot)} = \Phi(\lambda_1, \dots, \lambda_S)$  and their related distribution:  $T_{(\cdot)}^* = \Phi(\lambda_1^{*b}, \dots, \lambda_S^{*b})$ ,  $b = 1, \dots, B$ ;
3. compute the combined p-values  $\lambda_\Phi = \sum_{b=1}^B I(T_{(\cdot)}^{*b} \geq T_{(\cdot)})/B$ ;
4. reject the null hypothesis in (1) if  $\lambda_\Phi \leq \alpha$ .

### 3 A comparative simulation study

In the present section operating characteristics of procedure proposed in Section 2 are discussed compared with van Elteren test and the aligned rank test proposed in [6]. We performed 5000 Monte Carlo simulations based on  $B = 5000$  permutations for permutation tests. We considered the following simulation settings:

**Setting 1:**  $S = 3$ ;  $n_{h_s} = 12 \forall s = 1, \dots, S$  and  $h \in \{A, B\}$ , data generated from  $N(0 + \delta_h + \gamma_s, 1)$ ;

**Setting 2:**  $S = 3$ ;  $n_{h_s} = 12, \forall s = 1, \dots, S$  and  $h \in \{A, B\}$ , data generated from an ordinal variable with 10 categories;

**Setting 3:**  $S = 5$ ;  $n_{h_s} = 12 \forall s = 1, \dots, S$  and  $h \in \{A, B\}$  data generated from  $N(0 + \delta_h + \gamma_s, 1)$ ;

**Setting 4:**  $S = 5$ ;  $n_{h_s} = 12, \forall s = 1, \dots, S$  and  $h \in \{A, B\}$ , data generated from an ordinal variable with 10 categories;

where  $(\delta_A, \delta_B) = (0.50, 0.00)$  and for Setting 1:  $(\gamma_1, \gamma_2, \gamma_3) = (0.25, 0.5, 0.55)$  and for Setting 3:  $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = (0.50, 0.25, 0.15, 0.05, 0.00)$ .

Actually we started from sample size  $n_{h_s} = 5 \forall s = 1, \dots, S$  and  $h \in \{A, B\}$ , but the aligned rank test showed an anti-conservative behaviour so we decide to consider a sample size where nominal  $\alpha$  for all three procedures is respected, in order them to be comparable in power. Table 1 and Table 2 show rejection rates of the tests when treatment effect is constant across strata and when it varies across strata respectively.

For three testing procedures rejection rates are close to the nominal  $\alpha$  both in presence of normal and categorical variables. For comparisons under the alternative hypothesis we note that van Elteren test presents a lower power with respect to its



competitors. In particular NPC procedure presents an higher power among three compared tests. We can also note that power tends to increase when increasing the number of strata.

**Table 1** Rejection rates at significance level  $\alpha = 0.05$  over 5000 simulations, with treatment effect constant across strata.

		$A = B$		$A > B$	
		Normal	Ordinal Categorical	Normal	Ordinal Categorical
S=3	<i>NPC</i>	0.046	0.048	0.605	0.630
	<i>Align</i>	0.050	0.053	0.534	0.517
	<i>vE</i>	0.047	0.047	0.520	0.500
S=5	<i>NPC</i>	0.049	0.049	0.793	0.801
	<i>Align</i>	0.051	0.051	0.765	0.732
	<i>vE</i>	0.050	0.049	0.747	0.709

**Table 2** Rejection rates at significance level  $\alpha = 0.05$  over 5000 simulations, with treatment effect varying across strata.

		$A = B$		$A > B$	
		Normal	Ordinal Categorical	Normal	Ordinal Categorical
S=3	<i>NPC</i>	0.047	0.052	0.619	0.630
	<i>Align</i>	0.050	0.051	0.543	0.518
	<i>vE</i>	0.050	0.047	0.523	0.508
S=5	<i>NPC</i>	0.052	0.050	0.792	0.808
	<i>Align</i>	0.052	0.051	0.764	0.765
	<i>vE</i>	0.050	0.048	0.746	0.740

## 4 An application example

We consider an industrial problem where a company producing bicycles has to choose among 2 different types of paints, say  $A$  and  $B$ , for bicycles. In order to compare quality of competitive paints, for each paint were recorded performance on frame of the bicycle. Performances were recorded by an instrument which investigate the smooth surface of the piece, giving a continuous measure from 1 to 100 intended as "the largest the better". There are 5 machines painting components and a specific machine could influence performance of paint so that we have to take this aspect into consideration. For each paint samples of size  $n = \sum_{s=1}^S n_{h_s} = 25$  with  $n_{h_s} = 5$  for  $h \in \{A, B\}$  and  $\forall s = 1, \dots, S, S = 5$  has been collected. Data are shown in Table 3. After applying the NPC procedure adopting the difference of means as

test statistic, we obtain a global p-value of  $\lambda_{\Phi}^{B>A} = 0.004$  indicating a significant difference in performances between paint *A* and paint *B*, in the sense that paint *A* has greater performance with respect to paint *B*. An important feature of NPC procedure is that it is possible investigating which stratum mainly affect the global results. Note that we perform comparisons in the two directions (i.e.  $\delta_A > \delta_B$  and  $\delta_B > \delta_A$ ), for the sake of simplicity here show only comparisons of interest, i.e.  $\lambda_{(1)}^{B>A} = 0.004$ ,  $\lambda_{(2)}^{B>A} = 0.004$ ,  $\lambda_{(3)}^{B>A} = 0.362$ ,  $\lambda_{(4)}^{B>A} = 0.254$ ,  $\lambda_{(5)}^{B>A} = 0.247$ .

As we can see from partial results, 2 out of 5 strata seem to mainly affect the global result. Extension to ordinal or mixed data is straightforward.

**Table 3** Response data for example application

Stratum	Paint A	Paint B
1	96.8, 96.7, 96.7, 93.2, 94.4	98.6, 99.4, 99.4, 99.8, 98.4
2	85.2, 76.2, 83.1, 76.8, 76.8	95.8, 97.3, 97.9, 96, 97.6
3	100, 96.1, 100, 100, 99.4	100, 100, 100, 99.4, 99.6
4	95, 93.8, 95, 94.3, 94.3	95.1, 94, 96, 94.3, 94.6
5	92.1, 97.5, 98.2, 97.4, 97.6	94.2, 97.6, 98.2, 98.2, 98.6

## 5 Conclusions

In the last years some alternatives to the van Elteren test for stratified two-sample analysis have been proposed. In particular the aligned rank test is a potential choice given its good operating characteristics. In the present work we proposed a new nonparametric NPC-based stratified test and we compared its performance with that of van Elteren and aligned rank tests.

Among tests compared NPC presented higher power and it respects the nominal  $\alpha$ -level for very small sample size.

Moreover, extensions to multivariate observations, to  $C > 2$  samples, to repeated measurement data, can be obtained within our NPC-based approach and will be considered in future research.

## References

1. Bertoluzzo, F., Pesarin, F., Salmaso, L.: On multi-sided permutation tests. *Communication in Statistics: Simulation and Computation*, **42**(6), 1380–1390 (2013)
2. Boos D. D., Brownie, C.: A rank-based mixed model approach to multisite clinical trials: *Biometrics*, **48**, 61–72 (1992)
3. Brunner, E., Munzel, U., Puri, M. L.: Rank-score tests in factorial designs with repeated measures: *Journal of Multivariate Analysis*, **70**, 286–317 (1999)

4. Brunner, E., Puri, M. L., Sun, S.: Nonparametric methods for stratified two-sample designs with application to multiclinic trials. *Journal of the American Statistical Association*, **90**(431), 1004–1014,(1995)
5. Gould, A. L.: Multi-center trial analysis revisited: *Statistics in Medicine*, **17**, 1779–1797 (1998)
6. Mehrotra, D. V., Lu, X. and Li, X.: Rank-based analyses of stratified experiments: alternatives to the van Elteren test. *The American Statistician*, **64**(2), 121–130 (2010)
7. Pesarin, F.: *Multivariate permutation tests: with applications in biostatistics*. Wiley, Chichester (2001).
8. Pesarin, F., Salmaso, L., Carrozzo, E., Arboretti, R.: Union intersection permutation solution for two-sample equivalence testing. *Statistics and Computing*, **26**(3), 693–701 (2016)
9. Pesarin, F., Salmaso, L.: *Permutation tests for complex data: theory, application and software*. John Wiley sons, Chichester (2010).
10. van Elteren, P. H.: On the combination of independent two sample tests of Wilcoxon. *Bulletin of the Institute of International Statistics*, **37**, 351–361 (1960)



# Crowd and Minorities: Is it possible to listen to both? Monitoring Rare Sentiment and Opinion Categories about Expo Milano 2015

## *Opinione di massa ed opinione di nicchia: possiamo misurare entrambi? Monitoraggio di sentiment ed opinioni rare riguardo ad Expo 2015*

Marika Arena, Anna Calissano and Simone Vantini

**Abstract** The talk introduces a new aggregated classification scheme aimed to support the implementation of text analysis methods in contexts characterised by the presence of rare text categories. This approach starts from the aggregate supervised text classifier developed by Hopkins and King and moves forward relying on rare event sampling methods. In details, it enables the analyst to enlarge the number of text categories whose proportions can be estimated preserving the estimation accuracy of standard aggregate supervised algorithms and reducing the working time w.r.t. to unconditionally increase the size of the random training set. The approach is applied to study the daily evolution of the web reputation of Expo Milano 2015, before, during and after the event. The data set is constituted by about 900,000 tweets in Italian and 260,000 tweets in English, posted about the event between March 2015 and December 2015. The analysis provides an interesting portray of the evolution of Expo stakeholders' opinions over time and allow to identify the main drivers of Expo reputation. The algorithm will be implemented as a running option of the next release of R package ReadMe

**Key words:** Sentiment Analysis, Opinion Analysis, Rare Sampling Design, Expo Milano 2015

---

Marika Arena  
Department of Management, Politecnico di Milano. Via Lambruschini, 4/B, 20156 Milano, Italy.  
e-mail: marika.arena@polimi.it

Anna Calissano  
MOX- Department of Mathematics, Politecnico di Milano. Piazza Leonardo da Vinci 32, 20100 Milan, Italy. e-mail: anna.calissano@polimi.it

Simone Vantini  
MOX- Department of Mathematics, Politecnico di Milano. Piazza Leonardo da Vinci 32, 20100 Milan, Italy. e-mail: simone.vantini@polimi.it

## 1 Introduction

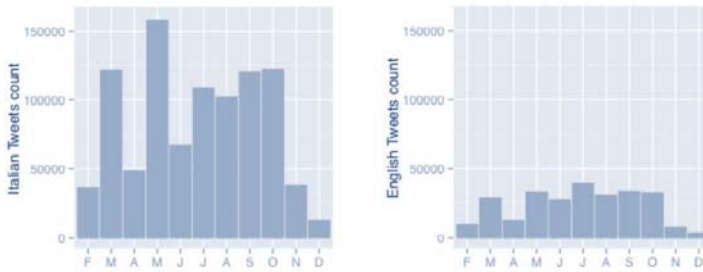
From the 1st of May 2015 to the 31st of October 2015, Milano hosted the 2015 World Exposition (Expo Milano 2015). Doubts and uncertainties characterized the event at the beginning. The enthusiasm of hosting a world fair was accompanied by controversies concerning its set up; the long-lasting discussion about the investments required to face its preparation alternated with the opportunity of exploiting positive externalities induced by the event. Discussions about corruption episodes were often on the news, and this cost overruns and delays. However, when the exposition started, initial skepticism gave way to growing curiosity and, in the end, turned out in an unexpected success. Milano Expo 2015 involved 140 countries and was visited by 21 millions of people, with 7 millions of foreign visitors and 2 millions of students [5]. “Feeding the Planet, Energy for Life” theme marks an opportunity to put the centrality of sustainability at the top of the political agenda and stimulated visitors with thought-provoking ideas coming from the pavilions of different countries. But how did the perception of Expo Milano 2015 evolve before, during, and after the event? Why was a changing dynamic registered? In the Talk we will answer these questions proposing a new aggregate supervised classification scheme [2]. The dataset used to train and test the model and to map precisely the reputation is Twitter.

## 2 Method and Case Study

In these talks, we will present in details the study about the web reputation of Expo Milano 2015, by analysing Twitter data through sentiment and opinion analysis. A soaring on-line discussions and web participation mirrored the bustle that surrounded the Expo, making social media an interesting channel for understanding what people were thinking and saying about the Expo. Among the existing social media platforms, we focus on Twitter because both it has a public philosophy and via API, Twitter offers a partial free download of its data, and it is a micro-blogging platform, where the users share in 140 characters their own opinion about specific topics. The sharpness of posts helps the sentiment analysis performances, conducted on sentence-level data-sets.

### 2.1 The Data Set

The tweets with tags related to Expo Milano 2015 were downloaded from the 17th of February 2015 to the 31st of December 2015. Both Italian and English written Tweets are analysed to cover the local and international nature of the public. Figure 1 shows the amounts of analyzed Tweets (here aggregated per month), in Italian and English, respectively.



**Fig. 1** Downloaded Tweets from the 17th February 2015 to 31st December 2015 via keywords concerning Expo Milano 2015. Data are here represented monthly aggregated, in both Italian and English languages.

To fully capture the evolution of sentiment about Expo, we have to deal with a critical methodological issue, i.e.: the management of rare categories in the data set. The broadness of Expo event involve many different agents on different topics. The mission of the Expo was educating the public, sharing innovation, promoting progress and fostering cooperation among participating countries, the event put together many different stakeholders, moved by diversified expectations and perceptions, resulting in a complex and varying arrangement of interests and feelings. This heterogeneity was reflected in the on-line discourse, that was characterised by some “mainstream” topics discussed by plenty of people and some “less represented” categories - hereafter named rare categories - related to issues discussed by fewer people, but still relevant to understand the multifaceted reputation of the event.

## 2.2 The Method

The presence of rare categories is particularly critical for the implementation of supervised sentiment classifiers, which nowadays are an essential instrument for performing sentiment analysis. As discussed in details in [2], supervised sentiment classifiers require a training set. As hypothesis, the language used in the training set is assumed to be representative of the entire text [4], and it is labelled through hand-coding to obtain a better interpretation of the sentiment [1]. When a corpus of texts is characterised by the presence of rare categories, there is a non-null probability of not gathering any text belonging to these rare categories in the training set, with the risk of losing some relevant pieces of information. Against this background, in this talk we present a new aggregated supervised classification scheme for sentiment and opinion analysis. This new classifier takes advantage of the integration of standard sentiment and opinion analysis techniques proposed by [1], with rare event sampling techniques [2]. The rare event sampling technique are strictly linked with the strate-

gies known as *choice-based*, and *case-control* sampling [6]. In particular, we focus on the sampling solution proposed by [7]. The estimation of both broad-discussed and niche topics is now possible thanks to these new approach, contrary to current approaches which are able to deal with the former ones exclusively. In addition, this specific feature is particularly relevant from a managerial point of view because the identification and the analysis of rare categories could be used to anticipate future trends, and to identify and manage potential risks or opportunities. All the algorithm is run in R and it will be implemented as a running option of the next release of R package ReadMe.

In the talk, we will outline the state of the art about opinion mining, with particular attention to classification methods and, more specifically, aggregate supervised ones, that represent the starting point for this work. The proposed classification scheme will be illustrated in terms of sentiment categories definition, texts pre-processing, variables definition, classification scheme evaluation, and results computation. All the algorithm steps will be displayed on the analysis of the web reputation of Expo Milano 2015. The talk finish with the results of a statistical comparison performed between our classification scheme and other existing ones.

## References

1. Hopkins, Daniel J and King, Gary : A method of automated nonparametric content analysis for social science. *American Journal of Political Science* Vol. 54, 1, Wiley Online Library
2. Arena, Calissano and Vantini : Monitoring Rare Categories in Sentiment and Opinion Analysis - Expo Milano 2015 on Twitter Platform. <https://mox.polimi.it/publications/>
3. Hopkins, Daniel and King, Gary and Knowles, Matthew and Melendez, Steven (2010) ReadMe: Software for automated content analysis *Institute for Quantitative Social Science*
4. Hand, David J. (2006) Classifier technology and the illusion of progress. *Statistical Science* 21(1):115.
5. Expo 2015 S.p.a. <http://www.expo2015.org/>
6. Breslow, Norman E (1996) Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association, Taylor & Francis Group* Vol. 91,433
7. King, Gary and Zeng, Langche (2001) Logistic regression in rare events data *Political analysis* Vol 9, 2, SPM-PMSAPSA



# Using administrative data for statistical modeling: an application to tax evasion

## *L'uso di dati amministrativi per la modellizzazione statistica: un'applicazione all'evasione contributiva*

Maria Felice Arezzo and Giuseppina Guagnano

**Abstract** Administrative data, gathered by public authorities with a general aim of control, are very precious sources of information because they allow to study phenomena that would remain otherwise unknown. On the other side, administrative data strictly contain the information they were collected for, and to be used for statistical purposes they need to be integrated. This work shows the potentials of the integration of three data sets for statistical modeling: the audits carried out in Italy in 2005 by the National Institute of Social Security on building and construction companies, the ASIA archive of Istat and the "Studi di Settore" of the Italian Revenue Agency.

**Abstract** *I dati amministrativi, raccolti dalle istituzioni pubbliche per scopi generalmente di controllo, sono fonti informative estremamente preziose in quanto permettono spesso di studiare fenomeni che in altro modo non potrebbero essere conosciuti. D'altro canto, proprio perchè rispondono a finalità specifiche, le indagini amministrative non contengono informazioni aggiuntive rispetto a quelle per le quali sono state pensate. Il lavoro illustra le potenzialità dell'integrazione di tre basi dati da fonte differente: le ispezioni INPS, l'archivio ASIA dell'Istat e gli Studi di settore dell'Agenzia delle entrate. La sperimentazione è stata condotta sulle imprese che operano nel settore delle costruzioni.*

**Key words:** Administrative data, Sample selection, Response-based sampling

---

Maria Felice Arezzo

Sapienza University of Rome, Address, e-mail: mariafelice.arezzo@uniroma1.it

Giuseppina Guagnano

Sapienza University of Rome, Address e-mail: giuseppina.guagnano@uniroma1.it

## 1 Introduction

Administrative data are archives of great interest as they often contain information available only to public authorities responsible for the control of some phenomena. Almost always, though, these files do not contain information other than those for which they were collected (a typical example are the socio-economic characteristics of the individuals), as the purpose underlying their gathering is not statistical modeling. For this very same reason, administrative data require, on the one side, a throughout pretreatment and validation process and, on the other, the development of statistical methodologies that allow for the drawing of valid inferences.

The purpose of our work is to draw the entire “production chain”: a) the creation of a dataset with all relevant variables, b) the evaluation of the dataset quality, c) the development of a statistical method suitable for the data at stake.

The case study is on the detection of the firms which evade worker contributions because they employ off-the-book workers (i.e. employee who are completely unknown to fiscal authorities)

## 2 Creation of the data set

Our starting point is an administrative dataset on the audits carried out in Italy in 2005 by the National Institute of Social Security (INPS henceforth) on building and construction companies (NACE section: F). It amounts to a total of 31,658 inspections on 28,731 firms. The global amount of firms operating in the building industry in Italy in the same year was  $N = 595,226$ . Audits data allow to observe the compliant/non-compliant behavior.

Following the idea that the risk of a non-compliant behavior can be predicted by the economic characteristics of the firm, we integrated the information of audits with two other sources of data. The first is the ASIA archive owned by the National Institute of Statistics (ISTAT). It contains data on the legal structure, turnover and number of employee and is a high quality source of data as the information are validated through a very careful process. The second, owned by the Italian Revenue Agency, is the so called ‘Studi di Settore’ (SS in the following) archive. It contains an exhaustive list of information on corporate organization, firm structure, management and governance.

The three data sets were merged using VAT numbers and/or tax codes. Surprisingly the match rate was only 51% meaning that the number of firms in the merged archive is 14,651.

The original variables were used to build economic indicators which can be grouped in the following different firm’s facets: a) 9 indicators for economic dimension, b) 13 for organization, c) 6 for structure, d) 6 for management, e) 11 for performance f) 38 for labor productivity and profitability g) 3 for contracts award mode h) 7 variables for location and type. The final dataset had 93 independent variables observed on 14,651 building companies with a match rate of 51%. The

**Table 1** Datasets characteristics

Data Owner	Content	Individual	Dimension
INPS	Inspections outputs (2005)	Inspection	31,658 inspections on 28,731 firms
Revenue Agency	Studi di settore (2005).Models: TG69U, TG75U (SG75U),TG50U (SG50U and SG71U), TG70U	Firm	Universe of firms with at most 5 million euros of income
ISTAT	Asia Archives (2005)	Firm	Universe of firms

variable to be predicted is named  $Y$  and it takes value 1 if in a firm there is at least one off-the-book worker and 0 otherwise. In the following we will refer to the final dataset as the integrated db because it gathers and integrate information from different sources.

### 3 The assessment of the integrated dataset

As we said, the matching rate was 51% which means that we had information on the features of interest for (roughly) half of the firms in original INPS database. We studied inspection coverage and the risk of non complying for different turnover class and corporate designation typologies and over the territory. The idea was to verify if a whole group of firms (for example all the companies in a geographical region) was lost because of the merging process.

We checked for: Regions (20 levels), Number of employee (9 classes), Legal structure (5 levels), Turnover (11 classes); we then made sure that during the matching procedure, no whole groups of individuals were lost.

### 4 The Model

Under a statistical point of view, there are two main methodological issues arising from the type of data we use. The first is the non-randomness of the inspections and the second is that the fraction of inspected firms in the population is low.

SELECTION BIAS IN THE SAMPLE OF INSPECTED FIRM. To detect undeclared work, an inspector audits firms. Inspected firms are not randomly chosen; they are chosen because the inspector thinks that there are some off-the-book workers and s/he has strong incentives to target the “right” firms (i.e. the irregular ones). We can think of the decision to inspect a firm as a rational process in which the inspection is made if the utility to inspect,  $U^A$ , (i.e. find undeclared workers and get a benefit) is higher than the utility of non-inspect,  $U^{\bar{A}}$ . Moreover we can observe the status of

the  $i$ -th firm (regular or not) only if it has been inspected, otherwise a censoring process intervenes. It is obvious that there is a strong selection bias in the sample of inspected firms.

As it is well known, [3] proposed a useful framework for handling estimation when the sample is subject to a selection mechanism. In the original framework, the outcome variable is continuous and can be explained by a linear regression model (called *output equation*), with a normal random component; in addition to the output equation, a *selection equation* describes the selection rule by means of a binary choice model (probit).

In our framework the output equation defines the compliance decision, so the dependent variable is binary, and the selection equation refers to the decision of inspecting a firm. Just as the inspection decision, the evasion is based on a rational process and it happens if the utility of evading,  $U_i^{\bar{C}}$ , is greater than the utility of complying  $U_i^C$ . The corresponding econometric model, in its general form, is:

$$Y_i^* = U_i^{\bar{C}} - U_i^C = \mathbf{X}_{1i}\beta + \varepsilon_{1i} \quad (1a)$$

$$A_i^* = U_i^A - U_i^{\bar{A}} = \mathbf{X}_{2i}\theta + \varepsilon_{2i} \quad (1b)$$

where  $\mathbf{X}_i = (\mathbf{X}_{1i}, \mathbf{X}_{2i})$  is a vector of exogenous variables (namely,  $\mathbf{X}_{1i}$  for  $Y_i$  and  $\mathbf{X}_{2i}$  for  $A_i$ ), containing all the relevant covariates.

Since we cannot observe directly the utilities (neither those determining compliance, nor those governing the decision to inspect), we assume that if in equation (1a)  $Y_i^* > 0$ , the firm does not comply, otherwise it does. Let's define a dummy variable  $Y_i$  which we can observe and that denotes the alternative selected:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Similarly, we can define an observable dummy variable  $A_i$  for the inspections, such that:

$$A_i = \begin{cases} 1 & \text{if } A_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The p.d.f. of  $Y_i$  and  $A_i$  is Bernoulli with probability of success respectively equal to  $\gamma\pi$  and  ${}_A\pi$  and depending on  $\mathbf{X}_{1i}\beta$  and on  $\mathbf{X}_{2i}\theta$ . A selection bias exists if  $\text{corr}(\varepsilon_1, \varepsilon_2) = \rho$  is not null.

As it is known (see for example [2]), the likelihood function for the Heckman's selection model is:

$$L(\eta) = \prod_{i=1}^n \left[ 1 - {}_A\pi(\mathbf{X}_i) \right]^{1-A_i} \cdot \left[ f(Y_i|A_i = 1) \cdot {}_A\pi(\mathbf{X}_i) \right]^{A_i} \quad (4)$$

where  $\eta = (\beta, \theta, \rho)$  is the vector of parameters to be estimated.

**THE CASE-CONTROL SETTING.** In this sampling design [4], also known as response-based, samples of fixed size are randomly chosen from the two strata identified by the dependent variable  $A$ . In particular  $n_A$  units are drawn at random from the  $N_A$  cases and  $n_{\bar{A}}$  from the  $N_{\bar{A}}$  controls.

The likelihood function is the product of the two stratum-specific likelihoods and depends on the probability that the individual is in the sample, and on the joint density of the covariates:

$$\prod_{i=1}^{n_A} Pr(\mathbf{X}_i | A_i = 1, S_i = 1) \cdot \prod_{i=1}^{n_{\bar{A}}} Pr(\mathbf{X}_i | A_i = 0, S_i = 1). \quad (5)$$

The c-c design is particularly suited in our study because the probability that a firm is inspected is very low and therefore it is much more convenient to directly sample from the two strata (inspected/non-inspected).

**A BINARY CHOICE MODEL WITH SAMPLE SELECTION AND CASE-CONTROL SAMPLING SCHEME.** In the following we provide the likelihood function under the framework of interest, i.e. a sample selection mechanism with a severe censoring process. The interested reader can find the full proof and the simulation results in [1].

We make the following very general and non restrictive assumptions:

1. we have a set of fully informative and exogenous covariates  $\mathbf{X}_i = (\mathbf{X}_{1i}, \mathbf{X}_{2i})$ ;
2. conditional on the covariates, the probability that an observation is uncensored doesn't depend on its value, i.e.  $P(A_i = 1 | S_i = 1, \mathbf{X}_i, Y_i) = P(A_i = 1 | S_i = 1, \mathbf{X}_i)$ ;
3. the set of covariates  $\mathbf{X}_{1i}$ , specific for  $Y_i$ , and the set  $\mathbf{X}_{2i}$ , specific for  $A_i$ , may have common elements but they cannot fully overlap;
4. the probability of being in the sample does not depends neither on the covariates  $\mathbf{X}_i$  nor on  $Y_i$ . More precisely, letting  $S_i$  be a binary variable which takes value 1 if the  $i$ -th individual is in the sample and 0 otherwise, it is true that  $P(S_i = 1 | \mathbf{X}_i, Y_i, A_i = a_i) = P(S_i = 1 | A_i = a_i)$ .

Assumption (1) means that it does not exist correlation between the covariates and the residual terms in equations (1a) and (1b). Assumption 2 is justified because, as the covariates are informative, all the information brought by  $Y_i$  is contained in  $\mathbf{X}_i$ . Assumption (3) is necessary for parameters identification (exclusion conditions). Assumption (4) is typical in the response-based sampling framework and no further explanation is required.

Under the conditions stated, the likelihood function for a binary choice model with sample selection under a response-based sampling is:

$$L(\eta) = \prod_{i=1}^n f(\mathbf{X}_i | S_i = 1) \left\{ (1 - {}_A\pi(\mathbf{X}_{2i})) \cdot \frac{N}{N_{\bar{A}}} \right\}^{1-A_i} \quad (6)$$

$$\cdot \left\{ \left[ {}_Y\pi(\mathbf{X}_{1i}) \cdot \frac{N}{N_A} \frac{n_A}{n_{1A}} \right]^{y_i} \cdot \left[ (1 - {}_Y\pi(\mathbf{X}_{1i})) \cdot \frac{N}{N_A} \frac{n_A}{n_{0A}} \right]^{1-y_i} \cdot {}_A\pi(\mathbf{X}_{2i}) \right\}^{A_i}$$

where  ${}_A\pi(\mathbf{X}_{2i})$  is the probability that an observation is uncensored and  ${}_Y\pi(\mathbf{X}_{1i})$  is the probability of observing  $Y = 1$  given that the observation is uncensored; as already said,  $n_A$  is the number of units sampled from the  $N_A$  uncensored observations and  $n_{\bar{A}}$  is the number of units sampled from the  $N_{\bar{A}}$  censored observations;  $n_{yA}$  is the amount of units in the sample having  $Y = y$ , with  $y = 0, 1$ .

It's easy to understand that the likelihood (6) is a weighted version of (4), and the weights simply take into account the sampling design. Note also that in the maximization process the term  $f(\mathbf{X}_i | S_i = 1)$  is non influential, as it does not contain any information on the vector of parameters  $\eta$ , and that in our estimator the only quantities to be known at the population level are  $N_A$  and  $N$ .

## References

1. Arezzo, M.F., Guagnano, G. Response-based sampling for binary choice models with sample selection. Working Paper 149, Department Memotef - Sapienza University of Rome (2017).
2. Cameron, A.C., Trivedi, P.K.: Microeconometrics: Methods and Applications. Cambridge University Press, New York (2005)
3. Heckman, J.J.: Sample selection bias as a specification error. *Econometrica*, 47(1): 153-162 (1979).
4. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. John Wiley & Sons (2013)

# Are Numbers too Large for Kids? Possible Answers in Probable Stories

## *Sono troppo grandi i numeri per un bambino? Rispondono le storie di probabilità*

Monica Bailot, Rina Camporese, Silvia Da Valle, Sara Letardi, Susi Osti<sup>1</sup>

**Abstract** Regardless of calculus ability, children need to approach statistics and stochastic literacy as soon as possible, in order to build up their ability to deal with uncertainty when making judgements and decisions. Moreover, statistics and probability are mandatory in school curricula since primary school. Maths can be narrated, stories are engaging and playful hands-on activities help kids to learn. So then, why not convey statistics and probability by means of fables? Animated fables, where kids play roles and immerse themselves into stories that evolve through events and decisions based on numbers and statistics. The paper presents two fables on probability, large numbers and repeated observations. They are part of a larger set of StatFables that are being written and tested in schools and libraries.

**Abstract** *Indipendentemente dalle abilità di calcolo, è importante che i bambini entrino in contatto con i rudimenti di statistica e probabilità, per imparare a gestire l'incertezza nell'esprimere giudizi e prendere decisioni. Statistica e probabilità, inoltre, sono obbligatorie nei curricula scolastici sin dalla scuola primaria. La matematica si può narrare, le storie catturano l'attenzione e le mani in pasta aiutano ad imparare. Allora, perché non insegnare statistica e probabilità con le fiabe? Fiabe animate, di cui i bambini sono protagonisti, immersi in storie che evolvono attraverso eventi e decisioni basati su numeri e statistiche. L'articolo presenta due fiabe su probabilità, grandi numeri e osservazioni ripetute. Fanno parte di un insieme di StatFiabe in corso di scrittura e test in scuole e biblioteche.*

**Key words:** Kids, uncertainty, large numbers, probability, bayesian reasoning

### 1. StatStory One: 1, 10, 100, 1000 Nights of Silver Moon

---

<sup>1</sup> Istat . Italian Institute of Statistics, corresponding author [rina.camporese@istat.it](mailto:rina.camporese@istat.it)

In the village of WhoKnowsWhere, every night a witch throws into the sky the moon, symbolically represented by a coin. And every night the merchant Hamlet is impatient to see the face showed by the coin: the silver face of the moon will light up his journey to the city where he will sell his merchandise, the black side will force him to wait in a dark night. Every night a doubt, every night two possible outcomes. Imagine one, ten, a hundred, a thousand nights... how many of them will the merchant spend on the road? And how many will he have to wait for a better chance? This is the mystery the merchant must solve to free his daughter Ada from the cave where she is held in captivity by the witch.

Ada has a passion and a flair for maths and her father keeps bringing her math books to kill the boredom of solitary life. Another game she plays in her long and boring days in the cave is to pile up white and black pebbles counting propitious and inauspicious nights. After many days of observation Ada realizes that the two piles of stones are getting closer and closer in size and in number of pebbles. This discovery and the remembrance of the Law of Large Numbers she had read somewhere - a law which she initially thought was a funny idea popped out in the mind of a sloppy mathematician with weak practical sense - make her resolve the arcanum and break the spell.

## 2. Why Stories? Why Kids?

Why stories to help children (only children?) deal with numbers, formulas and probability? Because stories have always been a privileged way to transmit culture.

Linguists say that a mathematical formula is an extreme form of text (Sabatini, 2016). A formula, a theorem... are stories in a nutshell, a mathematical nutshell. When linguists could see an extreme form of text in a formula, we statisticians could see an entire story in it, therefore the story contained in a formula can be unveiled in a full verbal narration. Maths can be narrated and stories are engaging. So then, why not convey statistics and probability by means of tales?

If statistical information is communicated in mathematical formats people have troubles in correctly reading and interpreting it. When it comes to the ability to deal with data, uncertainty and mathematical representation of phenomena, the words innumeracy and statistical illiteracy are the most appropriate for the majority of the population (Till, 2014). Needless to say this is a major cultural issue in our society.

Regardless of calculus ability, children need to approach statistics and stochastic literacy as soon as possible, in order to build up their ability to deal with uncertainty when making judgements and decisions and to understand numerical information. That's why statistics and probability are mandatory in early school curricula.

Children aged eight to ten do have probabilistic intuitions and can develop secondary intuitions. They can also reason on proportions before being able to formally deal with fractions (Fischbein, 1970). Here is what Christoph Till wrote in 2014 “[...] *risk and decision making under uncertainty can be a prevailing, exciting and meaningful topic at the end of primary school with sustainable effects. [...] it is possible to foster elementary competencies for risk assessment and probabilistic*



*decision making in fourth class. [...] in a playful learning environment, as advised by results of cognitive psychologists. With these representations children can think probabilistically without the need of fractions or percentages. [...] As Gigerenzer (2011; 2013) has repeatedly pointed out elementary probability concepts should be taught in an informal and heuristic manner at an early stage.”*

## **2.1. Words in action: animated stories**

Kids learn by playing and immersing themselves into stories. They also find hands on activities extremely engaging; and this is true also for Maths, Stats and Probability (Martignon, 2009). Listening to stories can be fun and relaxing, but being the protagonists of an animated story can be even more exciting. That's why fables can be animated and kids can play roles and immerse themselves into stories that evolve through events and decisions based on numbers and statistics.

The story on the Law of Large Numbers is suitable for children aged four to ten; the next story on Bayesian reasoning is thought for children aged six to twelve. The action associated to them can be modulated depending on children's familiarity with numbers, fractions and spreadsheet. Kids are engaged in the story through telling small episodes, performing actions and posing questions.

Since questions are the real engine in the learning process and active participation enhances the engagement (Chavannes, 2016), children are guided to find out answers by carrying out experiments and reasoning on the results.

## **3. StatStory Two: The Witches of Bayes<sup>2</sup>**

“The Witches of Bayes” is an animated fable to be played by children from 9 to 12 years of age. The aim of the game is to become familiar with the Bayesian reasoning and the use of all the pieces of information available in order to make decisions.

A group of witches haunts the village of Bayes. Every day one of them, randomly chosen and unknown to the villagers, asks for a dish of food by placing her hat outside the cave. The witches have different tastes, some love the sweet and some the salty. If the dish offered meets with the taste of the witch, the day passes peacefully, otherwise there will be trouble for everyone. Every morning Coco Head, the head of the village, relies on the fate and throws a ritual coin with one face indicating salty and the other one sweet; based on the visible face he gives orders to the kitchen. The gloomy days, however, are numerous.

Nora, a young girl, notices that witches have different hats, some are black and some others are purple. It seems to her that purple hats do appear more often and that the witches rather prefer sweet dishes, but she is not certain. She also wonders whether there be a link between the colour of the hat and the taste of the witch. Then she

---

<sup>2</sup> Poster presented at SISBAYES Meeting in Rome, 7-8 February 2017.

discovers a parchment full of information, thanks to which she develops an alternative strategy to choose the dish. She tries to convince the village head to change the traditional method, but he is unshakable. Nora then asks for permission to serve the meal herself and she prepares in secret, when necessary, some dishes in alternative to those indicated by the ritual coin. After a while the villagers realize that the devastating raids of the witches are less frequent. Coco Head organizes a ceremony in honour of the god Bias to thank him for his increased magnanimity.

Nora, the protagonist, finds a way to decide despite the uncertainty, using everything she knows. Coco Head, on the contrary, does not change his actions when new elements arise; worse than that, he finds a way to reinforce his beliefs thanks to elements that should rather put them into question. He suffers from the status quo bias, i.e. his perception of risk, relative to changes, is amplified by the unjustified belief that a different choice can only make things worse. The ending is controversial to inflame the discussion.

The tale is not read nor told, but it is animated by the children with the help of hats, giant coins and bags through which the selection criteria of the dish are applied. The choices of Coco Head and Nora are simulated for the thirty-one days of a month and children evaluate which method can lead to more favorable results. After that, Nora's diary is studied, for she has been noting down good and bad days for ten years. Behaviours, methods and results are discussed together with the children. If they are familiar with fractions, they can also apply Bayes' theorem, looking into the story and the underlying data for the necessary information. If the kids are familiar with spreadsheet, Nora's diary can be "calculated" by simulating the sequence of results in a decade of choices, with one method or the other.

#### 4. What's around the stories

Authors in Istat are working on a set of activities devoted to promote statistical literacy through unconventional instruments and non-specialized languages meant to be used in schools, libraries and events such as the Festival of Statistics and the European Researchers' Night.

These two stories belong to a set about statistics, basic descriptive measurements and probability and are being tested in schools and libraries. They aim at conveying statistics and probability through words, by drawing a the path to numbers and formulas that passes through verbal narration of concepts and active experimentation thanks to playful activities. An apparent paradox of such activities is that, sometimes, they deal with mathematical concepts without showing any formula, sometimes not even a number, and that it done on purpose.

#### References

1. Chavannes I., Lezioni di Marie Curie. La fisica elementare per tutti, Dedalo ed. (2016)
2. Fischbein E., Pampu I., Mânzat I., Comparison of Ratios and the Chance Concept in Children, Child

- Development, Vol. 41, No. 2, pp. 377-389 (1970)
3. Gigerenzer, G., Gray, M., Better Patients. Better Doctors. Better Decisions. Envisioning Health Care 2020. Cambridge: MIT Press (2011)
  4. Martignon, L., Krauss, S., Hands on activities for fourth graders: A tool box for decision – making and reckoning with risk, International Electronic Journal of Mathematics Education 4 (3), pp. 227258 (2009)
  5. Sabatini F., Lezione di italiano: grammatica, storia, buon uso, Mondadori ed. (2016)
  6. Till C., Fostering Risk Literacy in Elementary School, Mathematics Education, 9(2), 83-96 (2014)



# A polarity-based strategy for ranking social media reviews

## *Una strategia basata sulla polarità per ordinare le recensioni sui social media*

Simona Balbi, Michelangelo Misuraca and Germana Scepi

**Abstract** The Opinion Mining methods are widely used to analyse and classify the choices, preferences and behaviours of consumers through the opinions gathered on the Web. On social media like TripAdvisor such opinions are usually expressed with a score and a short text. This paper proposes a strategy for ranking reviews using a scale based jointly on the rating and on the text of the reviews.

**Abstract** *I metodi di Opinion Mining sono oggi ampiamente utilizzati per analizzare e classificare le scelte, le preferenze e il comportamento dei consumatori attraverso opinioni raccolte sul web. Sui social media come TripAdvisor tali opinioni vengono solitamente espresse con un punteggio e con un breve testo. In questo lavoro si propone una strategia per ordinare le diverse recensioni con una scala di misura basata sia sul punteggio sia sul testo scritto.*

**Key words:** Textual Data, Opinion Mining, Ranking

## 1 Introduction

With the rapid expansion of social media, it is more and more widespread the practice of sharing opinions on the Web. The ways for expressing those opinions are many: numbers, texts, emoticons, images, videos, audios. There are often a joint use of these communication tools. It is becoming a habit for users to evaluate the products/services they buy/use, by describing their personal feelings and judgments.

We can find online websites specialised in one or more topics, where people can give their opinion using an evaluation scale (e.g., from “terrible”=1 to “excellent”=5), visualised by bullets or stars, and combined with a written description.

---

Simona Balbi, Germana Scepi

Università Federico II di Napoli e-mail: simona.balbi@unina.it, germana.scepi@unina.it

Michelangelo Misuraca

Università della Calabria, Arcavacata di Rende e-mail: michelangelo.misuraca@unical.it

As this practice is nowadays considered the core of many marketing strategies, there is a large interest on how to extract knowledge from such a kind of information.

Opinion mining procedures have been developed with the main goal of understanding the mood in a text, transforming it in a numerical value. The basic idea is identifying positive, negative, or neutral viewpoints. Researchers involved in defining proper methods for mining opinions on the Web are mainly computer scientists and computational linguists. They often claim to use statistical techniques.

The main point we are interested in this paper is that we often see the lack of a statistical perspective. Statisticians are professionally involved into the problem of quantifying something that is not quantitative in itself. Furthermore, the implications in the choice of a scale, or in the choice of a weighting system, or in the choice of the proper method for analysing those unconventional data pertain to statisticians.

Here we focus our attention on the so called rating-inference problem [6], and its implications when we refer to “reviews and ratings” social media like TripAdvisor. In this kind of media we usually find ratings in a 1-to-5 stars system, together with written judgments. The challenge is stimulating for a statistician: on one hand, we have a judgment in a 5-point scale; on the other hand we have a (usually) short text. We propose a two-step strategy for taking into account jointly the two assessments and defining a unique rating.

The paper is organised as follows. Section 2 defines the theoretical framework. Section 3 considers the case study. The proposed strategy is presented in Section 4, while the main results of applying the strategy on TripAdvisor reviews are discussed in Section 5.

## 2 Theoretical framework

Sentiment analysis (SA), also known as opinion mining (OM), refers to the analysis of people’s opinions, attitudes, or emotions, in a written text. Note that SA is generally used in industry, while both SA and OM are used in academia. In the following, we interchangeably use the two terms. Opinions are usually published in specialised websites, devoted to peculiar topics like cinema, e-commerce, and so on.

The main goal of SA is to classify documents on the basis of their “polarity”. The term polarity is used in linguistics for distinguishing affirmative and negative forms. For a wide review of the different methods of SA refer to [1] [7]. In literature there are three different steps in determining the polarity:

1. the subjectivity/objectivity of a text (SO-polarity): decide if a text has a factual nature or expresses an opinion on its subjective matter.
2. the positivity/negativity of a text (PN-polarity): decide if a subjective text expresses a positive or negative opinion.
3. the positivity/negativity strength of a text (PN-strength): identify different grades of positive or negative sentiments in opinions.

These steps are sequentially ordered, but it is not mandatory to perform all three.

Focusing on the unit of the analysis, we can consider different levels: a document-level, a sentence-level, an aspect-level. The first two levels are usually considered in the so called polarity-based SA, while the latter one is used in a topic-based perspective. The document-level aims at defining the polarity of each document, i.e. if it expresses a positive or a negative sentiment. In the sentence-level each document is segmented into sentences, and we want to determine the polarity of each sentence. The PN-polarity is quantified by considering a score of -1, 0 and 1 for negative, neutral and positive sentiment, respectively [2]. Some authors have proposed different scoring systems by defining the polarity not only in terms of sign but also taking into account the PN-strength of the sentiment [5]. The aspect-level SA aims at quantifying specific aspects and it allows to obtain fine-grained results. The aspect-level SA requires a greater computational complexity.

In this paper, we aim at determining the PN-polarity of a document, by considering a sentence-level approach. This is the first step of a mixed strategy that uses both textual and numerical information.

### 3 The Uffizi Gallery on TripAdvisor

In the last decades several private and public institutions operating in the field of cultural heritage, like museums, have looked at the visitors from a visitor satisfaction perspective. The so called museum audience is became strategically central, because it has a major connection to museums' sustainability. In this framework, it is more and more important to collect and analyse data coming from different sources. Together with classical sample surveys, carried out on a limited number of visitors, it is possible to use secondary data available on the Web. This huge amount of online data can be seen in a big data frame, as they have different natures and are available in real-time. In this paper, we study the audience of the Uffizi Gallery by analysing a set of reviews published on TripAdvisor.

TripAdvisor is a social media specialised in tourism reviews about both businesses and attractions. According to the most general classification of social media, it can be defined as a "reviews and ratings" media. It has been founded in U.S. in February 2000. Since mid-2010 is both an online service on the Web and a mobile application on portable devices. It has been one of the first websites to implement user generated content.

We use a scraping approach by launching a custom crawler on February 11<sup>th</sup> 2017. In this way we retrieved 9639 reviews written in English and posted on TripAdvisor from February 27<sup>th</sup> 2003 up to February 10<sup>th</sup> 2017. The crawler has also provided some meta-information about the author of each review (e.g., location, contribution level on TripAdvisor, number of submitted reviews) and about the review itself (e.g., date, rating, device used for publishing the review). Here in the following we only focus our attention on the reviews and the corresponding ratings. We decide not to perform any lexical pre-treatment on the reviews. Only the parts

not in English have been deleted, because some reviews also contained sentences in the mother-tongue language of the author.

## 4 A two-step strategy for a polarised rating/ranking

The rating scale used by TripAdvisor is an ordinal scale. In details, the ratings from 1 to 5 are associated with the terms *terrible*, *poor*, *average*, *very good* and *excellent*, respectively, and a corresponding number of bullets. In Fig. 1 it is possible to see the rating distribution of the Uffizi Gallery updated at April 5<sup>th</sup> 2017. The ordinal rating can be seen as a global and comparable measure of the experience, while the textual description is an evaluation highlighting which aspects are positive and negative. Therefore, we propose a two-step strategy for the computation of a polarised rating of a review by combining the rate and the sentiment in the text.

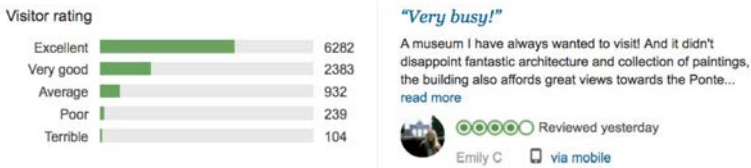


Fig. 1 Visitor rating distribution at April 5<sup>th</sup> 2017

### Step 1: Computing the reviews' polarities

In order to compute the polarity of the reviews, we follow an SA sentence-level approach. This level seems to be more suitable, because in these texts each sentence includes an opinion of the contributor on the different aspects of the offered service.

The polarity scores have been calculated by using the R package *sentimentr*. The equation used in this package is based on the concept of valence shifters [8]. It is a procedure allowing to capture the polarity of a sentence by considering the context of use of its terms. The polarity of each term is weighted by taking into account negators (e.g., "never", "none"), amplifier and de-amplifier (e.g., "very", "few"), adversative and contrasting conjunctions (e.g., "but", "however"). This weighting system allows to emphasise or dampen the positivity and negativity of the terms, and obtain a more proper measure of the sentence sentiment.

Each review  $d_i$  (with  $i = 1, \dots, n$ ) is segmented into a set  $S_{d_i}$  of  $q_i$  sentences  $\{s_{i1}, \dots, s_{ij}, \dots, s_{iq_i}\}$ , by considering as separators only full stops, question marks and exclamation points. Each sentence  $j$  is represented as a sequence of its  $p_j$  terms  $\{w_{ij1}, \dots, w_{ijk}, \dots, w_{ijp_j}\}$ . Each term  $w_{ijk}$  in the sentence  $s_{ij}$  is compared with a lexicon of polarised terms, with a score  $r_{w_{ijk}}$  of -1 for negative terms and 1 for positive terms, respectively. The terms not included into the lexicon are assumed to be neutral, with a score  $r_{w_{ijk}}$  equal to 0.



The polarity score of each sentence depends on the dictionary of polarised terms used into the analysis, while the PN-polarity of the whole document depends on the polarities of its sentences. Different dictionaries are available. It is possible to consider manually created resources or automatically and partially automatically created resources. There are many papers in literature dealing with the problem of choosing one dictionary [4]. We use the Jockers dictionary, a lexicon of more than 10000 terms developed by the Nebraska Literary Lab for the R package *syuzeh* [3].

The final polarity score  $r_{s_{ij}}$  of each sentence is computed as the sum of its weighted term scores  $r^*_{w_{ijk}}$  (taking into account the shifters) on the square-root of the sentence length:

$$r_{s_{ij}} = \frac{\sum_{k=1}^{p_j} r^*_{w_{ijk}}}{\sqrt{p_j}} \quad (1)$$

As we are interested in computing a polarity score for the whole review, we compute the score  $r_{d_i}$  of each document by a down-weighted zeros average of its sentence polarities. In this averaging function the sentences with neutral sentiment have minor weight:

$$r_{d_i} = \frac{\sum_{j=1}^{q_i} r_{s_{ij}}}{\tilde{q}_i + \sqrt{\log(2 - \tilde{q}_i)}} \quad (2)$$

where  $\tilde{q}$  is the number of sentences with a positive or negative polarity. The logic of down-weighting neutral sentences is that they have less emotional impact in the review than the polarised ones.

### Step 2: Computing the polarised rating

The new score for each contributor is obtained by summing the original rating with the polarity score of the review. Because of the unboundedness of the polarity scores, we bring all values into a range [0,1]. For each category  $c_h$  in the rating system (with  $h = 1, \dots, H$ ), the  $\hat{r}_{d_i}$  rescaled scores are computed as:

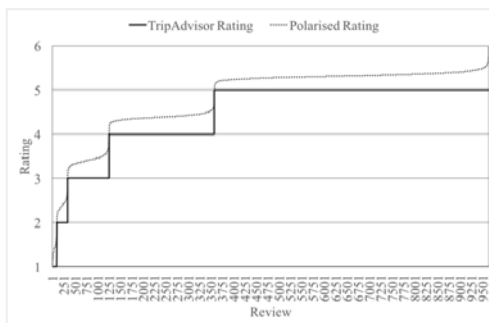
$$\hat{r}_{d_i} = \frac{r_{d_i} - \min_{d_i \in c_h} r_{d_i}}{\max_{d_i \in c_h} r_{d_i} - \min_{d_i \in c_h} r_{d_i}} \quad (3)$$

The resulting scoring system has a range [1,  $H+1$ ], where 1 expresses the strongest criticism and  $H+1$  expresses the strongest appreciation. The polarised rating can be interpreted as a ranking, because the new score allows the sorting of the reviews. Users can not only browse and read the reviews by rating, but also with respect to the sentiment.





**Fig. 4** Distributions of the TripAdvisor ratings and the polarised ratings



In Fig. 4 we show the distribution of the original ratings and the distribution of the polarised ratings. The new scale introduces a useful gradation in the judgments. Here in the following we can see two examples of reviews rated 1 by the contributor, and rated 1.0 and 1.9 by the polarised rating, respectively:

**Review #2061:** *I'm not sure why this museum is so famous, the truth is: it's extremely boring, full of statues and religious paintings, all the same, not even the building is nice!! The line up is insane, even if you buy tickets in advance, it's ridiculous, lots of people! Worthless!!! Save yourself the trouble, go browse Florence, so much to see outside. Totally waste of time and energy, nothing interesting, we were in and out!!! Horrible!!*

**Review #1121:** *Buy your tickets online beforehand otherwise you will wait a long time in a queue. There is a very good rooftop cafe with reasonably priced food and drinks. Some spectacular photo opportunities through the windows overlooking Florence.*

## References

1. Bing, L.: Sentiment Analysis: mining opinions, sentiments, and emotions. Cambridge University Press (2015)
2. Bing, L., Minqing, H., Junsheng C.: Opinion Observer: Analyzing and Comparing Opinions on the Web. In: Proceedings of the 14<sup>th</sup> IW3C2 conference, pp. 342–352 (2005)
3. Jockers, M.L.: Syuzhet: Extract sentiment and plot arcs from text (2017) Available via <https://github.com/mjockers/syuzhet>
4. Marquez, F.B., Mendoza, M., Poblete, B.: Meta-level sentiment models for big social data analysis, Knowledge-Based Systems. **69**, pp. 86–99 (2014)
5. Nielsen, F.Å.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In: Rowe, M., Stankovic, Dadzie, A., Hardey, M. (eds) Proceedings of the Workshop on Making Sense of Microposts: Big things come in small packages, Vol. 718, pp. 93–98 (2011)
6. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43<sup>rd</sup> Annual Meeting on Association for Computational Linguistics (ACL05), pp. 115–124 (2005)
7. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. In: Foundations and Trends in Information Retrieval, Vol. 2, pp. 1–135 (2008) doi: 10.1561/1500000011
8. Polanyi, L., Zaenen, A.: Contextual valence shifters. In: Shanahan, J., Qu, Y., Wiebe J. (eds) Computing Attitude and Affect in Text: Theory and Applications. The Information Retrieval Series, Vol. 20, pp. 1–9. Springer, Dordrecht (2004)

# Monitoring the spatial correlation among functional data streams through Moran's Index

## *Monitoring della correlazione spaziale tra data stream funzionali attraverso il Moran Index*

A. Balzanella, S.A. Gattone, T. Di Battista, E. Romano, R. Verde

**Abstract** This paper focuses on measuring the spatial correlation among functional data streams recorded by sensor networks. In many real world applications, spatially located sensors are used for performing at a very high frequency, repeated measurements of some variable. Due to the spatial correlation, sensed data are more likely to be similar when measured at nearby locations rather than in distant places. In order to monitor such correlation over time and to deal with huge amount of data, we propose a strategy based on computing the well known Moran's index on summaries of the data.

**Abstract** *Il presente articolo è incentrato sul misurare la correlazione spaziale tra data stream acquisiti da una rete di sensori. In molti campi applicativi, mediante l'utilizzo di sensori è possibile effettuare, ad elevata frequenza, misurazioni ripetute di fenomeni reali. Spesso, a causa della posizione geografica dei sensori, è presente una correlazione spaziale tra le osservazioni. In particolare, sensori spazialmente vicini registrano dati tra loro più simili rispetto a quanto rilevato da sensori lontani. Al fine di monitorare tale correlazione nel tempo, tenendo conto dell'elevata numerosità delle registrazioni effettuate dai sensori, si propone una strategia basata sul calcolare il ben noto indice di Moran su sintesi dei dati.*

**Key words:** Data stream mining, Functional Data Analysis, Moran Index

---

A. Balzanella, E. Romano, R. Verde  
University of Campania Luigi Vanvitelli, Italy e-mail: antonio.balzanella@unicampania.it, elvira.romano@unicampania.it, rosanna.verde@unicampania.it

S. A. Gattone, T. Di Battista  
University G. d'Annunzio of Chieti-Pescara, Italy, e-mail: gattone@unich.it, dibattis@unich.it

## 1 Introduction

Functional Data Analysis (FDA) has become a topic of interest in Statistics due to the increasing ability to measure and record over a continuous domain results of natural phenomena [6]. In environmental sciences, monitoring a physical phenomenon in different places of a geographic area is becoming very common due to the availability of sensor networks which can perform, at a very high frequency, repeated measurements of some variable. We can think, for instance, at temperature monitoring, seismic activity monitoring, pollution monitoring, over the locations of a geographic space. In this context one works with data having complex characteristics including spatial dependence structures. Often, the data acquisition is performed by sensors having limited storage and processing resources. Moreover, the communication among sensors is constrained by their physical distribution or by limited bandwidths. Finally, the recorded data relate, often, to highly evolving phenomena for which it is necessary to use algorithms that adapt the knowledge with the arrival of new observations. The data stream mining framework offers a wide range of specific tools for dealing with these potentially infinite and online arriving data. An overview of recent contributions is available in [2].

An emerging challenge, in this context, is the monitoring of the spatial dependence among sensor data. The First Law of Geography, also frequently known as Tobler Law [5], states that "everything is related to everything else, but near things are more related than distant things". This law finds its major developments in Geostatistics but is still valid in the framework of data stream mining, when the data is collected by spatially located sensors. For instance, surface air temperatures streams, are more likely to be similar when measured at nearby locations rather than in distant places.

Measuring the spatial dependence among fast and potentially infinite data streams is a very challenging task. This is due to a set of stringent constraints: i) the available time for processing the incoming observations is small and constant; ii) the allowed memory resources are orders of magnitude smaller than the total size of input data; iii) only one scan of the data is feasible; iv) the communication between the sensors should be very limited.

This paper introduces a new strategy for monitoring the spatial dependence over time which adapts the classic Moran's index to the challenge of functional data stream processing.

We assume that sensors do not communicate with each other but only with a central node. Thus, a first part of the processing is performed at the sensors while a second part is performed at the central computation node using the output of the sensors. In particular, each data stream recorded by a sensor is processed individually through two summarization steps. The first one, splits the incoming data stream into non overlapping windows and provides a compact representation of the observation in each window. The second step, performs on each data stream a CluStream ([1]) algorithm adapted for working on functional data subsequence. CluStream groups the incoming data into homogeneous micro-clusters and represent these through prototypes.

With the flowing of data, each sensor performs two kinds of data transmission to the central computation node. The first one is a snapshot of the micro-cluster centroid at predefined time stamps. The second one, which is performed at each window, consists in sending the identifier of the micro-cluster to which the subsequences have been allocated. In this way, the communication between the sensors and the central node requires a low bandwidth as well as low memory resources. Only few micro-cluster prototypes are stored for each data stream at the central node and the sensor data are replaced by the micro-cluster centroid to which they have been allocated by the CluStream.

The central processing node is, still, used for measuring the spatial dependence among the streams computing the Moran's index on the micro-cluster centroids.

The next sections provide the details of the processing setup.

## 2 Sensor data summarization through on-line clustering

Let  $Y = \{Y_1(t), \dots, Y_i(t), \dots, Y_n(t)\}$  be a set of  $n$  functional data streams.  $Y_i(t)$ ,  $t \in T$ , denotes a function defined on an interval with  $T \subseteq \mathfrak{R}$ . Each functional data stream  $Y_i(t)$  is made by observations recorded by a sensor located at  $s_i \in S$ , with  $S \subset \mathfrak{R}^2$  be the geographic space.

We assume that the potentially infinite data is recorded on-line so that we can keep into memory only subsets of the streams. Thus, the analysis is performed using the observations in the most recent batch and some synopsis of the old data, no longer available.

In reality, we observe the data at a grid of  $N$  points,  $t_1, \dots, t_N$ . The functional data analysis viewpoint may be described by the following non-parametric model:

$$Y_{ij} = f_i(t_j) + \varepsilon_{ij} \quad (1)$$

where  $f_i(t)$  is the underlying signal curve,  $\varepsilon_{ij}$  is an observation noise with mean zero and null covariance and  $Y_{ij}$  denote the observed noisy data,  $i = 1, \dots, n$  and  $j = 1, \dots, N$ .

We split the incoming data streams into non overlapping windows identified by  $w = 1, \dots, \infty$ . A window is an ordered subset of  $T$ , having size  $b$  which frames, for each  $Y_i(t)$ , a data batch  $Y_i^w(t) = \{Y_i(t)\}_{t=j}^{j+b}$ .

A CluStream ([1]) algorithm, suitably adapted for working with the functional subsequences  $Y_i^w(t)$  of the data stream  $Y_i(t)$  is used for providing a fast to compute summarization of the stream (more details will be provided in the extended version of this manuscript).

The intuition that underlies the method, is to represent the incoming data through the center of low variability (micro)clusters. In order to have a high representativity of the input data, the number of clusters to keep updated is not specified a priori but only a threshold on their maximum number is fixed, to manage the memory resources.

As mentioned above, the data structure we use for data summarization is named micro-cluster. For each stream  $Y_i(t)$ , we keep a set  $\mu C_i = \{\mu C_1^k, \dots, \mu C_i^k, \dots, \mu C_i^K\}$  of micro-clusters, where  $\mu C_i^k$  records the following information:

- $\bar{Y}_i^k(t)$ : the cluster centroid;
- $n_i^k$ : number of allocated functions;
- $\sigma_i^k(t)$ : Standard deviation;
- $Sw_i^k$ : Sum of window Id;
- $SSw_i^k$ : Sum of squared window Id.

Whenever a new window  $w$  of data is available, CluStream allocates the subsequence  $Y_i^w(t)$  to an existing micro-cluster or generates a new one. The first preference is to assign the data point to a currently existing micro-cluster. If we choose the squared  $L^2$  distance as our dissimilarity metric between two functions defined as

$$d^2(Y_i, Y_i') = \int_0^T [Y_i(t) - Y_i'(t)]^2 dt, \quad (2)$$

then,  $Y_i^w(t)$  is allocated to the micro-cluster  $\mu C_i^k$  if

$$d^2(Y_i^w(t), \bar{Y}_i^k(t)) < d^2(Y_i^w(t), \bar{Y}_i^{k'}(t)) \quad (3)$$

and

$$d^2(Y_i^w(t), \bar{Y}_i^k(t)) < u \quad (4)$$

with  $k \neq k'$  and  $k = 1, \dots, K$ .

The threshold value  $u$  allows to control if  $Y_i^w(t)$  falls within the maximum boundary of the micro-cluster, which is defined as a factor of the standard deviation of the subsequences in  $\mu C_i^k$ . In order to take into account the functional nature of the data, a pre-smoothing step may be applied before clustering [3].

The allocation of a subsequence to a micro-cluster involves the updating of its information. The first update is the increasing by 1 of  $n_i^k$ . Then, it is necessary to update micro-cluster centroid and standard deviation. Finally, it is necessary to update the sum and the sum of squares of the window identifiers considering the time window  $w$ .

If  $Y_i^w(t)$  is outside the maximum boundary of any micro-cluster because of the evolution of the data stream, a new micro-cluster is initialized setting the  $Y_i^w(t)$  as centroid and  $n_i^k = 1$ . The functional standard deviation  $\sigma_i^k(t)$  is defined in a heuristic way by setting it to the pointwise squared Euclidean distance to the closest cluster.

The proposed procedure, performed in a parallel way on all the streams, permits to keep, at each time instant, a snapshot of the data behavior. This is due to the availability of the set of subsequences used as representatives.



### 3 Moran's index on data stream summaries

Moran's index ([4]) is a widely used measure for testing the global spatial autocorrelation in spatial data. It is based on cross-products of the deviations from the mean and is calculated for the  $n$  observations of a variable  $X$  at locations  $i, i'$ , as:

$$I = \frac{n}{\sum_i \sum_{i'} a_{i,i'}} \frac{\sum_i \sum_{i'} a_{i,i'} (x_i - \bar{x})(x_{i'} - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad (5)$$

where the weights  $a_{i,j}$  define the relationships between locations in the geographic area.

Morans index is similar, but not equivalent, to a correlation coefficient. It varies from  $-1$  to  $+1$ . In the absence of autocorrelation and regardless of the specified weight matrix, the expectation of Morans I statistic is  $-1/(n-1)$ , which tends to zero as the sample size increases.

According to the processing setup introduced above, at the central computation node it is kept a snapshot of micro-cluster centroids of each stream. Every time a new window becomes available, it is possible to measure the spatial autocorrelation by receiving at the central node, from each data stream, the identifier of the micro-cluster to which the subsequence of the window has been allocated. This approach, allows to measure the spatial dependence of the data in a window by using the micro-cluster centroids rather than the raw sensor data.

In this sense, the Moran's index can be computed by:

$$I = \frac{n}{\sum_i \sum_{i'} a_{i,i'}} \frac{\sum_i \sum_{i'} a_{i,i'} \int_0^T [\bar{Y}_i^k(t) - \bar{Y}(t)][\bar{Y}_{i'}^k(t) - \bar{Y}(t)]}{\sum_i \int_0^T [\bar{Y}_i^k(t) - \bar{Y}(t)]^2} \quad (6)$$

where  $\bar{Y}_i^k(t)$  and  $\bar{Y}_{i'}^k(t)$  are the micro-cluster centroids to which, respectively, the subsequences  $Y_i^w(t)$ ,  $Y_{i'}^w(t)$ , have been allocated and  $\bar{Y}(t)$  is the average subsequence.

The proposed Moran's index can be used for obtaining a different measure of the spatial dependence at every time window  $w$ , starting from the micro-cluster identifiers sent by the sensors to the central communication node.

### 4 Conclusions and perspectives

In this paper we have introduced an approach for measuring the spatial autocorrelation among functional data streams recorded by sensors. Since the main spatial dependence measures require a high computational effort, we have proposed to perform a data summarization and to compute the spatial autocorrelation on the summaries rather than on the original data. Preliminary tests on simulated data confirm the effectiveness of the proposed summarization strategy in keeping track of the spatial correlation structure.

## References

1. Aggarwal C. C., Han J., Wang J., Yu P. S.: A framework for clustering evolving data streams. In: VLDB 2003: Proceedings of the 29th international conference on Very large data bases, pp. 812. VLDB Endowment, 2003.
2. Garofalakis M., Gehrke J., Rastogi R: Data Stream Management: Processing High-Speed Data Streams. Springer, New York (2016).
3. Hitchcock, D.B., Casella, G., Booth, J.G.: Improved Estimation of Dissimilarities by Presmoothing Functional Data. *J. Am. Stat. Assoc.* 101 (473), pp 211-222 (2006).
4. Moran, P.A.P.: Notes on continuous stochastic phenomena. *Biometrika* 37, pp 17-23 (1950).
5. Tobler, W: A computer movie simulating urban growth in the Detroit region. In: *Economic Geography* Vol. 46(2) pp 234-240 (1970)
6. Wang, J.L., Chiou, J.M., Muller, H.G.: Functional Data Analysis. *Annu. Rev. Statist.* 3, pp 257-295 (2016).

# User query enrichment for personalized access to data through ontologies using matrix completion method

Oumayma Banouar and Said Raghay

**Abstract** Current information systems provide transparent access to multiple, distributed, autonomous and potentially redundant data sources. Their users may not know the sources they questioned, nor their description and content. Consequently, their queries reflect no more a need that must be satisfied but an intention that must be refined. The purpose of the personalization is to facilitate the expression of users' needs. It allows them to obtain relevant information by maximizing the exploitation of their preferences grouped in their respective profiles. In this work, we present a matrix completion method that minimize the nuclear norm to construct our users' profiles. Then we expose their query enrichment process expressed in SPARQL to interrogate data sources described by ontologies.

**Abstract** *Attuali sistemi informativi permettono un più facile accesso a fonti di dati multiple, distribuite, autonome e potenzialmente ridondanti. I loro utenti potrebbero non conoscere le fonti hanno messo in discussione, né la loro descrizione e ne' il contenuto. Di conseguenza, le query riflettono non più un bisogno che deve essere soddisfatto ma un'intenzione che deve essere raffinata. Lo scopo della personalizzazione è quello di facilitare l'espressione delle esigenze degli utenti. Esso consente loro di ottenere informazioni pertinenti, massimizzando l'espressione delle loro preferenze raggruppate nei rispettivi profili. In questo lavoro, presentiamo un metodo per il completamento di una matrice che minimizza la norma nucleare per costruire i profili dei nostri utenti. Poi, si mostra il processo di arricchimento delle query espresse in SPARQL, per interrogare le fonti di dati descritti da ontologie.*

**Key words:** Personalization, User profile construction, Matrix completion, Enrichment, Ontologies.

---

<sup>1</sup> Oumayma BANOUAR, LAMAI- Faculty of Science and Technics University Cadi Ayyad, Marrakesh, Morocco ; o.banouar@edu.uca.ma:

Said RAGHAY, LAMAI- Faculty of Science and Technics University Cadi Ayyad, Marrakesh, Morocco ; s.raghay@uca.ma:

## 1 Introduction

The multiplicity of data sources, their scalability and the increasing difficulty to control their descriptions and their contents are the reasons behind the emergence of the need of users' requests personalization. A major limitation of these systems is their inability to classify and discriminate users based on their interests, their preferences and their query contexts. They cannot deliver relevant results according to their respective profiles[1]. Consequently, the execution of the same request expressed by different users over an ontology-based mediation system will necessarily not provide the same results. We will talk here about a personalized access to data sources using ontologies. A user accessing an information system with the intention of satisfying an information need, may have to reformulate the query issued several times and sift through many results until a satisfactory, if any, answer is obtained. This is a very common experience. A critical observation is that: different users may find different things relevant when searching, because of different preferences, goals etc. Thus, they may expect different answers to the same query. The personalization of a query uses the user profile to rephrase his request by integrating elements of his interests or his preferences. Storing user preferences in a user profile gives a retrieval system the opportunity to return more focused, personalized and hopefully smaller answers.

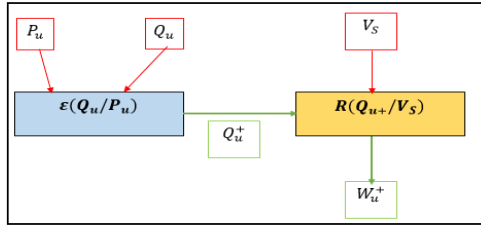
The objective of the query personalization process is to enhance the user query with his related preferences stored in his profile. This process focuses on the system user, enables the exploitation of what is called personal relevancy[2] instead of consensus relevancy. In the first one, the information system computes relevancy based on each individual's characteristics, unlike the second one where it presumes that the relevancy computed for the entire population is relevant for each user. This work presents a matrix completion method based on the optimisation of the nuclear form of the matrix that represents the preferences of our users over items. It then enrich the user query expressed in SPARQL to be evaluated over ontologies.

The remaining of this paper is organised as follows. Sections 2 and 3 present our proposed approach where section 4 discusses our experimental results. Finally, we conclude by exposing the next challenges for data management using learning methods in an information retrieval context.

## 2 Proposed approach

In our work, the enrichment and the rewriting process are dependent. These two algorithms add predicates to the user query. Profile predicates for enrichment and semantic links for rewriting.[1] It identifies the contributing sources in the execution of the user query and uses their definitions to reformulate it. The user expresses his query according to the terms of a global schema that procures a transparent access to multiple data sources. The rewriting process transforms it in order to evaluate it on

User query enrichment for personalized access to data through ontologies using matrix completion method 111  
the different data sources schemas. However, the user query rewritten is an enriched one. It contains predicates reflecting the preferences of the user over the research domain. The user profile regroups these preferences.



**Figure 1: Enrichment-rewriting process**

With:  $P_u$ : User profile,  $Q_u$ : User query,  $V_s$ : Data sources descriptions,  $Q_u^+$ : Enriched user query,  $W_u^+$ : Enriched user query rewritten.

The first process -the personalization process (enrichment process)- integrates elements of the centre of interests or preferences of the user in his query. Based on [2], the user profile is composed of a set of weighted predicates. The weight of a predicate expresses its relative interest to the user. It is specified by a real number between 0 and 1. The phase of profile construction should rely on a machine learning method.

Once the user query enriched the second process, take place. It is the query rewriting process. It depends on the way the system defines mappings. Our system adopts the Local-As-View approach. It defines each sources relationships by a query in terms of the virtual or global schema obtained by a global ontology. This mediation approach facilitates the incorporation of the dynamicity sources. Indeed, changing a source means changing a single query.

Each data sources integrated in the system disposes of a local schema describing its structure and content. This schema is obtained through a local ontology. The adoption of a Local- As-View approach for mediation assumes that the system defines every data source according to the terms of the global schema procured by a global ontology. This system frees the user from having to locate sources that are relevant to a query; interacts with each one in isolation, and combines data from multiple sources. The users do not ask queries in terms of the schemas in which the data is stored, but in terms of the mediated (global) schema. The mediated schema is a set of relations designed for a specific data integration application, and contains the salient aspects of the domain under consideration. The tuples of its relations are not actually stored in the data integration system. Instead, it includes a set of sources descriptions that provide semantic mappings between the relations in the sources' schemas and the relations in the mediated schema. An ontology-based mediator purveys information and provides mutual access to knowledge. [1]

The users' profiles construction is the key personalization enabler and the useful tactic in data integration tasks dealing with irrelevancy problem. It takes elements of the user preferences as input and determines his profile as output. A user profile is a

set of weighted elements that defines preferences of its owner over items. Machine learning approach enable the possibility to manipulate profiles automatically as much as possible.

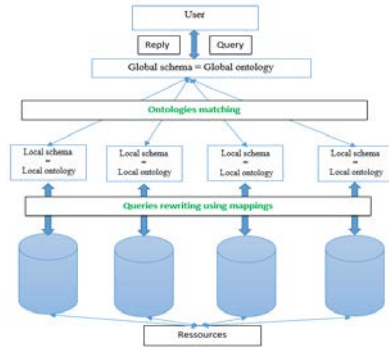


Figure 2: Ontology-based mediation architecture

### 3 Matrix completion for personalized access to data

Our proposed enrichment query process relies on the three main following steps:  
 - A learning process to identify users and preferences clusters; A predictive method using clusters found in step 1: A user query enrichment using the predicted preferences in step 2.

We start our learning process by a Singular Value Decomposition of our matrix  $R(m,n)$  modelling the users of our systems as its rows and the preferences as its columns and its transpose  $R'$ . The objective is to reduce the dimensionality of the matrix. This process apply the K-means algorithm twice. The first one, to obtain the users clusters while the second to obtain the preferences clusters. After the clustering of the items and users, the prediction process starts in the aim to complete the ratings given by a user for a corresponding item. For a given user, respectively an item, we identifies clusters in which the selected user, respectively the preference, belongs. The predicted score or rate is the result of Singular Value Thresholding SVT algorithm [3] applied on the matrix containing rates that users in the selected user cluster given to preferences in the selected preference cluster . It has as an objective function

$$\text{Minimize } \|R\|_* + \frac{1}{2} m_u \|R\|_F^2$$

$$\text{Subject to } \Omega x X=b$$

Where:  $\|R\|_*$  is the nuclear norm that sums the amplitude of the nonvanishing singular values and  $\|R\|_F$  is the Frobenius norm of the matrix  $R$ .

The adopted algorithm takes as parameters three mandatory elements.

-  $\Omega$  the set of locations corresponding to the observed entries. It might be defined in three forms. The first one as a sparse matrix where only the elements different of 0 are to take into account. The second one as a linear vector that contains the position of the observed elements. And the third one where  $\Omega$  is specified as indices  $(i,j)$  with  $(i,j) \in N$ .

-  $b$  the linear vector which contains the observed elements.

-  $m_\alpha$  the smoothing degree.

The application of the SVT algorithm in blocks procures in some cases certain results that are out of range. In this case, we use an aggregation process to predict the following rates. It is equal to the mean of all rates found by intersection between the cluster to which the user belongs and the cluster that contains the preference. For each user, the preferences with their weights are classed to select the  $K$  predicates with the heights weights to integrate to the initial user query for its evaluation over an ontology describing a data source.

## 4 Experimental results

The evaluation of the approach is done using the MovieLens dataset. The MovieLens dataset consists of: 1- 100 000 ratings from 943 users on 1682 films from 1 to 5. 2- Each user has rated at least 20 movies. 3- The data sets are 80% 20% splits into training and test data.

We performed the first step of our approach to detect 10 clusters, 5 for users and another 5 for films according to the rating scale. This step has as complexity of  $O(nkt)$  where  $n$  refers to the number of data objects while  $t$  is the number of iteration,  $k$  of course is the number of classes generated. The second step that corresponds to the predictive method allowed us to recover the initial matrix  $R$  as the matrix  $\tilde{R}$  the which dimension is  $943 \times 1682$  from only 100 000 known data that corresponds to almost 6.5% of global data. In the objective to demonstrate the efficiency of the combination between the aggregation method and the SVT algorithm per blocks, we applied several methods of Low-Rank Matrix Recovery and Completion over our experimental data. These methods minimize the nuclear norm of our users-preferences matrix in the aim to recover the missing data with a precise rank. We cite Augmented Lagrange multiplier method ALM [4], Accelerated Proximal Gradient method APG[5], Dual Method DM [6] and Fixed-Point Continuation method FPC[7]. Only SVT, FPC and ALM algorithms recovered the matrix with the desired rank 943.

The following table presents a comparison of the results obtained according to four metrics: Mean Absolute Error MAE, Root Mean Square Error MSE, Relative recovery error, Relative recovery in the spectral.

**Table 1:** Experimental results

<i>Method</i>	<i>MAE</i>	<i>RMSE</i>	<i>E<sub>1</sub></i>	<i>E<sub>2</sub></i>
Proposed approach	1.105209e-02	1.822554e-02	1.451178e-01	3.788417e-02
SVT algorithm	3.045597e-02	3.570902e-02	2.031278e-01	1.207209e-01
FPC algorithm	3.563525e-02	4.086687e-02	2.173032e-01	1.294834e-01
ALM algorithm	7.431944e-02	1.978355e-01	4.781151e-01	2.570103e-01

## 5 Conclusion

A major limitation of the ontology based information retrieval systems is their inability to deliver pertinent results according to the users preferences. Indeed, they depend on the users' queries, which are insufficient for giving a complete picture about what the users are looking for. In fact, these systems return the same result regardless of who submitted the query. In addition, the same user query is not essentially the same intent. In this work, we presented a construction profile process that is considered to enrich the user query expressed in SPARQL. It is based on three main steps which are: A learning process to identify users and preferences clusters. A predictive method using clusters found in step 1 that is based on the SVT algorithm for nuclear norm minimization and an aggregation function. A user query enrichment using the predicted preferences in step 2. The next challenge now is to increase the pertinence and the precision of information retrieval process by updating automatically the profile after each user-system interaction. We talk here about two operations namely: the user construction from a small-observed set of data and the user profile overloading after a modification of a user preference.

## References

- [7] E. T. HALE, W. YIN and Y. ZHANG, "Fixed-point continuation for  $\ell_1$ -minimization: Methodology and convergence", *SIAM Journal on Optimization*, Vol. 19, no 3, p. 1107-1130, 2008.
- [2] G. Koutrika and Y. Ioannidis, "Personalizing queries based on networks of composite preferences", *ACM Transactions on Database Systems*, Vol 35, pp. 1-50, 2010.
- [3] J. Cai, J.E. Candès, C. Zuowei, "A singular value thresholding algorithm for matrix completion", *SIAM Journal on Optimization*, Vol. 20, no. 4, pp. 1956-1982, 2010.
- [1] O. Banouar and S. Raghay, "User profile construction for personalized access to multiple data sources through matrix completion method", *IJCSNS*, vol. 16, no. 10, pp. 51-57, 2016.
- [4] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices," *UIUC Technical Report UILU-ENG-09-2215*, November 2009.
- [5] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast Convex Optimization Algorithms for Exact Recovery of a Corrupted Low-Rank Matrix", *UIUC Technical Report UILU-ENG-09-2214*, August 2009.
- [6] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast Convex Optimization Algorithms for Exact Recovery of a Corrupted Low-Rank Matrix", *UIUC Technical Report UILU-ENG-09-2214*, August 2009.



# The Trieste Observatory of cardiovascular disease: an experience of administrative and clinical data integration at a regional level

## *L'Osservatorio delle malattie cardiovascolari di Trieste: un'esperienza di integrazione di dati amministrativi e clinici a livello locale*

Giulia Barbati<sup>1</sup>, Francesca Ieva<sup>2</sup>, Francesca Gasperoni<sup>2</sup>, Annamaria Iorio<sup>3</sup>, Gianfranco Sinagra<sup>1</sup>, Andrea Di Lenarda<sup>4</sup>

**Abstract** The Trieste Observatory of Cardiovascular Diseases has been established in 2009 with the aim to integrate administrative and clinical data sources in order to conduct epidemiological studies based on real-world population. Our interests are focused on two main areas: from an epidemiological point of view, the aim is minimizing various source of bias in the design of the study, that commonly could arise in observational settings and are moreover a crucial point when using administrative data sources. Methodological research comparing different approaches to the analysis of longitudinal measures and their impact on time to event data, considering recurrent and competing events, represents the specific statistical domain of interest.

**Abstract** *L'Osservatorio delle malattie cardiovascolari di Trieste è stato istituito nel 2009, con l'obiettivo di integrare fonti di dati amministrativi e clinici per condurre studi epidemiologici basati su popolazioni del mondo reale. Il nostro*

---

<sup>1</sup> G. Barbati, Department of Medical Sciences, University of Trieste, Italy; email: [gbarbati@units.it](mailto:gbarbati@units.it)

<sup>2</sup> F. Ieva & F. Gasperoni, MOX-Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Italy; email: [francesca.ieva@polimi.it](mailto:francesca.ieva@polimi.it)

<sup>3</sup> A. Iorio, Cardiology Unit, Papa Giovanni XXIII Hospital, Bergamo, Italy

<sup>4</sup> A. Di Lenarda, Cardiovascular Center, Trieste, Italy

*interesse è focalizzato in due aree: dal punto di vista epidemiologico, minimizzare le possibili distorsioni nel disegno dello studio, particolarmente frequenti in ambito osservazionale ed in particolare utilizzando fonti amministrative. Dal punto di vista statistico, l'interesse è centrato sul confronto tra metodologie appropriate per l'analisi di dati longitudinali nel contesto di dati di sopravvivenza ed eventi ricorrenti e rischi competitivi.*

**Key words:** Administrative Health Data, Epidemiology, Cardiovascular Diseases, Population Attributable Fraction, Multi-State Models

## 1 Introduction

At the Outpatient Clinic of the Cardiovascular Center and at the Cardiovascular Department of the University Hospital of Trieste, the “Trieste Observatory of Cardiovascular Diseases” has been established in 2009. The Observatory is based on the integration at a regional level of administrative and clinical data sources.

The administrative source is the Regional Epidemiological Repository (RER) of the Friuli-Venezia-Giulia region, that includes several databases such as: Registry of Births and Deaths, Hospital Discharge (SDO), Laboratory tests performed in the public hospital as in- or out-patients, reimbursable by the National Health System (NHS), Public Drug Distribution System database, District Healthcare Services (Intermediate and Home Care). The clinical source is the electronic-chart (Electronic Health Recording, HER, Cardionet®) a cardiological medical record that includes medical information and history as collected by cardiologists during routine clinical practice (both in ambulatory visits and during cardiological hospitalizations). Diagnostic codes, laboratory tests, procedures (as for example echocardiography and electrocardiogram) and cardiological drugs prescription are recorded at each visit. Medical records are routinely reviewed by clinicians in each clinical evaluation to update medical history, diagnostic procedures, and treatment. This integrated database covers the Trieste population, i.e., 237,000 inhabitants. The RER is implemented in a SAS system, and the e-chart Cardionet® has been fully integrated in the same Data Warehouse via a dynamic single anonymous identification code. Previously, data were extracted by means of ad-hoc queries in a Business Object system, with the possibility to identify patients. Of note, population-based research of cardiovascular diseases is feasible in the area of Trieste because public health system care is largely dominant (87.1% of all cardiovascular ambulatory clinical evaluations, based on administrative reports). This is one of the first attempt in Italy, to the best of our knowledge, of a systematic integration between administrative and EHR system at regional level.

## 2 Applications

We have focused our interest on evaluating characteristics and outcome of “real-world” Heart Failure (HF) patients. HF prevalence steeply increases with aging, from less than 1% in the population aged between 20 and 39, to more than 20% in individuals over 80 [1, 2]. Population based studies report that one-year mortality rate ranges from 35% to 40% [3, 4] and more than 50% of patients are readmitted to hospital between six and twelve months after the first diagnosis [5]. In this epidemiological scenario, elders with HF are representative of the growing segment living longer with chronic health conditions prone to multiple transitions from hospital to home that negatively affect their quality of life and consume substantial healthcare resources [6]. Moreover, HF is a highly clinical variable syndrome that occurs across the entire range of left ventricular ejection fraction (LVEF), from patients with preserved LVEF (HFpEF: LVEF  $\geq$  50%) to those with reduced LVEF (HFrEF: LVEF  $<$  50%). In line with the aging of population, there is an increase in concomitant noncardiac conditions affecting chronic HF patients. These comorbidities frequently complicate management and may contribute to adverse outcomes. Given this public health issue, there is an urgent need to rearrange HF healthcare systems in order to improve evidence-based practice and create seamless care systems. To this purpose, we recently conducted two studies using data from the Observatory: in the first one, the aim was exploring the differential prevalence and the attributable risk of noncardiac comorbidities on outcomes between HFrEF and HFpEF patients in a large contemporary, community-based population. In the second one, we investigated clinical factors contributing to lengthen hospital stays and to increase multiple readmission rates to both hospitals and community services as Integrated Home Care (IHC) activations and Intermediate Care Unit (ICU) admissions.

### ***2.1 Prevalence and Prognostic Impact of Noncardiac Comorbidities in Heart Failure Outpatients: selection of the cohort, covariates and outcomes definition***

The cohort of consecutive ambulatory HF patients that attended the Outpatient Clinic of the Cardiovascular Center of Trieste between November 2009 and December 2013 was selected, and the first visit in the period was considered as a starting point (index visit). To identify the cohort we used firstly the EHR with diagnostic codes presenting clinical findings compatible with HF. The diagnosis of HF was made by using criteria of the European Cardiology Society [7]. Patients were divided into two groups according to the LVEF: preserved (LVEF  $\geq$  50%) and reduced (LVEF $<$ 50%). Clinical variables, including cardiac and noncardiac comorbidities, were determined according to the data of EHR integrated with diagnoses based on ICD-9CM derived from previous hospitalizations, laboratory data, and/or specific treatment of chronic illnesses. On the basis of the Charlson comorbidity index [8], we included the following noncardiac comorbidities: peripheral artery disease (PAD),

cerebrovascular accident, dementia, chronic obstructive pulmonary disease (COPD), rheumatological disorders, acquired immunodeficiency syndrome, peptic ulcer disease, diabetes mellitus, liver disease, malignancy, chronic kidney disease (CKD), psychiatric disorders, and anemia. According to Ather et al [9], we also included obesity and hypertension, because of their prognostic significance in HF patients. For each patient, the total number of comorbidities was calculated. Body mass index was calculated as the ratio of weight to square height ( $\text{kg}/\text{m}^2$ ), and obesity was defined if the body mass index was  $\geq 30 \text{ kg}/\text{m}^2$ . Hypertension was defined with a systolic blood pressure of  $\geq 140 \text{ mmHg}$  and/or a diastolic blood pressure of  $\geq 90 \text{ mmHg}$  at the index visit, and/or with a history of hypertension. Renal failure was defined in case of an estimated glomerular filtration rate (GFR)  $< 60 \text{ ml}/\text{min}/1.73\text{m}^2$ . Anemia was defined according to World Health Organization criteria (Hb  $< 13\text{g}/\text{dL}$  in men and  $12 \text{ g}/\text{dL}$  in women). Study outcomes of interest included death from any cause, all-cause hospitalization, HF hospitalization, and noncardiovascular hospitalization. Deaths were collected from the regional Registry of Birth and Deaths. First all-cause hospitalization, HF hospitalization, and noncardiovascular hospitalization were collected from the Hospital Discharge Registry. The principal discharge diagnosis for each hospitalization was assessed using primary ICD-9CM code, which is assigned by clinical personnel after discharge, and reflects the main reason for admission. The administrative censoring date was December 31th, 2014.

## ***2.2 Prevalence and Prognostic Impact of Noncardiac Comorbidities in Heart Failure Outpatients: statistical methods***

To examine the relationship between noncardiac comorbidities and outcomes, we estimated population attributable fractions (PAF) of each noncardiac comorbidity expressed by percentage in the overall HF population and in the LVEF subgroups. This proportion of incidence would not occur if the factor were eliminated [10]. The estimated PAF was reported with corresponding 95% confidence intervals (CIs). The unadjusted PAF in the exposed group ( $PAF_{exp}$ ) was calculated using the following formula:  $PAF_{exp} = (RR-1)/RR$ , where  $RR$  is the relative risk of event computed for the exposed group with respect to the unexposed group. The unadjusted PAF in the overall population ( $PAF_{pop}$ ) was calculated using the following formula:  $PAF_{pop} = p * (RR-1) / (p * (RR-1) + 1)$ , where  $p$  is the prevalence of exposure in the population. The corresponding adjusted estimates for both measures were derived from a logistic regression model adjusted for age and sex. In order to assess the interaction between LVEF groups and comorbidities (both individually, and as a sum of comorbidities per patients) hazard ratios of the interaction terms in Cox models adjusted for sex and age were calculated. To examine the effect of comorbidity load on all-cause mortality, HFrEF and HFpEF populations were divided into groups with different comorbidity loads (absence, 1, 2,  $\geq 3$  comorbidities); event curves for each comorbidity group were estimated using the Kaplan-Meier method within each LVEF group. Covariates for multivariable models of mortality were selected on the basis of a backward stepwise algorithm in a Cox proportional hazards model. The model included demographic, medical history, laboratory values, and the interaction between comorbidity load (absence, 1, 2,  $\geq 3$ ) and LVEF groups.

### ***2.3 Multi state modelling of Heart Failure care path: selection of the cohort, covariates and outcomes definition***

Between 2009 and 2014, a cohort of patients discharged with HF diagnosis hospitalized in the Trieste area was identified. HF diagnosis included ICD-9CM codes for HF (428:x) and hypertensive HF (402:01, 402:11, 402:91) according to the National Outcome Evaluation Program (in Italian PNE-Programma Nazionale Esiti) made by the National Agency of Regional Health-Care Services (in Italian AGENAS - AGENzia NAzionale per i servizi Sanitari regionali). Patients were classified as Worsening Heart Failure (WHF) or De Novo on the base of the presence of at least one HF hospitalization in the 5 years preceding the index admission, which is the first admission during the study period. The administrative censoring date was September 30th, 2015. For each cohort member, data included gender and age, length of stay, department of admission and discharge, diagnostic code at discharge, stay into Emergency/Intensive Care Units during the hospitalization, cardiological evaluation before the hospitalization (when performed), laboratory tests, LVEF (when performed). The Charlson Comorbidity Index was calculated using hospital diagnosis based on ICD-9CM that occurred within five years before the first admission and integrated with laboratory data and diagnosis recorded at the index admission. In particular, for the diagnosis of diabetes mellitus we integrated information about glycosylated haemoglobin at admission and the recorded diagnosis of diabetes mellitus in the previous 5 years. Similarly, to assess the presence of a chronic kidney disease, we integrated the creatinine value at admission to compute the estimated glomerular filtration rate (eGFR)  $< 60$  ml/min with the reported diagnosis of a chronic kidney disease in the previous 5 years. Study outcomes of interest included death for any cause, all-cause rehospitalization, and transitions in IHC/ICU. Deaths were collected from the regional Registry of Birth and Deaths. All-cause hospitalizations and admissions in IHC/ICU were collected respectively from the Hospital Discharge Registry and the District Healthcare Services database. The principal discharge diagnosis for each hospitalization was assessed using primary ICD-9CM code. Each cohort member was followed from the starting date (i.e. discharge from the index admission) until the end of the study or the date of death.

### ***2.4 Multi state modelling of Heart Failure care path: statistical methods***

The first multi-state model (hereafter Model 1) replicates a dynamic similar to the one described in [11] for repeated hospitalizations only (we are omitting community services in this case), i.e. a multi-state model fitting a cox-type regression for each transition. It provides a convenient description of the admission-discharge dynamics, pointing out which covariates act in certain transitions and how they affect the relative risk as well as the risk (i.e. the instantaneous probability) of moving from one status to another one. This model accounts for patient specific risk profile (distinguishing covariates acting on different transitions) as well as clinical information. The second model (hereafter Model 2) is still a multi-state model where patients are assumed to be in one of the following five states: in hospital, in ICU, in IHC, out (of hospital, or ICU or IHC) and dead. Through this model, we would like to detect what is the impact of patient characteristics on the risk of moving among these states. Both models include the adverse outcome of death as absorbing state. The death of a patient is a competing event with respect to all the other transitions.

## 4 Results

### 4.1 Prevalence and Prognostic Impact of Noncardiac Comorbidities in Heart Failure Outpatients

A total of 2772 patients met the pre-defined HF criteria during the study period. Of these, 209 (13%) patients were excluded because quantitative LVEF had not been documented, and 98 (4%) were excluded because of left side severe primary valvular disease. Thus, a total of 2314 patients met study selection criteria. Of these, 1373 (59%) patients were identified as HFpEF (i.e., LVEF  $\geq$ 50%) and 941 (41%) patients were identified as HFrEF. Overall, mean age was  $77 \pm 10$  with a substantial proportion of female patients (43%), significant background prevalences of ischemic heart disease (46%), hypertension (80%), and atrial fibrillation (54%). During median follow-up of 31 [IQRs 16 - 41] months, 472 (20%) patients died. Overall, there was a high morbidity burden, with first hospitalizations from any cause in 1533 pts (66%), hospitalizations for HF in 510 (22%), hospitalizations for noncardiovascular cause in 1422 (61%). Among all noncardiac comorbidities, anemia, CKD, COPD, diabetes mellitus, and PAD were all strongly associated with mortality in the overall HF population (adjusted HR [95% CI]: anemia=1.9 [1.5-2.4]; CKD=1.7 [1.3-2.1]; COPD=1.6 [1.3-1.9]; diabetes=1.4 [1.2-1.7]). Similar findings were observed for all-cause hospitalization, noncardiovascular, and HF hospitalizations (data not shown). Considering PAF (*PAF<sub>exp</sub>* and *PAF<sub>pop</sub>*) for all-cause mortality, anemia, CKD, diabetes mellitus, COPD, showed the highest quantitative contribution (*PAF<sub>pop</sub>* [95% CI]: anemia=14% [10-17]; CKD=17% [11-23]; COPD=14% [11-16]; diabetes=11% [8-15]). All other noncardiac comorbidities showed a PAF below 10%. Findings were similar for all-cause hospitalization, with exception of PAD which showed a high contribution only for all-cause hospitalization. For each LVEF groups, the noncardiac comorbidities presented similar quantitative contribution (data not shown). Concordantly, for all-cause mortality, noncardiac comorbidities had no significant interactions by LVEF, confirming no differences in their prognostic impact. This was confirmed to be similar for all-cause, HF, and noncardiovascular hospitalizations (data not shown). When we grouped HF patients according to comorbidities burden, the presence of  $\geq 3$  comorbidities was related with increased risk (HR 2.3, 95% CI: 2.1-3.5) for all-cause mortality. This trend was similarly observed in both LVEF groups ( $p=0.81$  for interaction). In multivariable Cox models, an increasing number of noncardiac comorbidities was associated with a higher risk for all-cause mortality (HR 1.25; 95% CI 1.1 - 1.3), all cause hospitalization (HR 1.17; 95% CI 1.12 - 1.23), HF hospitalization (HR 1.28; 95% CI 1.19-1.38), noncardiovascular hospitalization (HR 1.16; 95% CI 1.1 247 1-1.22). The multivariable Cox model revealed no significant difference in mortality rates between LVEF groups (HR 0.95; 95% CI: 0.63 to 1.42). This trend was confirmed also for morbidity outcomes (data not shown).

### 4.2 Multi state modelling of Heart Failure care path

A total of 4904 patients hospitalized with primary HF diagnosis between 2009 and 2014 were identified. The mode of clinical presentation of patients was De Novo HF, 4129 patients (84%), and WHF, 775 patients (16%). Overall, the mean age was  $81 \pm 10$  with a substantial proportion of female patients (55%), and significant background prevalences of non cardiac comorbidities. 2923 (71% out of 4129) De Novo patients had a previous hospitalization for

any cause. Indeed, more than half of the cohort (61%) had a renal disease and the median of LVEF, when recorded, was 53% (30% with LVEF < 40%; 13% with LVEF 40-49%; 57% with LVEF  $\geq$ 50%). Comorbidity burden was high, with the median of Charlson score of 2 (40% with Charlson score  $\geq$ 3). The rate of admission in cardiological ward (CW) was 23% at the first hospitalization. The median follow-up was 26 months, IQR [11-48]. In Model 1, a significant effect of aging and increasing of comorbidity burden on the rehospitalization risk was observed. Likewise, a relevant impact of WHF was observed in all readmission rates. No significant role of gender emerged. The probability of being discharged from hospital, i.e. shortening the Length Of Stay (LOS) was inversely related to age and Charlson score. Conversely, a direct relation with admission in cardiological ward was observed. When we considered the effect of covariates on risk of mortality related to readmission, the hospitalization in cardiological ward was protective up to the second hospitalization, while, after that, this effect was nullified. Ongoing, the aging and increasing of Charlson score were still associated with in and out of hospital death through all readmissions, whereas almost any adverse effect on death was observed for clinical condition of WHF. In Model 2 we observed that the aging and higher Charlson index increased the risk of being readmitted to hospital, ICU and IHC. When we considered the covariates effect on time spent in different states of Model 2, aging process was directly related to time spent in hospital, while it was inversely related to time spent in IHC. Furthermore, we noticed a higher risk of admission in ICU for female patients. WHF condition increased the risk of being readmitted to hospital and it behaved as a protective factor for death outside.

## 5 Conclusions and Perspectives

In the first study, we confirm in a contemporary community-based population that noncardiac chronic illnesses confer significant risk for mortality and hospitalization in HF patients. For the first time, we demonstrate the effect of noncardiac comorbidities, by estimating associated attributable risks in an HF community setting within each LVEF phenotype. Remarkably, the adverse impact of noncardiac chronic diseases appears similarly significant, irrespective of LVEF groups. Of all individual noncardiac comorbidities, CKD, anemia, diabetes mellitus, COPD, and PAD showed the highest significant association with mortality and morbidity.

In the second application, for the first time, we demonstrate the effect of certain clinical conditions in a community setting on multiple readmissions by including intermediate care states (IHC/ICU). These findings significantly enhance our understanding of clinical pattern of patients with HF for adverse prognosis and have implications for the management approach to HF patients. From a policy perspective, identification of patients at high risk for multiple readmission must encourage the implementation of appropriate preventable intervention strategies [12]. Future developments in data sources integration will include the individual linkage of socio-economic indicators from the ISTAT Census 2011 at a regional level, and the development of persistence/adherence indicators to the therapy prescribed at the cardiological visits and after episodes of hospitalization. From an epidemiological point of view, efforts will be addressed in minimizing various source of bias in the study design, that commonly could arise in observational studies

and are moreover a crucial point when using administrative data sources. From a statistical point of view, methodological research comparing different approaches to the analysis of longitudinal measures and their impact on time to event data, considering recurrent/competing events, will be the area of interest [13,14,15].

## References

1. Lloyd-Jones D, Adams R, Brown T, et al. Executive summary: heart disease and stroke statistics, *Circulation*. 2010;121:948-54.
2. Senni M, Tribouilloy CM, Rodeheffer RJ, Jacobsen SJ, Evans JM, Bailey KR, et al. Congestive heart failure in the community. *Archives of Internal Medicine*. 1999;159(1):29-34.
3. Stewart S, MacIntyre K, MacLeod M, Bailey A, Capewell S, McMurray J. Trends in hospitalization for heart failure in Scotland, 1990-1996. An epidemic that has reached its peak? *European Heart Journal*. 2001;22(3):209-217.
4. Levy D, Kenchaiah S, Larson MG, Benjamin EJ, Kupka MJ, Ho KK, et al. Long-term trends in the incidence of and survival with heart failure. *New England Journal of Medicine*. 2002;347(18):1397-1402.
5. Tuppini P, Cucher A, de Peretti C, Fagot-Campagna A, Danchin N, Juilliere Y, et al. Two-year outcome of patients after a first hospitalization for heart failure: A national observational study. *Archives of cardiovascular diseases*. 2014;107(3):158-168.
6. Naylor MD, Brooten DA, Campbell RL, Maislin G, McCauley KM, Schwartz JS. Transitional care of older adults hospitalized with heart failure: a randomized, controlled trial. *Journal of the American Geriatrics Society*. 2004;52(5):675-684.
7. McMurray JJ V, Adamopoulos S, Anker SD, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart. *Eur Heart J*. 2012; 33(14):1787-847.
8. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373-83.
9. Ather S, Chan W, Bozkurt B, et al. Impact of noncardiac comorbidities on morbidity and mortality in a predominantly male population with heart failure and preserved versus reduced ejection fraction. *J Am Coll Cardiol*. 2012;59(11):998-1005.
10. Azimi SS, Khalili D, Hadaegh F, Yavari P, Mehrabi Y, Azizi F. Calculating population attributable fraction for cardiovascular risk factors using different methods in a population based cohort study. *J Res Health Sci*. 2015 Winter;15(1):22-7.
11. Ieva F, Jackson CH, Sharples LD. Multi-State modelling of repeated hospitalisation and death in patients with Heart Failure: the use of large administrative databases in clinical epidemiology. *Statistical methods in medical research*. 2015 pii: 0962280215578777.
12. Driscoll A, Meagher S, Kennedy R, Hay M, Banerji J, Campbell D, et al. What is the impact of systems of care for heart failure on patients diagnosed with heart failure: a systematic review. *BMC Cardiovascular Disorders*. 2016;16(1):195. doi:10.1186/s12872-016-0371-7.
13. Van Houwelingen, H. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* 34, 70-85.
14. Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* 67, 819-829.
15. R.B. Geskus, *Data Analysis with Competing Risks and Intermediate States*, 2015, Chapman and Hall/CRC.



# Marginal modeling of multilateral relational events

## *Modelli marginali per eventi relazionali multilaterali*

Francesco Bartolucci, Stefano Peluso and Antonietta Mira

**Abstract** We implement the methodology of marginal modeling of relational events involving groups of actors, as developed in [3]. Current relational data analyses suffer from the representation of an event through edge variables, with potential loss of information when the events generate a set of multiple relations rather than bilateral connections or ties. To fully exploit the informational content of relational events, we model an event as a binary vector of response variables representing actors participating to the event. Univariate and bivariate distributions of the events are modeled through marginal parameters having a clear social interpretation.

**Abstract** *Implementiamo il metodo proposto da [3] per la modellizzazione marginale di eventi relazionali che coinvolgono gruppi di attori di diversa numerosità. Diversamente dagli attuali metodi statistici disponibili in letteratura, nel modello marginale proposto, l'evento è rappresentato tramite un vettore binario che indica gli attori partecipanti o meno all'evento stesso. Presentiamo una parametrizzazione delle distribuzioni marginali univariate e bivariate facilmente interpretabile in termini di comportamento individuale e collettivo degli attori.*

**Key words:** Marginal Models, Multilateral Events, Relational Data, Social Networks

---

Francesco Bartolucci

Department of Economics, University of Perugia, Italy, e-mail: francesco.bartolucci@unipg.it

Stefano Peluso

Department of Statistical Sciences, Università Cattolica del Sacro Cuore e-mail: stefano.peluso@unicatt.it

Antonietta Mira

InterDisciplinary Institute of Data Science, Università della Svizzera Italiana, Switzerland, and Department of Sciences and High Technology, Università degli Studi dell'Insubria, Italy, e-mail: antonietta.mira@usi.ch

## 1 Introduction

In many relational contexts, a set of events is observed, with each event involving a group of actors; this gives rise to a set of multiple relations rather than to single bilateral connections. In these applications, the interest is not only in studying the relation between units, taking into account that the same event may involve more than two actors, but also to model the tendency of each actor to be involved in different events, not excluding cases of events participated by a single actor.

Common strategies to analyze social networks originated from a sequence of events are based on models for edge variables of type  $Y_{ij}$  associated to each possible pair of actors  $(i, j)$  in a network of  $n$  actors:  $Y_{ij}$  is equal to 1 if there is a connection between actors  $i$  and  $j$  and to 0 otherwise. Then, standard models for cross-sectional social networks, such as exponential random graph models (ERGM) or latent space/blocks model, may be fitted to draw conclusions [for a review see 10]. When the edge variables are time or event specific, more sophisticated strategies are based on models for network dynamics, of which at least three approaches may be highlighted: actor-oriented models [11], dynamic ERGMs [9], and hidden Markov models [12, 2].

To fully exploit the information provided by each relational event, we propose an approach based on the direct representation of the outcome of an event  $e$  by a vector of response variables  $\mathbf{Z}^{(e)} = (Z_1^{(e)}, \dots, Z_n^{(e)})$ , with  $Z_i^{(e)}$  equal to 1 if unit  $i$  is involved in the event and 0 otherwise. Following [3], we then formulate a statistical model for the response vectors  $\mathbf{Z}^{(e)}$ , with parameters having meaningful interpretations from a social behavior perspective. In particular, the distribution of  $\mathbf{Z}^{(e)}$  is parametrized through a marginal model [1, 4] based on the specification of its univariate and bivariate marginal distributions. The parameters involved in the univariate marginal distributions represent the general tendency of an actor to be involved in an event. The parameters involved in the second-order marginal distributions account for the concordance between behaviors of two actors, that is, the tendency to be jointly involved (or not) in the same event.

In our approach, we take advantage of the availability of a typically long series of events to estimate in a reliable way individual parameters associated to the actors in the network. In particular, we rely on a fixed-effects composite likelihood approach [8], which is rather straightforward to use even with large networks.

We pay particular attention to the representation of the results. In this regard we introduce parametrizations that allow us to represent *trajectories* of the tendency to be involved in an event and of the concordance of behaviors with other units, highlighting their evolution over time. Furthermore, we can express the association parameters based on the euclidean distance between two actors in terms of suitably estimated subject-specific latent vectors. Then, a “map” may be visualized, with close units characterized, as in the latent space model of [7], by a high chance to be tied.

## 2 Model Description and Inference

In the present section we briefly outline the modeling framework introduced in [3]. Let  $n$  be the number of actors and  $r$  the number of events, with the binary variable  $Z_i^{(e)}$  and the binary vector  $\mathbf{Z}^{(e)}$  defined in Section 1. Also define the bivariate probability vector

$$\mathbf{p}_{ij}^{(e)} = \begin{pmatrix} p(Z_i^{(e)} = 0, Z_j^{(e)} = 0) \\ p(Z_i^{(e)} = 0, Z_j^{(e)} = 1) \\ p(Z_i^{(e)} = 1, Z_j^{(e)} = 0) \\ p(Z_i^{(e)} = 1, Z_j^{(e)} = 1) \end{pmatrix}$$

and the corresponding frequency vector

$$\mathbf{w}_{ij}^{(e)} = \begin{pmatrix} I(z_i^{(e)} = 0, z_j^{(e)} = 0) \\ I(z_i^{(e)} = 0, z_j^{(e)} = 1) \\ I(z_i^{(e)} = 1, z_j^{(e)} = 0) \\ I(z_i^{(e)} = 1, z_j^{(e)} = 1) \end{pmatrix},$$

for  $i = 1, \dots, n-1$ ,  $j = i+1, \dots, n$ , and  $e = 1, \dots, r$ , with  $z_i^{(e)}$  denoting the observed value of  $Z_i^{(e)}$  and  $I(\cdot)$  denoting the indicator function. A marginal parametrization for the distribution of the random vector  $\mathbf{Z}^{(e)}$  is based on effects of the following type, for a suitable series of subsets of response variables with index in  $A$ :

$$\log \frac{p(\mathbf{Z}_A^{(e)} = \mathbf{z})}{p(\mathbf{Z}_A^{(e)} = \mathbf{0})} = \mathbf{g}_A(\mathbf{z})' \boldsymbol{\gamma}_A. \quad (1)$$

In more detail, we specify first- and second-order effects for all the individuals and events as

$$\eta_i^{(e)} = \log \frac{p(Z_i^{(e)} = 1)}{p(Z_i^{(e)} = 0)}, \quad i = 1, \dots, n, \quad (2)$$

which is a particular case of (1) when  $A = \{i\}$ , and

$$\eta_{ij}^{(e)} = \log \frac{p(Z_i^{(e)} = 0, Z_j^{(e)} = 0)p(Z_i^{(e)} = 1, Z_j^{(e)} = 1)}{p(Z_i^{(e)} = 0, Z_j^{(e)} = 1)p(Z_i^{(e)} = 1, Z_j^{(e)} = 0)}, \quad i, j = 1, \dots, n, i \neq j, \quad (3)$$

which is obtained from (1) with  $A = \{i, j\}$ . As already mentioned in Section 1, the marginal logit  $\eta_i^{(e)}$  and the log-odds ratio  $\eta_{ij}^{(e)}$  are interpreted, respectively, as a measure of tendency of subject  $i$  to be involved in event  $e$  and as a measure of the concordance between the behaviors of subjects  $i$  and  $j$  to collaborate in event  $e$ .

Parameter estimation is performed through numerical maximization of the pairwise composite log-likelihood function:

$$p\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{e=1}^r [\mathbf{w}_{ij}^{(e)}]' \log \mathbf{p}_{ij}^{(e)}, \quad (4)$$

where  $\boldsymbol{\theta}$  collects all individual parameter vectors and every probability vector  $\mathbf{p}_{ij}^{(e)}$  is obtained from suitable elements of  $\boldsymbol{\theta}$  by an explicit formula elaborated by [6]. This numerical maximization exploits an expression for the derivative of  $p\ell(\boldsymbol{\theta})$  with respect to  $\boldsymbol{\theta}$  that is rather easy to be computed.

For large networks we also propose an estimation algorithm that explicitly groups individuals in separate clusters. Each cluster includes actors that are homogenous in terms of tendency to be involved in an event and tendency to be involved (or not) in the same event (concordance). This method is based on the classification version of the pairwise log-likelihood function defined above, formulated following the approach adopted by [5] in a simpler context.

### 3 Application

The approach is illustrated by some applications in which trajectories of the tendency to be involved in an event and concordance with other units are of interest, with special focus on the closeness between pairs of units and on its suitable depiction.

In more detail, we apply our method to a temporal network of e-mail exchanges between users affiliated with a large European research institution.<sup>1</sup> The network was generated using anonymized information about all incoming and outgoing e-mails between members of the research institution. The e-mails represent communication between institution members only, and the dataset does not contain incoming messages from or outgoing messages to the rest of the world.

In our approach to analyze these data, each e-mail is seen as a multilateral event  $e$  involving a different number of recipients from which an appropriate realization of the vector  $\mathbf{Z}^{(e)}$  is easily obtained assigning value 1 to the elements corresponding to the receivers, while all other elements are equal to 0. We also have four sub-networks corresponding to the communication between members of four different departments at the institution. The whole network counts 986 nodes, with 332,334 temporal edges (time-stamped directed edges), spanned over 209,508 e-mails sent in 803 days. The relevance of studying not only bilateral relations between recipients is manifest in our application, since more than 18,000 e-mails are sent to more than two addresses, and more than 170,000 e-mails have a unique address, with a maximum of 39 receivers per e-mail. We report the diagram for the number of recip-

<sup>1</sup> the data are freely available at <https://snap.stanford.edu/data/email-Eu-core.html>.

ients per e-mail in Figure 1. Subject heterogeneity is clear: the number of received e-mails per subject ranges from 0 to 4,637, with mean and median equal, respectively, to 332 and 138 e-mails, and standard deviation 478.6, clearly evidencing a network composed of actors with very different social behaviors.

The application is based on two phases: fixed-effects estimation and clustering. In the first step, we assume polynomials of order 2 for the effect of time, and we estimate the corresponding parameter vectors. This first result allows the derivation of trajectories in terms of tendency to be recipient of an e-mail in a certain period. Periods can be created arbitrarily, merging sending times falling in the same specified period, or can be fixed to avoid the presence in the same period of e-mails involving multiple ties among users, or can coincide with the effective times the e-mails are sent. The other interpretation we can draw from the estimates of the fixed-effects parameters is in terms of tendency to be recipient (or not) with other users in the same e-mails. We are able to represent these tendencies and their evolution over time.

In the second phase of our procedure we cluster the users according to their behaviors in terms of tendency to receive e-mails and of tendency to receive (or not) common e-mails. The two clusters are not necessarily overlapping, since they express two different features of the user: a user can be very active in receiving e-mails but always involving the same restricted number of collaborators.

### Number of recipients per email

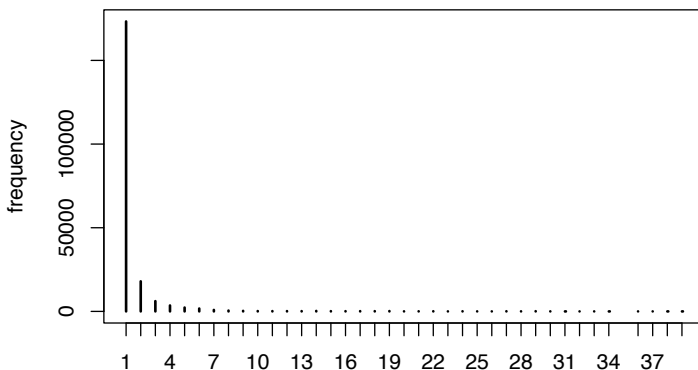


Fig. 1 Distribution of the number of recipients per e-mail

## References

1. Bartolucci, F., Colombi, R., Forcina, A.: An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica* **17**, 691–711 (2007)
2. Bartolucci, F., Marino, M.F., Pandolfi, S.: Composite likelihood inference for hidden Markov models for dynamic networks. Tech. report MPRA n. 67242 (2015)
3. Bartolucci, F., Peluso, S., Mira, A.: Modeling relational events via marginal models with individual-specific effects (2017). Mimeo
4. Bergsma, W., Croon, M.A., Hagenaars, J.A.: Marginal models: For dependent, clustered, and longitudinal categorical data. Springer Science & Business Media, NY (2009)
5. Choi, D.S., Wolfe, P.J., Airoldi, E.M.: Stochastic blockmodels with a growing number of classes. *Biometrika* **99**, 273–284 (2012)
6. Dale, J.R.: Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**, 909–917 (1986)
7. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97**, 1090–1098 (2002)
8. Lindsay, B.G.: Composite likelihood methods. *Contemporary Mathematics* **80**, 221–39 (1988)
9. Robins, G., Pattison, P.: Random graph models for temporal processes in social networks\*. *Journal of Mathematical Sociology* **25**, 5–41 (2001)
10. Snijders, T.A.: Statistical models for social networks. *Annual Review of Sociology* **37**, 131–153 (2011)
11. Snijders, T.A., Van de Bunt, G.G., Steglich, C.E.: Introduction to stochastic actor-based models for network dynamics. *Social Networks* **32**, 44–60 (2010)
12. Yang, T., Chi, Y., Zhu, S., Gong, Y., Jin, R.: Detecting communities and their evolutions in dynamic social networks: A Bayesian approach. *Machine Learning* **82**, 157–189 (2011)

# New Insights on Students Evaluation of Teaching in Italy

## *Nuove analisi sulle valutazioni degli studenti in Italia*

Bassi Francesca, Grilli Leonardo, Paccagnella Omar, Rampichini Carla and Varriale Roberta

**Abstract** This work presents new analyses on the relationship between student evaluation of teaching and student, teacher and course specific characteristics, exploiting the richness of information collected by a new survey carried out among professors of the University of Padua. Data collected in this survey are able to highlight teacher needs, beliefs and practices of teaching and learning. This allows to introduce in the study some *subjective* traits of the teachers. The role of these new variables in explaining student evaluations is deeply investigated.

**Abstract** *In questo lavoro vengono presentate delle nuove analisi sulla relazione fra le opinioni espresse dagli studenti per la valutazione della qualità della didattica universitaria e caratteristiche specifiche del corso, degli studenti e dei docenti, sfruttando la ricchezza di informazioni raccolte per mezzo di una nuova indagine realizzata tra i docenti dell'Università di Padova. Questa indagine è in grado di evidenziare i bisogni, le credenze e le pratiche dei docenti legate alle loro attività didattiche, permettendo di introdurre nelle analisi un insieme di caratteristiche soggettive dei docenti. Il loro ruolo viene quindi approfonditamente studiato nelle successive analisi.*

**Key words:** Record linkage, student evaluation of teaching, teacher opinions

Bassi Francesca  
Department of Statistical Sciences, University of Padua - Italy, e-mail: bassi@stat.unipd.it

Grilli Leonardo  
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence - Italy, e-mail: grilli@disia.unifi.it

Paccagnella Omar  
Department of Statistical Sciences, University of Padua - Italy, e-mail: omar.paccagnella@unipd.it

Rampichini Carla  
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence - Italy, e-mail: rampichini@disia.unifi.it

Varriale Roberta  
ISTAT (Italian National Statistical Institute) - Italy, e-mail: varriale@istat.it

## 1 Introduction

Students' opinions and judgements of teaching performances play a substantial role in higher education, particularly as instruments for gathering information on the quality of education and evaluating university courses [1, 8]. The relationship between student-, teacher-, course-specific characteristics and student evaluation of teaching (SET) is the topic of a huge amount of works in the literature (see an extensive review provided by [6]). However, findings concerning the relationship between SETs and the characteristics of courses, students and teachers are sometimes contradictory. Thus, these characteristics usually explain only a small portion of the total variance in SETs scores [5].

It is generally accepted that a multilevel analysis of the students' ratings is a satisfactory approach for investigating teaching evaluations, because of the hierarchical nature of the data (i.e. university students nested into classes) [3].

This work aims at enriching the multilevel literature on the student evaluation of teaching proposing some original analyses based on a wider set of teacher-specific characteristics, including also teachers' opinions on their teaching activities. This work exploits an innovative and original dataset available at the University of Padua, obtained after linkage of survey and administrative data coming from three different sources: first, the conventional survey on the student evaluation of teaching carried out among university students; second, administrative data related to the main features of the teachers and the didactic activities (DAs) they are involved; third, a new CAWI survey carried out by means of the research project PRODID (Teacher professional development and academic educational innovation). It started at the University of Padua in 2013, with the aim of developing strategies to support academic teachers and enhance their teaching competences. A specific questionnaire was then developed and addressed to all professors involved in almost all didactic activities of the University. This new survey collected opinions, beliefs and needs of the professors, with regard to their teaching activities developed in their classes.

This work is organised as follows. Section 2 introduces the data of this analysis, while the empirical application (model specification and results) is described in Section 3. Section 4 ends the paper, highlighting the main conclusions and some suggestions for future works.

## 2 The data

This work investigates data obtained by merging three different datasets coming from the University of Padua. The reference is the 2012-2013 academic year.

The first one is the standard online survey carried out by the University to measure students' opinions on the didactic activities. It involves all students who have been attending lessons of any degree courses of the Athenaeum. Students were asked to express their level of satisfaction on a scale from 1 to 10 (being 1 the lowest level) to a set of 18 items (seven if the student attended less than 30% of the lessons).



The second one is the administrative dataset that collects information on the teachers and the didactic activities of all Padua academic institutions.

The third one is an innovative dataset, collected by means of a new online survey aiming at providing a picture of the teaching experiences developed in the university classrooms. Indeed, the University of Padua in 2013 promoted the *PRO-DID* project (Teacher professional development and academic educational innovation - "Preparazione alla professionalità docente e innovazione didattica") with the purpose of developing an integrated system to improve teaching competences and academic innovation. The PRODID project promoted a research-based approach to creating training programs, faculty learning communities, pilot experimental contexts where teaching innovation could be tested and monitored ([2]). Following an evidence-based approach, the project aimed at highlighting teachers' needs, beliefs and practices of teaching and learning, which may constitute a privileged context for the development of innovative teaching activities within the institution.

The final questionnaire was developed according to the Framework of Teaching of [7] and was composed by three sections. The first section focuses on *practices* developed by the Padua professors in their teaching activities. The teacher is thought as a facilitator of the learning processes and for this reason the section asks for each DA (at most three) about the application (or not) of some specific practices in his/her activities. Eight items are collected. Six indicators are then constructed and five of them are obtained considering separately as dummy variables the first five items: *implementation of practices for actively getting involved students; proposal of external contributions (i.e. stakeholder); monitoring students learning during the course by means of specific tests/other ways; assessment of students learning using various types of examination; modification of teaching practices according to SET*. The sixth indicator is calculated summarising in a single dummy variable the last three items of the section (*reporting at least one activity involving technology practices*), since these three questions collect similar information on these practices. The second section deepens teachers' *beliefs* about teaching in higher education. By means of 20 questions, in a scale from 1 (fully disagree) to 7 (fully agree), some general dimensions are investigated: the Person as Teacher, Expert on Content Knowledge, Facilitator of Learning Processes and Scholar/Lifelong Learner. Considering also some questionnaire validation analyses (a factor analysis in particular), six factors are defined (they substantially replicate the aforementioned dimensions), calculated as the average values of the answers within each factor. These factors may be summarised as other subjective characteristics of the teachers: i) *passion for teaching*; ii) *passion for research*; iii) *feeling the need of support for improving teaching activities*; iv) *will to change teaching activities according to students needs*; v) *features of teaching and learning methods*; vi) *features of teaching and evaluation activities*. The third section focuses on teachers' *needs*, that are collected through some open-ended questions (however, they are not exploited in this analysis).

The PRODID questionnaire was addressed to all teaching staff of the University of Padua involved in any DA during the academic year 2012-2013; the response rate of this survey was slightly lower than 50%.

In this analysis we consider only students who attended at least 50% of lessons, involved in courses of the bachelor degree and enrolled in any undergraduate programmes, but Medicine. In the end, we excluded courses with a number of units smaller than five (in order to avoid comparisons based on too few ratings). According to these criteria, the linkage of the different sources led to a final dataset composed by 23605 complete records, based on students' evaluations.

### 3 The analysis

The analysis of the dataset described in the previous Section is based on the estimation of a multilevel random intercept model [4], where the level-1 dependent variable is the overall level of satisfaction (based on Item 14). Level-2 units are the DAs of each teacher. This choice follows from the fact that, within each course, the student is asked to evaluate the activities of each professor having a minimum number of hours taught in the course. The student degree is not a further level, but it is controlled by means of fixed effects. The total number of level-2 units is equal to 590, while 40 is the average number of observations per group.

In general, the rating of a student to a given item for a certain course may depend on course-related factors (class size and heterogeneity, course difficulty and so on), student-related factors (gender, age and so on) and teacher-related factors (age, gender, personal traits and so on) [6]. According to the aims of this work and the features of our dataset, the set of our explanatory variables may be divided in:

- Course characteristics: compulsory course, total number of hours, more than one teacher involved, location (in Padua or outside), shared course.
- Student - general characteristics: gender, age.
- Student - university career: year of enrolment, average (per year) number of passed exams, average grade of the exams in the referred academic year.
- Teacher - general characteristics: gender, age.
- Teacher - university career: academic position.
- Teacher - DA characteristics: proportion of the total number of hours within DA.
- Teacher - subjective characteristics: according to Section 2, the six indicators of teaching practices and the six factors of teacher beliefs.

This specification allows to particularly investigate the role of *objective* teacher characteristics and the one of *subjective* teacher characteristics.

#### 3.1 Main results

Results from the estimation of the random intercept model described in previous Section is reported in Table 1.

On the one hand, student characteristics are strongly related to the overall satisfaction rating of the DA, particularly those related to the academic experience of these students. The main features of the courses play a weak role instead.

On the other hand, there are some interesting results on the relationship between SETs and teacher characteristics. *Objective* teacher traits are weakly related with SET ratings: age is the only variable reporting a strong statistically significant estimate (the older, the better the teacher is evaluated, *ceteris paribus*). *Subjective* features of the teachers are also related to SET scores, but in some particular ways. Two indicators of *practices* and even four factors of *beliefs* are statistically significant. In particular, looking at these teacher beliefs, interesting relationships appear for those factors related to the sensitivity and the aptitude of teaching. For instance, according to the PRODID questionnaire the factor "Feeling the need of support to improve teaching activities" may highlight those teachers who feel some difficulties or inadequacies in their teaching activities/performances and for this reason they need help from experts. Students are able to perceive such difficulties and then reporting a lower evaluation of the course (other things being equal). On the contrary, students recognise those teachers with a high passion for teaching or the will to propose suitable and helpful instruments in their DAs: such traits may be able to enhance the transmission of knowledge from the teacher to the student.

It is worth noting the different relationships that come to light between SET evaluations and the *passion for teaching* and *passion for research* dimensions.

## 4 Conclusions

Exploiting the richness of information provided by an innovative survey on teaching experiences and beliefs of professors working at the University of Padua, the role of the teacher perceptions and needs on their activities is deeply investigated. Findings clearly show that *subjective* characteristics of the teachers play an important role in explaining SET ratings. However, this solution should be improved taking into account the fact that the sample of professors, who completed the PRODID questionnaire, is likely to be not randomly selected.

This work may be seen as a first step for enhancing the relationship between quality of a course (or university) and students' opinions. Indeed, teaching is a complex and multidimensional concept, so a future research strand could be the analysis of a multidimensional indicator of course quality, based on a battery of items.

## References

1. Emerson, J.D., Mosteller, F., Youtz, C.: Students can help improve college teaching: a review and an agenda for the statistics profession. In: Rao, C.R., Székely, G.J. (eds), *Statistics for the 21st century: methodologies for applications of the future*. Marcel Dekker, New York (2000)

**Table 1** Estimates of the random intercept model on the students' overall satisfaction

Group characteristics	Variable	Point estimate
Course	Compulsory course	-0.036
	Number of hours	0.432 *
	More than one teacher	-0.096
	Location of courses in Padua	-0.875 *
	Shared course	-0.126
Student - general	Female	-0.030
	Age	0.304 ***
Student - career	Second year of enrolment	-0.216 ***
	Third year of enrolment	-0.140 *
	Average number of passed exams (whole career)	0.088 **
	Average grade of passed exams (in 2012/13)	0.338 ***
Teacher - general	Female	-0.169 *
	Age	-0.185 ***
Teacher - career	Full professor	0.017
	Associate professor	0.077
Teacher - DA	Proportion of hours in DA	0.231
Teacher - subjective (practices)	Practices for actively getting involved students	-0.110
	Proposal of external contributions	0.192 **
	Monitoring students learning during the course	0.003
	Assessing students learning using different types of exam	-0.194 **
	Modification of teaching practices according to SET	-0.038
Teacher - subjective (beliefs)	Reporting at least 1 activity involving technology practices	0.053
	Passion for teaching	0.128 ***
	Passion for research	-0.049
	Need of support for improving teaching activities	-0.110 ***
	Will of changing teaching activities with students needs	0.075 *
	Features of teaching and learning methods	0.137 ***
	Features of teaching and evaluation activities	-0.007
	constant	6.152 ***
	ICC	21.2%

Note: \*\*\* = 1% of level; \*\* = 5% of level; \* = 10% of level

- Felisatti, E., Serbati, A.: The professional development of teachers: from teachers practices and beliefs to new strategies at the university of Padua. Proceedings of the ICED conference Educational development in a changing world, Stockholm, 16-18 June 2014 (2014)
- Rampichini, C., Grilli, L., Petrucci A.: Analysis of university course evaluations: from descriptive measures to multilevel models. *Statistical Methods & Applications* **13**, 357-373 (2004)
- Snijders, T.A.B., Bosker, R.J.: *Multilevel Analysis. An introduction to basic and advanced multilevel modelling*. Sage, London (2012)
- Spooren, P.: On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching. *Studies in Educational Evaluation* **36**, 121-131 (2010)
- Spooren, P., Brockx, B., Mortelmans, D.: On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research* **83**, 598-642 (2013)
- Tigelaar, D.E.H., Dolmans, D.H.J.M., Wolfhagen, I.H.A.P., Van Der Vleuten, C.P.M.: The development and validation of a framework for teaching competencies in higher education. *Higher Education* **48**, 253-268 (2004)
- Zabaleta, F.: The use and misuse of student evaluations of teaching. *Teaching in Higher Education* **12**, 55-76 (2007)

# Bayesian Quantile Regression using the Skew Exponential Power Distribution

Bernardi Mauro and Marco Bottone and Petrella Lea

**Abstract** Traditional Bayesian quantile regression relies on the Asymmetric Laplace distribution (ALD) due primarily to its satisfactory empirical and theoretical performances. However, the ALD displays medium tails and is not suitable for data characterized by strong deviations from the Gaussian hypothesis. In this paper, we propose an extension of the ALD Bayesian quantile regression framework to account for fat tails using the Skew Exponential Power (SEP) distribution. Linear and Additive Models (AM) with penalized spline are used to show the flexibility of the SEP in the Bayesian quantile regression context. Lasso priors are used to account for the problem of shrinking parameters when the parameters space becomes wide. We propose a new adaptive Metropolis–Hastings algorithm in the linear model, and an adaptive Metropolis within Gibbs one in the AM framework. Empirical evidence of the statistical properties of the model is provided through several examples based on both simulated and real datasets.

**Abstract** L'analisi Bayesiana per la regressione quantile si basa sull'uso della distribuzione Laplace asimmetrica come strumento inferenziale. Tale distribuzione pur fornendo performances soddisfacenti non ha un comportamento soddisfacente nel caso in cui il fenomeno sotto indagine presenti code con andamento diverso da quello gaussiano. In questo paper, per tener conto di code pesanti del fenomeno, proponiamo l'uso della distribuzione Skew Exponential Power (SEP) in un contesto di regressione quantile. Considereremo modelli lineari e modelli additivi attraverso l'uso di spline per effettuare l'inferenza bayesiana. Una distribuzione lasso a priori sui parametri del modello viene proposta per tener conto del problema della con-

---

Bernardi Mauro

Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, Padua e-mail: mauro.bernardi@unipd.it

Bottone Marco

Bank of Italy DG Economics, Statistics and Research, Italy e-mail: marco.bottone@bancaditalia.it

Petrella Lea

MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano 9, Rome e-mail: lea.petrella@uniroma1.it

trazione del numero degli stessi laddove lo spazio parametrico diventi elevato. Per effettuare l'inferenza bayesiana viene proposto un nuovo algoritmo adattivo di tipo Monte Carlo Markov Chain e analisi di simulazioni verranno proposte per validare il modello considerato.

**Key words:** Bayesian quantile regression; Skew Exponential Power; Additive Model.

## 1 Introduction

Quantile regression has become a very popular approach to provide a more complete description of the distribution of a response variable conditionally on a set of regressors. Since the seminal work of [1], several papers have been proposed in literature considering the quantile regression analysis both from a frequentist and a Bayesian points of view. Specifically, let  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_T)$  be a random sample of  $T$  observations, and  $\mathbf{X}_t = (1, X_{t,1}, \dots, X_{t,p-1})'$ , with  $t = 1, 2, \dots, T$  equal to the associated set of  $p$  covariates. Consider the following linear quantile regression model

$$Y_t = \mathbf{X}_t' \beta_\tau + \varepsilon_t, \quad t = 1, 2, \dots, T,$$

where  $\beta_\tau = (\beta_{\tau,0}, \beta_{\tau,1}, \dots, \beta_{\tau,p-1})'$  is the vector of  $p$  unknown regression parameters, varying with the quantile  $\tau$  level. As usual,  $\varepsilon_t$  represents the error term that, in the specific case of quantile regression, has the  $\tau$  quantile equal to zero and constant variance. This assumption allows us to interpret the regression line as the  $\tau$  conditional quantile of  $Y$  given the set of explanatory variables  $\mathbf{X} = \mathbf{x}$ , i.e.  $Q_\tau(Y | \mathbf{X} = \mathbf{x}) = \mathbf{x}' \beta_\tau$ . In what follows we omit the subscript  $\tau$  for simplicity. The estimation procedure of the  $\tau$ -th regression quantile in the frequentist approach is based on the minimization of the following loss

$$\min_{\beta} \sum_t \rho_\tau(y_t - \mathbf{x}_t' \beta)$$

with  $\rho_\tau(u) = u(\tau - I(u < 0))$ . From a Bayesian point of view [8] introduces the ALD as likelihood function to perform the inference. For a wide and recent Bayesian literature on quantile regression and ALD see for example [7], and [3]. Although the ALD is widely used in the Bayesian framework it displays medium tails which may give misleading informations for extreme quantile in particular when the data are characterized by the presence of outlier and heavy tails. The absence for the ALD of a parameter governing the tail fatness may influence the final inference. To overcome this drawback we propose an extension of the Bayesian quantile regression using the Skew Exponential Power (SEP) distribution proposed by [2]. The SEP distribution, like the ALD, has the property of having the  $\tau$ -level quantile as the natural location parameter but it also has an additional parameter governing the decay of the tails. Using the proposed distribution in quantile regression we are able to robustify the inference in particular when outliers or extreme values are observed.

When dealing with model building the choice of appropriate predictors and consequently the variable selection issue plays an important role. In this paper, we approach this problem, by considering the Bayesian version of Lasso penalization methodology introduced by [6] both for the simple linear regression quantile and for the non linear additive models (AM) with Penalized Spline (P-Spline) functions. To implement the Bayesian inference we propose a new adaptive Metropolis Hastings algorithm in the linear model, and an Adaptive Metropolis within Gibbs one in the AM framework for an efficient estimate of the penalization parameter and the P-Spline coefficients. We show the robust performance of the model with simulation studies.

## 2 Model and Inference

In their paper [2], the authors propose a parametrization of the SEP, that allows to consider the location parameter as the  $\tau$ -level quantile. With their parametrization the SEP density function can be written as:

$$f(y, \mu, \sigma, \tau, \alpha) = \begin{cases} \frac{1}{\sigma} \kappa(\alpha) \exp\left\{-\frac{1}{\alpha} \left(\frac{\mu-y}{2\tau\sigma}\right)^\alpha\right\}, & \text{if } y \leq \mu \\ \frac{1}{\sigma} \kappa(\alpha) \exp\left\{-\frac{1}{\alpha} \left(\frac{y-\mu}{2(1-\tau)\sigma}\right)^\alpha\right\}, & \text{if } y > \mu, \end{cases} \quad (1)$$

where  $y \in \mathbb{R}$ ,  $\mu \in \mathfrak{R}$  is the location parameter,  $\sigma \in \mathfrak{R}^+$  and  $\alpha \in (0, \infty)$  are the scale and shape parameters, respectively,  $\tau \in (0, 1)$  is the skewness parameter while  $\kappa = \left[2\alpha^{\frac{1}{\alpha}} \Gamma\left(1 + \frac{1}{\alpha}\right)\right]^{-1}$  and  $\Gamma(\cdot)$  is the complete gamma function. It can be showed that  $\mu$  is the  $\tau$  quantile and that the ALD is a particular case with  $\alpha = 1$ . Several model specifications can be obtained using the SEP likelihood by specifying a given function for the location parameter.

In this paper we consider both the linear quantile regression framework

$$\mu = \mu(\mathbf{x}_t) = \mathbf{x}_t^T \beta \quad (2)$$

where  $\mathbf{x}_t$  is a set of exogenous covariates than the Additive Models within a robust semi-parametric regression framework:

$$\mu = \mu(\mathbf{x}_t, \mathbf{z}_t) = \mathbf{x}_t^T \beta + \sum_{j=1}^J f_j(z_{t,j})$$

where  $\mathbf{x}_t^T \beta$  is the parametric component while  $\mathbf{z}_t = (z_{t,1}, \dots, z_{t,J})^T$  is an additional set of covariates and each  $f_j(z_{t,j})$  is a nonparametric continuous smooth function. To implement the Bayesian analysis we assume that  $f_j(z_{t,j})$ , can be approximated using a polynomial spline of order  $d$ , with  $k + 1$  equally spaced knots. Let's consider more specifically the linear case where the likelihood function can be easily computed starting from (1) by using  $\mu$  as in (2).

The Bayesian inferential procedure requires the specification of the prior distribution for the unknown vector of parameters  $\Xi = (\beta, \gamma, \sigma, \alpha)$ . Here in order to account for sparsity within the quantile regression model, we generalize the prior proposed in Park and Casella for the  $\beta$  parameter, assuming the hierarchical structure given below. The prior distribution is given by:

$$\pi(\Xi) = \pi(\beta | \gamma) \pi(\gamma) \pi(\sigma) \pi(\alpha),$$

with

$$\begin{aligned} \pi(\beta | \gamma) &\propto \prod_{j=1}^p L_1(\beta_j | 0, \gamma_j) \\ \pi(\gamma) &\propto \prod_{j=1}^p \mathcal{G}(\gamma_j | \psi, \varpi) \\ \pi(\sigma) &\propto \mathcal{IG}(a, b) \\ \pi(\alpha) &\propto \mathcal{B}(c, d) \mathbb{1}_{(0,2)}(\alpha), \end{aligned}$$

where  $\beta \in \mathbb{R}^p$ . Here  $(\psi, \varpi, a, b, c, d)$  are given positive hyperparameters and  $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_p)$  are the parameters of the univariate Laplace distribution:

$$L_1(\beta_j | 0, \gamma_j) = \frac{\gamma_j}{2} \exp\{-\gamma_j |\beta_j|\} \mathbb{1}_{(-\infty, +\infty)}(\beta_j).$$

with zero location and  $\gamma_j$  scale parameter. Here  $\mathcal{G}$ ,  $\mathcal{IG}$  and  $\mathcal{B}$  denote the Gamma, Inverse Gamma and Beta distributions, respectively. Given its characteristics, the Laplace distribution is the Bayesian counterpart of the Lasso penalization methodology introduced by [6] to achieve sparsity within the classical regression framework. By shrinking each regression parameter in a different way, we overcome problems that may arise in the presence of regressors with different scales of measurement. The Bayesian inference is performed by building an Adaptive Independent Metropolis Hastings MCMC algorithm using the location–scale mixture representation of the the Laplace distribution, see for example [9].

### 3 Simulation Studies

We have performed several simulation studies to highlight the improvements of our model specification with respect to the well known ALD model tool. In particular the first simulation experiment is built in order to show the robustness properties of the proposed methodology for quantile estimation when the joint distribution of the couple  $(Y_t, \mathbf{X}_t)$ , for  $t = 1, 2, \dots, T$ , is contaminated by the presence of outliers. The second study shows the effectiveness of the shrinkage effect, obtained by imposing the Lasso–type prior, used when the multiple quantile linear model is of key concern. The last experiment aims at highlighting the ability of the model to adapt



to non-linear shapes, when data come from heterogeneous fat-tailed distributions. All of the simulation studies showed the improvement in performances of the model proposed in this paper with respect to the ALD quantile regression commonly used in literature. Here we present only the second experimental study. In particular we carry out a Monte Carlo simulation study specifically tailored to evaluate the performance of the model when the Lasso prior is considered for the regression parameters. The simulations are similar to the one proposed in [4] and [5]. In particular, we simulate  $T = 200$  observations from the linear model  $Y_i = \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i$ , where the true values for the regressors are set as follows:

$$\text{Simulation 1.} \quad \boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)' ,$$

$$\text{Simulation 2.} \quad \boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)' ,$$

$$\text{Simulation 3.} \quad \boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)' ,$$

The first simulation corresponds to a sparse regression case, the second to a dense case, and the third to a very sparse case. The covariates are independently generated from a  $\mathcal{N}(0, \Sigma)$  with  $\sigma_{i,j} = 0.5^{|i-j|}$ . Two different distributions for the error terms generating process are considered for each simulation study. The first is a Gaussian distribution  $\mathcal{N}(\mu, \sigma^2)$ , with  $\mu$  set so that the  $\tau$ -th quantile is 0, while  $\sigma^2$  is set as 9, as in [4]. The second distribution is a Generalized Student't  $\mathcal{G}\mathcal{S}(\mu, \sigma^2, \nu)$  with two degrees of freedom, i.e.  $\nu = 2$ ,  $\sigma^2 = 9$  and  $\mu$  set so that the  $\tau$ -th quantile is 0. For three different quantile levels,  $\tau = (0.10, 0.5, 0.9)$  we run 50 simulations for each vector of parameters ( $\boldsymbol{\beta}$ ) and each distribution of the error term. Table 1 reports the median of mean absolute deviation (MMAD), i.e.  $\text{median}\left(\frac{1}{200} \sum_{i=1}^{200} |x_i' \hat{\boldsymbol{\beta}} - x_i' \boldsymbol{\beta}|\right)$ , and the median of the parameters  $\hat{\boldsymbol{\beta}}$ , over 50 estimates. Results for the first simulation are reported, since results from the other two simulations are qualitatively similar. The proposed Bayesian quantile regression method based on the SEP likelihood performs better in terms of MMAD for both distributions of the error term. This is evidence that the presence of the shape parameter  $\alpha$  in the likelihood better capture the behavior of the data. The estimated shape parameter is indeed greater and lower than one in the Gaussian and Generalized Student't cases, respectively; this provides a more reliable estimation of the vector  $\boldsymbol{\beta}$ , regardless of the tail weight of the error term distribution. These results are reinforced in the second and third simulation (not reported here) in which we exaggerate the density and the sparsity of the predictors structure. Furthermore, the proposed robust method reduces the bias of estimated  $\boldsymbol{\beta}$  for all quantile confidence levels. Regarding the shrinkage ability of the proposed estimator, when the true parameters are zero, the SEP distribution performs better than the ALD in identifying the parameters .

Error distribution	Par.	ALD			SEP		
		$\tau = 0.10$	$\tau = 0.50$	$\tau = 0.90$	$\tau = 0.10$	$\tau = 0.50$	$\tau = 0.90$
Gaussian	MMAD	1.0131	1.1008	1.0579	0.9096	1.0955	0.9708
	$\beta_1$	3.1323	3.2209	3.2145	3.0744	3.0036	3.2127
	$\beta_2$	1.6408	1.4786	1.6165	1.7656	1.4833	1.6800
	$\beta_3$	0.0444	0.0294	0.0267	0.0428	0.0228	0.0186
	$\beta_4$	0.0453	0.0243	0.0235	0.0248	0.0191	0.0156
	$\beta_5$	1.2731	1.2379	1.3471	1.3969	1.8405	1.4702
	$\beta_6$	0.0185	0.0161	0.0205	0.0124	0.0127	0.0128
	$\beta_7$	0.0112	0.0106	0.0120	0.0067	0.0063	0.0095
	$\beta_8$	0.0073	0.0078	0.0064	0.0038	0.0047	0.0051
Generalized Student t	MMAD	0.5163	0.1807	0.4685	0.4777	0.1789	0.4275
	$\beta_1$	3.0630	2.9884	2.9874	3.0826	2.9877	2.9934
	$\beta_2$	1.0484	1.3700	1.1366	1.0952	1.3951	1.2110
	$\beta_3$	0.0304	0.0144	0.0325	0.0252	0.0135	0.0412
	$\beta_4$	0.0258	0.0181	0.0162	0.0263	0.0163	0.0138
	$\beta_5$	1.7012	1.9036	1.7701	1.7558	1.9111	1.8052
	$\beta_6$	0.0128	0.0085	0.0137	0.0074	0.0072	0.0136
	$\beta_7$	0.0055	0.0057	0.0101	0.0052	0.0066	0.0082
	$\beta_8$	0.0067	0.0009	0.0002	0.0051	0.0011	-0.0021

**Table 1** Multiple regression simulated data example 1. MMADs and estimated parameters for Simulation 1 under the SEP and ALD assumption for the quantile error term.

## 4 Conclusion

We show how to implement the Bayesian quantile regression when the SEP distribution is considered. Linear and Additive Models (AM) with penalized spline are used with Lasso priors to account for the problem of shrinking parameters. Empirical analysis highlights how the SEP quantile regression better capture the behaviour of the data when outliers or heavy tails are concerned.

## References

1. Koenker, R., Basset, G.: Regression quantiles. *Econometrica* **46** (1978) 33-50.
2. Zhu, D., Zinde-Walsh, V.: Properties and estimation of asymmetric exponential power distribution. *Journal of Econometrics* **148** (2009) 86-99.
3. Bernardi, M., Gayraud, G., Petrella, L.: Bayesian tail risk interdependence using quantile regression. *Bayesian Analysis* **10** (2015) 553-603.
4. Li, Q., Xi, R., Lin, R.: Bayesian regularized quantile regression. *Bayesian Analysis* (2015) **5** 533-556.
5. Alhamzawi, R., and Benoit, D.F.: Bayesian adaptive lasso quantile regression *Statistical Modelling*, **12** (2012) 279-297.
6. Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **58**, (1996) 267-288.
7. Lum, K. and Gelfand A.: Spatial quantile multiple regression using the asymmetric laplace process. *Bayesian Analysis*, **7** (2012) 235-258.
8. Yu, K. and Moyeed, R. A. (2001). Bayesian Quantile Regression. *Statist. Probab. Lett.*, **54**, 437-447.
9. Kozumi, H. and Kobayashi, G. (2001). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, **81**, 1565-1575.

# Bayesian Factor–Augmented Dynamic Quantile Vector Autoregression

## *Stima bayesiana di modelli quantilici dinamici autoregressivi fattoriali*

Bernardi Mauro

**Abstract** This paper introduces a novel Bayesian model to estimate multi–quantiles in a dynamic framework. The main innovation relies on the assumption that the  $\tau$ –th level quantile of a vector of response variables depends on macroeconomic variables as well as on latent factors having their own stochastic dynamics. The proposed framework can be conveniently thought as a factor–augmented vector autoregressive extension of traditional univariate quantile models. We develop sparse Bayesian methods that rely on state space representation and data augmentation approaches that efficiently deal with the estimation of model parameters and the signal extraction from latent variables.

**Abstract** *Questo lavoro introduce un nuovo metodo per la stima di quantili dinamici multipli. L'innovazione consiste nell'assumere che il quantile di livello  $\tau$  di un vettore di variabili risposta dipenda da fattori macroeconomici e da fattori latenti aventi una loro dinamica. Il modello proposto può essere convenientemente pensato come l'estensione dei modelli quantilici univariati tradizionali per un modello fattoriale autoregressivo vettoriale aumentato con l'introduzione di fattori latenti. La stima dei parametri e l'estrazione del segnale latenti sono effettuati proponendo un algoritmo Gibbs sampler con una distribuzione a priori che introduce stima sparsa dei parametri.*

**Key words:** Quantile vector autoregression, Bayesian inference, Asymmetric Laplace, factor models, sparse estimation.

## 1 Introduction

Quantile regression models have been becoming increasingly popular because of their attractive characteristics of modelling the quantile of a response variable as

---

Bernardi Mauro

Department of Statistical Sciences, University of Padova e-mail: mauro.bernardi@unipd.it

a function of some covariates. Indeed, quantile models provide a more complete picture of the conditional distribution of the response variable than the traditional regression approach without relying on strong assumptions about the form of the error term. However, despite their obvious powerfulness, quantile methods have been mostly confined on modelling univariate response variables, see, e.g., Koenker (2005). In this paper we extend univariate quantile regression models to deal with multivariate response variables. Specifically, we model the marginal quantiles of a multivariate random variable as a function of macroeconomic variables and of latent factors having their own stochastic dynamics. Dynamic latent quantiles have been introduced by De Rossi and Harvey (2009) and extended to the bivariate Bayesian framework by Bernardi et al. (2015). We develop Bayesian methods that rely on state space representation and data augmentation approaches that efficiently deal with sparse estimation of model parameters and the signal extraction from latent variables. A multivariate Asymmetric Laplace distribution is imposed to the error term of the measurement equation in order to model  $\tau(0, 1)$ -th level quantile of each marginal random variable. When dealing with multivariate latent models the curse of dimensionality prevents any parametric inferential procedure. To overcome this problem we rely on sparse methods and in particular on the spike-and-slab (Mitchell and Beauchamp 1988 and George and McCulloch 1993) Least Absolute Shrinkage and Selection Operator (LASSO) prior of Tibshirani (1996).

The remainder of the paper is organised as follows. Section 2 introduces the multivariate Asymmetric Laplace distribution and its main properties. Section 3 introduces the dynamic factor-augmented quantile model and Section 4 deals with Bayesian inference and signal extraction.

## 2 Multivariate Asymmetric Laplace distribution and quantiles

In this Section we first introduce the multivariate Asymmetric Laplace (AL) distribution and its stochastic representation which will be useful to develop the data-augmentation Gibbs sampler scheme. Then we prove that the multivariate AL distribution characterises the univariate marginal quantiles.

**Definition 1.** Consider a  $p$ -dimensional random vector  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_p) \in \mathbb{R}^p$  from a multivariate Asymmetric Laplace (AL) distribution (see Kotz et al. 2001). The density of  $\mathbf{Y} \sim \text{AL}_p(\vartheta, \xi, \Sigma)$  is given by

$$f_{\mathbf{Y}}(\mathbf{y} \mid \vartheta, \xi, \Sigma) = \frac{2 \exp\{\xi' \Psi^{-1} \mathbf{D}^{-1}(\mathbf{y} - \vartheta)\}}{(2\pi)^{p/2} |\Sigma|^{1/2}} \left[ \frac{\delta(\mathbf{y}, \vartheta, \Sigma)}{2 + \xi' \Psi^{-1} \xi} \right]^{v/2} \times K_v \left( \sqrt{(2 + \xi' \Psi^{-1} \xi) \delta(\mathbf{y}, \vartheta, \Sigma)} \right), \quad (1)$$

where  $\vartheta \in \mathbb{R}^p$  and  $\xi \in \mathbb{R}^p$  are a  $p$ -dimensional vector of location and shape parameters, respectively. Moreover,  $\mathbf{D} = \text{diag}\{\sigma_1, \sigma_2, \dots, \sigma_p\}$  with  $\sigma_j > 0$ , for

$j = 1, 2, \dots, p$  and  $\Psi$  is a correlation matrix, such that  $\Sigma = \mathbf{D}\Psi\mathbf{D}$ ,  $\nu = \frac{2-p}{2}$ ,  $\delta(\mathbf{y}, \vartheta, \Sigma) = (\mathbf{y} - \vartheta)' \Sigma^{-1} (\mathbf{y} - \vartheta)$  is the squared Mahalanobis distance between  $\mathbf{y}$  and  $\vartheta$  and  $K_\nu(\cdot)$  is the modified Bessel function of the third type with index parameter  $\nu$ .

The multivariate AL distribution can be represented as a Gaussian location–scale mixture with the Exponential distribution acting as mixing random variable.

**Proposition 1.** *Let  $\mathbf{Y} \sim \text{AL}_p(\vartheta, \xi, \Sigma)$  as introduced in Definition 1, then*

$$\mathbf{Y} = \vartheta + \mathbf{D}\xi W + \Sigma^{\frac{1}{2}} \sqrt{W} \mathbf{Z}, \quad (2)$$

where  $\mathbf{Z} \sim \text{N}_p(\mathbf{0}, \mathbf{I}_p)$ ,  $W \sim \text{Exp}(1)$  with  $Z_j \perp\!\!\!\perp W$  for  $j = 1, 2, \dots, p$ , see, e.g., Kotz and Nadarajah (2004). It follows from equation (2) that  $\mathbf{Y} | W = w \sim \text{N}_p(\vartheta + \mathbf{D}\xi w, \Sigma w)$  and that the unconditional distribution of  $\mathbf{Y}$  is given by equation (1), see Kotz et al. (2001).

The following remark instead characterises the behaviour of the marginal distributions of each component  $Y_j$ , for  $j = 1, 2, \dots, p$ .

*Remark 1.* Let  $\mathbf{Y} \sim \text{AL}_p(\vartheta, \xi, \Sigma)$  as introduced in Definition 1, then

$$Y_j \sim \text{AL}_1(\vartheta_j, \xi_j, \sigma_j^2), \quad (3)$$

where  $\vartheta_j \in \mathbb{R}$  is the  $j$ -th element of  $\vartheta$ ,  $\xi_j \in \mathbb{R}$  is the  $j$ -th element of  $\xi$ , and  $\sigma_j^2 \in \mathbb{R}^+$  is the  $j$ -th element of the diagonal of the scale matrix  $\mathbf{D}$ , see Kotz et al. (2001).

The next proposition characterises the AL distribution as the natural candidate for modelling the innovation term in multivariate quantile models.

**Proposition 2.** *Let  $\mathbf{Z} = (Z_1, Z_2, \dots, Z_p) \in \mathbb{R}^p$  with  $\mathbf{Z} \sim \text{D}$ , where  $\text{D}$  is an unknown probability density function, if we assume the AL as misspecified density for  $\mathbf{Z}$ , i.e.,  $\mathbf{Z} \sim \text{AL}_p(\vartheta, \xi, \Sigma)$ , then*

$$\mathbb{P}(Z_j < \vartheta_j) = \tau, \quad (4)$$

if and only if

$$\xi_j = \frac{1 - 2\tau}{\tau(1 - \tau)} \quad (5)$$

$$\sigma_j^2 = \frac{2\delta_j}{\tau(1 - \tau)}, \quad (6)$$

for  $j = 1, 2, \dots, p$ , where  $\sigma_j^2$  is the  $j$ -th diagonal element of the matrix  $\mathbf{D}$ ,  $\delta_j \in \mathbb{R}^+$  and  $\tau \in (0, 1)$  is the quantile confidence level.

*Proof.* Following Kotz et al. (2001) the marginal distribution of  $Z_j$ , for  $j = 1, 2, \dots, p$  is Asymmetric Laplace, i.e.,  $Z_j \sim \text{AL}(\vartheta_j, \xi_j, \sigma_j^2)$ , and imposing the conditions (5)–(6) the result follows immediately, see, e.g., Yu and Moyeed (2001).

### 3 Dynamic latent factor–augmented quantile model

In this section, we introduce the dynamic latent factor–augmented quantile model. Let  $\mathbf{y}_t = (y_{1,t}, y_{2,t}, \dots, y_{d,t})' \in \mathbb{R}^d$  and  $\mathbf{x}_t = (x_{1,t}, x_{2,t}, \dots, x_{p,t})' \in \mathbb{R}^p$  be random vectors, we assume that  $(\mathbf{y}_t', \mathbf{x}_t')$  is a linear function of latent factors

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} \lambda & \beta \\ \mathbf{0}_{(p \times s)} & \mathbf{I}_p \end{bmatrix} \begin{bmatrix} \chi_t \\ \mathbf{x}_t \end{bmatrix} + \begin{bmatrix} \mathbf{I}_d \\ \mathbf{0}_p \end{bmatrix} \varepsilon_t, \quad t = 1, 2, \dots, T, \quad (7)$$

where  $\lambda$  is a  $(d \times s)$  matrix of loadings of the stochastic latent factors  $\chi_t = (\chi_{1,t}, \chi_{2,t}, \dots, \chi_{s,t})$ ,  $\beta$  is the  $(d \times p)$  matrix of loadings of the observed factors  $\mathbf{x}_t$ ,  $\mathbf{I}_d$  and  $\mathbf{I}_p$  denotes an identity matrices of dimension  $d$  and  $p$ , respectively, and  $\mathbf{0}_{(p \times s)}$  and  $\mathbf{0}_p$  denote zero matrices of dimension  $p \times s$  and  $p \times p$ , respectively. The stochastic term  $\varepsilon_t = (\varepsilon_{1,t}, \varepsilon_{2,t}, \dots, \varepsilon_{d,t})$  follows a multivariate Asymmetric Laplace distribution defined in equation (1), i.e.,  $\varepsilon_t \sim \text{AL}_d(\vartheta, \xi, \Sigma)$ . The state space formulation is completed by specifying a dynamic evolution for the latent and observed factors  $\chi_t$  and  $\mathbf{x}_t$

$$\begin{bmatrix} \chi_{t+1} \\ \mathbf{x}_{t+1} \end{bmatrix} = \mu + \begin{bmatrix} \Phi_1 & \mathbf{0} \\ \mathbf{0} & \Phi_2 \end{bmatrix} \begin{bmatrix} \chi_t \\ \mathbf{x}_t \end{bmatrix} + \eta_t, \quad t = 1, 2, 3, \dots, T-1, \quad (8)$$

where  $\eta_t \sim \text{N}_{d+p}(\mathbf{0}_{s+p}, \Omega_t)$  with

$$\Omega_t = \begin{bmatrix} \Omega_t^{1,1} & \Omega_t^{1,2} \\ \Omega_t^{2,1} & \Omega_t^{2,2} \end{bmatrix}, \quad (9)$$

positive definite matrix of order  $d+p$  with  $\Omega_t^{1,1} \in \mathcal{M}^{(s,s)}$ ,  $\Omega_t^{2,2} \in \mathcal{M}^{(p,p)}$ ,  $\Omega_t^{1,1} \in \mathcal{M}^{(s,p)}$ ,  $\Omega_t^{2,1} = \Omega_t^{1,2'}$ , and  $\Phi_1$  and  $\Phi_2$  are  $(s \times s)$  and  $(p \times p)$  transition matrices and  $\mu = (\mu_1', \mu_2')' \in \mathbb{R}^{d+p}$  with  $\mu_1 \in \mathbb{R}^d$  and  $\mu_2 \in \mathbb{R}^p$ . Here  $\mathcal{M}^{(p,q)}$  denotes the space of matrices of dimension  $p \times q$ . The transition equation (8) specifies first order vector autoregressive processes (VAR) for both the latent and observed factors. Indeed, alternative more flexible autoregressive specifications can be imposed by exploiting the companion form representation of VAR models, see, e.g., Harvey (1989). Furthermore, without loss of generality, we assume the VAR dynamic for  $\mathbf{x}_t$  to be stationary, i.e., all the eigenvalues of the matrices  $\Phi_j$ , for  $j = 2$  are outside the unit circle.

Concerning the specification of the initial states, we assume  $\chi_1 \sim \text{N}(\hat{\chi}_{1|0}, \mathbf{P}_{1|0})$ , where the variance–covariance matrix  $\mathbf{P}_{1|0}$  can be diffuse to handle the presence of non stationary elements of the latent states  $\chi_t$ . Moreover, given the imposed stationary dynamic evolution of the observed states  $\mathbf{x}_t$ , we assume  $\mathbf{x}_1 \sim \text{N}(\hat{\mathbf{x}}_{1|0}, \mathbf{V}_{1|0})$ , where  $\hat{\mathbf{x}}_{1|0} = (\mathbf{I}_p - \Phi_2)^{-1} \mu_2$  and  $\text{vec}(\mathbf{V}_{1|0}) = (\mathbf{I}_{p^2} - \Phi_2 \otimes \Phi_2)^{-1} \text{vec}(\Omega_{2,2})$  and  $\Omega_{2,2}$  is the square matrix of dimension  $p \times p$  of the long–run matrix  $\Omega_{2,2} =$

$\lim_{t \rightarrow \infty} \Omega_t^{2,2}$ . We name the model defined in equations (7)–(8), the Factor–Augmented Quantile Vector Autoregression (FAQVAR) model.

#### 4 Bayesian inference for the FAQVAR

The theory underlying the signal extraction and the Bayesian posterior computation and simulation of the quantiles can be stated for a generic model in state space form (see, e.g., Harvey 1989 and Durbin and Koopman 2012), where, without loss of generality, we assume the scale parameter of the AL distribution depends on time. Signal extraction and posterior mode computation of the latent generalised quantile are based on the Kalman filter (Kalman and Bucy 1961) and the associated smoother algorithm, see, e.g., De Jong and Shephard (1995) and Durbin and Koopman (2002). Let

$$\mathbf{y}_t^\dagger = \begin{bmatrix} \mathbf{y}_t \\ \mathbf{x}_t \end{bmatrix}, \quad \boldsymbol{\chi}_t^\dagger = \begin{bmatrix} \boldsymbol{\chi}_t \\ \mathbf{x}_t \end{bmatrix}, \quad (10)$$

then

$$\mathbf{y}_t^\dagger = \mathbf{Z}\boldsymbol{\chi}_t^\dagger + \mathbf{H}\boldsymbol{\varepsilon}_t \quad (11)$$

$$\boldsymbol{\chi}_{t+1}^\dagger = \boldsymbol{\mu} + \mathbf{T}\boldsymbol{\chi}_t^\dagger + \boldsymbol{\eta}_t, \quad t = 1, 2, \dots, T-1 \quad (12)$$

$$\boldsymbol{\chi}_1^\dagger \sim \mathcal{N}(\hat{\boldsymbol{\chi}}_{1|0}^\dagger, \mathbf{P}_{1|0}^\dagger), \quad (13)$$

where the selection matrix  $\mathbf{H}$ , and the measurement and transition matrix ( $\mathbf{Z}, \mathbf{T}$ ) are defined as

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_d \\ \mathbf{0}_p \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} \lambda & \beta \\ \mathbf{0}_{(p \times s)} & \mathbf{I}_p \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \mathbf{0}_{(s \times p)} \\ \mathbf{0}_{(p \times s)} & \boldsymbol{\Phi}_2 \end{bmatrix}, \quad (14)$$

and  $\boldsymbol{\chi}_{1|0}^\dagger = (\hat{\boldsymbol{\chi}}'_{1|0}, \hat{\mathbf{x}}'_{1|0})'$ ,  $\mathbf{P}_{1|0}^\dagger = \begin{bmatrix} \mathbf{P}_{1|0} & \mathbf{0}_{(s \times p)} \\ \mathbf{0}_{(p \times s)} & \mathbf{V}_{1|0} \end{bmatrix}$ , where  $\lambda, \beta, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2$  and  $\hat{\boldsymbol{\chi}}'_{1|0}, \hat{\mathbf{x}}'_{1|0}, \mathbf{P}_{1|0}$  and  $\mathbf{V}_{1|0}$  have been defined in the previous section.

The linear state space model introduced in equations (11)–(12) for modelling time–varying conditional quantiles is non–Gaussian because of the assumption made on the measurement innovation terms. In those circumstances, optimal filtering techniques used to analytically marginalise out the latent states based on the Kalman filter recursions can not be applied (see Durbin and Koopman 2012). However, exploiting the stochastic representation of the AL distribution in terms of location–scale continuous mixture of Gaussian in Proposition 1, the non–Gaussian state space model defined in equations (11)–(12), admits as conditionally Gaussian and linear state space representation. More specifically, equations (11)–(12) become:

$$\mathbf{y}_t^\dagger = \mathbf{a}\boldsymbol{\omega}_t + \mathbf{Z}\boldsymbol{\chi}_t^\dagger + \mathbf{H}\boldsymbol{\varepsilon}_t^\dagger \quad (15)$$

$$\boldsymbol{\chi}_{t+1}^\dagger = \boldsymbol{\mu} + \mathbf{T}\boldsymbol{\chi}_t^\dagger + \boldsymbol{\eta}_t, \quad t = 2, 3, \dots, T \quad (16)$$

$$\boldsymbol{\chi}_1^\dagger \sim \mathbf{N}\left(\boldsymbol{\chi}_{1|0}^\dagger, \mathbf{P}_{1|0}^\dagger\right), \quad (17)$$

where  $\mathbf{a} = \begin{bmatrix} \mathbf{D}\boldsymbol{\xi} \\ \mathbf{0}_{(p \times 1)} \end{bmatrix}$ , with  $\mathbf{D} = \{\sigma_{1,1}, \sigma_{2,2}, \dots, \sigma_{d,d}\}$ ,  $\boldsymbol{\varepsilon}_t^\dagger \sim \mathbf{N}_d(\mathbf{0}_d, \boldsymbol{\omega}_t \boldsymbol{\Sigma})$ ,  $t = 1, 2, \dots, T$ , are i.i.d. are Gaussian innovations,  $\boldsymbol{\omega}_t \sim \text{Exp}(1)$  independent of  $\boldsymbol{\varepsilon}_t$ , for  $t = 1, 2, \dots, T$  and  $\boldsymbol{\Sigma} = \tilde{\mathbf{D}}\boldsymbol{\Psi}\tilde{\mathbf{D}}$ , with  $\tilde{\mathbf{D}} = \sqrt{\delta}\mathbf{D}$  and  $\delta = \frac{2}{\tau(1-\tau)}$ . We assume the following prior distributions for the parameters  $\boldsymbol{\Xi} = (\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma_j, j = 1, 2, \dots, p, \boldsymbol{\Psi}, \boldsymbol{\mu}, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \boldsymbol{\Omega}_t)$

$$\boldsymbol{\Lambda} \sim \mathbf{N}_{(d \times s)}(\boldsymbol{\mu}_\lambda^0, \boldsymbol{\Sigma}_\lambda^0 \otimes \boldsymbol{\Sigma}_\lambda^1) \quad (18)$$

$$\boldsymbol{\beta} \sim \mathbf{N}_{(d \times s)}(\boldsymbol{\mu}_\beta^0, \boldsymbol{\Sigma}_\beta^0 \otimes \boldsymbol{\Sigma}_\beta^1) \quad (19)$$

$$\sigma_{j,j}^{-1} \sim \prod_{j=1}^d \text{IG}(a_0, b_0), \quad j = 1, 2, \dots, p \quad (20)$$

$$\boldsymbol{\psi}_{i,j} \sim C \prod_{i,j=1,2,\dots,d}^{i < j} \left\{ (1 - \pi) \mathbf{N}(0, v_0^2) \mathbb{1}_{(-1,1)}(\boldsymbol{\psi}_{i,j}) + \pi \mathbf{N}(0, v_1^2) \mathbb{1}_{(-1,1)}(\boldsymbol{\psi}_{i,j}) \right\} \quad (21)$$

$$\boldsymbol{\mu} \sim \mathbf{N}(\boldsymbol{\mu}_\mu^0, \boldsymbol{\Sigma}_\mu^0) \quad \boldsymbol{\Phi}_1 \sim \mathbf{N}(\boldsymbol{\mu}_{\phi_1}^0, \boldsymbol{\Sigma}_{\phi_1}^0) \quad (22)$$

$$\boldsymbol{\Phi}_2 \sim \mathbf{N}(\boldsymbol{\mu}_{\phi_2}^0, \boldsymbol{\Sigma}_{\phi_2}^0) \quad \mathbb{1}_{M_E}(\boldsymbol{\Phi}_2) \quad \boldsymbol{\Omega}_t \sim \text{IW}(c_0, \boldsymbol{\Omega}_{t-1}), \quad (23)$$

which are Normal, Inverse Gamma and Inverse Wishart respectively, with densities

$$\pi(x) \propto x^{-(a_0+1)} \exp\left\{-\frac{b_0}{x}\right\} \quad (24)$$

$$\pi(\mathbf{M}) \propto |\mathbf{C}_0|^{c_0} |\mathbf{M}|^{-(c_0 + \frac{K+1}{2})} \exp\{-\text{tr}(\mathbf{C}_0 \mathbf{M}^{-1})\},$$

and  $\sigma_{j,j} = \{\sigma_{j,j}, j = 1, 2, \dots, d\}$  and  $\sigma_{i,j} = \{\sigma_{i,j}, i, j = 1, 2, \dots, d, i < j\}$ , where  $\sigma_{i,j}$  denotes the  $(i, j)$ -th entry of the matrix  $\boldsymbol{\Sigma}$ . Standard Gaussian priors are imposed on the loading factors  $(\boldsymbol{\Lambda}, \boldsymbol{\beta})$  even if shrinkage Lasso could be used instead. Concerning the prior specification of the variance–covariance matrix  $\boldsymbol{\Sigma}$ , we follow the same approach of Wang (2015) which extends the spike–and–slab approach of Mitchell and Beauchamp (1988) and George and McCulloch (1993) to positive–definite matrices, which has recently received much attention as a viable alternative to Lasso prior to introduce sparsity in large dimensional regression models as well as to model variance–covariance matrices in Gaussian graphical models. Specifically, we impose an Inverse Gamma prior on the main diagonal elements of  $\boldsymbol{\Sigma}$  in equation (20) and a spike–and–slab prior for the off–diagonal elements in equation (21). The values of  $v_0$  and  $v_1$  are further set to be small and large, respectively and the term  $C$  represents the normalising constant that ensures the integration of the



density function  $\pi(\Sigma)$  over the space of positive–definite matrices is one, and it depends on  $\{\sigma_{i,j}, i, j = 1, 2, \dots, p+d, \pi, v_0, v_1\}$ . Prior in equation (21) can be defined by introducing binary latent variable  $\mathbf{Z} \equiv (Z_{i,j})_{i<j} \in \mathcal{Z} \equiv \{0, 1\}^{(p+d)(p+d-1)/2}$  and the corresponding hierarchical model

$$\pi(\Sigma | \mathbf{Z}) = \prod_{i<j} \mathbf{N}(\sigma_{i,j} | 0, v_{z_{i,j}}^2) \quad (25)$$

$$\pi(\mathbf{z}) = \prod_{i<j} \left( \pi^{z_{i,j}} (1 - \pi)^{1-z_{i,j}} \right), \quad (26)$$

where  $v_{z_{i,j}} = \begin{cases} v_0 & \text{if } z_{i,j} = 0 \\ v_1 & \text{if } z_{i,j} = 1. \end{cases}$  The rationale behind using  $\mathbf{Z}$  for structure learning

is as follows. For an appropriately chosen small value of  $v_0$ , the event  $z_{i,j} = 0$  means that  $\sigma_{i,j}$  comes from the concentrated component  $\mathbf{N}(0, v_0^2)$ , and so  $\sigma_{i,j}$  is likely to be close to zero and can reasonably be estimated as zero. For an appropriately chosen large value of  $v_1$ , the event  $z_{i,j} = 0$  means that  $\sigma_{i,j}$  comes from the diffuse component  $\mathbf{N}(0, v_1^2)$  and so  $\sigma_{i,j}$  can be estimated to be substantially different from zero. Because zeros in  $\Sigma$  determine missing edges in graphs, the latent binary variables  $\mathbf{Z}$  can be viewed as edge-inclusion indicators. Given data  $\mathbf{y}^\dagger$ , the posterior distribution of  $\mathbf{Z}$  provides information about graphical model structures. The next proposition characterises the full conditional distribution of the parameters  $(\Psi, \sigma)$ .

**Proposition 3.** *Give the latent indicators  $\mathbf{Z}$  and the latent variables  $\varpi$ , the conditional posterior distribution of  $\tilde{\Sigma} = \mathbf{H}\Sigma\mathbf{H}'$  can be factorised as follows*

$$\begin{aligned} \pi(\Psi, \sigma | \mathbf{Y}, \mathbf{W}, \mathbf{Z}) &\propto |\Sigma|^{\frac{T}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S}\tilde{\mathbf{D}}^{-1}\tilde{\Psi}^{-1}\tilde{\mathbf{D}}^{-1}) \right\} \\ &\times \exp \left\{ -\frac{1}{2} \text{tr}(\vartheta' \mathbf{Y}' \tilde{\mathbf{D}}^{-1} \tilde{\Psi}^{-1}) \right\} \\ &\times \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{W} \otimes \xi \xi' \tilde{\Psi}^{-1}) \right\} \\ &\times \prod_{i<j} \exp \left\{ -\frac{\psi_{i,j}^2}{2v_{z_{i,j}}^2} \right\} \prod_{j=1}^d \exp \left\{ -\frac{b_0 \sigma_{j,j}^2}{2} \right\}, \quad (27) \end{aligned}$$

where  $\vartheta = \mathbf{1}'_T \xi$ ,  $\mathbf{W} = \text{diag}\{\varpi_1, \varpi_2, \dots, \varpi_T\}$ ,  $\tilde{\Psi} = \mathbf{H}\Psi\mathbf{H}'$  and  $\mathbf{S} = \tilde{\mathbf{Y}}\mathbf{W}^{-1}\tilde{\mathbf{Y}}'$ , with

$$\tilde{\mathbf{Y}} = [\mathbf{y}_1 - \mathbf{Z}\chi_1 \quad \mathbf{y}_2 - \mathbf{Z}\chi_2 \quad \dots \quad \mathbf{y}_T - \mathbf{Z}\chi_T] \quad (28)$$

is the  $d \times T$  matrix of observations  $\mathbf{y}_t - \mathbf{Z}\chi_t$ ,  $t = 1, 2, \dots, T$  stacked by row. Then

$$\pi\left(\sigma_{j,j}^{-1} | \mathbf{Y}, \mathbf{W}, \mathbf{Z}, \Psi, \sigma_{-j}\right) \propto \mathbf{N}(\tilde{\mu}_\sigma, \tilde{\tau}_\sigma) \mathbb{1}_{(0,\infty)}\left(\sigma_{j,j}^{-1}\right), \quad (29)$$

with  $\tilde{\mu}_\sigma = \tilde{\tau}_\sigma \left( b_0 + \frac{a_{2,2}}{\sqrt{\delta}} \right)$  and  $\tilde{\tau}_\sigma = \frac{\delta}{2s_{2,2}}$ , for  $j = 1, 2, \dots, d$  and

$$\pi(\Psi_j | \mathbf{Y}, \mathbf{W}, \mathbf{Z}, \Psi_{-j}, \sigma) \propto N_{d-1} \left( \tilde{\boldsymbol{\mu}}_{\Psi_j}, \tilde{\boldsymbol{\tau}}_{\Psi_j} \right) \mathbb{1}_{(-1,1)}(\Psi_{i,j}), \quad (30)$$

with

$$\tilde{\boldsymbol{\mu}}_{\Psi_j} = \tau_{\Psi_j} \left[ \mathbf{s}'_{1,2} \tilde{\mathbf{D}}_{1,1}^{-1} \tilde{\Psi}_{1,1}^{-1} \tilde{\mathbf{D}}_{1,1}^{-1} + \mathbf{a}'_{1,2} \tilde{\mathbf{D}}_{1,1}^{-1/2} \tilde{\Psi}_{1,1}^{-1} \tilde{\mathbf{D}}_{1,1}^{-1/2} + \mathbf{b}'_{1,2} \tilde{\Psi}_{1,1}^{-1} \sum_{t=1}^T \tilde{\boldsymbol{\omega}}_t \right] \quad (31)$$

$$\begin{aligned} \tilde{\boldsymbol{\tau}}_{\Psi_j}^{-1} = & \tilde{\mathbf{D}}_{1,1}^{-1} \tilde{\Psi}_{1,1}^{-1} \mathbf{S}_{1,1} \tilde{\Psi}_{1,1}^{-1} \tilde{\mathbf{D}}_{1,1}^{-1} + \tilde{\mathbf{D}}_{1,1}^{-1/2} \tilde{\Psi}_{1,1}^{-1} \mathbf{A}_{1,1} \tilde{\Psi}_{1,1}^{-1} \tilde{\mathbf{D}}_{1,1}^{-1/2} \\ & + \mathbf{V}^{-1} + \tilde{\Psi}_{1,1}^{-1} \mathbf{B}_{1,1} \tilde{\Psi}_{1,1}^{-1} \sum_{t=1}^T \tilde{\boldsymbol{\omega}}_t. \end{aligned} \quad (32)$$

## References

- Bernardi, M., Gayraud, G., and Petrella, L. (2015). Bayesian tail risk interdependence using quantile regression. *Bayesian Analysis*, 10(3):553–603.
- De Jong, P. and Shephard, N. (1995). The simulation smoother for time series models. *Biometrika*, 82:339–350.
- De Rossi, G. and Harvey, A. (2009). Quantiles, expectiles and splines. *Journal of Econometrics*, 152:179–185.
- Durbin, J. and Koopman, S. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89:603–615.
- Durbin, J. and Koopman, S. (2012). *Time series analysis by state space methods*. Oxford University Press.
- George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.
- Harvey, A. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge.
- Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of Fluids Engineering*, 83(1):95–108.
- Koenker, B. (2005). *Quantile Regression*. Cambridge University Press, Cambridge.
- Kotz, S., Kozubowski, T., and Podgorski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Progress in Mathematics Series. Birkhäuser Boston.
- Kotz, S. and Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.
- Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.
- Wang, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Anal.*, 10(2):351–377.
- Yu, K. and Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54:437–447.

# Does data structure reflect monuments structure? Symbolic data analysis on Florence Brunelleschi Dome

## *La struttura dei dati riflette le caratteristiche strutturali dei monumenti? Analisi di dati simbolici relativi al monitoraggio della Cupola del Brunelleschi a Firenze*

Bruno Bertaccini, Giulia Biagi, Antonio Giusti and Laura Grassini<sup>1</sup>

### **Abstract.**

The paper describes the work in progress about the analysis of the behaviour of the web cracks on the Brunelleschi's Dome of Santa Maria del Fiore in Florence. The web cracks in the Dome have always given rise to concern about the stability of the monument. The mechanical and electronic instruments have generated more than 6 million measurements, and the analyses performed so far, showed a steady increase in the size of the main cracks and, at the same time, a relationship with the environmental variables. The paper provides a continuous monitoring through those (big) data with the methods of the Symbolic Data Analysis techniques.

**Abstract.** *Il contributo presenta l'attività in corso circa l'analisi del comportamento dell'insieme di fessure presente sulla cupola del Brunelleschi di Santa Maria del Fiore a Firenze. La "ragnatela" di crepe nella cupola ha sempre dato adito a preoccupazioni circa la stabilità del monumento. Gli strumenti meccanici ed elettronici installati sulle fessure hanno generato oltre 6 milioni di misurazioni, e le analisi effettuate finora, hanno mostrato un costante aumento delle dimensioni delle principali crepe e hanno evidenziato, allo stesso tempo, un rapporto con le variabili*

---

<sup>1</sup> Dipartimento di Statistica, Informatica, Applicazioni "Giuseppe Parenti", Università degli Studi di Firenze, grassini@disia.unifi.it

*ambientali. Il presente contributo intende fornire, attraverso tecniche di analisi di dati simbolici, un'analisi di monitoraggio nel tempo del dataset a disposizione.*

**Key words:** Symbolic Data Analysis, Big data, Brunelleschi's Dome.

## 1. Introduction

In Gothic style to the design of Arnolfo di Cambio and completed in 1436 with the dome engineered by Filippo Brunelleschi, Santa Maria del Fiore does not need to be mentioned, except to state that by 1418 all that was left to finish was the dome. Weighing 37,000 tons and using more than 4,000,000 bricks, Brunelleschi's dome was the greatest architectural feat in the Western world.

First cracks in the dome appeared at the end of the 15th century, because "*the weight of the upper dome and of the lantern (at the top of the dome) exceeds the resistance of the base of the monument*". In 1695, a first commission with the task to investigate on the stability of the Dome was established, but nothing concrete has been done so far. To date Brunelleschi's dome is the only large dome of the Renaissance that had not yet been protected by actions of containment (rigid structures). For this reason, the monitoring system installed in the Dome, with more than 160 instruments (e.g., mechanical and electronic deformometers, thermometers, piezometers), is currently one of the most accurate control systems installed on a historical and architectural monument.

Cracks are now present in all eight webs, mainly in the fourth and sixth webs, both at the opposite of the nave. In web 4, currently the main crack shows an average increase of 5.5 mm/century. 23 deformometers are currently monitoring webs 4 and 6 (13 in web 4 and 10 in web 6) since 1988. With 4 measurements per day, there is a lot of produced data which requires a multifaceted approach to deal with the long term patterns, seasonal and climatic reactions, impact of other temporary factors (for example: earthquakes).

Those data are already analysed by various researchers (see for example: [1], [2]). In particular, Bertaccini in 2015 [2], tried to explain the complex interrelationships between deformometer measures through a SEM model, in order to represent the so called *breathing mechanism* of the Dome over time: the cracks tend to expand and shrink cyclically according to seasonal, climatic factors, some in a concordant way and others not.

In this paper, we present a descriptive analysis of that large amount of data, using interval valued variables, according to the Symbolic Data Analysis (SDA) approach [3]. SDA has already been used in studies concerning the health monitoring of civil engineering structures with the generally aim to provide tools and instruments to monitor the behaviour of a structure and detect or announce any abnormal behaviour ([4], [5], [6]), and a first study of the Brunelleschi's Dome data was presented in [16].

The aims of this work are:

- 1) to compute variability of measurements over time, by trying to discover any underlying trend;
- 2) to explore the relationships between the various cracks as discussed also in [2], relationships which resembles the *breathing* mechanism of the Dome over time.

The common theoretical framework for both point 1 and 2 is the properties of decomposition of covariance according to [7], [8], [9]. In fact, the covariance between two interval valued variables can be decomposed into the sum of the within component, related with the size of the intervals, and the between component, which is simply the covariance between the intervals midpoints.

Data are provided from the arithmetic mean of the intra-day measurements of 23 deformometers installed on webs 4 and 6. An interval variable reporting the 10th and 90th percentile of the daily average within each week, since 1988 and July 2007, has been defined. In this way, we drastically reduce data size and allow for situations in which daily measures, for temporary faults, are less than 4.

The paper is structured as follows. Sections 2 recall some basic algebra for interval valued data. Section 3 describes the empirical analysis, and Section 4 contains some final remarks.

## 2. Arithmetic of interval symbolic data and statistical parameters

Before analysing the statistical indices applied on symbolic interval data, it is useful to recall some algebra related to interval analysis [10].

Let us consider two intervals  $X = [x_1, x_2]$  and  $Y = [y_1, y_2]$  where  $[x_1, x_2]$  and  $[y_1, y_2]$  are respectively the minimum and the maximum of each interval.

The addition and subtraction between two intervals are respectively:

$$[x_1, x_2] + [y_1, y_2] = [x_1 + y_1, x_2 + y_2]$$

$$[x_1, x_2] - [y_1, y_2] = [x_1 - y_2, x_2 - y_1]$$

Moreover, the linear combination of two intervals with coefficients  $a, b$  is:

$$a[x_1, x_2] + b[y_1, y_2] = \begin{cases} [ax_1 + by_1, ax_2 + by_2] & \text{if } a > 0, b > 0 \\ [ax_1 + by_2, ax_2 + by_1] & \text{if } a > 0, b < 0 \\ [ax_2 + by_2, ax_1 + by_1] & \text{if } a < 0, b > 0 \\ [ax_2 + by_1, ax_1 + by_2] & \text{if } a < 0, b < 0 \end{cases} \quad (1)$$

Let us consider, now, a set of observations on  $X$  and  $Y$  represented by  $n$  intervals:

$$X_i = [x_{i1}, x_{i2}] \quad Y_i = [y_{i1}, y_{i2}], \quad i=1, \dots, n.$$

The centre of each interval  $i$  is the midpoint of the interval. For  $X$ :

$$\bar{X}_i = \frac{x_{i1} + x_{i2}}{2} \quad (2)$$

and the overall mean is:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n \bar{X}_i = \frac{1}{2n} \sum_{i=1}^n (x_{i1} + x_{i2}) \quad (3)$$

Several definitions of sample variance are introduced in the SDA field. Billard [7] suggests the following:

$$S_x^2 = \frac{1}{3n} \sum_{i=1}^n (x_{i1}^2 + x_{i1}x_{i2} + x_{i2}^2) - \frac{1}{4n^2} \left[ \sum_{i=1}^n (x_{i1} + x_{i2}) \right]^2 \quad (4)$$

It is derived by [11] under the assumption of uniform distribution within each interval. It is proved that (4) can be rewritten as

$$nS_x^2 = \frac{1}{3} \sum_{i=1}^n [(x_{i1} - \bar{X}_i)^2 + (x_{i1} - \bar{X}_i)(x_{i2} - \bar{X}_i) + (x_{i2} - \bar{X}_i)^2] + \sum_{i=1}^n \left( \frac{x_{i1} + x_{i2}}{2} - \bar{X} \right)^2 \quad (5)$$

Expression (5) shows that the total deviance  $nS_x^2$  can be decomposed as the sum of two components: (1) the internal variations of the data ( $SSW$ ) and (2) the between variations ( $SSB$ ), expressed by the comparison between the interval means and the overall mean:

$$SSW_x = \frac{1}{3} \sum_{i=1}^n [(x_{i1} - \bar{X}_i)^2 + (x_{i1} - \bar{X}_i)(x_{i2} - \bar{X}_i) + (x_{i2} - \bar{X}_i)^2] = \frac{1}{12} \sum_{i=1}^n (x_{i2} - x_{i1})^2 \quad (6)$$

$$SSB_x = \sum_{i=1}^n \left( \frac{x_{i1} + x_{i2}}{2} - \bar{X} \right)^2 = \sum_{i=1}^n (\bar{X}_i - \bar{X})^2 \quad (7)$$

Finally, also the sample symbolic covariance between two interval variables can be expressed as the sum of within and between components [7, 9]:

$$CODEVT = nCov(X, Y) = \sum_{i=1}^n \frac{(x_{i2} - x_{i1})(y_{i2} - y_{i1})}{12} + \sum_{i=1}^n (\bar{X}_i - \bar{X})(\bar{Y}_i - \bar{Y}) \quad (8)$$

where the left and right elements in the sum are, respectively, the within ( $CODEVW$ ) un-centred codeviance on the ranges, and the between sample co-deviance ( $CODEVB$ ) between  $X$  and  $Y$ . These expressions are derived under the assumptions of uniform distributions within the intervals. From (8), we see that the within component ( $CODEVW$ ) is not a true covariance matrix of the ranges because it is not computed on ranges centred on the mean.  $CODEVW$  is always positive and its

magnitude depends on the intervals' ranges. Therefore, the within codeviance incorporates information about the size of the individuals' rectangles. The between component is the co-deviance of the centres, in the classical framework.

The codeviances, *CODEVT*, are always larger than the classical codeviance matrix based on midpoints, which coincides with the between part and is used in the centres method. It follows that there are fewer negative terms in *CODEVT* than in *CODEVB*, and that the sign of *CODEVB* may be negative while the sign of *CODEVT* positive.

From (8), we can derive the symbolic correlation between two interval symbolic variables. A crucial advantage of the symbolic covariance between two intervals is that it fully utilizes all information in the data.

### 3. Data description and analysis

The data taken into account in the analyses are provided by part of the electronic monitoring system installed by ISMES in 1987. That system consists, among the others, of 66 deformometers, 56 thermometers, and two piezometers. This system records data every 6 hours starting on January 8, 1988. Data used in this work end on July 31, 2007 [12, 1], which means 35,100 measures per deformometer, for approximately 6 million measurements.

Each deformometer records the deformation of the building: since 1988, it has been recording the growth (with positive values) and the reduction (with negative values) of a crack. The minimum and maximum values over a given time period define an interval valued variable. The centre of that variable is the average growth or decrease of the crack with respect to the 1988 status, under the hypothesis of uniform distribution within the interval.

Available data are affected by the presence of missing data and outliers [2], mostly due to storms and blackouts that often cause calibration problems. In order to analyse a complete data matrix, we used the cleaned data provided and used in [2].

The localization of the deformometers in the structure determine their behaviour: those located near the tambour are more stable and exhibit less variability and a more similar time pattern.

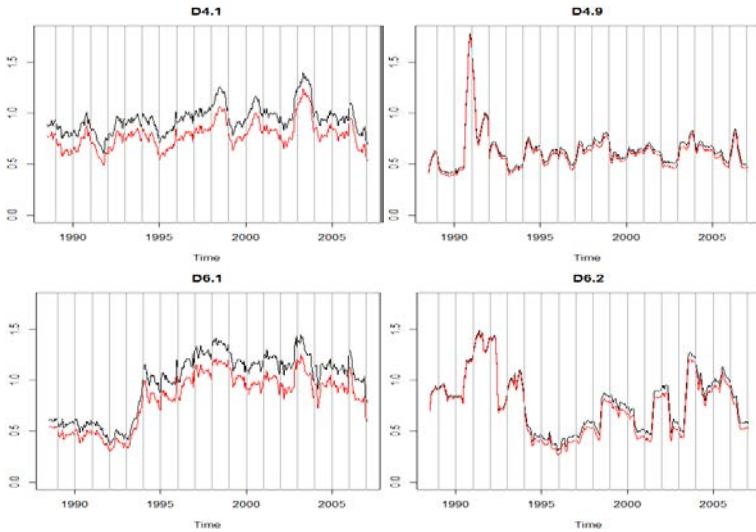
Web 6 exhibits, on the average, a larger variance than web 4, although the contribution of the within component is much lower, on the average (0.37% for web 6, 4.94% for web 4).

However, the cracks behaviour changes in time and it requires to take into account the dynamic nature of data. Therefore, we have computed a moving variance, by using a moving window of 52 weeks (one year). The idea is to explore the different time variability of the cracks, by selecting the year as the basic time window.

The use of adaptive methods and the choice of an annual period are not new in structural health monitoring and is practiced also with multivariate methods (for example, moving and recursive principal component analysis [13], [14], [15]).

Moreover, moving or rolling statistical indices are commonly computed to assess the constancy of model parameters. In this case, we wish to assess the importance of the between components of the variance.

Figure 1 shows the time patterns of the moving symbolic variance and the variance between, for some measurements. We can appreciate the different relevance of the within component, which is higher for some cracks (see graphs on the left) than others (see graphs on the right), depending on the placement of the cracks.



**Fig. 1:** Moving symbolic variance (black) and moving variance between (red) for some deformometers' measurements of webs 4 and 6

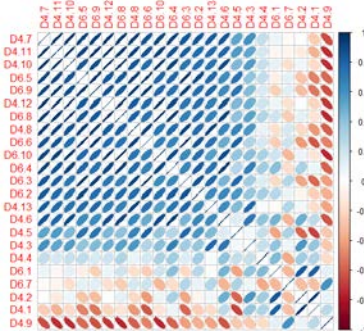
The exploration of the relationships among deformometers' measures is accomplished by the symbolic covariance matrix (see Figure 2a), in which we can appreciate the presence of strict positive relationships and not only within the same web. Only the deformometer D4.9 shows a negative correlation but it is mainly due to the abnormal behaviour in the early stages (Fig. 3). The variables are ordered to enlighten the similar behaviours. The determinant of the correlation matrix is almost zero, confirming the high linear relationships. However, there are 10 cases with a positive symbolic correlation and a negative correlation between midpoints, although the size of the correlation is very low. This huge presence of positive correlations may be determined by the positive contribution of the within covariance. For this reason, we provided the between correlation matrix as well (see Figure 2b).

From both matrices we see that the *breathing mechanism* of the Dome is mainly (about completely) characterized by a harmonious movement of the cracks as the

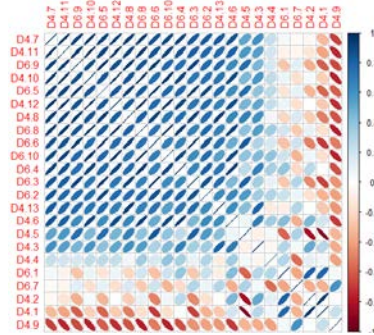


positive correlations are definitely dominant and stronger than the negative correlations.

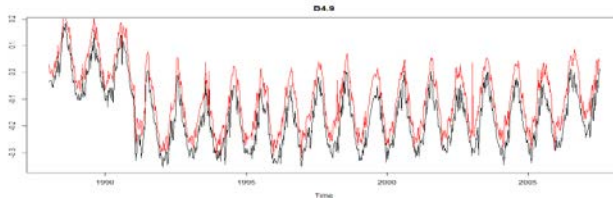
Excluding deformometer D4.9, negative correlations occur specifically for the cracks D4.1, D4.2, D6.1, D6.7, which tend to counteract the behaviour of most cracks.



**Fig. 2a:** Symbolic correlation matrix (order: first principal component)



**Fig. 2b:** – Between correlation matrix (order: first principal component)



**Fig. 3:** Time plot of the min and max values for D4.9

#### 4. Final remarks

The findings are consistent with the ones of [2]: the structure of the Dome may be assimilated to what in physics is defined as a “closed system”, in which the structural constraints define the relationship of forces between the various cracks, which in turn are subjected to the action of meteorological and seismic variables. To date, in literature a study that involves all variables simultaneously detected by the monitoring system is missing. A joint analysis of all available data determinates serious computational burdens and estimation and interpretation problems, due to the complexity of the relationship between the variables and the number of measures acquired per day.

With a reduced computational effort and working on a dataset reduced at less than the 15% of the one based on the daily averages, the Symbolic Data Approach conserves the same structure in the data. We are confident that the methods proper

of the Symbolic Data Analysis will permit to solve all those critical aspects that until now have prevented a comprehensive description of the mechanisms of the static-structural evolution of the monument, and to simulate the possible “reactions” to environment changes of exceptional nature.

## References

- 1 G. Bartoli, A. Chiarugi and V. Gusella, “Monitoring systems on historic buildings: the Brunelleschi Dome”, *Journal of structural engineering*, 1997.
- 2 B. Bertaccini, “Santa Maria del Fiore Dome Behavior: Statistical Models for Monitoring Stability”, *International Journal of Architectural Heritage: Conservation, Analysis, and Restoration*, 9:1, 25-37, 2015.
- 3 L. Billard and E. Diday, *Symbolic data analysis: Conceptual statistics and data mining*, Chichester: Wiley, 2006.
- 4 A. Cury, C. Crémone and E. Diday, “Application of symbolic data analysis for structural modification assessment”, *Engineering Structures*, vol. 32, 762–775, 2010.
- 5 A. Cury and C. Crémone, “Assignment of structural behaviours in long term monitoring: application to a strengthened railway bridge”, *Structural Health Monitoring*, vol. 11:4, 422–441, 2012.
- 6 J. Santos, C. Crémone, A. Orcesi, P. Silveira and L. Calado, “Static-based early-damage detection using symbolic data analysis and unsupervised learning methods,” *Frontiers of Structural and Civil Engineering*, vol. 9:1, 1-16, 2015.
- 7 L. Billard, “Some Analyses of Interval Data”, *Journal of Computing and Information Technology*, 16:4, 225-233, 2008.
- 8 J. Le-Rademacher and L. Billard, “Symbolic Covariance Principal Component Analysis and Visualization for Interval-Valued Data”, *Journal of Computational and Graphical Statistics*, 21:2, 413-432, 2012.
- 9 K. Košmelj, J. Le-Rademacher and L. Billard, “Symbolic Covariance Matrix for Interval-valued Variables and its Application to Principal Component Analysis: a Case Study”, *Metodološki zvezki*, 11:1, 1-20, 2014.
- 10 G. Alefeld and G. Mayer, “Interval analysis: theory and applications”, *Journal of computational and applied mathematics*, 121, 421-464, 2000.
- 11 P. Bertrand and F. Goupil, “Descriptive statistics for symbolic data”, in H-H. Boch and E. Diday (eds), *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Berlin, Springer-Verlag, 106-124, 2000.
- 12 C. Blasi and G. Bartoli, “Il sistema di monitoraggio della cupola di Santa Maria del Fiore: problematiche relative al funzionamento degli strumenti e alla gestione dei dati”, in *Monitoraggio delle strutture dell'ingegneria civile.*, Udine, CISM, 183-202, 1995.
- 13 D. Posenato, K. P. D. Inaudi and I. and Smith, “Methodologies for model-free data interpretation of civil engineering structures”, *Computers and Structures*, vol. 88:7/8, 467-482, 2010.
- 14 W. Li, H. Yue and S. Valle-Cervantes, “Recursive PCA for adaptive process monitoring”, *Journal of Process Control*, 10, 471-486, 2000.
- 15 X. Wang, U. Kruger and G. Irwin, “Process Monitorin Approach using Fast Moving Window PCA”, *Industrial & Engineering Chemistry Research*, vol. 44:5, 5691-5702, 2005.
- 16 B. Bertaccini, G. Biagi, A. Giusti and L. Grassini, “Symbolic data analysis approach for monitoring the stability of monuments”, in M. Pratesi and C. Perna (eds.), *Proceedings of the 48th Scientific Meeting of the Italian Statistical Society*, 1-6, 2016.

# A latent markov model approach for measuring national gender inequality

## *Modello latent markov per la misura delle disequità di genere nazionali*

Gaia Bertarelli and Franca Crippa and Fulvia Mecatti

**Abstract** Gender inequality - both in space and time - is a latent trait, namely only indirectly measurable through a collection of observable variables and indicators purposively selected. Even if composite indicators are normally used by social-scientists, when measuring gender-gap they are known to have case-specific technical limitations. In this paper we propose an innovative approach based on a multivariate Latent Markov model (LMM) for the analysis of gender inequalities as measured by the aforementioned indicators.

**Abstract** *La Statistica di Genere si occupa di sviluppare metodologie atte a cogliere disparità e differenze nella situazione delle donne e degli uomini in tutti gli aspetti della vita. Negli ultimi anni le disponibilità di dati per l'analisi di genere è aumentata poiché sempre più paesi stanno adottando survey specifiche. Gli strumenti più comuni nella letteratura della statistica di genere sono gli indicatori compositi che tuttavia presentano note limitazioni metodologiche. Vogliamo proporre un approccio innovativo alla statistica di genere basato su un modello latent markov multivariato per le analisi delle disuguaglianze.*

**Key words:** Gender Statistics, Clustering, GID-Database OECD, latent variable.

## 1 Background and Introduction

Gender equality is a recognized goal of modern democracies and an objective for global civilization since the effects of policies and actions capable at reducing gen-

---

Gaia Bertarelli  
University of Perugia, e-mail: [gaia.bertarelli@unipg.it](mailto:gaia.bertarelli@unipg.it)

Franca Crippa  
University of Milano - Bicocca, e-mail: [franca.crippa@unimib.it](mailto:franca.crippa@unimib.it)

Fulvia Mecatti  
University of Milano - Bicocca, e-mail: [fulvia.mecatti@unimib.it](mailto:fulvia.mecatti@unimib.it)

der disparities would actually benefit the society as a whole, both women *and* men. The availability of good quality data for engendered statistical analysis at the national level has increased since the 90's. Gender statistics based on household surveys and administrative records are becoming widely available. Gender inequality is a latent trait, namely only indirectly measurable through a collection of observable variables and indicators purposively selected as micro-aspects contributing to the latent macro-dimension. This is one of the main reasons for the popular use of *composite* indicators as current gender statistics indicators, *i.e.* aggregations - usually linear combinations - of a collection of simple indicators each singled out for assessing a puctual micro-aspect of the latent gender dimension. Several world rankings, based upon national gender composite indicators, are periodically released by supranational agencies (see for instance [2] for a compareate review). Even if normally used by social-scientists, such gender-gap measures are known to have case-specific technical limitations [3], which often lead to internal inconsistency since the ranking of a single country can vary in relation to the indicator considered. Moreover, a significant amount of the literature criticizes the use of composite indicators on the ground of trivial marginalization and arbitrariness [4]. In this paper we propose an innovative approach to gender inequality measure based on a multivariate Latent Markov model (LMM).

## 2 Data

We focus on two inequality indexes, the Gender Inequality Index (GII) and the Global Gender Gap Index (GGGI). A main reason for selecting them is their recentness, whose the aforementioned technical issues ask for advanced knowledge. The GII, introduced by UNDP in 2010, measures gender inequalities in three aspects of human development: reproductive health, empowerment and economic status. The GGGI was first introduced by the World Economic Forum in 2006 as a framework for capturing the magnitude of gender-based disparities and for tracking their progress. Three basic concepts underlie the GGGI. First, the index focuses on measuring gaps rather than levels. Second, it captures gaps in outcome variables rather than in input variables. Third, it ranks countries according to gender gaps rather than women's empowerment. It measures four aspects: economic participation and opportunity, educational attainment, health and survival and political empowerment. Rankings based on these indicators are different from each other as well as not constant over time, as a consequence of different choices in both measurable variable selection and aggregation system. In this paper we consider a multivariate model of latent markov type, able to receive as input both indexes as well as a set of covariates. An improved gender inequality measure is expected as a result. A preliminary univariate analysis is conducted for the period 2010-2016 able to assess possibly measurement errors in GGGI and GII. After considering constitutional gender equity (see <http://constitutions.unwomen.org/en>) and social structure

as covariates in the latent model component, we introduce time use in the measurement part.

### 3 Model

LMMs (see [1] for a general review), are a class of statistical models for longitudinal data which assume the existence of a latent process which affects the distribution of the response variables. The existence of two processes is assumed: an unobservable finite-state first-order Markov chain  $U_i^{(t)}$ ,  $i = 1, \dots, n$  and  $t = 1, \dots, T$  with state space  $\{1, \dots, m\}$  and an observed process  $Y_i^{(t)}$ ,  $i = 1, \dots, n$  and  $t = 1, \dots, T$ , where  $Y_i^{(t)}$  denotes the response variables for area  $i$  at time  $t$  and similiary for  $U_i^{(t)}$ . We assume that the distribution of  $Y_i^{(t)}$  depends only on  $U_i^{(t)}$ : the latent process fully explains the observable behaviour of an item together with possibly available covariates. Therefore it is important to distinguish between two components: the measurement model, which concerns the conditional distribution of the response variables given the latent process, and the latent model, which concerns the distribution of this latent process.

The unknown vector of parameters  $\phi$  in a LMM includes both the parameters of the Markov chain  $\phi_{lat}$  and the vector of parameters of the state-dependent distribution  $\phi_{obs}$ . The *measurement model* involves  $\phi_{obs}$  and it can be written as  $Y_i^{(t)}|U_i^{(t)} \sim f(y, u, \phi_{obs})$ . The *latent model* includes the parameters  $\phi_{lat}$  of the Markov chain which are the elements of the transition probability matrix  $\Pi = \{\pi_{u|\bar{u}}\}$ , with  $u, \bar{u} = 1, \dots, m$ ; where  $\pi_{u|\bar{u}} = P(U_i^{(t)} = u|U_i^{(t-1)} = \bar{u})$  is the probability that area  $i$  visits state  $u$  at time  $t$  given that at time  $t - 1$  it was in state  $\bar{u}$ , and the vector of initial probabilities  $\pi = (\pi_1, \dots, \pi_u, \dots, \pi_m)'$  where  $\pi_u = P(U_i^{(1)} = u)$  is the probability of being in state  $u$  at the initial time for  $u = 1, \dots, m$ . In this work we consider homogeneous LMMs.

LMMs can assess the presence of measurement errors or account for unobserved heterogeneity between areas in the analysis including covariates in the measurement model which do not completely explain the heterogeneity in the response variables. In LMMs the effect of the unobservable variable has its own dynamics. Moreover, a latent clustering of the population of interest can be pointed out. Our proposal is based on adapting the LMM to the gender statistics framework by interpreting national gender gap as the latent status of interest and using the distributions of the GGGI and GII as response variables. This methodology is derived by integrating into the same LMM both the selected composite indicators and a set of available observable covariates of any and possibly mixed nature. Our methodology organizes countries in ordinal clusters representing of the severity of gap. The classification is produced taking into account the values of the considered covariates and this overcomes the so called "world-at-two-speed" effect, i.e gender inequalities due to the denial of basic human rights (under-developing or in transition countries) or due to uneven opportunities between men and women (developed countries with gender

equality stated by law) ([2])) which is evident especially in the GII's distribution. However, looking at the temporal distributions, it seems that this gap goes to dwindle with time. Because of this, a longitudinal analysis is appropriated. We conduct a two-step analysis. At the beginning we apply a LMM with only spacial and gender constitutional equality covariates on the latent model in order to identify clusters of countries actually comparable under the "two-speed" effect mentioned above. Then we apply a LMM within each cluster considering social and economic covariates in the measurement model to detect main differences and variability within the same group.

## 4 Expected Results

We propose to integrate into the same LMM both the selected composite indicators and a set of available observable covariates of any and possibly mixed nature, categorical, ordinal and quantitative, fully exploiting the multidimensional latent nature of gender imbalance. The model would provide an organization of the countries in a (optimal) number of ordered cluster. The classification is produced taking into account the values of the considered covariates and this overcomes the so called "world-at-two-speed" effect which is evident especially in the GII's distribution. Moreover the proposed methodology deals with the forecasting of the future response and the path prediction.

## References

- [1] Francesco Bartolucci, Alessio Farcomeni, and Fulvia Pennoni. *Latent Markov models for longitudinal data*. CRC Press, 2012.
- [2] Fulvia Mecatti, Franca Crippa, and Patrizia Farina. A special gen (d) re of statistics: roots, development and methodological prospects of gender statistics. *International Statistical Review*, 80(3):452–467, 2012.
- [3] Iñaki Permanyer. The measurement of multidimensional gender inequality: continuing the debate. *Social Indicators Research*, 95(2):181–198, 2010.
- [4] Martin Ravallion. On multidimensional indices of poverty. *The Journal of Economic Inequality*, 9(2):235–248, 2011.

# Eurostat's methodological network: Skills mapping for a collaborative statistical office

Agne Bikauskaite and Dario Buono

**Abstract** Collaboration, interaction and exchange of knowledge among staff are important components for development and enriching of scientific intelligence within a statistical office. Eurostat's methodological network has been built as a skills mapping tool aiming identify in-house competencies for innovation and affordability of diffusion of knowledge, promotion and modernisation of collaboration on methodological issues, and processes within statistical office. We mainly focus on staff's knowledge and working and academic experience in methodological areas, domains and tools on statistics and econometrics. Quantitative network analysis metrics are used to measure the strengths of existing methodological competencies within Eurostat, to identify groups of people for collaboration in providing results on specific tasks, or characterise areas that are not fully integrated into methodological network. By combining network visualisation and quantitative analysis, we able easily assess competency level for each dimension of interest. Network analysis helps us in making decisions related to improvement of staff communication and collaboration, by building mechanisms for information flows, filling competency gaps. Data represented as mathematical graph makes readily visible general view, absorbs its structure, permits us to focus on persons, competencies and relations between them. Modernisation of ways of working leads to a more cost effective use of existing resources.

**Key words:** complex network, data analysis, network visualization, bipartite graphs, network projection, ego network, network analysis

## 1 Introduction

Collaboration, interaction and exchange of knowledge among staff are important components for development and enriching of scientific intelligence within a statistical office, especially when this exchange happens across areas of interest by both interacting sides. Methodological network has been built as a skills mapping tool aiming identify in-house competencies for innovation and affordability of

---

<sup>1</sup>

Agne Bikauskaite; email: agne.bikauskaite@ext.ec.europa.eu

Dario Buono, Eurostat; email: dario.buono@ec.europa.eu

diffusion of knowledge and information, and promotion and modernisation of collaboration on methodological issues and processes within statistical office. We mainly focus on staff's knowledge and working and academic experience in methodological areas, domains and tools on statistics and econometrics. This paper provides a set of mathematical network analysis measures from basic ones as size and degree to more complex as clustering coefficient and their correlation with degree that evaluates and makes better understandable the methodological knowledge network structure.

## 2 Methods

Quantitative network metrics are used to measure the strengths of existing methodological competencies within statistical office, to identify groups of people for collaboration in providing results on specific tasks, or characterise areas that are not fully integrated into methodological network. Network analysis helps us in making decisions related to improvement of staff communication and collaboration, by building mechanisms for information flows, filling competency gaps. By combining network visualisation and quantitative analysis, we can easily assess competency level for each dimension of interest.

### 2.1 *Bipartite graph*

Network data consists of a set of elements with relations on those elements and it may be represented as a graph. Our research subjects, individuals, form links which characterise their competencies in statistics and econometrics. Formally we have a graph  $G = (V, E)$ , where  $G$  is a relational structure consisting of set of vertices  $V$  and set of edges  $E$  [2]. We say that a graph is bipartite when the vertex set  $V$  is divided into two finite, disjoint  $V_1 \cap V_2 = \emptyset$  sets [4]. When  $V_1$  composed of the first mode vertices and  $V_2$  of the second mode vertices, we have the bipartite graph  $G = (V_1, V_2, E)$  where ties map the elements of different modes only.

### 2.2 *Network analysis*

In order to understand organisational methodological network and its structure network analysis statistical models have been employed. Data arranged as person by skill matrix  $A$  of size  $n_{V_1} \times n_{V_2}$ , where the rows correspond to methodological



$$A_{ij} = \begin{cases} 1, & \text{if person } i \text{ has a link to methodological skill } j; \\ 0, & \text{otherwise.} \end{cases}$$

The two most basic parameters of the graph are the number of vertices  $n = n_{V_1} + n_{V_2}$ , where  $n_{V_1} = |V_1|$  and  $n_{V_2} = |V_2|$ , and the number of edges  $m = |E|$ . [3]

Degree of the vertex helps to identify the best known competencies, and to diagnose critical areas within the methodological network. The average degree of sets of vertices  $V_1$  corresponding to survey respondents and  $V_2$  characterising listed methodological competencies are commonly used summarizing how well connected the network is, and is defined as proportions of number of links the network and number of nodes [1]

$$k_{V_k} = \frac{m}{n_{V_k}}, \text{ where } k = 1, 2.$$

While the average degree of overall network is obtained from the total numbers of nodes and edges by following equation [1]

$$k = \frac{2m}{n_{V_1} + n_{V_2}}.$$

The density  $\delta$  of the bipartite graph  $G$  measures average ratio of the actual degree of the nodes in the network and the maximum possible degree, which corresponds to the number of nodes in the set of different mode nodes

$$\delta(G) = \frac{m}{n_{V_1} n_{V_2}}.$$

This index is equal to 1 for the fully connected network (i.e.  $G$  has one component) and takes value of 0 when network is fully disconnected (i.e.  $G$  is composed entirely of isolates).

The clustering coefficient which concerns link correlation gives an idea of how compact is the network. The clustering coefficient of a node  $i$  is the proportion of links between the nodes within its neighbourhood divided by the number of edges that could possibly exist between the nodes [4]

$$cc_{ijl} = \frac{q_{ijl}}{(k_j - \eta_{ijl}) + (k_l - \eta_{ijl}) + q_{ijl}}$$

here  $j$  and  $l$  are a pair of neighbours of node  $i$ ,  $q_{ijl}$  is the number of squares which include these three nodes, and  $\eta_{ijl} = 1 + q_{ijl} + \theta_{jl}$  with  $\theta_{jl} = 1$  if  $i$  neighbours  $j$  and  $l$  are connected with each other and 0 otherwise.

Existing correlation of links allows us to sustain collaboration between methodological network members, while otherwise would not be able to function. If persons  $i$  and  $k$  form links to common competencies  $j$  and  $l$ , then efficient cooperation between them is more likely possible.

### 3 Results

The methodological knowledge network of this study case is simple, undirected, unweighted, static, and structured as bipartite graph, which consist of 117 vertices connected by 595 edges. The competencies degree of staff participated in the survey ranges from 3 to 11 which a mean of 8.88. While the degree of competencies nodes ranges from 0 to 39, with a mean of 10.2, what indicates, that each competence from the list has been indicated as well known by 10 respondents on average.

Degree sequence of competencies in statistics and econometrics indicates that most of methodological network members are familiar to Data Analysis and Time Series, highly competent in Social Statistics and National Accounts, and experienced in R and SAS statistical analysis software. While the biggest gap within methodological network observed of experts on Micro-data access and Statistical confidentiality, knowledgeable in Transport and Energy statistics, and capable to work with Hadoop tool. Other competencies from defined list are more or less covered and known by methodological network members.

The standard density measure gives a value 0.17, which shows a fairly sparse network with presence of 17 per cent of the possible links for average node. However, in this particular case the standard denominator is clearly not appropriate defining methodological network members' competencies. Due to restriction of choice of maximum 11 dimensions out of 50 possible, it cannot be interpreted as actual possible density. Using modified denominator, network obtain density of 0.79, which indicates high competency level of methodological network members.

In network studied, the clustering coefficient of competencies vertices set is not so high, above 20 per cent. The moderate correlation between clustering coefficient and degree is detected.

Data represented as mathematical graph makes readily visible general view, absorbs its structure, and permits us to focus on persons, competencies and relations between them. We distinguish the two node sets by colours, so that nodes of the same type have the same colour. The vertices of staff willing collaborate are coloured in green, blue, and red depending on the type of interest in involvement, while the set of yellow vertices corresponds to 28 methodological areas, 12 statistical domains and 10 tools. The size of the label and vertex is proportional to its degree.

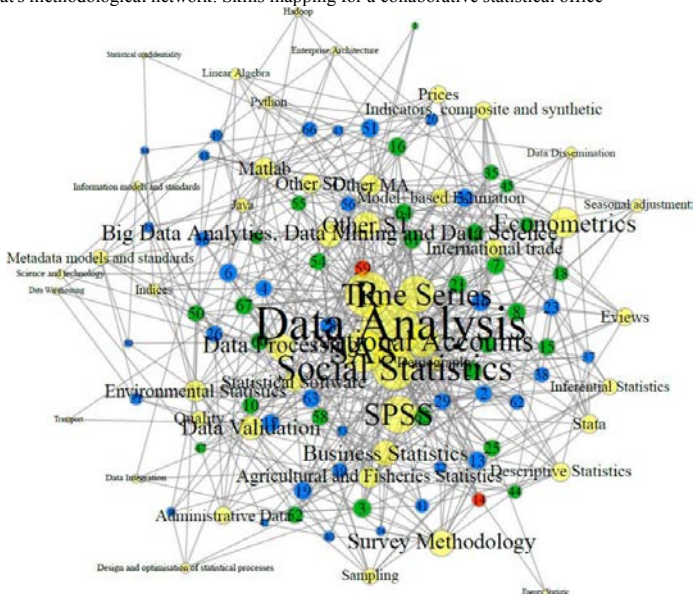


Figure 1: Organisational methodological knowledge network

In order to simplify visualisation, for deeper analysis of existing knowledge features and easier identification of clusters of correlated areas, methodological network has been divided into sub-networks by different breakdowns. Projections into one mode networks to grasp weighted relations between the same set of vertices had been made available as well by multiplying matrix  $A$  and its transpose  $A'$ . Analysing sub-networks we notice the tendency of increase of the density when average degree decreases. Overlapping of the structure of the nodes is very small, what points that there is large community of the methodological network members with knowledge and skills in different variation of areas.

## 4 Conclusions and discussion

In this study we map and evaluate existing methodological skills within the statistical office applying network analysis techniques. Networks as analytical and visualisation tools provide a number of useful outcomes. By detecting and then mapping methodological skills within organisation we are able to understand, spread, monitor and maintain existing skills, to develop tools for better knowledge accessibility and modernise information diffusion ways.

Obtained results provide quantitative evidence that methodological network members are qualified in different areas, given measures ensure possibility of well collaboration performance within the statistical office. Network is highly connected, significant gap of competencies is detected only in one methodological area from the defined competencies list of interest.

We can outline the importance of detecting and monitoring existing knowledge and skills within modern statistical office. Two employees could affect each other only if they know about each other and that common competencies are available between them, as efficient communication and collaboration within the organisation is possible only when we know with whom we could potentially contact. As well modernisation of the statistical office's ways of working leads to a more cost effective use of existing resources. Network is a key source in promoting and supporting of knowledge diffusion and expanding, enriching professional and personal skills and filling knowledge gaps within statistical office.

## References

1. S. P. Borgatti, M. G. Everett (1997). Network analysis of 2-mode data. *Social networks*, 19, 243-269
2. C. T. Butts (2008). Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, 11, 13-41
3. R. A. Hanneman, M. Riddle (2005). Introduction to social network methods. <http://faculty.ucr.edu/~hanneman/nettext/index.html>
4. M. Latapy, C. Magnien, N. Del Vecchio (2008). Basic notions for the analysis of large two-mode networks. *Social Networks*, 30, 31-48

# Big Data and Population Processes: A Revolution?

## *“Big Data” e processi di popolazione: una rivoluzione?*

Francesco C. Billari and Emilio Zagheni

**Abstract** We first discuss the centrality of data paradigms in demography, documenting their rise and fall over time also making use of Google Books Ngram Viewer. We then move on to discuss the undergoing “Data Revolution” in demography, with a focus on emerging forms of big data access and on the use of digital breadcrumbs.

**Abstract** *Il contributo discute la centralità dei paradigmi basati su dati in demografia, documentando il loro emergere e declino anche usando informazioni derivate da Google Books Ngram Viewer. Successivamente si discute l’attuale data revolution in demografia, focalizzando l’attenzione sulle forme emergenti di accesso ai “bid data” e sull’uso di “briciole di pane” digitali.*

**Key words:** computational demography, Big Data, digital demography

### 1 Four demographic data paradigms?

In Kuhn’s well-known discussion of the role of paradigms and “normal science” in scientific progress, the “normal” data to be used within a group of scholars are central to a paradigm. Discussions and debates take place in relation when paradigms are challenged: “The pre-paradigm period, in particular, is regularly marked by frequent and deep debates over legitimate methods, problems, and standards of

---

Francesco C. Billari

Department of Policy Analysis and Public Management; Carlo F. Dondena Centre for Research on Social Dynamics and Public Policies; Bocconi Institute for Data Science and Analytics; Università Bocconi via Röntgen 1, 20136 Milano, Italy, e-mail: francesco.billari@unibocconi.it

Emilio Zagheni

Department of Sociology, University of Washington, 211 Savery Hall, Box 353340, Seattle, WA 98195-3340, USA, e-mail: emilioz@uw.edu

solution” [19]. Given the data-intensive nature of demography, we characterize paradigms in demography by referring to the “normal” data used within a given paradigm.

We now illustrate four data paradigms in demography<sup>1</sup>. In order to illustrate the rise and fall of data paradigms we use the approach based on the prevalence of combination of terms (*Ngrams*) in books indexed in the Google corpus and accessible through Google Books Ngram Viewer<sup>2</sup>. [2].

### 1.1 *Census and administrative records*

That the study of population processes needs “Big Data” should come as no surprise. Indeed, data on population processes have always been “Big”, relatively to the epoch. It is useful to shortly recall here instances from the history of population research, taking into account that, historically, governments, churches and local authorities were the monopolists of data collection, curation and storage: census and administrative records are the paradigmatic data in this first demographic data paradigm.

In addition to Malthus’ ground-breaking work on the relationship between population change and economic development, which has been linked to the emergence of the modern population census, demographic research originates historically from the creative and innovative use of data originally collected for other purposes. Graunt’s 1662 *Natural and Political Observations Made upon the Bills of Mortality* are considered the founding essay for demography, as well as for epidemiology and statistics. The patient and pioneering analysis of the bills, which were published at a weekly rate, together with the low technological level then available, tells us that Graunt’s population data were already “Big”, relative to the epoch.

In historical demography, Henry and his colleagues pioneered the systematic linkages of parish registers to reconstruct population processes. They started from a village to extending this reconstruction to broader geographical areas, and generalized the effort through the careful preparation and analysis of linked data [9]. The family reconstructions of Henry and colleagues were again already “Big data” for the epoch.

The use of population-wide individual-level register data has become the marker, and the comparative advantage of Nordic demographers, with later efforts to link individual-level Census records and other registers extending to other countries

---

<sup>1</sup> This section is loosely inspired by Gray’s idea of a fourth paradigm [15]. Our characterization of four paradigms in demography differs from the one of Courgeau and Frank, who refer to the micro-macro perspective through which relationships between the individual- and population-level have been seen over time [6].

<sup>2</sup> <https://books.google.com/ngrams>. Figures 1, 2, 3 and were generated using Google Books Ngram Viewer with the English language 2012 corpus, using data between 1800 and 2008. The robustness of results was checked against case sensitiveness. Héran [14] carried a detailed analysis of the “demographic vocabulary” using the same approach

such as Belgium and the Netherlands in particular. Register data with individual identifiers (PINs, i.e. Personal Identification Numbers) that allow to link multiple individuals and to follow individuals over time are “Big Data”. Systems of PINs have been implemented in Sweden in 1947, in Norway in 1961, in Finland in 1964, and in Denmark in 1968 [22]. This lead, after some time, to the abolition of population censuses, with register-based “Big Data” also replacing the census for population counts.



**Fig. 1** The rise and fall of the “Census” in English books indexed in Google books. Source: <https://books.google.com/ngrams>

The demographic data paradigm based on census and administrative records is interested only in macro-level outcomes. Even when individual-level data are used as the starting point, the main interest is to quantify population-level parameters. Formal demography has emerged, developing the mathematical basis of the measurement of population-level quantities and of the study population dynamics, to complement and to inspire data analyses. To quantify the rise – and fall – of this paradigm in a graph, we here consider the emergence of the “census” as referred in books is depicted in Figure 1, where the peak is in the mid-1970s.<sup>3</sup>

## 1.2 Theory-driven micro-level data

After World War II, sample surveys began to be widespread to study population processes, following up on earlier development during the 1930s. During this period, “Demographers at the Bureau of Census, in collaboration with applied statisticians, began to develop sampling methods for meeting demands for timely measures of unemployment levels” [29]. By the end of the 1950s, at least in the United States, sample surveys have become central in social science research, including population research. By the end of the 1960s statistical packages have become available to

<sup>3</sup> A similar trend over time could be found when restricting the search to “population census”.

analyze what were the “big Data” of that epoch. Theory-driven micro-level data are the paradigmatic ones in this second demographic data paradigm.

The World Fertility Survey (WFS), coordinated by the London office of the International Statistical Institute (ISI), is the first major attempt to field a comparable sample survey across a wide range of countries, with 61 countries fielding a WFS between 1973 and 1984. At the outset of the programme, Sprehe, on behalf of the ISI states the basic aim of the WFS: “to provide scientific information that will permit each participating country to describe and interpret its population’s fertility” [32]. All WFS have to include “independent variables” that measure factors affecting fertility, to be included in micro-level statistical analyses. The choice of these factors is based on existing fertility theories, and builds on earlier survey-based research.

Since the WFS, sample surveys have become a prime, if not the major, source for population research during the last quarter of the Twentieth Century, in particular for what concerns family and fertility research. Demographers have also engineered, through formal demography, ways to exploit limited and defective data in order to estimate population-level parameters from information available through, for instance, the Demographic and Health Surveys (DHS), the successor of the WFS for developing countries. This “demographic estimation” approach has been recently and systematically illustrated by Moultrie and colleagues [24].

While traditional formal demography remains significantly anchored at the macro-level within this paradigm, there is a parallel development in which micro-level outcomes become the target of demographic research. The emergence of micro-level data, and of micro-level outcomes, as a central target in the study of population processes, and therefore to the second data paradigm in demography, is linked to the role of how micro- and meso-level factors influence demographic choices. Statistics comes to support this micro-level focus, and the 1972 article by David Cox [7] provides an elegant and general regression-based approach to life-table, three centuries after Graunt.

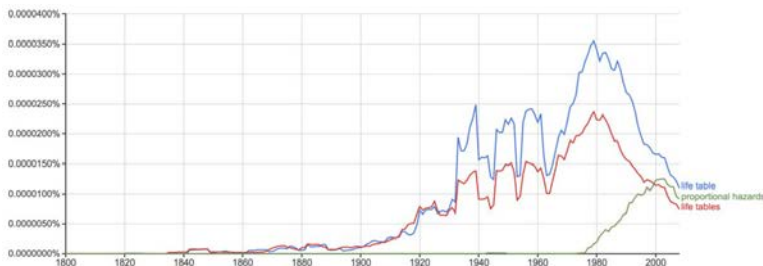
To quantify the rise – and fall – of this second paradigm, we look at two trends. First, the presence of the ngram “life table” or “life tables” as compared to “proportional hazards” in Figure 2. By the early 2000s “proportional hazards” basically reached the frequency of “life table” and seems to have started its decline. Second, in Figure 3, we show the rise and fall of the WFS as compared to the DHS and the Fertility and Family Survey between 1970 and 2008.

### ***1.3 Data-driven discovery meets theory-driven discovery***

The key critique to the second demographic data paradigm, is that it leads to forget population-level processes, the ultimate object of population scholars [21]. The idea that a multi-level paradigm should substitute the micro-based one has been suggested, among others, by Courgeau and Frank [6].

One way to see the link between demography targeting macro- and micro-level outcomes, as well as the link between data and theory in population research, is





**Fig. 2** Ngram prevalence for “life table” and of “proportional hazards” in English books indexed in Google books. Source: <https://books.google.com/ngrams>



**Fig. 3** Ngram prevalence for “World Fertility Survey”, “Demographic and Health Survey”, and “Fertility and Family Survey” in English books indexed in Google books. Source: <https://books.google.com/ngrams>

to cast a two-stage view of demographic research, distinguishing a *discovery* stage from an *explanation* stage [5]. The first–discovery–stage aims at the production of novel evidence at the population level. While description is often abhorred in the social sciences, research on population processes has shown that it is fundamental to anchor science to solid empirical bases. However, only novel evidence contributes to the cumulation of knowledge. The second–explanation–stage aims at developing accounts of demographic change and tests how the action and interaction of individuals generate what is discovered in the first stage.

The distinction between discovery and explanation does not refer to the fact that discovery on population processes should only be data-driven. However, discovery should not only be theory-driven, as for instance advocated by some social theorists [35] who do not give empirical discoveries a proper role in social science. The meeting between data-driven and theory-driven discovery has emerged in population research more recently, with a marriage between demography’s “powerful descriptive potential” and causal analysis [25]. A data example of this meeting is the effort to link administrative records with theory-based survey data. The Generations

and Gender Survey, for instance, in Nordic countries has exploited the available administrative records and linked them with the theory-based questionnaire [36].

In terms of methods, the key challenge for this third demographic data paradigm has been to link micro-level (data and theory) processes with macro-level population processes. The spread of computational, agent-based modeling has been seen as a potential solution. It is however too early to say whether this approach has made it to the core of a paradigm [2]. The systematic approach linking data- and theory-driven micro-founded simulation models with data at the population level has however become visible on demographic journals [3][17].

### ***1.4 A fourth paradigm?***

Are we at the dawn of a fourth demographic data paradigm? Yes and no: we are observing debates and trends that are typical of pre-paradigm shifts. The outcome is yet to be determined, but we identify three key crossroads.

*Centralization vs Decentralization.* Data collection and data interpretation, the essence of research, have been so far in the hands of a small minority of experts. Internet and the digital revolution have marked a discontinuity in practices of research. Everyone can potentially collect data for their own use or for research purposes, in various forms that include collecting genealogical family trees or using a mobile phone app to monitor health. The process is completely decentralized. However, corporations have emerged to tap into these new sources and bring them to a centralized repository. Similarly, the “open science” movement has brought non-professionals into the realm of research. Wikipedia is an example of a revolutionary form of mass-collaboration that goes beyond professional scientists to produce knowledge.

*Bias vs Variance.* Against the backdrop of decreasing survey response rates and the increasing availability of non-representative data, we are observing a strong interest in developing rigorous techniques to make sound inference from biased, non-probabilistic samples. For example, Wang et al. (2015)[37] showed that it is feasible to forecast election using data from surveys run on the videogame Xbox, if an appropriate approach that involves post-stratification is used.

*Re-purposing data vs Re-purposing methods.* Although there has always been data collected for goals other than research, today the mere scale of data that are available to anyone is so large that it is driving new directions of research. Re-purposing data might become the norm in social sciences and it may lead to the development of new methods, as well as re-purposing classical approaches to the new “Big Data” context.

## 2 Here comes the Data Revolution

The potential emergence of a fourth paradigm, has not gone completely unnoticed within the demographic research community. The International Union for the Scientific Study of Population (IUSSP) has joined the movement initiated by the United Nations towards a *Data Revolution*, i.e. a “new international initiative to improve the quality of statistics and information available to citizens” [1]<sup>4</sup>. We shortly address two aspects of this Data Revolution: “new” old Big Data and the so-called digital breadcrumbs.

### 2.1 “New” old Big Data

Ruggles [30] describes an “explosion” in the availability of “Big Microdata” for population research. The approach pushed by Ruggles and his colleagues at IPUMS, with a strong basis at the University of Minnesota, is to make micro-level population data from censuses and other sources as accessible as possible for other researchers. These data should allow unprecedented opportunities for population research over time and place, with rich geographical detail.

Other examples on how old big demographic data can be used in a new way use innovative approaches, including crowd-sourcing, to extract information from paper-based demographic documents, including hand-written ones [11]. In this case, the meeting between demographers and computer scientists has been fruitful and is promising in potentially delivering a wide range of (big) micro data about several sources.

### 2.2 Digital breadcrumbs

#### 2.2.1 Re-purposing data

The global spread of Internet and digital technologies, as well as the rapid diffusion of smartphones, have profoundly transformed our lives. As a consequence of the digital revolution, individuals leave an increasing quantity of traces online that can be analyzed to advance knowledge on population processes. Here we include examples of research that leveraged online data to study the three main components of demographic change: fertility, mortality and migration.

*Fertility.* Web searches represent the main online data source that has been used to study fertility. Reis and Brownstein (2010) show that the volume of Internet searches for abortion is inversely proportional to local abortion rates and directly proportional

---

<sup>4</sup> The two authors of this paper are also co-chairing the IUSSP Scientific Panel on “Big Data and Population Processes” established for the 2015–18 period.

to local restrictions on abortion [28]. Billari et al. (2013) show that Google searches for fertility-related queries, like ‘pregnancy’ or ‘birth’, can be used to predict fertility intentions and fertility rates several months ahead [4]. Ojala et al. (2017) use Google Correlate to detect evidence for different socio-economic contexts related to fertility (e.g., teen fertility, fertility of high income households, etc.) [26]. One of the most important messages of this line of literature is that combining traditional data sources with new data, like Web searches, can improve the predictive power of demographic models. However, that cannot be done in a naive way as correlations between aggregate Web searches and individual intentions may not persist for long periods of time. For example, the widely known Google flu approach to track influenza symptoms and detect potential outbreaks using Web searches [12] has been very useful and successful. However, at times, it also produced largely erroneous estimates, typically when the nature of the relationship between searches, news and behaviors changed [20]. Thus the results of these models have to be interpreted carefully and with caution.

*Mortality.* In the context of mortality analysis, the main source of online data results from decentralized collaborations that have produced genealogical data sets. For example, Fire and Elovici [8] use data collected from the WikiTree website to study correlations in lifespans among parents and children, as well as spouses. Similarly, Kaplanis et al. (2017) [16] leverage the data produced by enthusiasts of genealogy to evaluate population genetics theories on the dispersion of families. The key here is that (a) there are digital records that are left behind by people or institutions and (b) there is a critical mass of people who organize the data in meaningful ways for their own purposes and common goals.

*Migration.* Trends in international migrant flows have been estimated by tracking the locations, inferred from IP addresses, of users who repeatedly login into a Web service (e.g., Yahoo! [39, 34]). Geo-located Twitter tweets have been used to integrate the dimensions of internal and international migration [38] and to study global mobility patterns [13]. LinkedIn data have proven useful to evaluate trends in migration by educational attainment and sector of employment [33]. Google+ data, which provide pseudo migration histories, have proven useful to study how migrants connect countries within a network of flows. [23]

## 2.2.2 Re-purposing methods

Data science is about data, including re-purposing data. However, above all it is about the scientific use of data to advance knowledge. In this section we include a couple of examples of applications of classic social science methods and research design to the new data environment.

*Demographic calibration.* Non-representative digital breadcrumbs have to be calibrated against ‘ground truth’ data in order to evaluate biases and model them. Zagheni and Weber develop a method that combines the parsimonious perspective of model life tables, based on level and shape parameters, with standard calibration techniques [39]. The approach is inspired by calibration models for stochastic

microsimulations [31]. The underlying idea in the microsimulation literature is that simulations may generate estimates of quantities of interest that are biased. Identifying and modeling the bias is thus key to make statistical inference. We can consider social media and the Internet as “laboratories” that produce estimates of quantities of interest that are biased, but in a systematic way. Here, “systematic”, means that there are hidden, potentially stochastic rules that determine the relationship between the online data and the offline quantities of interest. Conditional on a model for the bias, statistical inference for the quantities of interest can be made using techniques like the Bayesian melding [27, 31].

*Difference-in-differences.* In some situations, “ground truth” data do not exist. Without any knowledge about the size and the direction of the bias, providing a reliable picture for the quantity of interest at one point in time is not possible. In these cases, instead of estimating the absolute value of variables of interest, a more modest task can be accomplished: estimating relative changes in quantities. This can be done using a difference-in-differences approach. A first demographic example includes estimating trends in migration patterns using geo-located Twitter data. [38] A second type of application relates to the evaluation of how shocks, like anti-immigrant laws, shape public sentiments about migration. [10]

### 2.2.3 Can formal demography make a comeback?

Can digital breadcrumbs offer new opportunities for formal demography? Online users form populations that can be analyzed using classic tools of formal demographic analysis. In turn, new types of population dynamics generate new questions that require new ways of formalization. Here we offer an illustrative example.

Consider users of a social media platform, like Twitter. The date when customers sign up for the service can be interpreted as a birth. The date when they stop tweeting for a long-enough interval of time can be interpreted as a death.

Figure 4 shows the age structure of a sample of active Twitter users (mid-2016). The histogram, which is equivalent to a population pyramid, reveals an age structure tilted towards ‘young’ users. In other words, it is a population that is growing rapidly. With these data only, we cannot say whether the growth in the population of Twitter users is driven by bots or real users, or whether it is related to ‘life course transitions’ of users, who may be quite active when they sign up, but then stop tweeting after a certain interval of time. However, we can use standard demographic techniques to estimate an approximate rate of growth in Twitter customers.

The problem can be stated as follows: given the number of individuals  $P_x$  at age  $x$  and  $P_y$  at age  $y$ , at time  $t$ , the goal is to find the rate at which the births were increasing between years  $t - x$  and  $t - y$ . It turns out that, under the assumption of exponential growth of births, the population rate of growth  $r$  is (see Keyfitz and Caswell [18]):

$$r = \frac{1}{y-x} \log\left(\frac{P_x L_y}{P_y L_x}\right) \quad (1)$$

where  $L_x$  and  $L_y$  are the fraction of people surviving  $x$  and  $y$  years, respectively.

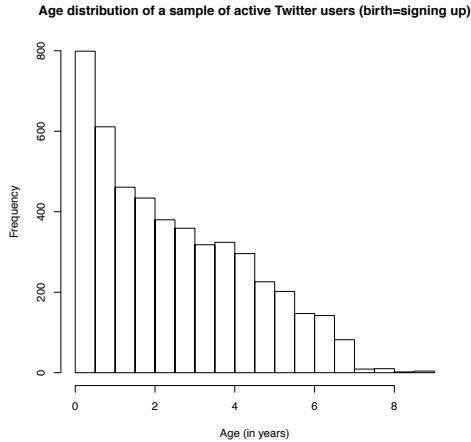
For the illustrative example based on a small sample of Twitter users, one obtains an estimated annual growth rate of the Twitter population equal to around 0.3. This is extremely fast growth, compared to rates for human populations.

### 3 Not a conclusion: The Data Revolution is not a dinner party

If we believe that a paradigm shift in the study of population processes, around the emergence of “Big Data”, is undergoing, it is by definition impossible to make firm conclusions. For sure, this “Data Revolution” will not be a dinner party.

Conventional wisdom will need to be challenged. Existing borders between disciplines might become a hindrance to scientific progress. Sticking to traditional approaches within the demographic research community might prevent further progress, or just let other, bolder, communities of scholars bring the advances needed to further our understanding of population processes. These challenges will need to be accompanied by new types of training for the younger generations of scholars—and perhaps even more relevantly, for the older generations. A fruitful way ahead is perhaps to combine traditional approaches with new one: counting and now-casting, indirect estimation and the used of non-representative Web-based data, official statistics and digital breadcrumbs.

A bit of patience, despite the speed of the field, is needed. Setbacks will happen and mistakes will be made within the “Data Revolution”. Trial and errors are needed.



**Fig. 4** Age distribution of a sample of active Twitter users (mid-2016), where ‘birth’ indicates the date when the user signed up. Source: own elaboration of data collected using the Twitter API.

Taking a very conservative stance that requires a new paradigm to have fully shown its potential in order to legitimize its approaches would however be an even bigger mistake. For the study of population processes, the Data Revolution is already here.

**Acknowledgements** This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n. 694262), project *DisCont - Discontinuities in Household and Family Formation*.

## References

1. The IUSSP on a Data Revolution for development. *Population and Development Review*, 41(1):172–177, 2015.
2. Jakub Bijak, Daniel Courgeau, Eric Silverman, and Robert Franck. Quantifying paradigm change in demography. *Demographic Research*, 30:911–924, 2014.
3. Jakub Bijak, Jason Hilton, Eric Silverman, and Viet Dung Cao. Reforging the wedding ring: Exploring a semi-artificial model of population for the united kingdom with gaussian process emulators. *Demographic Research*, 29:729, 2013.
4. Francesco Billari, Francesco D'Amuri, and Juri Marcucci. Forecasting births using Google. In *Annual Meeting of the Population Association of America*, 2013.
5. Francesco C. Billari. Integrating macro-and micro-level approaches in the explanation of population change. *Population Studies*, 69(sup1):S11–S20, 2015.
6. Daniel Courgeau and Robert Franck. Demography, a fully formed science or a science in the making? an outline programme. *Population-E*, 62(01):39–45, 2007.
7. D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.
8. Michael Fire and Yuval Elovici. Data mining of online genealogy datasets for revealing lifespans patterns in human population. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(2):28, 2015.
9. Michel Fleury and Louis Henry. *Des registres paroissiaux à l'histoire de la population: manuel de dépouillement et d'exploitation de l'état-civil ancien*. Institut National d'Études Démographiques, Paris, 1956.
10. René D Flores. Do anti-immigrant laws shape public sentiment?: A study of Arizona's SB 1070 using Twitter data. *American Journal of Sociology*, 2017.
11. Alicia Fornés, Josep Lladós, Joan Mas, Joana Maria Pujades, and Anna Cabré. A bimodal crowdsourcing platform for demographic historical manuscripts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 103–108. ACM, 2014.
12. Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.
13. Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41, 2014.
14. François Héran. The vocabulary of demography, from its origins to the present day: A digital exploration. *Population-E*, 70(3):497–536, 2015.
15. Tony Hey, Stewart Tansley, and Kristin M. Tolle. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft Research, Redmond, WA, 2009.
16. Joanna Kaplanis, Assaf Gordon, Mary Wahl, Michael Gershovits, Barak Markus, Mona Sheikh, Melissa Gymrek, Gaurav Bhatia, Daniel G MarArthur, Alkes Price, et al. Quantitative analysis of population-scale family trees using millions of relatives. *bioRxiv*, page 106427, 2017.

17. Ridhi Kashyap and Francisco Villavicencio. The dynamics of son preference, technology diffusion, and fertility decline underlying distorted sex ratios at birth: A simulation approach. *Demography*, 53(5):1261–1281, 2016.
18. Nathan Keyfitz and Hal Caswell. *Applied Mathematical Demography*, volume 47. Springer, 2005.
19. Thomas S. Kuhn. *The Structure of Scientific Revolutions. Second Edition, Enlarged*. University of Chicago Press, Chicago, IL, 1970.
20. David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of Google Flu: Traps in big data analysis. *Science*, 2014.
21. Ronald Lee. Demography abandons its core, 2001.
22. Torkild Hovde Lyngstad and Torbjørn Skardhamar. Nordic register data and their untapped potential for criminological knowledge. *Crime and Justice*, 40(1):613–645, 2011.
23. Johnatan Messias, Fabricio Benevenuto, Ingmar Weber, and Emilio Zagheni. From migration corridors to clusters: The value of Google+ data for migration studies. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 421–428. IEEE, 2016.
24. Tom Moultrie, Rob Dorrington, Allan Hill, Kenneth Hill, Ian Timæus, and Basia Zaba. *Tools for Demographic Estimation*. IUSSP, Paris, 2013.
25. Máire Ní Bhrolcháin and Tim Dyson. On causation in demography: Issues and illustrations. *Population and Development Review*, 33(1):1–36, 2007.
26. Jussi Ojala, Emilio Zagheni, Francesco C Billari, and Ingmar Weber. Fertility and its meaning: Evidence from search behavior. *Proceedings of the International Conference on Web and Social Media (ICWSM) 2017*, 2017.
27. David Poole and Adrian E Raftery. Inference for deterministic simulation models: the Bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255, 2000.
28. Ben Y Reis and John S Brownstein. Measuring the impact of health policies using Internet search patterns: the case of abortion. *BMC public health*, 10(1):514, 2010.
29. Peter H. Rossi, James D. Wright, and Andy B. Anderson. Sample surveys: History, current practice, and future prospects. In Peter H. Rossi, James D. Wright, and Andy B. Anderson, editors, *Handbook of Survey Research*, chapter 1, pages 1–20. Academic Press, New York, NY, 1983.
30. Steven Ruggles. Big microdata for population research. *Demography*, 51(1):287–297, 2014.
31. Hana Ševčíková, Adrian E Raftery, and Paul A Waddell. Assessing uncertainty in urban simulations using Bayesian melding. *Transportation Research Part B: Methodological*, 41(6):652–669, 2007.
32. J. Timothy Sprehe. The World Fertility Survey: An international program of fertility research. *Studies in Family Planning*, 5(2):35–41, 1974.
33. Bogdan State, Mario Rodriguez, Dirk Helbing, and Emilio Zagheni. Migration of professionals to the us: Evidence from LinkedIn data. In *Proceedings of the 6th International Conference on Social Informatics (SoCInfo)*, 2014.
34. Bogdan State, Ingmar Weber, and Emilio Zagheni. Studying inter-national mobility through IP geolocation. In *WSDM*, pages 265–274, 2013.
35. Richard Swedberg. *The art of social theory*. Princeton University Press, Princeton, NJ, 2014.
36. Andres Vikat, Zsolt Spéder, Gijs Beets, Francesco C Billari, Christoph Bühler, Aline Désesquelles, Tineke Fokkema, Jan M Hoem, Alphonse MacDonald, Gerda Neyer, et al. Generations and Gender Survey (GGS): Towards a better understanding of relationships and processes in the life course. *Demographic research*, 17:389–440, 2008.
37. Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.
38. Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. Inferring international and internal migration patterns from Twitter data. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 439–444. International World Wide Web Conferences Steering Committee, 2014.
39. Emilio Zagheni and Ingmar Weber. You are where you e-mail: using e-mail data to estimate international migration rates. In *WebSci*, pages 348–351, 2012.



# Bayesian Tensor Regression models

Monica Billio and Roberto Casarin and Matteo Iacopini

**Abstract** In this paper we introduce the literature on regression models with tensor variables and present a Bayesian linear model for inference, under the assumption of sparsity of the tensor coefficient. We exploit the CONDECOMP/PARAFAC (CP) representation for the tensor of coefficients in order to reduce the number of parameters and adopt a suitable hierarchical shrinkage prior for inducing sparsity. We propose a MCMC procedure via Gibbs sampler for carrying out the estimation, discussing the issues related to the initialisation of the vectors of parameters involved in the CP representation.

**Key words:** Tensor regression, Sparsity, Bayesian Inference, Hierarchical Shrinkage Prior

## 1 Introduction

The increasing availability of large sets of data presented in different formats (the most general class of examples includes all data that comes as images, such as EEG or the outcome of many other medical tests, video recordings and so on) has put forward some limitations of the existing multivariate econometric models. In the era of the so-called “*Big Data*”, the traditional mathematical representations of information in terms of matrices has some non-negligible drawbacks, the most remarkable of them being the difficulty of accounting for the structure within the data. When

---

Monica Billio  
Ca' Foscari University of Venice, e-mail: billio@unive.it

Roberto Casarin  
Ca' Foscari University of Venice, e-mail: r.casarin@unive.it

Matteo Iacopini  
Ca' Foscari University of Venice and Université Paris 1 - Panthéon-Sorbonne, e-mail: matteo.iacopini@unive.it

the information is available in the form of a collection of matrices, or in higher-order structures, such as tensors (e.g. in text or image processing), one approach to inference relies on vectorizing the object of interest by stacking all the elements in a column vector, which is then studied by means of multivariate analysis techniques. Though this way is well established in the literature, it is suited for dealing with low dimensional and unstructured arrays, but is not advisable in higher dimensions. By stacking all the elements of the object of interest in a long unidimensional array, we lose the structural information encrypted in the original shape of the variable. In other words, the physical features of the data matter since the value contained, for example, in a cell of a matrix is highly likely to depend on the values of a subset of the whole matrix; however, the process of vectorization does not allow to preserve this kind of information. Thus the introduction of novel methods able to treat 2-dimensional or multidimensional data as they are, that is, without modifying their shape by vectorization, still an open challenging question in statistics and econometrics.

Matrix models in econometrics have been employed over the past decade, especially in time series analysis where they have been widely used for the state space representation of these models. However it is only recently that the attention of the academic community has moved towards the study of matrix models. Continuing the stream of literature on time series models, [5] utilized these tools for studying dynamic linear models. Other fields of application include the analysis of Gaussian graphical models and the classification of longitudinal datasets.

A different stream of literature concentrates on tensor regression models and can be divided into two categories, according to the specification of the model. First of all, linear models as in [7] and [6] generally include in a regression function the scalar product between a tensor  $\mathcal{X} \in \mathbb{R}^{d_1 \times \dots \times d_p}$  and a tensor of coefficients  $\mathcal{B} \in \mathbb{R}^{d_1 \times \dots \times d_p}$ . More in detail, [6] propose a multivariate model with tensor covariate for longitudinal data analysis; whereas [7] uses a generalized linear model with exponential link and tensor covariate for analysing image data.

Following a different purpose, [3] generalizes the univariate or multivariate regression by allowing both the response and the covariate to be tensor-valued. He exploits the Tucker product, which has been originally developed as a tensor representation method, then follows the Bayesian approach for the estimation. From a frequentist perspective, the literature is still limited, partly due to the highly complex optimisation problems involved in the estimation process, which generally relies on iterative maximum likelihood procedures.

Motivated by the need for new methodologies able to deal directly with two- or higher-dimensional variables, we propose a new linear regression modelling framework well suited for data that are available in the shape of tensors, as both the response variable and the covariate are concerned. The general model we propose is shown to encompass both univariate and multivariate regression as special cases. Furthermore, we address the issue of dimensionality by the exploitation of a suitable parametrization which enables to achieve both parameter parsimony and to incorporate sparsity in the coefficients. For what concerns inference, the Bayesian approach is very appealing in this framework as it allows the necessary flexibility while re-

taining analytical and computational tractability. Therefore adopt this perspective and provide a Monte Carlo Markov Chain (MCMC) procedure for carrying out the estimation.

The main contribution of this paper is to provide a unifying framework for existing econometric models, which generalises to higher-dimensional and structured variables. From a computational perspective, we focus on the issues related to the initialization of the Gibbs sampler for the vectors of parameters involved in the CONDECOM/PARAFAC (CP) representation of the tensor of regression coefficients.

The remind of the paper is the following: in Section 2 we present the model and briefly discuss its most relevant characteristics. The inferential approach is then outlined in Section 3 and the results of the estimation process based on a simulated dataset are given in Section 4. Finally, we draw conclusions and give an outline of current research in Section 5.

## 2 Bayesian Tensor Regression Model

Define a tensor as a generalisation of a matrix into a  $D$ -dimensional space, namely:  $\mathcal{X} \in \mathbb{R}^{d_1 \times \dots \times d_D}$ , where  $D$  is the order of the tensor and  $d_j$  is the length of dimension  $j$ . Clearly, matrices, vectors and scalars are particular cases of tensor variables, of order 2, 1 and 0, respectively. The common operations defined on matrices and vectors in linear algebra can be applied also to tensors (henceforth, to be intended of order  $\geq 3$ ), via slight generalisations in their definition. Moreover other operators and representation can be defined on tensors which are not defined on lower dimensional objects. For a remarkable survey on this subject, see [4].

The general tensor linear regression model (see [1] for greater details) we present here can manage covariates and response variables in the form of vectors, matrices or tensors. It is given by:

$$\mathcal{Y}_t = \mathcal{A} + \mathcal{B} \times_{D+1} \text{vec}(\mathcal{X}_t) + \mathcal{C} \times_{D+1} \mathbf{z}_t + \mathcal{D} \times_n W_t + \mathcal{E}_t, \quad \mathcal{E}_t \stackrel{iid}{\sim} \mathcal{N}_{d_1, \dots, d_D}(0, \Sigma_1, \dots, \Sigma_D) \quad (1)$$

where the tensor response and errors are given by  $\mathcal{Y}_t, \mathcal{E}_t \in \mathbb{R}^{d_1 \times \dots \times d_D}$ ; while the covariates are  $\mathcal{X}_t \in \mathbb{R}^{d_1^X \times \dots \times d_M^X}$ ,  $W_t \in \mathbb{R}^{d_n \times d_2^W}$  and  $\mathbf{z}_t \in \mathbb{R}^{d_z}$ . The coefficients are:  $\mathcal{A} \in \mathbb{R}^{d_1 \times \dots \times d_D}$ ,  $\mathcal{B} \in \mathbb{R}^{d_1 \times \dots \times d_D \times p}$ ,  $\mathcal{C} \in \mathbb{R}^{d_1 \times \dots \times d_D \times d_z}$ ,  $\mathcal{D} \in \mathbb{R}^{d_1 \times \dots \times d_{n-1} \times d_2^W \times d_{n+1} \times \dots \times d_D}$  where  $p = \prod_i d_i^X$ . The symbol  $\times_n$  stands for the mode- $n$  product between a tensor and a vector, as defined in [4].

Notice that this model provides a generalization of several well-known econometric linear models, among which univariate and multivariate regression, VAR, SUR and Panel VAR models and matrix regression model (see [1] for formal proofs).

We focus on the particular case where both the regressor and the response variables are square matrices of size  $k \times k$  and the error term is assumed to be distributed

according to a matrix Normal distribution:

$$Y_t = \mathcal{B} \times_3 \text{vec}(X_t) + E_t \quad E_t \stackrel{iid}{\sim} \mathcal{N}_{k,k}(\mathbf{0}, \Sigma_c, \Sigma_r). \quad (2)$$

In this case the coefficient is a three dimensional tensor  $\mathcal{B} \in \mathbb{R}^{k \times k \times k^2}$ . Since the model is overparametrised, in order to provide a significant reduction of the number of parameters we assume a CONDECOM/PARAFAC (CP) representation (more details in [4] and [1]) for the tensor, as follows:

$$\mathcal{B} = \sum_{r=1}^R \mathcal{B}_r = \sum_{r=1}^R \beta_1^{(r)} \circ \dots \circ \beta_D^{(r)}, \quad (3)$$

where the vectors  $\beta_j^{(r)} \in \mathbb{R}^{d_j} \forall j = 1, \dots, D$  are also called margins of the CP representation and  $R$  is the CP-rank of the tensor. Since estimation of  $R$  is a NP-hard problem, in the following we are assuming a fixed value for it. The CP decomposition permits a significant reduction of the number of parameters of the tensor of coefficients. For a  $D$  order tensor with length  $d_i$  of dimension  $i$ , it decreases from  $\prod_{i=1}^D d_i$  to  $R \sum_{i=1}^D d_i$ . The corresponding gain we obtain by making this assumption in model (2) consists in a reduction of the complexity from  $\mathcal{O}(k^4)$  to  $\mathcal{O}(k^2(R+1))$ .

### 3 Bayesian Inference

We follow the Bayesian approach for inference, thus we need to specify a prior distribution for all the parameters of the model. The adoption of the CP representation for the tensor of coefficients is crucial from this point of view, as it allows to reduce the problem of specifying a prior distribution on a multi-dimensional tensor, for which very few possibilities are available in the literature, to the standard multivariate case. In fact it suffices to define a prior distribution for all the margins: this can be done in a very flexible way by using multivariate distributions. As a consequence, we are allowed to embed the prior knowledge of sparsity of the coefficient by the choice of a suitable hierarchical shrinkage prior.

Building from [2], we define a prior for each of the margins  $\beta_j^{(r)}$  of the tensor coefficient  $\mathcal{B}$  by means of the following hierarchy:

$$\pi(\beta_j^{(r)} | \mathbf{W}, \phi, \tau) \sim \mathcal{N}_{d_j}(\mathbf{0}, \tau \phi_r \mathbf{W}_{j,r}) \quad \forall r = 1, \dots, R \quad \forall j = 1, 2, 3 \quad (4)$$

$$\pi(w_{p,j,r}) \sim \mathcal{Exp}(\lambda_{j,r}^2/2) \quad \forall r = 1, \dots, R \quad \forall j = 1, 2, 3 \quad \forall p = 1, \dots, d_j \quad (5)$$

$$\pi(\phi) \sim \mathcal{Dir}(\alpha, \dots, \alpha) \quad \forall r = 1, \dots, R \quad (6)$$

$$\pi(\tau) \sim \mathcal{Ga}(a_\tau, b_\tau) \quad (7)$$

The idea behind this prior construction is to induce sparsity of the tensor  $\mathcal{B}$  in a flexible way via the introduction of hyperparameters accounting for the different

levels. The component  $W_{j,r}$  is a  $d_j \times d_j$  diagonal matrix whose entries  $(\{w_{p,j,r}\}_{p=1}^{d_j})$  represent the individual (local) share of the variance; instead  $\phi_r$  introduces sparsity at a medium level by shrinking all the  $r$ -th margins of the CP representation in eq. (3). Finally,  $\tau$  provides a global control on the variance, common for all the vectors.

We complete the prior specification by assuming two Inverse Wishart as prior distributions for the covariance matrices of the error term:

$$\pi(\Sigma_r) \sim \mathcal{IW}_k(\nu_r, \Psi_r) \quad (8)$$

$$\pi(\Sigma_c) \sim \mathcal{IW}_k(\nu_c, \Psi_c) \quad (9)$$

In compact notation, the joint prior distribution is given by:

$$\pi(\theta) = \pi(\mathcal{B}|\mathbf{W}, \phi, \tau)\pi(\mathbf{W})\pi(\phi)\pi(\tau)\pi(\Sigma_c)\pi(\Sigma_r). \quad (10)$$

Given a sample  $(\mathbf{Y}, \mathbf{X}) := \{Y_t, X_t\}_{t=1}^T$  and defining  $\mathbf{x}_t := \text{vec}(X_t)$ , the likelihood function of the model (2) is given by:

$$L(Y_1, \dots, Y_T | \theta) = \prod_{t=1}^T (2\pi)^{-\frac{k^2}{2}} |\Sigma_c|^{-\frac{k}{2}} |\Sigma_r|^{-\frac{k}{2}} \exp \left\{ -\frac{1}{2} \Sigma_c^{-1} (Y_t - \mathcal{B} \times_3 \mathbf{x}_t)' \Sigma_r^{-1} (Y_t - \mathcal{B} \times_3 \mathbf{x}_t) \right\}. \quad (11)$$

The details of the Gibbs sampler along with the analytical derivation of the full conditionals are given in [1].

## 4 Simulation

The model poses a problem for the initialisation of the vectors  $\{\beta_j^{(r)}\}_{j,r}$ , since there is no guidance in which could be a suitable starting value for the Gibbs sampler (which is known to be sensitive to the initial point). A naïve way to initialise the vectors is to use an accept/reject algorithm based on a proposal corresponding to the prior distribution. However this approach has been proven to converge slowly due to the low acceptance rate of good starting values. Instead, we address this issue by initialising the vectors  $\{\beta_j^{(r)}\}_{j,r}$  with the outcome of a simulated annealing algorithm. This is a stochastic optimisation algorithm close in spirit to the Metropolis-Hastings algorithm: by the choice of a tempering process, at the initial iterations it makes big moves on the domain, allowing exploration of the parameter space, while at successive steps the reduction of the temperature contracts the range where optima are looked for, until convergence. For a suitable tuning of the tempering scheme, this algorithm is able to provide the global optimum. In practice, it delivers good starting points for the Gibbs in fast computing time.

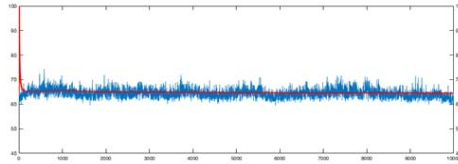
We performed a stimulation study by drawing a sample of  $T = 100$  couples  $\{Y_t, X_t\}_t$  of square matrices of dimension  $k = 10$ . The regressor is built by entry-wise independent AR(1) processes:

$$\begin{cases} x_{ij,t} - \mu = \alpha_{ij}(x_{ij,t-1} - \mu) + \eta_{ij,t} & \eta_{ij,t} \sim \mathcal{N}(0, 1) \\ y_{ij,t} = 10 + \beta_{ij}x_{ij,t} + \varepsilon_{ij,t} & \varepsilon_{ij,t} \sim \mathcal{N}(0, 1) \end{cases} \quad (12)$$

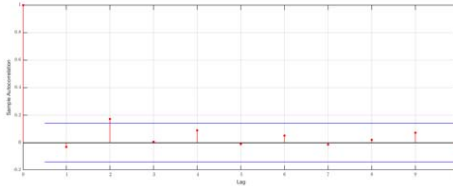
where  $\mathbb{E}[\eta_{ij,t}\eta_{kl,v}] = 0$ ,  $\mathbb{E}[\varepsilon_{ij,t}\varepsilon_{kl,v}] = 0$  and  $\mathbb{E}[\eta_{ij,t}\varepsilon_{kl,v}] = 0$ ,  $\forall (i, j) \neq (k, l)$  and  $\forall t \neq v$ ; moreover  $\mathbb{E}[\eta_{ij,t}] = 0$ . In addition the coefficients are drawn from  $\alpha_{ij} \sim \mathcal{U}(-10, 10)$  and  $\beta_{ij} \sim \mathcal{U}(-1, 1)$ .

We demeaned the simulated data, then we initialised the marginals of the tensor  $\mathcal{B}$  by simulated annealing and run the Gibbs sampler for  $N = 10000$  iterations. As an indicator of the goodness of fit of the estimated parameters, we computed the Frobenious norm between the original tensor and the one reconstructed via the posterior of the marginals. The outcome is shown in Fig.(1): in blue it is shown the trace plot, while the red curve is the progressive mean across iterations. In order to reduce the autocorrelation of the posterior sample, we performed thinning by keeping one observation every 50, the final autocorrelation function after this step is plotted in Fig. (2).

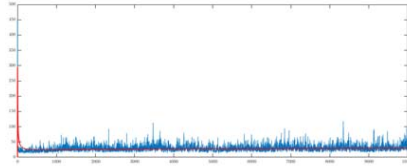
**Fig. 1** Trace plot (blue) and progressive mean (red) of  $\|\mathcal{B}^{true} - \mathcal{B}^{post}\|_2$ .



**Fig. 2** ACF of reconstructed tensor from  $\beta_1, \beta_2, \beta_3$  after thinning, by keeping one simulated value every 50.

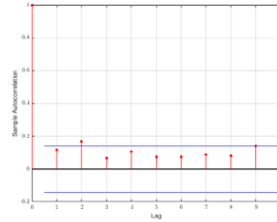


We report also the results for the parameter  $\tau$ , which drives the global component of the shrinkage, in Fig. (3): the trace plot and progressive mean indicate the convergence of the algorithm, however even in this case the autocorrelation between iterations is present, though lower than for the marginals. This is due to the fact that the  $\beta_1, \beta_2, \beta_3$  are sampled individually by approximating their joint distribu-



**Fig. 3** Trace plot (blue) and progressive mean (red) of  $\tau$ .

tion via the full conditionals of each single  $\beta_j^{(r)}$ . For the sake of comparison with the previous graph, in Fig. (4) we report the autocorrelation function of the posterior sample of  $\tau$  thinned by taking one value every 50. Major details on the results of the simulation are reported in [1].



**Fig. 4** ACF of  $\tau$ , after thinning by keeping one simulated value every 50.

We are currently working on simulations for models of bigger size as well as for applying this methodology for the study of temporal dynamics of real networks.

## 5 Conclusions

We propose a linear regression model for matrices which is a generalisation of standard econometric models and allows each entry of the covariate to exert a different effect on each entry of the response. The model is a reduced form of a general tensor regression, nonetheless all the analytical and computation results discussed are directly applicable to the more general form. In particular, the delicate issue of the initialisation of the sampler has been carried out by means of an efficient implementation of simulated annealing. The model has been tested on a synthetic dataset, showing good performance in the reconstruction of the true tensor of coefficients. We plan to apply the methodology to real network datasets in order to study their temporal dynamics.

**Acknowledgements** This research has benefited from the use of the Scientific Computation System of Ca' Foscari University of Venice (SCSCF) for the computational for the implementation of the inferential procedure.

## References

1. Billio, M., Casarin, R., Iacopini, M.: Bayesian Matrix Regression, Ca' Foscari University of Venice - Dipartimento di Economia, Venice (2017)
2. Guhaniyogi, R., Qamar, S., Dunson, D. B.: Bayesian Tensor Regression, arXiv preprint arXiv:1509.06490, (2015)
3. Hoff, P. D.: Multilinear Tensor Regression for Longitudinal Relational Data, *The Annals of Applied Statistics*, **9**, 1169–1193 (2015)
4. Kolda, T. G., Bader, B. W.: Tensor Decompositions and Applications, *SIAM Review*, **51**, 455–500 (2009)
5. Wang, H., West, M.: Bayesian Analysis of Matrix Normal Graphical Models, *Biometrika*, **96**, 821–834 (2009)
6. Zhang, X, Li, L., Zhou, H., Shen, D.: Tensor Generalized Estimating Equations for Longitudinal Imaging Analysis, arXiv preprint arXiv:1412.6592 (2014)
7. Zhou, H., Li, L., Zhu, H.: Tensor Regression with Applications in Neuroimaging Data Analysis, *Journal of the American Statistical Association*, **108**, 540–552 (2013)



# Bayesian nonparametric sparse Vector Autoregressive models

## *Modelli Autoregressive multivariati: un approccio Bayesiano nonparametrico con sparsità*

Monica Billio and Roberto Casarin and Luca Rossini

**Abstract** Seemingly unrelated regression (SUR) models are useful in studying the interactions among economic variables. In a high dimensional setting, these models require a large number of parameters to be estimated and suffer of inferential problems. To avoid overparametrization issues, we propose a hierarchical Dirichlet process prior (DPP) for SUR models, which allows shrinkage of coefficients toward multiple locations. We propose a two-stage hierarchical prior distribution, where the first stage of the hierarchy consists in a lasso conditionally independent prior of the Normal-Gamma family for the coefficients. The second stage is given by a random mixture distribution, which allows for parameter parsimony through two components: the first is a random Dirac point-mass distribution, which induces sparsity in the coefficients; the second is a DPP, which allows for clustering of the coefficients.

**Abstract** I modelli di regressione (SUR) sono utili per studiare le interazioni tra variabili economiche di interesse. Quando si lavora con grandi dimensioni, questi modelli richiedono la stima di un gran numero di variabili e soffrono di problemi inferenziali. Per evitare i problemi di sovrapparametrizzazione, noi proponiamo una prior di tipo Dirichlet gerarchico (DPP) per il modello SUR, dove si permette la contrazione dei coefficienti attraverso diverse posizioni. Noi proponiamo una prior gerarchica a due stages, dove il primo stage consiste in una lasso prior per i coefficienti dalla famiglia delle Normali-Gamma. Il secondo stage usa una mistura di distribuzioni random attraverso due componenti: la prima è una distribuzione di Dirac su un punto, che induce sparsità per i coefficienti; la seconda invece si basa su una DPP, che permette la clusterizzazione dei coefficienti.

**Key words:** Bayesian nonparametrics, Bayesian model selection, Shrinkage, Large vector autoregression.

---

Monica Billio and Roberto Casarin  
University Ca' Foscari of Venice, e-mail: billio@unive.it and e-mail: r.casarin@unive.it

Luca Rossini  
Free University of Bozen, e-mail: luca.rossini@unibz.it

## 1 Introduction

In the last decade, high dimensional models and large datasets have increased their importance in economics (e.g., see [8]). The use of large dataset has been proved to improve the forecasts in large macroeconomic and financial models (see, [1], [3], [5], [9]). For analyzing and better forecasting them, SUR/VAR models have been introduced [11, 12], where the error terms are independent across time, but may have cross-equation contemporaneous correlations. SUR/VAR models require estimation of large number of parameters with few observations. In order to avoid overparametrization, overfitting and dimensionality issues, Bayesian inference and suitable classes of prior distributions have been proposed.

In this paper, a novel Bayesian nonparametric hierarchical prior for multivariate time series is proposed, which allows shrinkage of the SUR/VAR coefficients to multiple locations using a Normal-Gamma distribution with location, scale and shape parameters unknown. In our sparse SUR/VAR (sSUR/sVAR), some SUR/VAR coefficients shrink to zero, due to the shrinking properties of the lasso-type distribution at the first stage of our hierarchical prior, thus improving efficiency of parameters estimation, prediction accuracy and interpretation of the temporal dependence structure in the time series. We use a Bayesian Lasso prior, which allows us to reformulate the SUR/VAR model as a penalized regression problem, in order to determine which SUR/VAR coefficients shrink to zero (see [10] and [7]).

As regards to the second stage of the hierarchy, we use a random mixture distribution of the Normal-Gamma hyperparameters, which allows for parameter parsimony through two components. The first component is a random Dirac point-mass distribution, which induces shrinkage for SUR coefficients; the second component is a Dirichlet process hyperprior, which allows for clustering of the SUR/VAR coefficients.

The structure of the paper is as follows. Section 2 introduces the vector autoregressive model. Section 3 describes briefly the Bayesian nonparametric sparse model. Section 4 presents some simulation results for different dimensions. Section 5 concludes.

## 2 The Vector Autoregressive model

Let  $\mathbf{y}_t = (\mathbf{y}'_{1,t}, \dots, \mathbf{y}'_{N,t})' \in \mathbb{R}^m$  be a vector-valued time series. We consider a VAR model of order  $p$  (VAR(p)) as

$$\mathbf{y}_t = \mathbf{b} + \sum_{i=1}^p B_i \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t, \quad (1)$$

for  $t = 1, \dots, T$ , where  $\mathbf{y}_t = (y_{1,t}, \dots, y_{m,t})'$ ,  $\mathbf{b} = (b_1, \dots, b_m)'$  and  $B_i$  is a  $(m \times m)$  matrix of coefficients. We assume that  $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \dots, \varepsilon_{m,t})'$  follows a independent and

identically distributed Gaussian distribution  $\mathcal{N}_m(\mathbf{0}, \Sigma)$  with mean  $\mathbf{0}$  and covariance matrix  $\Sigma$ .

The VAR( $p$ ) in 1 can be rewritten in a stacked regression form:

$$\mathbf{y}_t = (I_m \otimes \mathbf{x}_t') \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t, \quad (2)$$

where  $\mathbf{x}_t = (1, y_{t-1}', \dots, y_{t-p}')'$  is the vector of predetermined variables,  $\boldsymbol{\beta} = \text{vec}(B)$ , where  $B = (\mathbf{b}, B_1, \dots, B_p)$ ,  $\otimes$  is the Kronecker product and  $\text{vec}$  the column-wise vectorization operator that stacks the columns of a matrix in a column vector.

### 3 Bayesian nonparametric sparse VAR

In this paper we define a hierarchical prior distribution which induces sparsity on the vector of coefficients  $\boldsymbol{\beta}$ . In order to regularize (2) we incorporate a penalty using a lasso prior  $f(\boldsymbol{\beta}) = \prod_{j=1}^r \mathcal{NG}(\beta_j | 0, \gamma, \tau)$ , where  $\mathcal{NG}(\beta | \mu, \gamma, \tau)$  denotes the normal-gamma distribution with location parameter  $\mu$ , shape parameter  $\gamma > 0$  and scale parameter  $\tau > 0$ . The normal-gamma distribution induces shrinkage toward the prior mean of  $\mu$ , but we can extend the lasso model specification by introducing a mixture prior with separate location parameter  $\mu_j^*$ , separate shape parameter  $\gamma_j^*$  and separate scale parameter  $\tau_j^*$  such that:  $f(\boldsymbol{\beta}) = \prod_{j=1}^r \mathcal{NG}(\beta_j | \mu_j^*, \gamma_j^*, \tau_j^*)$ . In our paper, we favor the sparsity of the parameters through the use of carefully tailored hyperprior and we use a nonparametric Dirichlet process prior (DPP), which reduces the overfitting problem and the curse of dimensionality by allowing for parameters clustering due to the concentration parameter and the base measure choice.

In our case we define  $\boldsymbol{\theta}^* = (\boldsymbol{\mu}^*, \boldsymbol{\gamma}^*, \boldsymbol{\tau}^*)$  as the parameters of the Normal-Gamma distribution, and assume a prior  $\mathbb{Q}_l$  for  $\boldsymbol{\theta}_{ij}^*$ , that is

$$\beta_j \stackrel{\text{ind}}{\sim} \mathcal{NG}(\beta_j | \mu_j^*, \gamma_j^*, \tau_j^*), \quad (3)$$

$$\boldsymbol{\theta}_{ij}^* | \mathbb{Q}_l \stackrel{\text{i.i.d.}}{\sim} \mathbb{Q}_l, \quad (4)$$

for  $j = 1, \dots, r_l$  and  $l = 1, \dots, N$ .

Following a construction of the hierarchical prior similar to the one proposed in [4] we define the vector of random measures

$$\begin{aligned} \mathbb{Q}_1(d\boldsymbol{\theta}_1) &= \pi_1 \mathbb{P}_0(d\boldsymbol{\theta}_1) + (1 - \pi_1) \mathbb{P}_1(d\boldsymbol{\theta}_1), \\ &\vdots \\ \mathbb{Q}_N(d\boldsymbol{\theta}_N) &= \pi_N \mathbb{P}_0(d\boldsymbol{\theta}_N) + (1 - \pi_N) \mathbb{P}_N(d\boldsymbol{\theta}_N), \end{aligned} \quad (5)$$

with the same sparse component  $\mathbb{P}_0$  in each equation and with the following hierarchical construction as previously explained,

$$\mathbb{P}_0(d\boldsymbol{\theta}) \sim \delta_{\{(0, \gamma_0, \tau_0)\}}(d(\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\tau})),$$

$$\begin{aligned}
\mathbb{P}_l(d\boldsymbol{\theta}) &\stackrel{i.i.d.}{\sim} \text{DP}(\tilde{\alpha}, G_0), \quad l = 1, \dots, N, \\
\boldsymbol{\pi}_l &\stackrel{i.i.d.}{\sim} \mathcal{B}e(\boldsymbol{\pi}_l | 1, \boldsymbol{\alpha}_l), \quad l = 1, \dots, N, \\
(\boldsymbol{\gamma}_0, \boldsymbol{\tau}_0) &\sim g(\boldsymbol{\gamma}_0, \boldsymbol{\tau}_0 | \nu_0, p_0, s_0, n_0), \\
G_0 &\sim \mathcal{N}(\boldsymbol{\mu} | c, d) \times g(\boldsymbol{\gamma}, \boldsymbol{\tau} | \nu_1, p_1, s_1, n_1)
\end{aligned} \tag{6}$$

where  $\delta_{\{\boldsymbol{\psi}_0\}}(\boldsymbol{\Psi})$  denotes the Dirac measure indicating that the random vector  $\boldsymbol{\Psi}$  has a degenerate distribution with mass at the location  $\boldsymbol{\Psi}_0$ , and  $g(\boldsymbol{\gamma}_0, \boldsymbol{\tau}_0)$  is the conjugate joint prior distribution (see [6]). We apply the Gibbs sampler and the hyperparameters given in [2] for the posterior approximation.

## 4 Simulation Results

The nonparametric prior presented in Section 3 allows for shrinking the SUR coefficients. In order to assess the goodness of the prior we performed a simulation study of our Bayesian nonparametric sparse model. We consider different datasets with sample size  $T = 100$  from the VAR model of order 1:

$$\mathbf{y}_t = B\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \stackrel{i.i.d.}{\sim} \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma}) \quad t = 1, \dots, 100,$$

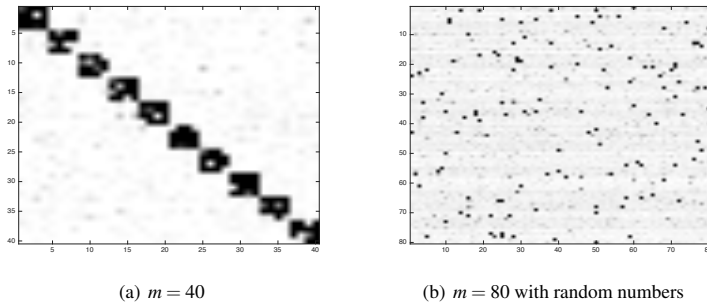
where the dimension of  $\mathbf{y}_t$  and of the square matrix of coefficients  $B$  can take different values:  $m = 20$  (small dimension),  $m = 40$  (medium dimension) and  $m = 80$  (large dimension). Furthermore, we choose different settings of the matrix  $B$ , focusing on a block-diagonal structure with random entries of the blocks:

- the block-diagonal matrix  $B = \text{diag}\{B_1, \dots, B_{m/4}\} \in \mathcal{M}_{(m,m)}$  is generated with blocks  $B_j$  ( $j = 1, \dots, m/4$ ) of  $(4 \times 4)$  matrices on the main diagonal:

$$B_j = \begin{pmatrix} b_{11,j} & \dots & b_{14,j} \\ \vdots & \vdots & \vdots \\ b_{41,j} & \dots & b_{44,j} \end{pmatrix},$$

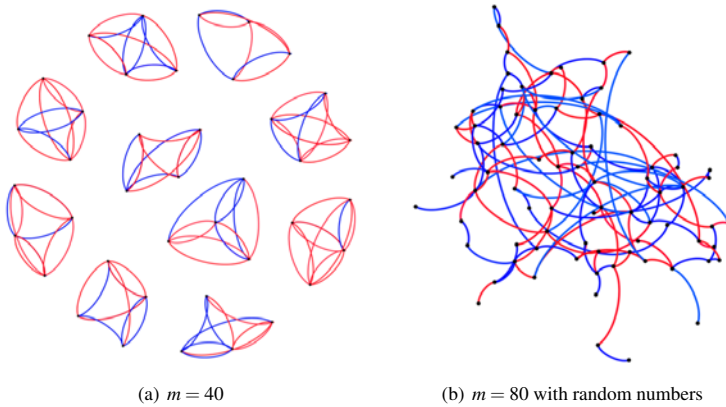
- where the elements are randomly taken from an uniform distribution  $\mathcal{U}(-1.4, 1.4)$  and then checked for the weak stationarity condition of the VAR;
- the random matrix  $B$  is a  $(80 \times 80)$  matrix with 150 elements randomly chosen from an uniform distribution  $\mathcal{U}(-1.4, 1.4)$  and then checked for the weak stationarity condition of the VAR.

Figure 1 exhibits the posterior mean of  $\Delta$ , which shows us the allocation of the coefficients between the two random measures  $\mathbb{P}_0$  and  $\mathbb{P}_l$ . In particular, we have that the white color indicates if the coefficient  $\delta_{ij}$  is equal to zero (i.e. sparse component), while the black one if the  $\delta_{ij}$  is equal to one, for nonsparse components. The definition of the pairwise posterior probabilities and of the co-clustering matrix for the atom locations  $\boldsymbol{\mu}$  allows us to build the weighted networks (see Figure 2),



**Fig. 1** Posterior mean of the matrix of  $\delta$  for  $m = 40$  (left) and for  $m = 80$  (right) with random element.

where the blue edges represent negative weights, while the red ones represent the positive weights. In each coloured graph the nodes represent the  $n$  variables of the VAR model, and a clockwise-oriented edge between two nodes  $i$  and  $j$  represents a non-null coefficient for the variable  $y_{j,t-1}$  in the  $i$ -th equation of the VAR.



**Fig. 2** Weighted network for  $m = 40$  (left) and for  $m = 80$  (right) with random elements, where the blue edges means negative weights and red ones represent positive weights.

## 5 Conclusions

This paper proposes a novel Bayesian nonparametric prior for SUR models, which allows for shrinking SUR coefficients toward multiple locations and for identifying groups of coefficients. We introduce a two-stage hierarchical distribution, which consists in a hierarchical Dirichlet process on the parameters of the Normal-Gamma distribution. The proposed hierarchical prior is used to propose a Bayesian nonparametric model for SUR models. We provide an efficient Monte Carlo Markov Chain algorithm for the posterior computations and the effectiveness of this algorithm is assessed in simulation exercises. The simulation studies illustrate the good performance of our model with different sample sizes for  $B$  and  $\mathbf{y}_t$ .

## References

1. M. Banbura, D. Giannone, and L. Reichlin. Large Bayesian vector autoregressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010.
2. M. Billio, R. Casarin, and L. Rossini. Bayesian Nonparametric sparse Seemingly unrelated regression model. <https://arxiv.org/abs/1608.02740>, 2017.
3. A. Carriero, T.E. Clark, and M. Marcellino. Bayesian VARs: Specification choices and forecast accuracy. *Journal of Applied Econometrics*, 30(1):46–73, 2015.
4. S. J. Hatjispyros, T. N. Nicolieris, and S. G. Walker. Dependent mixtures of Dirichlet processes. *Computational Statistics & Data Analysis*, 55(6):2011–2025, 2011.
5. G. Koop. Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, 28(2):177–203, 2013.
6. R. Miller. Bayesian analysis of the two-parameter gamma distribution. *Technometrics*, 22(1), 1980.
7. T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, June 2008.
8. S. L. Scott and H. R. Varian. Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2), 2013.
9. J. H. Stock and M. W. Watson. Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493, 2012.
10. R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.
11. A. Zellner. An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. *Journal of the American Statistical Association*, 57(298):500–509, 1962.
12. A. Zellner. *An introduction to Bayesian inference in econometrics*. New York Wiley, 1971.

# Using GPS Data to Understand Urban Mobility Patterns: An Application to the Florence Metropolitan Area

## *Analizzare i comportamenti di mobilità urbana attraverso i dati GPS: un'applicazione all'area metropolitana fiorentina*

Chiara Bocci, Daniele Fadda, Lorenzo Gabrielli, Mirco Nanni Leonardo Piccini

**Abstract** Big Data, originating from the digital breadcrumbs of human activities, let us observe the individual and collective behaviour of people at an unprecedented detail. In this paper we investigate the informative potential of the digital tracking that GPS-enabled devices can offer to academic research and to policy makers, with a specific attention for urban and metropolitan settings. The unstructured nature of the dataset requires a careful consideration and correction of possible biases which could lead to unreliable results. We use the 2011 census commuting matrix as a validation tool for our proposed methodology. GPS data contain information that would not be otherwise available, i.e. non-systematic mobility patterns. The produced estimates are then used to analyse mobility patterns within the Florence Metropolitan Area in a more exhaustive and detailed form.

**Abstract** *L'evoluzione tecnologica ha portato, nel corso degli ultimi anni, ad un notevole incremento dei dispositivi in grado di produrre e memorizzare tracce digitali dei nostri comportamenti quotidiani. In questo lavoro vogliamo indagare il potenziale informativo contenuto nelle tracce prodotte da apparecchi dotati di sistemi GPS per scopi di ricerca o di pianificazione delle politiche, con particolare riferimento all'ambito urbano e metropolitano. La natura spontanea e non strutturata dei dati richiede un'attenzione particolare alle possibili fonti di distorsione. Utilizziamo la matrice di pendolarismo del censimento 2011 come strumento di validazione dei risultati. I dati GPS contengono informazioni altrimenti di difficile reperibilità, come i comportamenti di mobilità non-sistematica. Le stime ottenute sono utilizzate per un caso di studio incentrato sull'Area Metropolitana Fiorentina.*

**Key words:** big data, mobility, urban planning, O-D matrix validation

---

Chiara Bocci, Leonardo Piccini  
IRPET, Firenze, Italy e-mail: name.surname@irpet.it

Daniele Fadda, Lorenzo Gabrielli, Mirco Nanni  
KDDLAB ISTI CNR, Pisa, Italy e-mail: name.surname@isti.cnr.it

## 1 Introduction

Technological evolution brought along, in recent years, a remarkable increase in the diffusion of devices that can record digital footprints of our behaviour on a daily basis, tracking a vast degree of activities. Constant and basically unintentional production of such tracks generates huge datasets that contain a precious quantum of information about socio-economic behaviour that may be extracted and used for socio-economic research and for policy analysis [1].

Big data sources may support policy makers in the ex-ante phase of policy implementation, by providing a more sophisticated depiction of the socio-economic environment and may be used for ex-post evaluation purposes in quasi-experimental design and counterfactual settings.

Literature on the matter and practical experiences have highlighted pros and cons of this approach [5]. Some of the pros include timeliness, cost effectiveness, spatial and temporal disaggregation, emergence of unexpected and/or unobservable phenomena. On the other hand, since the relative novelty of the methodologies used to deal with these data, extra carefulness needs to be used to acknowledge possible shortcomings in terms of quality, accessibility, applicability, relevance, privacy policy and ownership of the data, all of which may affect the quality of policy evaluation and appraisal. Nonetheless, we believe that big data sources can be successfully used to foster the capabilities of the public institutions to deal with complex problems, to plan effective policies and to evaluate the outcomes of their actions. To this extent, we propose a methodology that allows us to use data collected from GPS-enabled devices, installed on private vehicles for insurance purposes, to analyse and understand mobility patterns within a urban setting.

## 2 Research statement, objectives and data sources

The aim of the paper is to find a viable method to use GPS data to produce a non-biased Origin-Destination matrix for the selected study area, i.e. the Florence Metropolitan Area. Since the GPS dataset is derived from private car mobility, our focus will be mainly on this type of flows. However, since we want to use our estimates to assess the intensity and characteristics of the relations between different geographic zones within the metropolitan area, we need to find a way to correct for the different propension on public transport usage which we expect to observe across the different Origin-Destination pairs.

Typically, this kind of data is collected systematically every 10 years, during the nationwide official census. However, census data, while very rich with information and details, has two major drawbacks: the temporal lag between census, during which we have no information on mobility, and the focus on what we call systematic mobility, i.e. the mobility which happens almost every day and is mainly related to home-to-school or home-to-work trips, leaving out an increasingly relevant segment of non-systematic mobility, which, by its own nature, is difficult to capture with tra-



ditional methods. If our methodology is correct, we can thus increase our analytical capability with an informative base that can be updated almost continuously and that includes all mobility and not only the systematic one.

For this study we use GPS data that are provided by a leader company in the Insurance Telematics that deals with about the 2% of the total vehicles circulating in Italy. Our dataset counts about 150k private vehicles crossing Tuscany in the month of June 2011, and represents a primary source of information for studying the mobility behaviours. Data on vehicle fleet in Italy provided by the Italian Automobile Club (ACI), Census data provided by ISTAT, and trip duration and distances with different transportation means computed using Google services are used to re-scale the vehicle sample to the real mobility flows. Once we validate our data and estimate a reliable O-D matrix, we can use the data to carry out an extensive descriptive analysis of mobility patterns in our selected geographic area. To demonstrate the informative potential of this kind of data, we choose the Florence Metropolitan Area as a case study.

### 3 Estimating a detailed O-D Matrix using GPS data

As we previously discussed, GPS data contain an inherent bias: they account only for private cars usage (specifically, for the fraction of vehicles that have a GPS device installed for insurance purposes and that are being monitored by our provider).

Since we want to use GPS data for socio-economic analysis and for policy planning and evaluation, we need to find a way to scale back the flows that we observe towards our real population, which means accounting for (at least) three missing dimensions:

1. We observe vehicles, but we want to estimate the number of people actually travelling, which means accounting for average car occupation;
2. We observe a fraction of vehicles that is geographically heterogeneous, so we want to account for different market penetration by our provider;
3. We observe only private cars, so we want to account for an heterogeneous share of public transport users.

In order to estimate a complete O-D Matrix, we use the 2011 Census Origin-Destination Matrix as a validation tool. Such matrices are usually released with a territorial detail that corresponds to the administrative units of municipalities. These matrices contain information on municipality of origin, municipality of destination, time of departure, duration of the trip, mean of transport, gender and purpose of the trip (work- or school-related). A geographically more detailed matrix is also released by ISTAT, with a disaggregation to the census zones, but with less information on the characteristics of the trip (only the purpose). Since we want to be able to estimate an O-D Matrix to analyse urban areas, we want a sub-municipality disaggregation for larger municipalities. We therefore use the more detailed matrix for our validation. Since we also need at least the mean of transportation for our validation

we split our flows using the share of public transportation registered between the corresponding municipalities.

Our starting dataset is comprised of systematic trips observed over the month of June 2011 and aggregated by the 2011 Census zone partitions. If we hypothesise our data to be a random sample extracted from the population of all car movements happening within Tuscany borders during our time frame, we can estimate our target values (a census zone O-D matrix of people using all available means of transportations) with the following formula

$$XFlow_{i,j} = flow_{i,j} * car.pen_i * avg.occ_i * public.t.ind_{i,j}$$

where  $XFlow$  is our desired estimate from zone  $i$  to zone  $j$ ,  $flow$  is our observed flow from zone  $i$  to zone  $j$ ,  $car.pen$  is the market penetration of our data provider for the municipality within which zone  $i$  falls,  $avg.occ$  is the average occupancy rate for systematic mobility departing in municipality within which zone  $i$  falls (derived from census data),  $public.t.ind$  is a public transport accessibility index calculated between zone  $i$  and zone  $j$  (calculated using google services).

#### 4 Validating GPS data using the Census O-D Matrix

Once we estimate our O-D Matrix, we want to check how our estimates perform against our reference values, i.e. the 2011 census O-D Matrix. Literature on matrix comparison has produced different indicators that asses how similar two matrices are (see [2] to an detailed presentation and discussion of these indicators). Moreover, one recent thread of research has been trying to evaluate the similarity of O-D matrices by using image quality assessment techniques mutuated from image processing methodologies [6, 7]. We test the performance of our estimated matrix applying different measures: some classical statistical indicators (like the  $R^2$  association measure, the Root Mean Square Error ( $RMSE$ ) and the Pearson  $\chi^2$  test), the Geoffrey E. Havers ( $GEH$ ) statistic which evaluate the level of closeness of each pair of cell of the two matrices, and the recently proposed (and still under study) Mean Structural Similarity Index ( $MSSIM$ ) by Van vuren and Day-pollard [6], which compare two O-D matrices considering the means, variances and covariance of contiguous matrix cells evaluated within a moving block of cells in each matrix.

#### 5 Using the estimated Matrix for socio-economic analysis

Once we have validated the estimation, we can use our matrix to produce a variety of indicators that can help policy makers understand the connection and mobility patterns which operate within their territories. The case study area that we selected is the Florence Metropolitan Area.

### 5.1 Filling the gaps and assessing mobility patterns

We can use our methodology to estimate an O-D Matrix for subsequent years and compare the results as a time series. Moreover, since we can unpack our matrix in a spatially detailed manner, we can assess mobility patterns within the municipality boundaries. As an example, in Figure 1 we can determine the average speed of the observed trajectories as it varies hourly within the day and for different partition of the city (in this case, the 5 administrative neighbourhoods of the city of Florence).

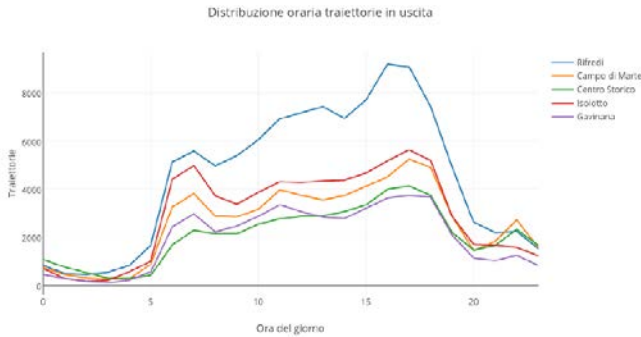


Fig. 1 Average speed by hour and zone of departure

### 5.2 The boundaries of the city

Generally the border of the city are measured looking at just census data i.e. population density in absolute terms, or the variation over time [4, 3]. We propose a clustering approach aimed at partitioning territories on the basis of human movements inferred using Big Data.

The aim of our work is to contribute to this debate, by providing a tool for policy makers to build a novel definition of regions, seen as functional areas. Focused on the Metropolitan Area of Florence, we aggregate territories that maximise internal traffic and minimise external one.

Given two generic nodes  $a$  and  $b$ , we define internal traffic, the sum of the flows from  $a$  to  $b$  and vice versa. For each pair  $a, b$  we calculate the distance matrix as the percentage of internal flow respect to the total flows. The clustering methods seeking the best partitioning minimises the distances contained in the matrix provided as input, so each pair of the distance matrix is calculated as  $d = 1 - \%internal\ flows$ .

We provide the distance matrix as input to DBSCAN and we evaluate the possible values of epsilon and we extract the territorial partition shown in Figure 2.

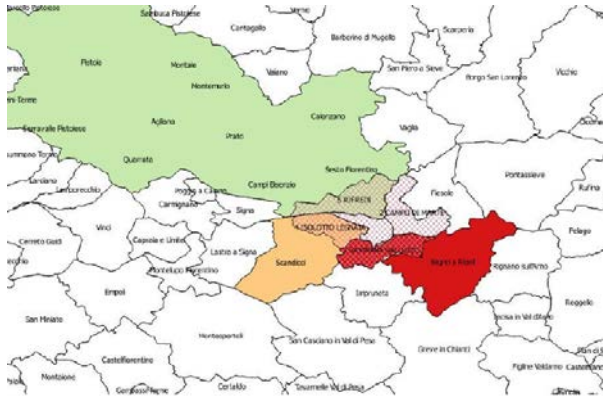


Fig. 2 Boundaries of Florence Metropolitan Area using GPS data

## 6 Conclusions and future research

The proposed methodology allows us to reliably use GPS data for urban mobility behaviour analysis, without relying on the snapshot provided every ten years by the national census. The informative potential of this source is very high and flexible. Future lines of research include expanding the methodology to validate non-systematic data and further validation using GSM data (call records from mobile phones usage).

## References

1. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., Trasarti, R.: Unveiling the complexity of human mobility by querying and mining massive trajectory data. *VLDB J* **20**(5), 695–719 (2011)
2. Hollander, Y., Liu, R.: The principles of calibrating traffic microsimulation models. *Transportation* **35**(3), 347–362 (2008)
3. ISTAT: La nuova geografia dei sistemi locali. ISTAT (2015)
4. OECD: Redefining Urban: a new way to measure metropolitan areas. OECD (2012)
5. Scannapieco, M., Virgillito, A., Zardetto, D.: Placing Big Data in official statistics: a big challenge. In: *NTTS 2013 Proceedings* (2013)
6. Van Vuren, T., Day-Pollard, T.: 256 shades of grey - comparing OD matrices using image quality assessment techniques. In: *2015 Scottish Transport Applications Research Conference Proceedings* (2015). URL <http://www.starconference.org.uk/star2015.html>
7. Zhou, W., Bovik, A., Sheikh, H.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* **13**(4), 600–612 (2004)

# Relative privacy risks and learning from anonymized data

## *Privacy e learning in dati anonimizzati*

Michele Boreale and Fabio Corradi

**Abstract** We consider group-based anonymized tables, a popular approach to data publishing. This approach aims at protecting privacy of the involved individuals, by releasing an *obfuscated* version of the original data, where the exact correspondence between individuals and attribute values is hidden. When publishing data about individuals, one must typically balance the *learner's* utility against the risk posed by an *attacker*, potentially targeting individuals in the dataset. Accordingly, we propose a MCMC based methodology to learn the population parameters from a given anonymized table and to analyze the risk for any individual in the dataset to be linked to a specific sensitive value when the attacker has got to know the individual's nonsensitive attributes. We call this *relative risk* analysis. Finally, we illustrate results obtained by the proposed methodology on a real dataset.

**Abstract** *Nel lavoro consideriamo tabelle anonimizzate realizzate per rendere disponibili informazioni sulla popolazione, nascondendo però l'attribuzione dei dati sensibili ai singoli rispondenti. Si valuta l'informazione sulla popolazione che rimane disponibile e il rischio di violare la privacy dei rispondenti, fornendo diverse forme di apprendimento e di valutazione. Vengono riportati i risultati di un esperimento condotto su un dataset reale.*

**Key words:** Privacy, anonymization, k-anonymity, MCMC methods.

---

Michele Boreale  
Università di Firenze, Dipartimento di Statistica, Informatica, Applicazioni. e-mail: michele.boreale@unifi.it

Fabio Corradi  
Università di Firenze, Dipartimento di Statistica, Informatica, Applicazioni. e-mail: corradi@disia.unifi.it

**Table 1** A table (left) anonymized according to Local recoding (center) and Anatomy (right).

ID	Nat.	ZIP	Dis.	ID	Nat.	ZIP	Dis.	GID	Nat.	ZIP	Dis.
1	Malaysia	45501	Heart	1	{M,J}	4550*	Heart	1	Japan	45502	Heart
2	Japan	45502	Flu	2	{M,J}	4550*	Flu	1	Malaysia	45501	Flu
3	Japan	55503	Flu	3	Japan	5550*	Flu	2	Japan	55504	Flu
4	Japan	55504	Stomach	4	Japan	5550*	Stomach	2	Japan	55503	Stomach
5	China	66601	HIV	5	{C,J}	66601	HIV	3	Japan	66601	HIV
6	Japan	66601	Diabetes	6	{C,J}	66601	Diabetes	3	China	66601	Diabetes
7	India	77701	Flu	7	{I,M}	77701	Flu	4	Malaysia	77701	Flu
8	Malaysia	77701	Heart	8	{I,M}	77701	Heart	4	India	77701	Heart

Original table                      Local recoding                      Anatomy

## 1 Introduction

It is a common practice to release datasets involving individuals in some anonymized form. The goal is to enable the computation of population characteristics with reasonable accuracy, at the same time preventing leakage of sensitive information about individuals in the dataset. We are interested in *group-based* techniques, put forward in Computer Science in the last 15 years or so: *k*-anonymity [5] and its variants, like  $\ell$ -diversity [2], and Anatomy [8]. Despite their weakness against attackers with strong background knowledge, these techniques are a common choice when it comes to table publishing [3]. In group-based methods, the anonymized or obfuscated version of a table is obtained by partitioning records in groups enjoying certain properties (see Section 2). Generally speaking, even knowing that an individual belongs to a group of the anonymized table, it will not be possible for an attacker to link an individual to a specific sensitive value in the group. Two examples of group based anonymization are in Table 1, adapted from [7]. The original table collects medical data from eight individuals; here *Disease* is considered as the only sensitive attribute. The central table is a 2-anonymous table, obtained by *local recoding*: within each group, the nonsensitive attributes are generalized so as to make them indistinguishable. This is an example of *horizontal* scheme. Generally speaking, each group in a *k*-anonymous table consists of *at least k* records, which are indistinguishable as far as the nonsensitive part is concerned. Finally there is an example of application of *Anatomy*: within each group, the nonsensitive part of the rows are *vertically* randomly permuted, thus breaking the link between sensitive and nonsensitive values.

We put forward a probabilistic model to reason about the *relative risk* posed by the release of anonymized datasets (Section 2), i.e. the leakage of sensitive information for an individual in the table, *beyond* what is implied for the general population. To see what is at stake here, consider the central table of Fig. 1. An adversary may reason that, with the exception of the first group, a Japanese is never connected to Heart Disease. This hint can become a strong evidence in a larger, real-world table. Suppose now that the attacker’s target, a Malaysian living at ZIP code 4550\*, is known to belong to the table, so he must be in the first group. On the basis of the

evidence about Japanese not suffering from Heart Disease, the attacker can then link with high probability his target to Heart Disease. Here, the attacker combines knowledge learned from the anonymized table and about his victim with the group structure of the table itself. To formally reason about this phenomenon, we will define the *relative* privacy risks by comparing two conditional probability distributions, encoding respectively: what can be learned about the population from the anonymized table; and what can be learned about a the victim, given knowledge of her/his non-sensitive attributes *and* presence in the table (Sections 3). Generalizing Kifer [1] and Wong et al. [7], we propose a MCMC to learn both the parameter's population and the attacker's probability distribution from the anonymized data (Section 4). We finally illustrate the results of an experiment on a real-world dataset (Section 5).

## 2 Group based anonymization schemes and the probabilistic model

Given a dataset of  $N$  individuals, let  $\mathcal{R}$  and  $\mathcal{S}$ , ranged over by  $r$  and  $s$ , be finite nonempty sets of *nonsensitive* and *sensitive* values. A *row* is a pair  $(s, r) \in \mathcal{S} \times \mathcal{R}$ .

In a group based scheme a cleartext *table* is an arrangement of a multiset of  $N$  rows, say  $d = (s_1, r_1), \dots, (s_N, r_N)$ , into a sequence of *groups*,  $t = g_1, \dots, g_k$ , where each group is a sequence  $g_j = (s_{j_1}, r_{j_1}), \dots, (s_{j_{n_j}}, r_{j_{n_j}})$ . Given a generic group  $g$ , its *obfuscation* is a pair  $g^* = (l, m)$ , where  $m = s_1, s_2, \dots$  is the sequence of sensitive values occurring in  $g$ , and  $l$ , called *generalized nonsensitive value*, is:

- a *superset* of  $g$ 's nonsensitive values for *horizontal* schemes (e.g. k-anonymity);
- the *multiset* of  $g$ 's nonsensitive values  $\{|r_1, r_2, \dots|\}$ , for *vertical* schemes.

Given a table  $t = g_1, \dots, g_k$ , an obfuscated table is a  $t^* = g_1^*, \dots, g_k^*$ , such that each  $g_j^*$  is an obfuscation of the corresponding group  $g_j$ . An *anonymization algorithm*  $\mathcal{A}$  is a – possibly probabilistic – mechanism that maps collections of  $N$  rows,  $d$ , into obfuscated tables,  $t^*$ .

Our model consists of the following random variables with the associated meaning.

- $\Pi$ , taking values in the set of full support probability distributions  $\mathcal{D}$  over  $\mathcal{S} \times \mathcal{R}$ : the (unknown) joint probability distribution of the population.
- $T = G_1, \dots, G_k$ , taking values in the set of tables. Each group  $G_j$  is in turn a sequence of  $n_j$  consecutive rows in  $T$ ,  $G_j = (S_{j_1}, R_{j_1}), \dots, (S_{j_1+n_j}, R_{j_1+n_j})$ ; the number  $k$  of groups is not fixed, but depends itself on the rows  $S_j, R_j$ ;
- $T^* = G_1^*, \dots, G_k^*$ , taking values in the set of obfuscated tables.

We assume that the above three random variables form a Markov chain  $\Pi \longrightarrow T \longrightarrow T^*$ . In other words, the joint probability density  $f$  of these variables can be factorized as:

$$f(\pi, t, t^*) = f(\pi)f(t|\pi)f(t^*|t). \quad (1)$$

We also assume the following.

- $\pi \in \mathcal{D}$  is encoded as a pair of  $(\pi_S, \pi_{R|S})$  such that  $f(s, r|\pi) \propto f(s|\pi_S)f(r|\pi_{R|s})$ . Here, each  $\pi_S$  is a distribution over  $\mathcal{S}$ , and each  $\pi_{R|S}$  is viewed as a collection of distributions over  $\mathcal{R}$ ,  $\pi_{R|S} = (\pi_{R|s})_{s \in \mathcal{S}}$ . We posit that the  $\pi_S$  and the  $\pi_{R|s}$ s are chosen independently, according to Dirichlet distributions of hyperparameters  $\alpha = (\alpha_1, \dots, \alpha_{|\mathcal{S}|})$  and  $\beta^s = (\beta_1^s, \dots, \beta_{|\mathcal{R}|}^s)$ , respectively. In other words

$$f(\pi) = \text{Dir}(\pi_S | \alpha) \cdot \prod_{s \in \mathcal{S}} \text{Dir}(\pi_{R|s} | \beta^s). \quad (2)$$

- The  $N$  individual rows composing the table  $t$ ,  $(s_1, r_1), \dots, (s_N, r_N)$  are assumed to be drawn i.i.d. conditionally to  $\Pi$ . This amounts to positing that:

$$f(t|\pi) \propto f(s_1, r_1|\pi) \cdots f(s_N, r_N|\pi). \quad (3)$$

### 3 The honest learner, the attacker and measures of relative risk

A *honest learner* is someone who, after observing  $T^* = t^*$ , updates his knowledge on the population parameters  $\pi$ . In addition an *attacker* also knows the nonsensitive value  $r_v$  of a victim in  $T$ . In what follows we shall fix once and for all  $t^*$  and  $r_v$  such that  $f(r_v, t^*) \triangleq f(r_v \text{ occurs in } T, T^* = t^*) > 0$ . Let  $p_L(s, r)$  be the joint probability distribution on the population that can be learned given from  $t^*$ . Formally, for each  $(s, r)$

$$p_L(s, r|t^*) \triangleq E_{\pi \sim f(\pi|t^*)}[f(s, r|\pi)] = \int_{\pi \in \mathcal{D}} f(s, r|\pi) f(\pi|t^*) d\pi. \quad (4)$$

Of course, we can condition  $p_L$  on any given  $r$  so also the victim's nonsensitive attribute  $r_v$  and obtain the corresponding distribution on  $\mathcal{S}$ .

$$p_L(s|r_v, t^*) \triangleq E_{\pi \sim f(\pi|t^*)}[f(s|r_v, \pi)] = \int_{\pi \in \mathcal{D}} f(s|r_v, \pi) f(\pi|t^*) d\pi. \quad (5)$$

Given knowledge of  $r_v$  and knowledge that the victim is in  $T$ , we can define the attacker's distribution on  $\mathcal{S}$  as follows. Let us introduce a random variable  $V$ , identifying the victim as one of the individuals in  $T$ . In other words,  $V$  is an index, which we posit is a priori uniformly distributed on  $1..N$ , and independent from  $\Pi, T$ . Recalling that each row  $(S_j, R_j)$  is identified by a unique index  $j$ , we can define the attacker's probability distribution on  $\mathcal{S}$ , after seeing  $t^*$  and  $r_v$ , as:

$$p_A(s|r_v, t^*) \triangleq f(S_V = s | R_V = r_v, t^*). \quad (6)$$

Theorem 1 provides  $p_A(s|r_v, t^*)$  only based on the marginals  $R_j$  given  $t^*$ .

**Theorem 1.** *Let  $T = (S_j, R_j)_{j \in 1..N}$  and  $s_j$  the sensitive value in the row  $j$  of  $t^*$ . Then*



$$p_A(s|r_v, t^*) \propto \sum_{j:s_j=s} f(R_j = r_v | t^*). \quad (7)$$

We now define some measures of relative privacy risk to be put at work in Section 5.

**Definition 1 (risk measures).** Let  $p$  a full support distribution on  $\mathcal{S}$  and  $(s, r)$  a row in  $t$ . We say this row is *at risk under  $p$*  if  $p(s) = \max_{s'} p(s')$ , and that its *risk level under  $p$*  is  $p(s)$ . For an individual row  $(s, r)$  in  $t$ , which is at risk under  $p_A(\cdot|r, t^*)$ , its *relative risk level* is  $\mathbf{R}(s, r, t, t^*) \triangleq \frac{p_A(s|r, t^*)}{p_L(s|r, t^*)}$ . For  $\ell \in \{L, A\}$ , let us define (using the multiset notation  $\{\{ \dots \}\}$ )  $N_\ell(t, t^*) \triangleq |\{(s, r) \in t : (s, r) \text{ is at risk under } p_\ell(\cdot|r, t^*)\}|$ . The *global relative risk* of  $t$  given  $t^*$  is:  $\mathbf{GR}(t, t^*) \triangleq \max\left\{0, \frac{N_A(t, t^*) - N_L(t, t^*)}{N}\right\}$ .

## 4 Gibbs sampling

For real world datasets, none of the distributions (4), (5) or (7) will be computable analytically. Nonetheless, we can build accurate estimations of these distributions from samples of the marginals of the density  $f(\boldsymbol{\pi}, t | t^*)$ , with  $t = g_1, \dots, g_k$  (note that here the sensitive values  $s_j$  are actually fixed and known from  $t^*$ ). This can be done using a Gibbs sampler, provided we can effectively sample from the full conditionals of  $\boldsymbol{\pi}$  and  $g_j$ , for  $1 \leq j \leq N$ . This is discussed below.

The Gibb's chain state sequence  $(\boldsymbol{\pi}^i, t^i)$ ,  $i = 0, 1, \dots$ , is defined in the usual way, starting from an initial state  $x_0 = (\boldsymbol{\pi}^0, t^0)$  and sampling in turn  $\boldsymbol{\pi}^i$  and each of the groups of  $t^i = g_1^i, \dots, g_k^i$  separately, from the respective full conditionals. From equations (1), (2) and (3), it is easy to check that:

$$f(\boldsymbol{\pi} | t, t^*) = f(\boldsymbol{\pi} | t) \quad (8)$$

$$f(g_j | t_{-j}, \boldsymbol{\pi}, t^*) \propto f(g_j | \boldsymbol{\pi}) f(g_j^* | t) \quad (1 \leq j \leq k). \quad (9)$$

Each of the above two relations enables sampling from the corresponding full conditional on the left-hand side. Indeed, (8) is a posterior Dirichlet distribution, from which effective sampling can be easily performed. Denote by  $\boldsymbol{\gamma}(t) = (\gamma_1, \dots, \gamma_{|\mathcal{S}|})$  the vector of the frequency counts  $\gamma_i$  of each  $s_i$  in  $t$ . Similarly, given  $s$ , denote by  $\boldsymbol{\delta}^s(t) = (\delta_1^s, \dots, \delta_{|\mathcal{S}|}^s)$  the vector of the frequency counts  $\delta_i^s$  of the pairs  $(r_i, s)$ , for each  $r_i$ , in  $t$ . Then, for each  $\boldsymbol{\pi} = (\boldsymbol{\pi}_S, \boldsymbol{\pi}_{R|S})$ , we have

$$f(\boldsymbol{\pi} | t) = \text{Dir}(\boldsymbol{\pi}_S | \boldsymbol{\alpha} + \boldsymbol{\gamma}(t)) \cdot \prod_{s \in \mathcal{S}} \text{Dir}(\boldsymbol{\pi}_{R|s} | \boldsymbol{\beta}^s + \boldsymbol{\delta}^s(t)).$$

Let us discuss now (9). Here we will confine ourselves to the important case when the following conditions are satisfied: (a) the obfuscation function is deterministic, so that  $f(g_j^* | t)$  equals 0 or 1; (b) the set  $\mathcal{G}_j$  of the  $g_j$ 's such that  $f(g_j^* | g_j, t_{-j}) = 1$  depends solely on  $g_j^* = (l_j, m_j)$ , and is given by

$$\mathcal{G}_j = \begin{cases} \{g = (s_1, r_1), \dots, (s_n, r_n) : r_\ell \in l_j \text{ for } 1 \leq \ell \leq n\} & \text{(horizontal schemes)} \\ \{g = (s_1, r_{i_1}), \dots, (s_n, r_{i_n}) : \text{for } r_{i_1}, \dots, r_{i_n} \text{ a permutation of } m_j\} & \text{(vertical schemes)}. \end{cases} \quad (10)$$

This assumption is exact in many important cases (e.g. Anatomy) and reasonable in the remaining ones. Under assumptions (a), (b) and (10) above, sampling from (9) amounts to drawing an element  $g_j \in \mathcal{G}_j$  with probability  $\propto f(g_j|\pi)$ . This can be achieved via different techniques in each of the two cases of interest, horizontal and vertical; the details are omitted here due to lack of space.

## 5 Experiments

We have put a proof-of-concept implementation of our methodology at work on a subset of the Adult dataset from the UCI machine learning repository [6]. The considered subset consists of 5692 rows, with the following categorical attributes: *sex*, *race*, *marital status*, *education*, *native country*, *workclass*, *salary class*, *occupation*, with *occupation* considered as the only sensitive attribute. Using the ARX anonymization tool [3], we have obtained three different anonymized versions of the considered dataset, enjoying  $k$ -anonymity for, respectively:  $k = 4$ ,  $k = 5$  and  $k = 10$ . The average size of the groups varied from 38 rows (for  $k = 4$ ) to 355 rows (for  $k = 10$ ). We run the Gibbs sampler on each of these three anonymized datasets. We obtained the following figures for the global relative risks (cf. Def. 1) of the three datasets:  $\mathbf{GR}_1 = 3.98\%$ ,  $\mathbf{GR}_2 = 1.7\%$  and  $\mathbf{GR}_3 = 1.86\%$ . In absolute terms, the fraction of rows of  $t^*$  correctly classified by the attacker ranged from 27.3% to 29.4%. The *maximal* relative risk level  $\mathbf{R}$  ranged from about 1.9 to 3.93.

All in all, these results indicate that, in each case the considered anonymized datasets imply a significant relative privacy risk, for an appreciable fraction of the rows.

## References

1. D. Kifer. Attacks on privacy and deFinetti's theorem. *SIGMOD 2009 Conference*: 127-138, 2009.
2. A. Machanavajjhala, J., Gehrke, D., and Kifer.  $\ell$ -diversity: privacy beyond  $k$ -anonymity. In *ICDE'06*: 24, 2006.
3. F. Prasser, F. Kohlmayer. Putting Statistical Disclosure Control Into Practice: The ARX Data Anonymization Tool. In: Gkoulalas-Divanis, Aris, Loukides, Grigorios (Eds.): *Medical Data Privacy Handbook*, Springer, November 2015. ISBN: 978-3-319-23632-2.
4. C.P. Robert, G. Casella. *Monte Carlo Statistical Methods*. 2/e, Springer, 2004.
5. L. Sweeney.  $k$ -anonymity: a model for protecting privacy. *International journal on uncertainty, Fuzziness and knowledge based systems* 10(5), 557-570, 2002.
6. UCI Machine Learning repository, Adult dataset. <https://archive.ics.uci.edu/ml/datasets/Adult>, 1996
7. R. Chi-Wing Wong, A. Wai-Chee Fu, Ke Wang, Ph. S. Yu, J. Pei. Can the Utility of Anonymized Data be Used for Privacy Breaches? In *TKDD'11* 5(3): 16:1-16:24, 2011.
8. X. Xiao, and T. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB'06*: 139-150, 2006.

# A stochastic volatility framework with analytical filtering

Giacomo Bormetti, Roberto Casarin, Fulvio Corsi and Giulia Livieri

**Abstract** Motivated by the fact that realized measures of volatility are affected by measurement errors, we introduce a new family of discrete-time stochastic volatility models having two measurement equations relating both the observed returns and realized measures to the latent conditional variance.

**Key words:** Bayesian Inference, Monte Carlo Markov Chain, High-frequency, Realized volatility, ARG, Stochastic volatility

## 1 Introduction

In this paper we introduce a new family of discrete-time Stochastic Volatility (SV) models, for the joint modelling of returns and realized measures of volatility. The proposed model is characterized by having two *measurement equations* for the latent volatility: (i) a Normal density for the daily returns and (ii) a Gamma density for the RV measure. We then term the general version of the proposed model as SV-ARG. A salient feature of the SV-ARG is that it allows for analytical filtering and smoothing recursions for the latent factor that guides the dynamics of daily returns. This permits us to develop an effective Bayesian inference procedure for both parameters and latent factor.

---

Giacomo Bormetti  
University of Bologna, e-mail: giacomo.bormetti@unibo.it

Roberto Casarin  
University Ca' Foscari of Venice, e-mail: r.casarin@unive.it

Fulvio Corsi  
University Ca' Foscari of Venice, e-mail: corsi@unive.it

Giulia Livieri  
Scuola Normal Superiore, Pisa, e-mail: giulia.livieri@sns.it

## 2 The model

Consider a financial log-return process  $r_t$ , a realized variance process  $y_t$  and a latent volatility process  $h_t$ . Let  $\mathcal{F}_t \doteq \sigma(r_t, y_t)$  be the  $\sigma$ -algebra containing the information about observable quantities (log-return and realized variance  $y_t$ ) available at time  $t$ , and  $\widetilde{\mathcal{F}}_t^H \doteq \sigma(\mathcal{F}_{t-1}, h_t)$ . We assume the following model for the dynamics of the log-returns:

$$r_t = \mu + \gamma h_t + \sqrt{h_t} \varepsilon_t, \quad \varepsilon_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1), \quad (1)$$

$t = 1, \dots, T$ , where  $\mu$  is the risk-free rate and  $\gamma$  is the market price of risk.  $\mathcal{N}(m, \sigma^2)$  indicates the univariate normal distribution with mean  $m$  and variance  $\sigma^2$ . The dynamics in Equation (1) differs from that employed in Corsi et al. (2013); Majewski et al. (2015) for daily log-returns inasmuch in these works authors consider as driving process for returns a realized measure of volatility. Specifically, they employ the continuous part of the realized variance, hereafter RV, defined as the sum of squared returns over non-overlapping intervals within a sampling period. We refer to Equation (1) as *return equation*.

Since the RV contains information on the latent volatility process, we follow authors in Hansen and Lunde (2006); Engle and Gallo (2006); Shephard and Shephard (2010); Takahashi et al. (2009) and introduce another measurement equation, termed *realized variance equation*, which relates the observed RV to the latent process  $h_t$ . Specifically, we assume that the realized variance  $y_t$  is sampled from a Gamma distribution

$$y_t | \widetilde{\mathcal{F}}_t^H \stackrel{i.d.}{\sim} \mathcal{G}(\alpha, h_t), \quad (2)$$

where  $\alpha \in \mathbb{R}_+$  is constant. In the previous equation,  $\mathcal{G}(k, \vartheta)$  denotes a Gamma distribution with positive shape,  $k$ , and scale parameter,  $\vartheta$ .

We assume that  $h_t$  follows an autoregressive gamma process with transition distribution (see Gouriéroux and Jasiak, 2006):

$$h_t | \widetilde{\mathcal{F}}_{t-1}^H, r_{t-1}, y_{t-1} \stackrel{d}{\sim} \mathcal{G}(\nu, \frac{\phi}{c} h_{t-1}, c). \quad (3)$$

In the previous equation,  $\mathcal{G}(\nu, \frac{\phi}{c} h_{t-1}, c)$  denotes the non-central gamma distribution with shape  $\nu > 0$ , scale  $c > 0$  and non-centrality  $\frac{\phi}{c} h_{t-1}$ . Using the Poisson mixture representation for the non-central gamma distribution (see Gouriéroux and Jasiak, 2006, for more details), we rewrite Equation (3) as

$$\begin{aligned} h_t | z_t &\stackrel{i.d.}{\sim} \mathcal{G}(\nu + z_t, c), \\ z_t | h_{t-1} &\stackrel{i.d.}{\sim} \mathcal{P}(\phi h_{t-1}), \end{aligned}$$

where, in general,  $\mathcal{P}(\nu)$  indicates the Poisson distribution with intensity parameter  $\nu$ . The latter representation is useful for both the characterization of  $h_t$  and the inference procedure. The stationarity conditions, the conditional moment generating

function of this process and its risk neutral dynamics are given in (Bormetti et al., 2016).

### 3 Analytical filtering and smoothing

Applying similar argument as in Creal (2015), we are able to provide analytical expressions for the: (i) conditional likelihood, (ii) Markov transition, (iii) initial distribution of  $z_t$ , (iv) filtering and the smoothing of the latent process  $h_t$ . In particular, the following two propositions hold.

**Proposition 1.** *For the SV-ARG model described in Equation (1), (2) and (3) the conditional likelihood,  $p(r_t, y_t | z_t, \theta)$ , the Markov transition,  $p(z_t | z_{t-1}, r_{t-1}, y_{t-1}, \theta)$ , and the initial distribution of  $z_t$ ,  $p(z_1; \theta)$ , are respectively given by:*

$$\begin{aligned} p(r_t, y_t | z_t; \theta) &= 2\eta(z_t, y_t; \theta) K_{\lambda(z_t)} \left( \sqrt{\psi \chi^{(t)}} \right) \left( \sqrt{\frac{\chi^{(t)}}{\psi}} \right)^{\lambda(z_t)}, \\ p(z_t | z_{t-1}, r_{t-1}, y_{t-1}; \theta) &\propto \mathcal{S} \left( \lambda(z_{t-1}), \chi^{(t-1)} \frac{\phi^{(d)}}{c}, \psi \frac{c}{\phi^{(d)}} \right), \\ p(z_1; \theta) &\propto \mathcal{N} \mathcal{B} \left( \mathbf{v}, \phi^{(d)} \right), \end{aligned}$$

with

$$\begin{aligned} \eta(z_t, y_t; \theta) &= \frac{\exp(\gamma \mu_{1t}) y_t^{\alpha_t - 1}}{\sqrt{2\pi}} \frac{1}{\Gamma(\alpha_t) \Gamma(v + z_t) c^{v+z_t}}, \\ \mu_{1t} &= r_t - \mu, \\ \alpha_t &= \alpha, \\ \lambda(z_t) &= v + z_t - \alpha_t - 1/2, \\ \chi^{(t)} &= \mu_{1t}^2 + 2\mu_{2t}, \\ \mu_{2t} &= y_t, \\ \psi &= \gamma^2 + \frac{2}{c}. \end{aligned}$$

*Proof.* See Bormetti et al. (2016).

**Proposition 2.** *Let  $\lambda(z_t)$ ,  $\chi^{(t)}$  and  $\psi$  be the quantities defined in Proposition 1. The marginal filtered,  $p(h_t | \mathbf{r}_{1:t}, \mathbf{y}_{1:t}, \mathbf{z}_{1:t}, \mathbf{x}_{1:t}; \theta)$ , and smoothed,  $p(h_t | \mathbf{r}_{1:T}, \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \mathbf{x}_{1:T}; \theta)$  distributions are*

$$\begin{aligned} p(h_t | \mathbf{r}_{1:t}, \mathbf{y}_{1:t}, \mathbf{z}_{1:t}, \mathbf{x}_{1:t}; \theta) &\propto \mathcal{G}ig \left( \lambda(z_t), \chi^{(t)}, \psi \right), \\ p(h_t | \mathbf{r}_{1:T}, \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \mathbf{x}_{1:T}; \theta) &\propto \mathcal{G}ig \left( \lambda(z_t) + z_{t+1}, \chi^{(t)}, \psi + 2 \frac{\phi^{(d)}}{c} \right), \end{aligned}$$

$t = 1, \dots, T$ .

*Proof.* See Bormetti et al. (2016).

## 4 Simulation results

For the SV-ARG model we simulate 50 data-series of 1,000 observations. For each data-series we run the Gibbs sampler in Borretti et al. (2016) for 100,000 iterations, discard the first 20,000 draws to avoid dependence from initial conditions, and finally apply a thinning procedure to reduce the dependence between consecutive draws. We test the efficiency of the algorithm in three different scenarios: LOW-PERSISTENCE ( $\beta = 0.3$ ), MEDIUM PERSISTENCE ( $\beta = 0.6$ ), and finally, HIGH PERSISTENCE ( $\beta = 0.9$ ). The true values for the other parameters used in the simulations are reported in Table 1 together with the grand average of the parameter posterior means along with their robust standard deviations. The results in Table 1 indicates the accuracy of the MCMC scheme is remarkable for all the scenarios (LOW PERSISTENCE, MEDIUM PERSISTENCE, HIGH PERSISTENCE). As regards the efficiency, the magnitudes of the inefficiency factor after applying a thinning procedure are below ten.

**Table 1** SUMMARY OUTPUT OF THE PARAMETER ESTIMATES FOR THE SV-ARG MODEL

$\theta$	TRUE	LOW PERSISTENCE		MEDIUM PERSISTENCE		HIGH PERSISTENCE	
		ESTIMATE	STD	ESTIMATE	STD	ESTIMATE	STD
$\mu$	0.0	0.0018	0.0118	-0.0051	0.0177	-0.0074	0.0358
$\gamma$	1.0	1.0552	0.0738	1.0523	0.0720	1.0685	0.0784
$\alpha$	0.8	0.8428	0.0572	0.8327	0.0575	0.8474	0.0647
$\nu$	0.8	0.8033	0.0371	0.7981	0.0394	0.8182	0.0576
$c$	1.0	0.9654	0.0938	0.9706	0.0909	0.9395	0.0790
$\beta$		0.3118	0.0595	0.6376	0.0746	0.9702	0.0839

## 5 Conclusions

Motivated by the presence of measurement errors in the empirically computed realized volatility measures we introduce a new family of discrete-time models. We derive the analytical filtering and smoothing and show that they can be used for efficient inference on the parameters and the latent volatility process.

**Acknowledgements** All authors warmly thank Drew D. Creal for helpful comments on the implementation of the algorithm for computing the Bessel function of the second kind and Dario Alitab for support during the development of the pricing code. The research activity of RC is supported by funding from the European Union, Seventh Framework Programme FP7/2007-2013 under Grant agreement FP7/2007-2013, and by the Italian Ministry of Education, University and Research (MIUR) PRIN 2010-11 Grant MISURA. GL acknowledges research support from the Scuola Normale Superiore Grant SNS\_14\_BORMETTI and CI14\_UNICREDIT\_MARMI. This research used the SCSCF multiprocessor cluster system at University Ca' Foscari of Venice.

## References

- Borometti, G., Casarin, R., Corsi, F., and Livieri, G. (2016). Smiles at errors: A discrete-time stochastic volatility framework for pricing options with realized measures. *Working Paper, University Ca' Foscari of Venice*.
- Chib, S., Nardari, F., and Shephard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics*, 108(2):281–316.
- Corsi, F., Fusari, N., and La Vecchia, D. (2013). Realizing smiles: Options pricing with realized volatility. *Journal of Financial Economics*, 107(2):284–304.
- Creal, D. D. (2015). A class of non-Gaussian state space models with exact likelihood inference. *Journal of Business & Economic Statistics*, (just-accepted).
- Engle, R. F. and Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, 131(1):3–27.
- Gouriéroux, C. and Jasiak, J. (2006). Autoregressive gamma processes. *Journal of Forecasting*, 25(2):129–152.
- Hansen, P. R. and Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, 24(2):127–161.
- Majewski, A. A., Borometti, G., and Corsi, F. (2015). Smile from the past: A general option pricing framework with multiple volatility and leverage components. *Journal of Econometrics*, 187(2):521–531.
- Shephard, N. and Sheppard, K. (2010). Realising the future: Forecasting with high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics*, 25(2):197–231.
- Takahashi, M., Omori, Y., and Watanabe, T. (2009). Estimating stochastic volatility models using daily returns and realized volatility simultaneously. *Computational Statistics & Data Analysis*, 53(6):2404–2426.





# Estimating Italian inflation using scanner data: results and perspectives

## *L'uso degli scanner data per la stima dell'inflazione: risultati e prospettive*

Alessandro Brunetti, Stefania Fatello, Federico Polidoro

**Abstract** Scanner data coming from the retail trade outlets of modern distribution represent a crucial challenge for the inflation indicators. Istat is actively participating in the European project aimed at obtaining and processing scanner data to compile HICP. Since the end of 2013 a stable cooperation has been set up among Istat, Association of modern distribution, retail trade chains and Nielsen in order to provide Istat with scanner data. For 2014, 2015 and 2016, scanner data of grocery products have been collected by Istat through Nielsen for about 1400 outlets of the main six retail trade chains for 37 provinces. For 2016 and 2017 scanner data of about 2100 outlets for the entire national territory will be available. In sight of the adoption on large scale of scanner data to estimate inflation, scheduled for January 2018, experimental HICPs/CPIs of one ECOICOP group (non-alcoholic beverages) for two provinces are compiled using scanner data. A comparison with the indices currently released is carried out, providing some preliminary evaluation about the impact of the new sources of data on inflation estimation. Issues concerning formula of indices are dealt with and those regarding missing observations, imputations and replacements are explored.

**Abstract** *Gli scanner data provenienti dai punti vendita della grande distribuzione organizzata costituiscono un'opportunità cruciale per la stima dell'inflazione. L'Istat partecipa al progetto europeo finalizzato all'acquisizione ed elaborazione di queste informazioni per il calcolo degli indici dei prezzi al consumo, avviando, dalla fine del 2013, una stretta collaborazione con l'Associazione Distribuzione Moderna, le catene della grande distribuzione e Nielsen. Per gli anni 2014, 2015 e 2016, l'Istat ha acquisito i prezzi dei prodotti grocery di circa 1400 punti vendita, delle principali sei catene della grande distribuzione organizzata, per 37 province. Per il*

---

<sup>1</sup> Alessandro Brunetti, Istat, [albrunet@istat.it](mailto:albrunet@istat.it)  
Stefania Fatello, Istat, [fatello@istat.it](mailto:fatello@istat.it)  
Federico Polidoro, Istat, [polidoro@istat.it](mailto:polidoro@istat.it)

*2016 e 2017 è prevista la fornitura di dati scanner riferiti a oltre 2100 negozi, a copertura dell'intero territorio nazionale. In vista dell'adozione su larga scala degli scanner data per la stima dell'inflazione, prevista dal 2018, il paper sviluppa un confronto tra indici ufficiali e indicatori sperimentali (basati sugli scanner data) di un gruppo ECOICOP (bevande analcoliche) di due province, fornendo una prima valutazione dell'impatto delle nuove fonti sulla stima dell'inflazione. L'analisi si concentra poi sulle formule utilizzate per il calcolo degli indici ed esplora i problemi relativi al trattamento delle mancate risposte, alle imputazioni e alle sostituzioni di prodotto.*

**Key words:** Scanner data, inflation, modern distribution, dynamic approach

## 1 Introduction (extended abstract)

Scanner data coming from the retail trade outlets of modern distribution represent a crucial challenge for the inflation indicators. Istat is actively participating in the European project aimed at obtaining and processing scanner data to compile HICP. Since the end of 2013 a stable cooperation has been set up among Istat, Association of modern distribution, retail trade chains and Nielsen in order to provide Istat with scanner data. For 2014, 2015 and 2016, scanner data of grocery products have been collected by Istat through Nielsen for about 1400 outlets of the main six retail trade chains for 37 provinces. For 2016 and 2017 scanner data of about 2100 outlets for the entire national territory will be available.

In sight of the adoption on large scale of scanner data to estimate inflation, scheduled for January 2018, experimental HICPs/CPIs of one ECOICOP group (non-alcoholic beverages) for two provinces (Rome and Turin) are compiled using scanner data. Experimental indices are calculated starting from unit values of, on average, more than 75,000 product-offers<sup>1</sup> available in about 300 outlets included in the samples selected for the two provinces for years 2015-2016 (in paragraph 2 a detailed description of the dataset used is provided). The aggregation process of elementary data is addressed in paragraph 3, in which particular attention is devoted to the formula used to calculate micro-indices. As discussed in the paper, the choice of the aggregation method, at the lowest level of calculation of indices, strictly depends on the approach used to define the sample of product offers (references) within the single outlets. Specifically, in what follows a dynamic approach to sampling is adopted: according to this methodology, the set of product-offers selected in each outlet varies from month to month. In this framework, monthly

---

<sup>1</sup> “product-offer” or “reference” mean a specific item, tagged by a GTIN code, sold in a specific outlet for which information on turnover and quantities are available.

indices at the very first step of aggregation are calculated by chaining monthly relatives based on the same sample over two adjacent months.

In the fourth paragraph of the paper, the results of the present analysis are discussed. A preliminary estimate of the impact of the new sources of data on inflation is provided by comparing the experimental indices of the two chief towns with the corresponding indicators currently released.

Conclusions of the paper focus the attention on the open issues (above all how to deal, in the monthly selection of references, with temporarily not available product-offers, replacements and seasonal goods), tracing in general the solution that Istat is going to adopt and the future development of the project.

## 2 Description of the dataset

Scanner data provided by the six main retail trade chains represent, at national level, about 57% of the turnover of modern distribution. Istat receives weekly data for each outlet distinguished by outlet-type (hypermarket and supermarket) for food and grocery products (excluding fresh with variable weight).

The analysis has been carried out on all outlets of the provinces of Rome and Turin delivered by Nielsen for two years 2015 and 2016. Table 1 contains the number of outlets by retail chain, outlet-type and province in each year considered.

**Table 1:** Number of outlets by retail chain, outlet-type and province (2015-2016)

Chain GDO	2015						2016					
	Rome			Turin			Rome			Turin		
	Hyper	Super	Hyper+Super	Hyper	Super	Hyper+Super	Hyper	Super	Hyper+Super	Hyper	Super	Hyper+Super
Conad	4	38	42	2	10	12	4	33	37	2	10	12
Coop Italia	7	8	15	6	19	25	7	8	15	6	19	25
Esselunga	-	-	-	3	-	3	-	-	-	3	-	3
Auchan	4	27	31	3	6	9	4	26	30	3	5	8
Carrefour Italia	4	71	75	13	25	38	4	71	75	13	26	39
Selex	-	26	26	1	34	35	-	26	26	1	7	8
<b>Total</b>	<b>19</b>	<b>170</b>	<b>189</b>	<b>28</b>	<b>94</b>	<b>122</b>	<b>19</b>	<b>164</b>	<b>183</b>	<b>28</b>	<b>67</b>	<b>95</b>

Scanner data coming from retail chains contain weekly data on turnover and quantities sold per item code or GTIN. GTIN (Global Trade Item Number) is the current name of the barcode, formerly known as EAN, and the most commonly used code when dealing with scanner data. GTIN identifies a unique product and consistently refers to the same product over time. Therefore unit value prices per item code can be derived as the average of prices actually paid by households for products. Per each GTIN weekly price (weekly unit values) can be calculated dividing weekly turnover with weekly quantities and monthly prices (monthly unit

value) can be calculated dividing monthly turnover with monthly quantities. CPIs/HICPs can be calculated using the first three full weeks of the month but in the following analysis the indices are calculated considering also two weeks or only one. The underlying hypothesis is that the missing weeks are estimated with the others weeks of which information are available.

To go up from each GTIN to the ECOICOP lowest level of classification, it was necessary to pass through the lowest level of ECR classification (ECR-market), that is the classification of products shared by the industrial and distribution companies and to which each GTIN code is attributed. Istat mapped ECR-markets (about 1600 voices) to Italian ECOICOP-6 level (consumption segments), so that GTINs are automatically classified within the ECOICOP-6 level.

In each consumption segment there are a variable number of ECR-markets with very different turnover shares. ECOICOP group 01.2 “Non-alcoholic beverages” has been chosen for the analysis and in table 2, for each ECOICOP-6 level belonging to this group, the corresponding number of ECR-markets and turnover shares of all outlets of Rome and Turin in the years 2015 and 2016 are reported. ECR-market within each specific outlet is defined the elementary aggregate (EA) and this is the lowest level with the elementary indices are calculated using scanner data. Table 3 shows the average number of GTIN within COICOP-6 level used to calculate elementary indices in both provinces for each year.

**Table 2:** Number of ECR-markets within COICOP-6 level with turnover shares by province (2015-2016)

<i>Coicop-6 level</i>	<i>Description</i>	<i>N° markets</i>	<b>2015</b>		<b>2016</b>	
			<i>Rome</i>	<i>Turin</i>	<i>Rome</i>	<i>Turin</i>
01.2.1.1.0	Coffee	13	24,0	27,0	24,2	27,5
01.2.1.2.0	Tea	12	5,7	6,2	5,9	6,3
01.2.1.3.0	Cocoa and powdered chocolate	3	1,5	1,4	1,4	1,4
01.2.2.1.0	Mineral or spring waters	15	31,8	28,0	32,4	28,9
01.2.2.2.1	Carbonated soft drinks	13	18,5	17,4	18,3	17,0
01.2.2.2.2	Other soft drinks	8	5,0	7,3	4,8	6,8
01.2.2.3.0	Fruit and vegetable juices	43	13,6	12,8	13,0	12,0
<b>Total</b>		<b>107</b>	<b>100,0</b>	<b>100,0</b>	<b>100,0</b>	<b>100,0</b>

**Table 3:** Number of GTINs within COICOP-6 level by province (2015-2016)

<i>Coicop-6 level</i>	<i>Description</i>	<b>2015</b>		<b>2016</b>	
		<i>Rome</i>	<i>Turin</i>	<i>Rome</i>	<i>Turin</i>
01.2.1.1.0	Coffee	8.722	5.360	9.190	5.010
01.2.1.2.0	Tea	7.077	5.039	7.239	4.467
01.2.1.3.0	Cocoa and powdered chocolate	953	666	984	609
01.2.2.1.0	Mineral or spring waters	6.132	3.253	6.196	3.048
01.2.2.2.1	Carbonated soft drinks	7.700	4.920	7.987	4.442
01.2.2.2.2	Other soft drinks	5.182	3.148	5.066	2.694
01.2.2.3.0	Fruit and vegetable juices	12.606	7.875	12.847	7.120
<b>Total</b>		<b>48.372</b>	<b>30.259</b>	<b>49.508</b>	<b>27.390</b>

### 3 Index calculation formulas

In compliance with the dynamic approach, a sample of those product-offers that are present in both the current and the preceding month is monthly selected in each outlet. Particularly, the dynamic basket of references is obtained by using a set of filters to select a matched sample each month comparing the current month with the preceding month. To this aim, the following three different filters are considered:

- A dump filter that removes products where strong decreases in price and turnover/quantities;
- An outlier filter that removes prices that drop/increase above certain thresholds
- A low-sales filter that filters out item codes with very low sales (the low-sales filter is empirically determined so that the selected item codes represent about 80% of turnover at the ECR-market level)

The EA index is calculated on the basis of the matched set of representative item codes, classified in a given ECR-market, that are actually sold in two subsequent periods in a given outlet. Specifically, in each ECR-market, in each outlet, an unweighted Jevons index is calculated over the current and preceding month as follows:

$$P_{jev}^{(m-1),t;m,t} = \left( \prod_{n \in S^{m-1,t}} \frac{p_n^{m,t}}{p_n^{m-1,t}} \right)^{1/\zeta(S^{m-1,t})}$$

where:

$p_n^{m,t} / p_n^{m-1,t}$  is the price relative between month  $m-1$  and  $m$ ,

$S^{m-1,t}$  is the set of representative items sold in month  $m-1$  and  $m$ ,

$\zeta(S^{m-1,t})$  is the number of representative item codes in  $S^{m-1,t}$ .

The EA chain-linked index is then as follows:

$$P_{jev}^{m,t} = \left( \prod_{n \in S^{0,t}} \frac{p_n^{1,t}}{p_n^{0,t}} \right)^{1/\zeta(S^{0,t})} \times \left( \prod_{n \in S^{1,t}} \frac{p_n^{2,t}}{p_n^{1,t}} \right)^{1/\zeta(S^{1,t})} \times \dots \times \left( \prod_{n \in S^{m-1,t}} \frac{p_n^{m,t}}{p_n^{m-1,t}} \right)^{1/\zeta(S^{m-1,t})}$$

It has to be noted that relaunches (new versions of the same item with some superficial differences and a new item code) and replacements (discounts that receive a new item code; products of a certain brand replaced by similar products of another brand) form a potential problem for this sampling method as the system does not automatically link a disappearing product-offer with its relaunch or replacement. For this reason, algorithms have to be implemented in order to detect and treat

relaunches and replacements appropriately (by combining old and new product-offers and then calculate unit value indices for the combination). Moreover, prices for item codes that are not present in subsequent periods should be imputed to ensure seasonal items re-enter the index at the correct time.

For the calculation of experimental indices commented in this paper, however, neither the treatment of relaunches/replacements nor the imputation of absent references have been taken into account.

Concerning the ECOICOP experimental indices (3 and 4 digits level), they are calculated as Laspeyres indices, through successive aggregations of EA indices:

- 1) Firstly, the ECR-market indices for different outlet type (hypermarket and supermarket) at the retailer's level are obtained as weighted arithmetic mean of EA chain-linked indices, with weights proportional to the share of turnover of the concerned outlets.
- 2) The ECR-market indices of the two outlet types (hypermarket and supermarket) are then compiled by aggregating the corresponding indices of the different retailers (weights proportional to turnover shares).
- 3) The provincial ECR-market indices are calculated by aggregating the ECR-market indices of hyper and supermarket (weights proportional to turnover shares).
- 4) Finally ECOICOP indices results from the aggregation of ECR-market indices (weights proportional to turnover shares at the 7 digits level, and HICP weights for 6 to 4 digits indices).

## 4 Results

Figure 1 and figure 2 show the first results of the present analysis for the group considered and one class (mineral waters, soft drinks, fruit and vegetable juices). The annual rates of change calculated with experimental scanner data indices are compared with the corresponding indicators calculated with territorial data collection (provincial HICPs for Rome and Turin are compiled for this purpose).

In both provinces the figures show a similar trend for the two annual rates of change compared. Two main evident differences emerge from a preliminary analysis. The first one is the more regular trend of the indices compiled with scanner data. The second one is the higher inflation registered by the experimental indices. As a matter of fact, annual rate of change on average in 2016 of prices of non-alcoholic beverages in Rome is +0.4% with scanner data and 0.0% with territorial data (in Turin +0.4% versus -0.3%) and the same results emerge for mineral waters, soft drinks, fruit and vegetable juices (respectively +0.6% versus -0.3% in Rome and +0.6% versus -0.2% in Turin).

Although this empirical evidence is circumscribed, it shows that the use of big amount of data (as those represented by scanner data) that better cover time (not just

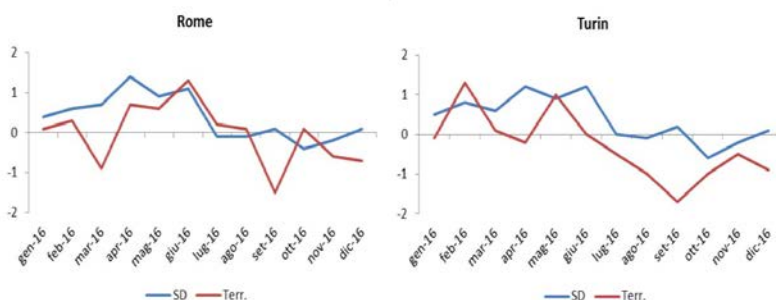
collected once a month as in the territorial data collection) and space (a huge number of outlets with respect to present sample currently used, all the GTINs belonging to a segment of consumption not just the most sold product-offer) should allow eliminate the volatility coming from the limits of the present territorial data collection.

Moreover the results obtained show that the sign of the impact on inflation estimation deriving from the adoption of scanner data, at least locally, could be different (up) from what is expected (down).

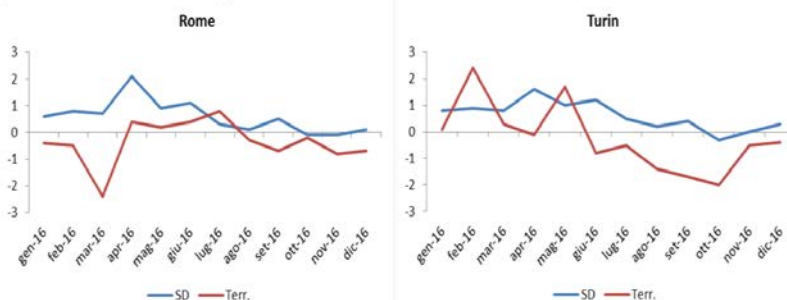
It is clear that there is not possibility to generalize the results obtained. The reason lies not only on the limited test (two ECOICOP aggregations, two towns) but on the issues concerning the use of scanner data whose solutions, for the time being, have not been implemented: replacements, treatment of relaunches and seasonal goods, imputations, combing indices calculated with scanner data coming from modern distribution with indices compiled for traditional distribution.

It is just worth to note that, for what concerns replacements and relaunches, the filter used in dynamic approach adopted in this test ensures good representativeness of the sample in terms of turnover.

**Figure 1:** Harmonized index of consumer price of non-alcoholic beverages. M/M-12 rates of changes - Year 2016



**Figure 2:** Harmonized index of consumer price of mineral waters, soft drinks, fruit and vegetable juices. M/M-12 rates of change - Year 2016



## 5 Concluding remarks and open issues

The potential improvements coming from the use of scanner data in inflation estimation makes this source a key point of the project of modernization of price statistics pursued at Italian and European level. Qualitative improvements emerge also from the analysis carried out in this paper and they depends on the characteristics of scanner data: wide temporal and spatial coverage, prices referred to actual transactions, information of quantities sold and turnover, the possibility of reducing the administrative burden which now, for example, weighs on the Municipal Offices of statistics that are involved in data collection of elementary price quotes in the field.

Together with these potentialities, also the issues have to be stressed in order to deal with them. Most of these issues, in addition to those ones represented by the dependency on the retailers for the data provision, are of methodological nature. As cited before, they regard relaunches, disappeared items and their replacements, treatment of seasonal goods and imputations of missing observations.

Since Istat is moving towards the adoption of a dynamic approach for the sampling of references and taking into account the big amount of data to be treated on monthly basis, automatic solutions have to be adopted to work on all the references that is not possible to match in the monthly sample (the disappeared references of the previous and the new references of the current month), evaluating their nature (seasonal or not), if they are relaunches, if it is better to impute the prices or not for the disappeared references, if and how to possibly replace them.

Dealing with issues is the main and crucial task that Istat has assumed for the next weeks and months toward the adoption of scanner data to calculate official HICPs/CPIs in 2018.

Intermediate results of this challenging work will be discussed in a workshop in Istat scheduled for the beginning of July.

## References

1. EUROSTAT, Practical Guide for Processing Supermarket Scanner Data, Draft March 2017
2. EUROSTAT, Scanner data recommendation, Meeting of the Price Statistics Working Group, Luxembourg 7-8 November 2016 (PSWG09.2016/10)
3. ISTAT, Workshop Scanner Data, Roma 1-2 ottobre 2015 (<http://www.istat.it/it/archivio/168890>):
4. Dalén J. (Eurostat), The Use of Unit Values in Scanner Data, Workshop Scanner data Vienna October 2014 ([http://www.statistik.at/web\\_en/about\\_us/events/scanner\\_data\\_workshop/index.html](http://www.statistik.at/web_en/about_us/events/scanner_data_workshop/index.html))
5. EUROSTAT, Compendium of HICP Reference Documents, 2013 edition.
6. Norberg, Anders, Sammar, Muhanad and Tongur, Can (2011), "A Study on Scanner Data in the Swedish Consumer Price Index" (<http://www.stats.govt.nz/ottawa-group-2011/agenda.aspx>)
7. ILO (2004), Consumer Price Index Manual. Theory and Practice. Available at <http://www.ilo.org/public/english/bureau/stat/guides/cpi/>
8. SCB (2001), "The Swedish Consumer Price Index. A handbook of methods". Available at <http://www.scb.se>



# Clustering of histogram data : a topological learning approach

## *Tecniche di raggruppamento di dati a istogramma: un approccio basato sull'apprendimento topologico*

Guénaél Cabanes, Younès Bennani, Rosanna Verde and Antonio Irpino

**Abstract** An histogram data is described by a set of distributions. In this paper, we propose a clustering approach using an adaptation of the Self-Organizing Map (SOM) algorithm. The idea is to combine the dimension reduction obtained with a SOM and the clustering of the data in this reduced space. The L2 Wasserstein distance is used to measure dissimilarity between distributions and to estimate local data densities in the original space. The main advantage of the proposed algorithm is that the number of clusters is found automatically. Applications on synthetic and real data sets demonstrate the validity of the proposed approach.

**Abstract** *I dati a istogramma sono descritti da distribuzioni (rappresentati in forma di istogramma). In questo lavoro, proponiamo un approccio di classificazione utilizzando un adattamento delle carte auto-organizzate (SOM). L'idea è di utilizzare una combinazione della riduzione dimensionale ottenuta con una SOM con la classificazione dei dati nello spazio ridotto. Per misurare la dissimilarità tra le distribuzioni viene utilizzata la distanza L2 di Wasserstein. La stessa viene utilizzata per la stima di densità di dati locali nello spazio originario. Il principale vantaggio dell'algoritmo è che il numero di gruppi viene determinato automaticamente. La validità dell'approccio proposto viene mostrata attraverso la sua applicazione su dati artificiali e reali.*

**Key words:** Clustering, Self-Organising Map, Histogram data, Wasserstein distance, Density measure

---

Cabanes and Bennani

LIPN-CNRS, UMR 7030, Université Paris13, 99 Avenue J-B. Clément, 93430 Villetaneuse, France, e-mail: cabanes@lipn.univ-paris13.fr

Verde and Irpino

Dip. Matematica e Fisica, Università della Campania "Luigi Vanvitelli", Viale A. Lincoln, 5, 81100 Caserta, Italy

## 1 Introduction

An histogram data is described by a set of distributions, represented by histograms variables. A histogram is constituted by a sequence of continuous intervals with associated a set of weights (e.g. the relative frequencies). They have been introduced in the context of Symbolic Data Analysis (SDA) by Bock and Diday, in the SDA reference book [1]. As histogram data-sets mostly from empirical distribution, the techniques recently developed for such data refer to distributional data. A field of research on distributional data analysis has also provided by the use of a suitable distance, the  $L_2$  Wasserstein distance, to compare distributions. This measure allows a different way of computing the distance between distributional data. The  $L_2$  Wasserstein distance can be decomposed in two components: the first related to the means (location parameter) and the second related to the higher moments (scale and shape parameters). In such way, the results of distributional data analysis take into account the main characteristics of the data. Such two components can be into consideration, separately, to analyse the influence of the size and shape of the data in the analysis.

SOM for symbolic data was firstly proposed by Bock [1] to visualise in a reduced subspace the structure of symbolic data. Further SOM method for particular symbolic data, the interval data, have been developed using suitable distances for interval data, like Hausdorff distance;  $L_2$  distance, adaptive distances [2]. In the analysis of histogram data, that represent another representation of symbolic data by empirical distributions, SOM has been proposed by [3] based on the Wasserstein  $L_2$  distance to clustering distributions. Adaptive Wasserstein distance has been also developed in this context to find, automatically, weights for the variables as well as for the clusters. However, the most part of these methods can provide a quantification and a visualization of symbolic data (intervals, histograms) but cannot be used directly to obtained a clustering of the data. The recent algorithm proposed by [4]: S2L-SOM learning for interval data, is a two-level clustering algorithm based on SOM that combine the dimension reduction by SOM and the clustering of the data in a reduced space in a certain number of homogeneous clusters. Here, we propose an extension of this approach to histogram data. In the clustering phase is used the  $L_2$  Wasserstein distance according to the dynamic clustering algorithm proposed by [5]. The number of cluster is not a priori fixed as parameter of the clustering algorithm but it is automatically found according to an estimation of local density and connectivity of the data in the original space, as in [6].

The paper is organized as follow. In section 2 we present the proposed approach. Section 3 shows the experimental protocol and the results obtained to validate our approach. Finally, a conclusion and future work is given in section 4.

## 2 DHSOM: a topological density-based clustering for Histogram data

### 2.1 Principles of the approach

A SOM consists of a set of artificial neurons that represent the data structure. These neurons are connected with their neighbours according to topological connections (also called neighbourhood connections). The input dataset is used to organize the SOM under topological constraints of the input space. Thus, a correspondence between the input space and the mapping space is built. Two observations, close in the input space, should activate the same neuron or two neighbouring neurons of the SOM. Each neuron is associated with a prototype and, to respect the topological constraints, neighbouring neurons of the Best Match Unit of a data (BMU, the most representative neuron) also update their prototype for a better representation of this data. This update is important because the neurons are close neighbours of the best neuron.

In DS2L-SOM [6], the prototypes of a SOM are enriched with local estimations of density and connectivity, allowing an estimation of the underlying distribution of the data. More specifically, we compute an estimation of the local density of the data: a measure of the data density surrounding the prototype. The local density is an information about the amount of data present in an area of the input space. We use a Gaussian kernel estimator [7] for this task. The connectivity measures how close are to prototypes for the data representation. The connectivity value of two prototypes is the number of data that are well represented by both of them (the two prototypes are the first two Best Match Unit for these data). From this estimation, it is possible to compute a clustering of the prototypes (as a representation of the data's clustering) as described in [6]. In that case, clusters are defined as regions of the representation space having a relative high density, separated with regions of relative low density. As in most density-based methods, the number of clusters is detected automatically.

To adapt the principles of DS2L-SOM to histogram data, we need a modified version of the Self-Organising Map and an adapted enrichment of the prototypes. We chose here a SOM algorithm for histogram data that have been proposed in [3], where each prototype is defined as an histogram and the distances between data and prototype are computed with the  $L_2$  Wasserstein distance. In addition, the estimation of the local densities and variabilities in DS2L-SOM are mainly based on the distance between the data and the prototype. By using the  $L_2$  Wasserstein distance in the enrichment step, the clustering of histogram data becomes possible. It is worth of notice that the density estimated by this metric, allows to keep into account the all information about the characteristics of the data.

## 2.2 SOM for histogram data

The adaptation of SOM for histogram data is based on two principle: each prototype is an histogram and the distances between observations and prototypes are computed with the  $L_2$  Wasserstein metric. In this paper we propose the use of a batch version of SOM adapted to histograms. The first step of algorithm is the competition step, where each observation is assigned to the neuron with the closest prototype (i.e. the *BMU*: Best Match Unit) according to the  $L_2$  Wasserstein distance. The second step is the Adaptation step, where each prototype is updated to minimise the average distance between the prototypes and the observations, weighted by the topological structure of the map. The function to minimize is the following:

$$R(w) = \sum_{k=1}^N \sum_{i=1}^M K_{iu^*(x^k)} \|w^i - x^k\|^2 \quad (1)$$

where  $x$  is an observation represented as an histogram,  $w$  is a prototype,  $N$  represents the number of learning samples,  $M$  the number of neurons in the map,  $u^*(x^k)$  is the neuron having the weight vector closest to the observation  $x^k$  (i.e. the best match unit: *BMU*), and  $K_{ij}$  is a positive symmetric kernel function: the neighbourhood function. The relative importance of a neuron  $i$  compared to a neuron  $j$  is weighted by the value of the kernel function  $K_{ij}$  which can be defined as:

$$K_{i,j} = \frac{1}{\lambda(t)} \times e^{-\frac{d_1^2(i,j)}{\lambda^2(t)}} \text{ with } \lambda(t) = \lambda_i \left( \frac{\lambda_f}{\lambda_i} \right)^{\frac{t}{t_{max}}}$$

$\lambda(t)$  is the temperature function modelling the topological neighbourhood extent.  $\lambda_i$  and  $\lambda_f$  are respectively initial and the final temperature.  $t_{max}$  is the maximum number allotted to the time.  $d_1(i, j)$  is the Manhattan distance defined between two neurons  $i$  and  $j$  on the map grid. To minimize eq.1, each prototype is updated to represent the barycentre of the observations, weighted by  $K_{ij}$ . The prototypes can be computed using a decomposition of the center and radius in each dimension. The weighted barycentre is then expressed as follow:

$$\bar{w}^j = \{([\bar{c}_1^j - \bar{r}_1^j; \bar{c}_1^j + \bar{r}_1^j], \pi_1^j) \dots ([\bar{c}_v^j - \bar{r}_v^j; \bar{c}_v^j + \bar{r}_v^j], \pi_v^j) \dots ([\bar{c}_h^j - \bar{r}_h^j; \bar{c}_h^j + \bar{r}_h^j], \pi_h^j)\} \quad (2)$$

where:

$$\bar{c}_v^j = \frac{\sum_{i=1}^N K_{ij} c_v^i}{\sum_{i=1}^N K_{ij}} \text{ and } \bar{r}_v^j = \frac{\sum_{i=1}^N K_{ij} r_v^i}{\sum_{i=1}^N K_{ij}} \quad (3)$$

The introduction of the  $L_2$  Wasserstein distance between histograms pass through the piecewise quantile functions. So that a linear combination of quantile function is again a quantile function only if the weights are positive. The complete algorithm is described in algorithm 1.

**Algorithm 1** SOM for histogram data

---

```

1: Define the topology of the SOM.
2: Initialize the prototypes  $w^j$ .
3: repeat
4:   for all histogram data  $x^k$  do
5:     Select the BUM  $u^*(x^k)$  according to the  $L_2$  Wasserstein distance;
6:   end for
7:   for all prototype  $w^j$  do
8:     Update  $w^j$  according to eq. 2 and 3.
9:   end for
10: until  $t = t_{max}$ 

```

---

**2.3 Prototypes Enrichment**

When the prototypes are computed, the model can be enriched with additional information associated to each prototypes, in order to improve the representation of the underlying structure of the data. Two information are computed in this step (algorithm 2). The connectivity between neurons is a measure of discontinuity in the topological space and allows to detect clusters separated by an empty region of the representation space. As this region are often not well represented by the prototypes, this assure the detection of cluster borders between two adjacent neurons. However, when the clusters' boundary is defined by a region of lower density between two regions of higher density, the connectivity is not sufficient and an estimation of local densities is necessary. The local density  $D_i$ , associated to each prototype  $w^i$ , is estimated as follow:

$$D_i = 1/N \sum_{k=1}^N \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{d_{W}^2(w^i, x^k)}{2\sigma^2}} \quad (4)$$

with:  $\sigma$  a bandwidth parameter chosen by user and  $d_{W}^2(w^i, x^k)$  the distances between the  $M$  prototypes  $w^i$  and the  $N$  histogram data  $x^k$ , computed with the  $L_2$  Wasserstein metric.

**Algorithm 2** Enrichment of prototypes

---

```

Input: The Wasserstein distance between each observation and each prototype
Output: A density value  $D_i$  for each neuron  $w^i$  and a connectivity value  $v_{i,j}$  for each pair of neurons  $i$  and  $j$ .
for all neuron  $i$  do
  Compute the local density  $D_i$  using eq. 4.
end for
for all data  $x^k$  do
  Find the two closest prototypes (BMUs)  $u^*(x^k)$  and  $u^{**}(x^k)$  using:

$$u^*(x^k) = \operatorname{argmin}_i d_{W}^2(w^i, x^k)$$

  Compute  $v_{i,j}$  = the number of data having  $i$  and  $j$  as two first BMUs.
end for

```

---

## 2.4 Clustering of prototypes

Various prototypes-based approaches have been proposed to solve the clustering problem [8, 9, 10]. However, the obtained clustering is never optimal, since part of the information contained in the data is not represented by the prototypes. Here we use the density and connectivity to optimize the clustering (see algorithm 3).

---

### Algorithm 3 Clustering of enriched prototypes

---

**Input:** the density values  $D_i$  and the connectivity values  $v_{i,j}$ .

**Output:** The clusters of prototypes.

1: Extract the sets of connected neurons  $P = \{C_i\}_{i=1}^L$ , such as:

$$\forall m \in C_i, \exists n \in C_i \text{ such as } v_{m,n} > \text{threshold}$$

2: In this paper  $\text{threshold} = 0$ .

3: **for all**  $C_k \in P$  **do**

4: Find the set  $M(C_k)$  of density maxima.

$$M(C_k) = \{w^i \in C_k \mid D_i \geq D_j, \forall w^j \text{ neighbor to } w^i\}$$

Prototypes  $w_i$  and  $w^j$  are neighbour if  $v_{i,j} > \text{threshold}$ .

5: Determine the merging threshold matrix:

$$S = [S(i,j)]_{i,j=1,\dots,|M(C_k)|} \text{ with } S(i,j) = \left( \frac{1}{D_i} + \frac{1}{D_j} \right)^{-1}$$

6: **for all** prototype  $w_i \in C_k$  **do**

7: Label  $w^i$  with one element  $\text{label}(i)$  of  $M(C_k)$ , according to an ascending density gradient along the neighbourhood. Each label represents a micro-cluster

8: **end for**

9: **for all** pair of neighbours prototypes  $(w^i, w^j)$  in  $C_k$  **do**

10: merge the two micro-clusters if:

$$\text{label}(i) \neq \text{label}(j), D_i > S(\text{label}(i), \text{label}(j)) \text{ and } D_j > S(\text{label}(i), \text{label}(j))$$

11: **end for**

12: **end for**

---

The main idea is that the core part of a cluster can be defined as a region with high density. Then, in most cases the cluster borders are defined either by low density region or “empty” region between clusters (i.e. large inter cluster distances) [11]. At the end of the enrichment process, each set of prototypes, linked together by connectivity value  $v > 0$ , defines well separate clusters (i.e. distance-defined). The estimation of the local density ( $D$ ) is then used to detect cluster borders defined by low density. Each cluster is defined by a local maximum of density). Thus, a “Watersheds” method [12] is applied on prototypes’ density for each well separated cluster to find low density area inside these clusters. Finally, for each pair of adjacent subgroups we use a density-dependent index [13] to check if a low density area is a reliable indicator of the data structure.

### 3 Experimental results

We generated six datasets with different number of clusters and dimensions. Each dataset contains 1000 observations, each observation is constituted by 2 or 10 histograms (i.e. 2 or 10 dimensions). For each histogram, 1000 values were generated using a Gamma distribution with 3 parameters: the mean value, the standard deviation and a shape parameter, controlling the skewness of the distribution. From this values, an equi-depth histogram is computed using the 10th percentiles of the values for each interval of the support of the histogram, for a total of 10 intervals per histogram, such that each interval has a weight  $\pi = 0.1$ . The observations are generated in respectively 3 or 5 clusters according to different parameters of the gamma distribution. In our proposal, the Wasserstein distance takes into account the different components of the distribution (mean, standard deviation and shape). We expect that the results are strongly depending on the distance. To validate our method ("Prop"), we compare the results with different strategies based on different dissimilarities. We tested measures using only the component  $c_i$  ("Center") and  $r_i$  ("Radius") in the Wasserstein distance decomposition. We also tested a distance based on the "means"  $\mu$  of the distributions, and a distance based on the standard deviation ("Std")  $\sigma$  of the distribution. Finally, we tested two distances between interval data computed from support values of the histograms. In the first case ("Int1") has been considering the lower and upper values over the distribution support to define the interval bounds:  $[min, max]$ . In a second case ("Int2"), we considered the mean and standard deviation of the distribution to compute the interval bounds:  $[\mu - \sigma, \mu + \sigma]$ .

The obtained results are shown in Table 1. The performance of the different approaches is evaluated using the adjusted Rand index. This index take values in  $[0, 1]$ , 1 being a perfect match with the expected clustering and 0 denoting a random solution.

Table 1: Adjusted Rand Index for each dataset and each approach. *Param* is the parameter defining the differences between the clusters, *k* is the number of clusters, *d* is the number of dimensions (i.e. the number of histogram per data)

Param	k	d	Prop	Center	Radius	Mean	Std	Int1	Int2
Mean	3	2	1.00	0.88	0.00	1.00	0.00	0.00	1.00
Mean	5	10	1.00	1.00	0.00	1.00	0.00	0.43	1.00
Shape	3	2	0.95	0.57	0.00	0.00	0.00	0.52	0.00
Shape	5	10	0.81	0.60	0.00	0.00	0.00	0.19	0.00
Std	3	2	1.00	1.00	0.00	0.00	1.00	0.98	1.00
Std	5	10	1.00	1.00	0.99	0.00	1.00	1.00	1.00

From the result we can see that the proposed method ("Prop") with the  $L_2$  Wasserstein is able to detect correctly the cluster separations (and therefore the correct number of clusters) for the 6 datasets. Our approach is the only able to detect correctly the difference in shape in the distributions, in addition to detect clusters with different means or standard deviations.

## 4 Conclusion

In the paper we proposed a two-level clustering method for histogram data. The approach takes into account all the information about size and shape of the distributional data in the analysis thanks to the  $L_2$  Wasserstein metric. This method is fast and doesn't require the number of clusters to be fixed by the user. Indeed, the cluster's boundaries are detected automatically based on an estimation of local densities and connectivities in the partitioning process. The core part of the cluster being defined as the region with higher density, the Wasserstein distances between histogram data allows to detect areas of low density between clusters.

The specificity of the presented strategy is that the density is different from the density of classical data. Indeed, the Wasserstein distance allows to compute the data density according to the characteristics of the data distributions, resulting in a richer model of the data structure. We have shown how the proposed method give better results than concurrent strategies.

## References

1. H.-H. Bock and E. Diday, Eds., *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data.* Springer Verlag, Heidelberg, 2000.
2. A. Irpino and R. Verde, "Dynamic clustering of interval data using a wasserstein-based distance," *Pattern Recogn. Lett.*, vol. 29, no. 11, pp. 1648–1658, 2008.
3. F. d. A. De Carvalho, A. Irpino, and R. Verde, *Batch self organizing maps for interval and histogram data*, isi ed. Curran Associates, Inc. (2013), 2013, pp. 143–154.
4. G. Cabanes, Y. Bennani, R. Destenay, and A. Hardy, "A new topological clustering algorithm for interval data," *Pattern Recognition*, vol. 46, no. 11, pp. 3030–3039, 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0031320313001520>
5. R. Verde and A. Irpino, "Dynamic clustering of histograms using wasserstein metric," in *Proceedings in Computational Statistics, COMPSTAT 2006*, A. Rizzi and M. Vichi, Eds., Compstat 2006. Heidelberg: Physica Verlag, 2006, pp. 869–876.
6. G. Cabanes, Y. Bennani, and D. Fresneau, "Enriched topological learning for cluster detection and visualization," *Neural Networks*, no. 32, pp. 186–195, 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0893608012000482>
7. B. Silverman, "Using kernel density estimates to investigate multi-modality," *Journal of the Royal Statistical Society, Series B*, vol. 43, pp. 97–99, 1981.
8. E. L. J. Bohez, "Two level cluster analysis based on fractal dimension and Iterated Function Systems (IFS) for speech signal recognition," *IEEE Asia-Pacific Conference on Circuits and Systems*, pp. 291–294, 1998.
9. M. F. Hussin, M. S. Kamel, and M. H. Nagi, "An efficient two-level SOMART document clustering through dimensionality reduction," in *ICONIP*, 2004, pp. 158–165.
10. E. E. Korkmaz, "A two-level clustering method using linear linkage encoding," in *International Conference on Parallel Problem Solving From Nature, Lecture Notes in Computer Science*, vol. 4193. Springer-Verlag, 2006, pp. 681–690.
11. A. Ultsch, "Clustering with SOM: U\*C," in *Proceedings of the Workshop on Self-Organizing Maps*, 2005, pp. 75–82.
12. L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, pp. 583–598, 1991.
13. S.-H. Yue, P. Li, J.-D. Guo, and S.-G. Zhou, "Using greedy algorithm: DBSCAN revisited II," *Journal of Zhejiang University SCIENCE*, vol. 5, no. 11, pp. 1405–1412, 2004.



# Measuring Wellbeing by extracting Social Indicators from Big Data

## *Estrarre Indicatori Sociali dai Big data per misurare il benessere*

Renza Campagni<sup>1</sup>, Lorenzo Gabrielli<sup>2,3</sup>, Fosca Giannotti<sup>2</sup>, Riccardo Guidotti<sup>3</sup>,  
Filomena Maggino<sup>4</sup>, Dino Pedreschi<sup>3</sup>

**Abstract** Traditionally, the construction of social indicators is based upon the availability of data collected on purpose (e.g., official statistics). It is a common view that constructing social indicators can benefit from the availability of new sources of data, that is, big data. One of the big challenges in dealing with new data sources is related to the possibility of describing complex social phenomena from different perspectives in order to enrich already used indicators and/or build new ones. However, this possibility introduces new issues in constructing indicators. Our study intends to explore how the classical methodology for social indicators construction should be re-considered in light of using data collected for other aims. In this perspective, the individual sales receipts, collected during the period 2007/13 and made available to our group by a big Italian chain of stores, allow us to explore not only a particular social phenomenon but also the methodological implications in dealing with big data. In particular, we try to (i) understand what kind of information can be extracted from these data, important and informative in constructing social indicators; (ii) study different families' behaviour in a crucial period, by detecting possible changes in people's lifestyle and eventually the role of the crisis of last years in these changes; (iii) elaborate and test a model aimed to extracting social indicators from big data. The study is enriched by the possibility to observe across different areas the groups behaviour (by referring to the territorial distribution of the stores) and to trace the individual spending behaviour over time, while ensuring the anonymity of the sensitive information eventually present in data. The starting stage of our study presented here can show some results obtained by analysing data with data mining clustering techniques, in order to identify some typical purchase behaviour but also to test if and how starting from this information it is possible to estimate other structural information.

**Key words:** Big Data, Social Indicators, Clustering, Classification

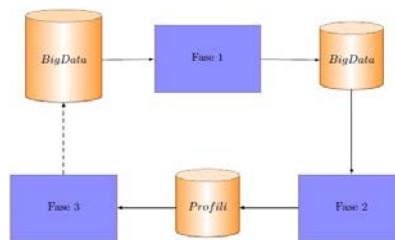
---

<sup>1</sup>Università degli Studi di Firenze, <sup>2</sup>ISTI-CNR di Pisa, <sup>3</sup>Università di Pisa, <sup>4</sup>Sapienza Università di Roma

## 1 Introduction

The aim of this project is to define new social indicators describing the customers behaviors, starting from the analysis of big data related to their purchases in stores of a chain of supermarket. Traditionally, the construction of social indicators is based on data collected specifically, as in the case of official statistics but, in recent years, with the availability of new data sources such as big data, it is spreading the need to build new social indicators from this information. One of the great challenges in deal with these new data sources is related to the possibility to describe complex social phenomena by different points of view. Figure 1 shows the phases of the project that aims to discover information than can be extracted from these data and that can serve to the construction of social indicators. By analyzing the behavior of different families in a crucial period, we can observe possible changes in the lifestyle of the people and the role of the crisis of recent years and we can develop and test a model aimed to define social indicators starting from big data. The study of social behaviors and lifestyles of families, for which several articles in the journal *Social Indicators Research* are published by Springer, can help in defining new social indicators, as discussed in [2] and [3].

The dataset labeled *BigData* at the top left of Figure 1 identifies the data of the starting process; the second *BigData* dataset represents the data after the selection phase, that is, data suitable for the project purpose. The output of the analysis phase is the third dataset labeled *Profili*, which corresponds to the groups of customers based on the classification emerged from the analysis phase, performed by using data mining techniques. The dashed arrow going from the definition of the indicators to the first dataset indicates the iterativity of the process that, in the case where the process has led to a good definition of the indicators, can indicate the application of these to another dataset. It can happen that a first iteration of the process does not produce any good indicators, then the process begins again from the selection of useful informations. The project therefore aims to explore how the classic methodological approach of the definition of social indicators may be reconsidered in light of using the data collected for other purposes, such as for example those administrative. Starting from transactional data, such as analyzed in [5] and [6], or by the



**Fig. 1** Flow of the entire iterative project.

receipt of customer purchases, collected during the period 2007 to 2013, the project aims to explore not only a social phenomenon particular, but also the methodological implications that are encountered in dealing with big data. Through the analysis of this information particular categories of consumers can be identified, classified also in accord to how customers changed their purchasing behaviors in the period of economic crisis that in recent years has involved Italy. This classification, together with the analysis of purchasing changes, can allow to estimate the particular social and/or economic hardships. The analysis can also be done by deepening the behavior of particular groups of customers in relation to the geographical component, that is, by referring to the territorial distribution of the stores, but also trying to trace the individual spending behavior over time and to verify whether and how the results obtained with it is possible to estimate other structural information (e.g., the size and structure of the family). We analyze data of our case study, concerning purchases in a single store of the chain, selected with the appropriate features (for example not affected by the seasonality problem), by using the software R [8] for traditional descriptive investigations and the software KNIME [9] for data mining analysis.

## **2 Understanding the customers behaviours**

### ***2.1 Defining Social Indicators***

The idea is to define indicators starting from the analysis of data stored for other purposes, with which we can have a higher freshness than that one obtained with official indicators; the purpose is to control important signals, resulting from changes in customers purchases behaviors, which can be important to predict changes in the macroeconomic framework. The methodology that we propose starts from grouping customers using clustering techniques, with the aim to identify, for each group of customers, typical characteristics of each cluster, defined by how much, what, and how they have purchased. These three analysis dimensions are translated at macroscopic level, respectively in the total expenditure, in the total quantity of products purchased and the number of times in which customers have been shopping. The choice of the temporal component, that is, the unit of observed period, is very important in this first phase of the project; in this analysis it is the year.

### ***2.2 Clustering and classification: customers profiles***

We explore, through the classification of the categories to which products belong, if when the values of the parameters listed in the previous section change, also the types of products purchased change. For example, by deepening the analysis on the categories of products purchased, we can find that, during the crisis, a customer

segment decreased the purchase of niche products, to the benefit of basic/low-level products. The study analyzes customer data concerning amounts, quantities and number of expenses aggregated on the year. Starting from amount and quantity we can deepen the analysis to the level of the category. Firstly, the three attributes of analysis are analyzed in yearly level one at a time, in such a way that, for example starting from the information on the total amount in each year, we can associate to each customer a sequence of seven values, which are the amounts in the seven years involved. We apply the  $K$ -means clustering algorithm to data organized as shown in Table 1, in order to identify  $K$  clusters, containing customers who had similar behaviors over the years for what concerns yearly amounts. A customer can be represented as a series of  $n$  points, one for each year; we can thus represent the behavior of a customer with a broken line describing its total annual expenditure trend. From this metaphor graphics, proposed by the authors in [1], the analysis shows customer groups that follow the same pattern in several years. Then the analysis is repeated by considering other information, that is, data relating to the quantity and to the number of expenses, as well as on the attribute given by the ratio amount/number of expenses. The results of these different analyzes can be interpreted and investigated together to understand, for example, what is the relationship between customers who have been assigned to a particular cluster for what concern the amounts or the relationship between those who have been assigned to the clusters obtained with the analysis of quantity. A peculiarity of the analysis with the  $K$ -means algorithm is the choice of the number  $K$  of clusters; for this, a choice of the value of this parameter based on different values SSE (Sum of Squared Error), as reported in [7], obtained from several values of  $K$ , was adopted, to locate on the curve, which has a decreasing trend when increases the value of  $K$ , the point of maximum bending, which we know it can give good results.

**Table 1** An example of yearly information concerning customer 10.

Customer_Id	Year0	Year1	...	Year(n-1)	Type
10	5	4	...	8	number of exp.
10	100	120	...	250	quantity
10	300.75	600.604	...	1050.10	amount
⋮	...	...	...	...	...

### 2.3 Clustering and classification: products profiles

From the analysis presented in the previous section can emerge some customers groups which may suggest analytical insights with the aim to find which products have led to a change of purchase behaviors of customers. The analysis focus moves

so by the customer to the product: it is therefore important to understand if and over what types of product, over the years affected by the economic crisis, purchasing behaviors changed. In particular, we apply clustering techniques to group customers to understand if there are changes in shopping cart. We would like to know what happens for what concern the typologies of products purchased when amounts, quantities or number of expenses significantly change. During crisis, it can happen that a group of customers reduced purchase of niche products, to the benefit of lower-end products. The goal is to find which are the products that can be considered sentry products; by keeping under control these products can help us to identify important signs of change in peoples lifestyle. We start from aggregated data, by selecting products being to 75<sup>0</sup> percentile, that is, products that are been bought at least from the 75% of customers; for each year and for each category we have the purchased quantity, as shown in Table 2.

**Table 2** An example of product quantities aggregated on the year.

Category	Year0	Year1	...	Year(n-1)
bread	5460	6745	...	18271
dried fruit	2900	3036	...	4194
⋮	...	...	...	...
potatoes	5971	5910	...	5553
⋮	...	...	...	...

The analysis begins by performing a clustering step of products data concerning the first year, illustrated by *K-Means* node in the figure; then the resulting model of this first step is used to group data of the others years (illustrated by *Clustering Assigner* nodes in the figure). At the end of the process to each product is associated the list of clusters that it passed through over the years. Table 3 illustrates this result where the first clustering step was performed by finding  $k$  clusters. The value of  $k$  was chosen by considering the trend of the SSE value. Our methodology adopts a convection regarding the name of the resulting clusters: we assign them numerical labels so that the cluster corresponding to the lowest values of the amount of product purchased is the one with the number 0, the one corresponding to the highest values both the one with the number  $n$ , in the interval between these minimum and maximum values there are any other cluster labels.

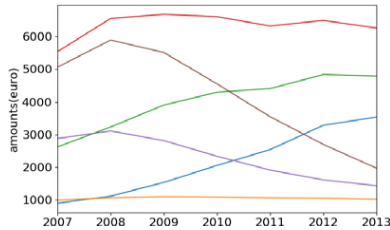
The ultimate aim is to know if there are any products that could reveal interesting behavior hidden in purchases that customers made. We can assign labels to clusters obtained by numbering them from ones containing products purchased in smaller quantities to those purchased in greater quantities; this agreement helps us to identify those groups of products that, over the years under analysis, have remained constant or have been purchased gradually more or, on the contrary, they have been bought for less.

**Table 3** An example of products trend over the years regarding the products purchased by customers.

Category	Year0	Year1	...	Year(n-1)
bread	cluster(k-2)	cluster(k-2)	...	cluster(k-1)
dried fruit	cluster(k-1)	cluster(k-1)	...	cluster(k-1)
⋮	...	...	...	...
potatoes	cluster(k-1)	cluster(k-1)	...	cluster(k-1)
⋮	...	...	...	...

### 3 Case study

We observed purchases of about 13000 customers during 2007-2013 by analyzing several attributes describing the way in which they have been shopping in a store of a big supermarket. The dataset on which has been held the first stage of the analysis that we performed has 39192 lines, corresponding to 13064 customers for seven years from 2007 to 2013. For each customer and for each year, we have information about the number of expenses, quantity of products purchased and amounts spent. A preliminary analysis of the data showed that the information useful for this first study on the purchasing behaviors are those relating to customers who in each year have a total amount of spending under 10000 euros. From the initial dataset we removed 1048 customers, obtaining a final dataset of 10095 customers.



**Fig. 2** : Trajectories of the centroid clustering on annual data, by amount, with  $K = 6$ .

According to the previous sections, we performed three  $K$ -means clustering analysis, respectively, on data concerning amounts spent, quantities of products purchased and the number of expenses; Figure 2 shows the results obtained from the analysis about the amounts of expenses, with  $K = 6$ . We observed and investigated some interesting behaviours-groups respect to the annual amounts: LC Low Constant (yellow line, 2485 customers) representing the group of customers who in the years spent low constant amounts; LG Low Growing (light blue line, 1580 customers) representing the group of customers who made purchases by spending low increasing

amounts. At least, MG Medium Growing (green line, 1527 customers) that represents the group of customers who made purchases by spending medium increasing amounts. According to the section 2.3, for each of the groups of customers selected by the result of the previous phase, we chose the products that were purchased from most customers, that is those being to 75<sup>0</sup> percentile. With this choice, for each of the three behaviours-groups we selected, we are dealing with about 100 categories of products; it is important to note that we are using a merchandise classification reaching down to the details of the product category, in the food sector, related to 400 products categories. By analyzing the trend of products categories purchased from LC customer, we obtained that many products quantities remain constant in the period, but we observe a particular behaviour of some **senry** products: *elaborate red meat* and *slice salumi takeaway* decrease, while *internal production bread* increases. The trend of products categories purchased from LG customer put in evidence the same **senry** products, but with some differences: *internal production bread* increases, *slice salumi takeaway* decreases, *elaborate red meat* decrease, but in this case in a *lightly* way. For the MG group of customers, the trend of products categories purchased instead shows a different behaviour: *elaborate red meat* decrease, *internal production bread* increases, *slice salumi takeaway* remains constant and *savory snacks* decreases. We used the colors green, red and yellow, only visible in the electronic version of this paper, to correlate the results to what we can observe through some corresponding color maps. We precise that color maps can be produced by considering only the values of selected purchased products for each customers group under analysis. We validate the results of the analysis on customer profiles, by considering the entropy of price. In particular, for each customer, has been calculated the variation in price of goods in the basket. The entropy of price, calculated between 0 and 1, suggests how much an individual has a stable expenditure (low entropy) or variable (high entropy). For the three groups of customers corresponding to LC, LG and MG cluster, we obtained that, for what concerns the LC group, we measured a higher but constant variability, which can mean a search for continuous offers. For LG and MG groups we found an increase in the annual expenditure, showing an almost constant trend index of entropy, that means a less attention to the price of products purchased. We also calculated the entropy of price for the customers groups corresponding to the red line in Figure 2 and for the decreasing lines brown and violet; for the red group we observed a minor change in the price, indicating a customer loyalty on the chain. Brown and violet groups, related to customers who, due to the crisis, spend less annually, show an increase in entropy, signifying a change of the basket in terms of price; with high probability there is more attention to the price than to the quality of purchased products. We combined and compared our results with some official statistics, that are in accord with our results; in particular, such as in [10, 11, 12], we analyzed the trends of some official indicators and indexes related to the city and to the region to which data analyzed in our case study refer. The period corresponding to the object of our analysis and that is monitored only by two temporal points (census surveys of 2010 and 2011), presents no big changes in the indicators considered, unless a slight increase in the employment rate; besides, we observe that there are no information

about the indicators considered in the period between the two censuses, where there could be fluctuations are important.

## 4 Conclusions and future works

In This study, that has to be seen as a phase in the definition of indicators that can measure the wellbeing, we explored how customers change their buying patterns and we found out important signals putting in evidence a crisis that is also reflected in purchasing of essential goods. Sometimes customers opted to buy cheaper products, in other cases someone decided to reduce the purchase of certain products for the benefit of others. We are interested to understand the reason for which customers behaviours change: it can be because the shops network change or because people generally start to eat less a food, for example the meat.

**Acknowledgements** This work is supported by the European Community's H2020 Program under the scheme 'INFRAIA – 1 – 2014 – 2015: Research Infrastructures', grant agreement #654024 'SoBigData: Social Mining & Big Data Ecosystem'. (<http://www.sobigdata.eu>).

## References

1. R. Campagni, D. Merlini, and M.V. Verri. An Analysis of Courses Evaluation Through Clustering. In *Zvacek S., Restivo M., Uhomoihi J., Helfert M.. Computer Supported Education*, pages 211–224 Springer International Publishing, ISBN:978-3-319-25767-9, 2015.
2. F. Maggino. Measuring wellbeing of nations: challenges, needs and risks in defining indicators. *Social Indicators Research journal*. Springer, 2015, vol. 6, pages 7–11, ISSN:2037-4186.
3. F. Maggino. Introduzione. In *P. Corvo, G. Fassino. Quando il cibo si fa benessere. Alimentazione e qualità della vita.*, 2015, pages 9–10 Franco Angeli, ISBN:8891713155.
4. F. Maggino. The good society: defining and measuring wellbeing, between complexity and limits. *JOURNAL DE CIENCIAS SOCIALES.*, 2014, vol. 1, pages 20–36, ISSN:2362-194X Accesso ONLINE all'editore.
5. D. Pennacchioli, M. Coscia, S. Rinzivillo, D. Pedreschi, and F. Giannotti. Explaining the product range effect in purchase data. In *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, pages 648–656, 2013.
6. D. Pennacchioli et al. The retail market as a complex system.. *EPJ Data Science.*, 2014, pages 3–33.
7. P. N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.
8. I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Third Edition, Morgan Kaufmann, 2011.
9. KNIME. <https://tech.knime.org/files/KNIME/quickstart.pdf>.
10. ISTAT [http://ottomilacensus.istat.it/download-dati/confini\\_09.csv](http://ottomilacensus.istat.it/download-dati/confini_09.csv).
11. ISTAT [http://dati.istat.it/DCCN\\_TNA\\_Data\\_3c97fbf3-e523-41ca-8e87-7c3f0f6d8730.csv](http://dati.istat.it/DCCN_TNA_Data_3c97fbf3-e523-41ca-8e87-7c3f0f6d8730.csv).
12. ISTAT [http://dati.istat.it/DCCN\\_TNA\\_Data\\_06d9c343-9694-42ca-9575-318e38a96060.csv](http://dati.istat.it/DCCN_TNA_Data_06d9c343-9694-42ca-9575-318e38a96060.csv).



# Assessing Selectivity in the Estimation of the Causal Effects of Retirement on the Labour Division in the Italian Couples

## *Selettività nella Stima degli Effetti Causali del Pensionamento sulla Divisione del Lavoro nelle Coppie Italiane*

Maria Gabriella Campolo and Antonino Di Pino<sup>1</sup>

**Abstract** Analysing the data on the Use of Time in Italy, it emerges that the influence of the latent bargaining process between partners affects for selectivity the estimation of the effect of the man's retirement on the housework time of the woman in the family. We apply a proper estimation procedure in order to estimate the causal effects of retirement on the labour division between partners controlling for the selectivity of bargaining process. The results of a sensitivity analysis confirm the robustness of our estimates.

**Abstract** *Dai dati dell'Indagine Istat sull'Uso del Tempo in Italia emerge che il processo latente di contrattazione fra i partner comporta una stima distorta dell'effetto del pensionamento dell'uomo sulla riduzione del lavoro domestico della donna. Tuttavia, adottando una particolare procedura di matching come stimatore, si possono controllare le stime degli effetti causali del pensionamento dell'uomo dalla selettività del processo di contrattazione. La robustezza delle stime ottenute è confermata dall'analisi sulla sensitività dei risultati.*

**Key words:** Causal Effects, intra-household work allocation, bargaining process, sensitivity of matching estimators

---

<sup>1</sup> Maria Gabriella Campolo, Università di Messina; email: mgcampolo@unime.it  
Antonino Di Pino, Università di Messina; email: dipino@unime.it

## Introduction

The retirement of the male partner in older Italian couples does not seem to lead to a more equitable distribution of housework between partners [7,2,3]. One of the possible explanations is that Italian married men have a strong bargaining power and leave most of the housework to their wives, even after retirement. However, being the bargaining between partners strongly dependent on their latent cultural and psychological characteristics, the influence of the bargaining process on the relationship between retirement and partners' housework division is generally misspecified. Misspecification of bargaining involves a "reverse-causality" effect, that is the latent endogenous influence of bargaining leads to an overstatement of the effect of the man's retirement on the time devoted to housework by a woman with a higher bargaining power and, conversely, leads to an understatement of the effect of the man's retirement on the housework time of a woman with lower bargaining power. In this study, we try to solve this problem taking into account the extent to which the endogenous component of the bargaining process influences the causal effects of retirement on the housework time of both partners and correcting the estimation results accordingly.

To estimate causal effects of retirement, we apply a propensity-score matching procedure to compare the housework time of the couples in which the male partner is retired and the housework time of the couples in which the male partner is not retired. In order to control the estimated propensity to retire by the latent influence of bargaining, we perform a Bivariate-Probit (*Biprobit*) regression model [6], in which the two binary response-variables are given, respectively, by the decision to retire of the male partner (Retirement Equation) and by the satisfaction expressed by the woman (if she is satisfied or not) with labour division within the couple (Satisfaction Equation). The woman's satisfaction with labour division is here assumed as a proxy of the woman's bargaining power. A Maximum Likelihood estimator allows us to correct *Biprobit* estimates for the influence of the correlation between the error terms of the two equations such as a Seemingly Unrelated Regression model (*SUR*), being this correlation assumed as a measure of the endogenous influence of the bargaining power of the woman on the man's decision to retire.

The result to be obtained after the correction of the estimated propensity scores for the cross-correlation in the error terms is that differences in latent characteristics of individuals who decide to retire (treated) and of individuals that choose not to retire (untreated) can be considered not relevant for the propensity score estimation and for the results of matching ("ignorability" condition). In order to verify if the ignorability condition does hold, we perform a sensitivity analysis to verify the extent to which both estimated propensity scores and estimated causal effects change as a consequence of the influence on matching results of a simulated "confounding" covariate, introduced in the regressors set of *Biprobit*.

In the next paragraph we explain how the matching procedure here adopted can be used as an estimator, and we discuss the properties of the estimated parameters in evaluating the causal effects of retirement on partners' working activity. In the third

Section we discuss the estimation results and show how a marked reduction of the woman's domestic activity as a causal effect of the retirement of the male partner is registered prevalently in the couples where the woman is generally satisfied with housework division in the family. Finally, the results of the sensitivity analysis show that controlling matching estimates for heterogeneity due to the latent bargaining process leads to obtain robust findings [8].

## Data and Methods

We compare the value of the observed housework time of a woman,  $y_{1i}$ , whose partner is retired, with the housework time of a woman with the same characteristics, but observed in a counterfactual condition (the male partner is not retired), given by  $y_{0i}$ . We will denote a woman who experienced the partner's retirement as "treated", and a woman who did not experience this event as "untreated". The parameters here considered to evaluate the causal effects of retirement of man on women's domestic work are the Average Treatment Effect (*ATE*), the Average Treatment Effect on Treated (*ATT*), and the Average Treatment Effect on Untreated (*ATU*). We assume that the observed variables  $Z_i$ , influencing the propensity to retire are the same for treated and untreated, while the decision to retire is indicated by the binary dummy  $R_i(0;1)$ , with  $R_i = 1$  signalling if the male partner is retired. Adopting the simple matching estimator of *ATE* based on propensity score, we compute

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n [(\hat{y}_{1i} - \hat{y}_{0i})Z_i] \quad (1)$$

The estimator (1) can be easily modified conditioning the differences  $y_{1i} - y_{0i}$ , respectively, to  $R_i = 1$  (*ATT*) and to  $R_i = 0$  (*ATU*).

The estimated propensity of the male partners to retire, obtained by the *Biprobit* estimation, are used to perform the matching procedure. That is, we match women of the two groups (partner retired or not). Modelling both retirement equation and perceived fairness equation as a *Biprobit* model, we assume that the decision of man to retire,  $R_i$  and the perceived-fairness of woman,  $S_i$ , are specified as follows:

$$R_i^* = \mathbf{z}'_{ri} \boldsymbol{\beta}_r + u_{ri} \quad (2)$$

With  $R_i = 1$  (partner retired) if  $R_i^* > 0$ , and  $R_i = 0$  (partner not retired)

$$S_i^* = \mathbf{z}'_{si} \boldsymbol{\beta}_s + u_{si} \quad (3)$$

With  $S_i = 1$  (woman satisfied with the housework division) if  $S_i^* > 0$ , and  $S_i = 0$  (woman dissatisfied).  $\mathbf{z}'_{ri}$  and  $\mathbf{z}'_{si}$  are, respectively, row vectors of the matrices  $Z_r$  and  $Z_s$  of the observable variables conditioning, respectively, the propensity of man to retire and the perceived fairness of woman.  $\boldsymbol{\beta}_r$  and  $\boldsymbol{\beta}_s$  are vectors of coefficients. We assume that the error terms  $u_{ri}$  and  $u_{si}$  are normally distributed  $N(0, \boldsymbol{\Sigma})$ . The correlation between the error terms can be considered as a measure of the latent influence of bargaining on the decision of man to retire. The Equations (2) and (3) are estimated performing a Maximum Likelihood (*ML*) procedure (as in a *SUR* model) taking into account the endogenous influence of bargaining measured by the errors correlation. We use the predicted propensities to retire, provided by *Biprobit*

regression, to apply the Simple Matching Estimator (*SME*), above reported by Eq. 1. In addition, in order to avoid the selectivity bias due to the unbalancing in the observed covariates,  $Z_r$  and  $Z_{-s}$ , which condition the matching, we perform also a “Bias-Corrected” matching estimator (*BCME*) and a “Stratification-Matching” (*SM*) estimator [8,1,9]. For this study we use cross-sectional microdata selected from the 2008-2009 ISTAT Survey on Time Use in Italy, in which the use of time is surveyed with the diary method. The selected sample is composed of No. 3,126 elderly women living with their male partners (aged 50-66), equitably distributed by area of residence. Male partners of the selected couples are employed (equal to 2,096) or retired (equal to 1,030).

## Estimation Results

Estimation results obtained performing *Biprobit* model (Table 1) show that the retirement of man, more frequent in the North-Centre of Italy, is negatively related to his education level. In addition, the retirement of man is positively related with the retirement of the woman.

Table 1: *Biprobit* estimation results

<i>Dependent variables:</i>	<i>S:</i> <i>Woman's satisfaction in housework division (dummy: 1= satisfied)</i>	<i>R:</i> <i>Man's retirement decision (dummy: 1= retired)</i>
<i>Explanatory variables:</i>	coef.	coef.
Intercept	2.96*	-14.05***
Education of woman (years of schooling)	0.01	0.02
Education of man (years of schooling)	0.01	-0.06***
Religiosity: 1 if the woman attends church <sup>a</sup>	0.11*	
Children living in the family: 1 yes <sup>a</sup>	0.05	0.11
Worried: 1 if the woman feels in trouble for his work <sup>a</sup>	-0.27***	-0.15*
Woman's Economic Dependency <sup>a</sup>	0.07	0.27*
Age of woman	-0.12*	0.09
Age <sup>2</sup> of woman	0.001*	0.0001
Age of man	0.01	0.20***
Area of residence: 0= North-Centre; 1= Southern regions <sup>a</sup>	-0.10*	-0.32***
Help received in paid form: 1 = yes <sup>a</sup>	-0.08	-0.43**
Health: 1 = Sick <sup>a</sup>		0.17*
Woman retired: 1= yes <sup>a</sup>		0.48***
Retirement Eligibility of man ( <i>Eligibility</i> ) = 1 if he is 58 years old, at least <sup>a</sup>		0.49***
Eligibility *(Age-58)		0.09
[Eligibility *(Age-58) <sup>2</sup>		-0.08
[Eligibility *(Age-58) <sup>3</sup>		0.01*

Note: Estimated correlation between error terms,  $\rho = 0.19$  (LR test on  $\rho = 0$ :  $\text{Chi}^2 = 26.29$ ;  $p < 0.0001$ ). P-value: \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ ; <sup>a</sup> Dummy Variable

Note that the estimated correlation between the error terms of both Retirement and Satisfaction equations ( $\rho = 0.189$ ) is positive and significant. This confirms that an endogenous relationship between retirement of man and bargaining occurs.

Our results indicate a partial reallocation of the intra-household housework time in favour of the woman. In fact, the estimation of the treatment parameters show a reduction of the commitment of woman in domestic work and a simultaneous increase of the commitment of the male partner (cf. Table 2). More in detail, the estimation results on the full sample show a modest reduction of the housework time of woman as an effect of the retirement of the male partner. Considering the average of the results obtained by applying the three estimators (*SME*, *BCME* and *SM*), we have, as an estimated *ATE*, a reduction in the woman's housework of about 17-25 minutes per day and a reduction of about 17-25 minutes as an estimated *ATT*. Contextually, the man's housework increases by an average of 90 minutes per day as reported by the *SME* estimates of both *ATE* and *ATT* (without bias correction). However, markedly lower results have been obtained applying *BCME* and *SM* estimators (approximately 60-70 minutes). The estimates reported in table 3 show that women with higher bargaining power (satisfied with labour division) generally obtain a higher reduction of the housework time than women with lower bargaining power (dissatisfied with labour division). The reduction of housework time estimated by *SM* indicates a difference of about 20 minutes in a day in favour of the women with higher bargaining power.

Table 2 - Estimated causal effects of man's retirement on housework time (minutes in a day).

Matching estimators:		ATT	SE	ATU	SE	ATE	SE	HI
Domestic work of woman	SME	-19.94	8.72	-24.89	8.38	-22.52	8.37	0.41
	BCME	-17.38	14.36	-15.83	15.33	-16.66	13.18	-0.07
	SM	-25.03	22.20	-25.44	22.70	-25.16	20.37	0.01
Domestic work of man	SME	91.55	7.31	86.22	6.92	88.77	6.95	0.53
	BCME	58.38	12.60	85.94	13.44	71.13	11.59	-1.50
	SM	72.33	17.09	63.40	16.95	67.88	15.13	0.37

Note: *HI* = Heterogeneity Indicator:  $HI = (ATT - ATU) / SE_{(ATT - ATU)}$

Table 3 - Estimated causal effects on housework time (minutes in a day) by satisfaction of the woman with labour division within the family

Matching estimators:	Woman satisfied						Woman dissatisfied						
	ATT	SE	ATU	SE	ATE	SE	ATT	SE	ATU	SE	ATE	SE	
Domestic work of woman	SME	-18.17	13.37	-23.23	15.06	-20.77	11.75	-14.39	14.31	-23.55	12.28	-19.24	13.08
	BCME	-30.98	15.61	-24.56	14.99	-27.68	13.65	-7.9	17.64	-22.85	18.25	-15.82	16.08
	SM	-37.05	28.98	-36.09	26.93	-36.61	27.58	-6.95	35.54	-10.67	34.79	-8.36	32.03
Domestic work of man	SME	96.5	10.58	101.55	9.03	99.1	8.73	74.39	12.64	65.41	12.61	69.63	10.13
	BCME	63.73	12.53	64.91	12.51	64.34	11.18	48.71	17.07	52.58	15.49	50.76	14.53
	SM	64.14	24.36	65.73	23.98	64.95	20.94	55.64	27.3	59.73	29.61	57.41	26.53

### 1.1 Sensitivity Analysis

We follow a parametric approach to sensitivity analysis in order to test the effect of a confounder variable on the estimated propensity scores and on the treatment parameters, such as *ATT* [4,5]. In particular, we replicate estimates of both propensity scores and *ATT*, by including a simulated endogenous confounder in the

regressors set of the *Biprobit* regression, with the purpose to violate the condition of Conditional Independence assumption (*CIA*). In doing this, a confounder covariate,  $U_i$ , is simulated using the predicted values of a linear regression of the outcome variable  $y_i$ , (housework time) on the covariates which condition the propensity score. We draw random values from the confidence intervals of the regression coefficients in order to generate the confounder and replicate the estimation procedure (No. of replications = 1000). In Table 3 we report the results of sensitivity analysis.

Table 4 - Sensitivity analysis of *SME* estimation using a confounder variable

	Confounder in Retirement eq		Confounder in Satisfaction eq		Confounder in both eqs	
	mean	SE	mean	SE	mean	SE
Wilcoxon Signed-Ranks Test on matched-pairs of propensity score values	0.189	0.016	-0.189	0.228	0.192	0.018
<b>ATT for woman's domestic work</b>	<b>-24.696</b>	<b>0.118</b>	<b>-20.456</b>	<b>0.045</b>	<b>-24.662</b>	<b>0.120</b>
Student-t test on paired causal effects	-0.233	0.013	0.081	0.013	-0.225	0.013

Note: value for the *ATT* estimates using *SME*: **Mean= -19.935**; **SE=9.135**; 95% Conf. Interval (-37.875;-1.995)

Sensitivity analysis confirms the robustness of matching procedure using *Biprobit*. The average *ATT* computed on No. 1000 replications of *SME* procedure using confounder does not differ substantially with respect to the estimated value without confounder. In particular, the treatment parameter *ATT* does not change significantly as a consequence of the endogeneity of the woman's perceived fairness (endogeneity of bargaining), simulated by introducing the confounder in the Satisfaction equation.

## References

1. Abadie, A. and Imbens, G.: Bias-Corrected Matching Estimators for Average Treatment Effects. *J. Bus. Econ. Stat.* (2011) doi:10.1198/jbes.2009.07333
2. Caltabiano, M., Campolo, M.G., Di Pino, A.: Retirement and intra-household labour division of Italian couples: A new simultaneous equation approach. *Soc. Indic. Res.* 128(3), 1217–1238 (2016)
3. Ciani, E.: Retirement, pension eligibility and home production. *Labour Econ.* (2016) doi:10.1016/j.labeco.2016.01.004
4. Copas, J.B., Li, H.G.: Inference for non-random samples. *J. Roy. Stat. Soc. B.* 59(1), 55–95 (1997)
5. Frank, K.A.: Impact of a confounding variable on the inference of a regression coefficient. *Sociol. Method. Res.* 29(2), 147–194 (2000)
6. Greene, W.H.: *Econometric Analysis*. Prentice Hall, New Jersey (2003)
7. Mills, M., Mencarini, L., Tanturri, M.L., Begall, K.: Gender equity and fertility intentions in Italy and the Netherlands. *Demogr. Res.* 18(1), 1–26 (2008)
8. Rosenbaum P.R., Rubin D.B.: Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* 79, 516–524 (1984)
9. Xie, Y., Brand, J.E., Jann, B.: Estimating Heterogeneous Treatment Effects with Observational Data. *Sociol. Methodol.* 42(1), 314–347 (2012)

# Composite indicators for ordinal data: the impact of uncertainty

## *Indici composti per dati ordinali: l'impatto dell'incertezza*

Stefania Capecchi and Rosaria Simone

**Abstract** Composite indicators are becoming one of the most prominent analysis tools, especially in social sciences where the need arises to compare and rank groups of respondents by managing huge and diversified amounts of data. The aggregation of information is a powerful yet incomplete operation since it usually disregards of accounting for *uncertainty*. Uncertainty is here meant as the inherent indeterminacy of any decision process, specifically with reference to the discrete-choice process yielding interviewees to provide an ordinal evaluation out of their latent perception. The class of CUB mixture models for ordinal data is grounded on the probabilistic specification of this component, thus establishing a direct control for heterogeneity. Empirical evidence and methodological studies set this framework as an effective statistical modeling among well-known consolidated theories. In this setting, our contribution proposes a technique to build model-based composite indicators that discloses the role of uncertainty also at an aggregated level. The presentation is lead by applications to real data and comparisons with existing methods.

**Abstract** *Gli indici composti sono spesso annoverati tra i più rilevanti strumenti di analisi, specialmente nell'ambito delle scienze sociali in cui emerge la necessità di confrontare e classificare gruppi di rispondenti gestendo grandi quantità di dati. L'aggregazione delle informazioni è un'operazione complessa e può risultare incompleta se prescinde dalla considerazione dell'incertezza. Qui per incertezza si intende l'intrinseca indeterminazione di ogni processo decisionale. In riferimento al processo di scelta discreta che porta gli intervistati ad esprimere una valutazione ordinale della loro percezione latente, la classe dei modelli CUB per dati ordinali è basata sulla specificazione probabilistica di questa componente, permettendo così anche un controllo diretto dell'eterogeneità. Evidenza empirica e studi metodologici rendono questa modellistica un'alternativa efficace tra teorie ben consolidate. In questo contesto, il nostro contributo propone una tecnica di costruzione di indici*

---

Department of Political Sciences, University of Naples Federico II, Via Leopoldo Rodinò 22, I-80138 Naples, Italy  
e-mail: stefania.capecchi@unina.it e-mail: rosaria.simone@unina.it

*compositi basata su modelli che rivelano il ruolo dell'incertezza anche a livello aggregato. La presentazione è guidata da applicazioni a dati reali e confronti con metodi esistenti.*

**Keywords: Uncertainty; Model-based Composite Indicators; Ordinal Data; CUB Models**

## 1 Introduction

Composite indicators are one of the most prominent tools in social sciences able to synthesize several information on a specific topic [3], as in well-being measurements, for instance [7]. Indeed, in the Big Data era, the introduction of composite indicators allows to summarize efficiently complex and multi-dimensional issues and to reduce the size of the available list of indicators [12].

Generally, composite indicators concern official data [8] and, in this respect, several proposals are collected [9] and promoted [6]; thus, uncertainty and sensitivity analysis are suggested for ensuring robustness and reliability of the results [12]. Then, the main burden is to discard as little information as possible in the synthesis.

In this area, our contribution concerns indicators computed on the basis of ordinal data arising from surveys where people are asked to manifest their perception with a rating over a set of discrete choices [13]. These data are frequent in several scientific fields as, for instance:

- *University evaluation*: scores are collected for investigating characteristics of both teaching and structures.
- *Elderly well-being*: ratings concern medical, physical and mental abilities.
- *Customer satisfaction*: many aspects of the relationship between clients and company are examined to investigate loyalty, for instance.

In the present work we assume that respondents express their ratings according to CUB mixture models [10, 2, 5]. This modelling approach prescribes that responses stem from the combination of two main components driving the decision process, named as *feeling* and *uncertainty*, and it has been successfully applied to analyze ratings on opinions, judgments and preferences in several disciplines. Here we propose a strategy to maintain these two main components also at an aggregated level by presenting a model-based composite indicator. Due to space constraints, we defer any unspecified detail to references and we limit to recall that the characterizing uncertainty parameter ( $\pi$ ) may be interpreted as a distance from a completely random choice, and that each CUB model may be uniquely represented as a point in the unit square.

The paper is organized as follows: in the next section, the new proposal is introduced and in Section 3 some empirical evidence is discussed. Final considerations about further generalizations and developments are summarized in the concluding remarks.



## 2 A new proposal

Consider a questionnaire designed to measure a latent trait (such as teacher’s performance, customer satisfaction, etc.) via  $K$  observable variables  $R_1, \dots, R_K$  collected on an ordinal scale (say, with  $m$  categories). Let  $\|r_{ik}\|$ , for  $i=1,2,\dots,n$  and  $k = 1,2,\dots,K$  the matrix of the responses given to  $K$  items by the  $n$  respondents. Then,  $\mathbf{r}_i = (r_{i,1}, \dots, r_{i,K})$  is the row-vector of observations on the  $i$ -th subject,  $i = 1, \dots, n$ .

According to a model-based approach, we assume that a CUB model fits the data in an effective parametric way; thus,  $R_k \sim \text{CUB}(\hat{\pi}_k, \hat{\xi}_k)$ , for  $k = 1, \dots, K$ , where:

$$Pr(R_k = r | \pi_k, \xi_k) = \pi_k \binom{m-1}{r-1} \xi_k^{m-r} (1 - \xi_k)^{r-1} + (1 - \pi_k) \frac{1}{m}, \quad r = 1, \dots, m.$$

Then, we propose a weighted CUB model  $\tilde{R} \sim \text{CUB}(\tilde{\pi}, \tilde{\xi})$ :

$$\tilde{\pi} = \sum_{k=1}^K w_k \hat{\pi}_k, \quad \tilde{\xi} = \sum_{k=1}^K w_k \hat{\xi}_k \tag{1}$$

as a 2-dimensional composite indicator for the latent trait (Composite Indicator CUB model, CI-CUB, for short). This choice allows to take both uncertainty and feeling into account by assigning higher weights to the most relevant items (as meant, for instance, by PCA).

Customarily, composite indicators are computed on individual basis by ranging the data matrix per rows. Classical proposals include some average operations (arithmetic, geometric, harmonic) of the individual ratings, or the selection of the first component of a principal component analysis (PCA) performed to the data matrix, say  $Y_1$ :

$$Y_1 = a_1 R_1 + \dots + a_K R_K,$$

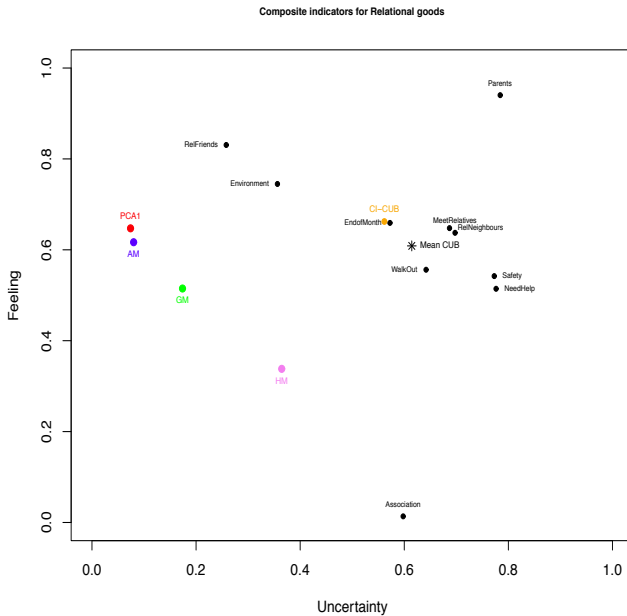
with weights  $a_1, \dots, a_K$  such that:  $\sum_{k=1}^K a_k^2 = 1$ , and set  $w_k = a_k^2$ .

Then, in order to get an overall assessment of the latent trait under investigation, their distributions should be suitably taken into account. For instance, a new variable ranging from 1 to  $m$  can be obtained for each of them with a (uniform) discretization over  $m$  categories. In this way, comparisons with the CI-CUB proposal can be enhanced by fitting a CUB model to each of the resulting variables and the corresponding estimated parameter vectors  $(\hat{\pi}, \hat{\xi})$  can be considered as model-based composite indicator itself. For instance, if we consider the discretized version of the arithmetic mean  $R_A$ , we assume that  $R_A \sim \text{CUB}(\hat{\pi}_A, \hat{\xi}_A)$  is an adequate model for the arithmetic mean composite indicator. Similarly, for the first PCA component and the other average operators.

### 3 Empirical evidence

We experiment the proposed approach (1) on the data set `relgoods` referred to a survey on well-being and relational goods (available in the package `CUB` in R). As for any measure of subjective perception, data related to personal awareness always come with some remarkable caveats. Many studies have discussed the reliability of self-reported measures, even with respect to frame-of-reference effects and adaptations to life events: see [11], among others. Moreover, to detect progress and human (well-being and/or) “good-life” it is necessary to build a set of reliable indicators to make the information understandable to stakeholders [7].

Figure 1 displays a multiplot of CUB models of the selected items, the estimated CUB models derived from the arithmetic (*AM*), geometric (*GM*), harmonic (*HM*) averages, and from the first component of the *PCA1*, as explained in Section 2. This scatterplot is obtained by plotting estimated uncertainty  $1 - \pi$  against estimated feeling  $1 - \xi$  for each model (weights for the CI-CUB here have been derived from the first principal component). In addition, the mean CUB model (*MeanCUB*) obtained by simply averaging both the estimated  $\pi$ 's and  $\xi$ 's parameters (constant weights) is shown.



**Fig. 1** MultiCUB for relational goods and related composite indicators

Differently from other proposals, From this visual inspection, the CI-CUB proposal affords a more adequate aggregation of information since it coherently preserves both uncertainty and feeling. Indeed, it reports the explicit evidence of a possible heterogeneity in the responses. Conversely, *PCA1*, *AM*, *GM* and *HM* give a rather biased synthesis of the data since the model-based versions of these composite indicators is farther from the estimated data models and loose to catch the uncertainty component.

### 4 Comparing groups and individuals

Customarily, composite indicators are exploited to compare and rank different groups with respect to the investigated phenomenon (countries, departments of universities, teachers, clusters of respondents identified by covariates, and so on). To pursue this task according to the CI-CUB proposal for ordinal responses, assume that data are gathered into  $H$  groups.

Then, for  $h = 1, \dots, H$ , consider the methodology described above to build a CI-CUB  $\tilde{R}_h \sim \text{CUB}(\hat{\pi}_h, \hat{\xi}_h)$ :

- Let  $\mathbf{r}_i^{(h)} = (r_{i,1}^{(h)}, \dots, r_{i,K}^{(h)})$  denote the vector of observations of the  $i$ -th subject (unit) within the  $h$ -th group,  $i = 1, \dots, n_h, h = 1, \dots, H$ .
- For every  $h = 1, \dots, H$  and  $k = 1, \dots, K$  fit a CUB model to responses  $R_k$  in the  $h$ -th group, resulting in  $R_k^{(h)} \sim \text{CUB}(\pi_k^{(h)}, \xi_k^{(h)})$ .
- Then, for a suitable system of weights, consider the *CI-CUB* model  $\tilde{R}^{(h)} \sim \text{CUB}(\tilde{\pi}^{(h)}, \tilde{\xi}^{(h)})$ , with:

$$\tilde{\pi}^{(h)} = \sum_{k=1}^K w_k \hat{\pi}_k^{(h)}; \quad \tilde{\xi}^{(h)} = \sum_{k=1}^K w_k \hat{\xi}_k^{(h)}.$$

Since there is not a unique ordering of two-dimensional vectors, if the (latent) phenomenon under investigation is positive in the direction of the scale, we suggest to compare and rank the  $H$  groups according to the composite indicator  $Pr(\tilde{R}^{(h)} \geq m^* | \tilde{\pi}^{(h)}, \tilde{\xi}^{(h)})$ , where  $m^* \leq m$  is a threshold category lower-bounding the *positive* responses. The reverse direction applies if the (latent) phenomenon under investigation is negative in the direction of the scale. If, instead, an individual composite indicator is more suitable for the analysis, a model-based proposal stemming from the setting here developed is to consider:

$$I_i = \sum_{k=1}^K w_k Pr(R_k = r_{i,k} | \hat{\pi}_k, \hat{\xi}_k), \quad i = 1, \dots, n.$$

## 5 Conclusions

As confirmed in the selected case-study, commonly acknowledged choices to build a social indicator based on averages and PCA are highly data-dependent. Then, if the first principal component is not really explanatory (for instance, the first principal component captures only about 27% of the variability for the selected items of the case study), the resulting index cannot be assumed for the latent trait under examination. In addition, they completely loose to account for the uncertainty component, thus the standard approaches waste an important amount of information. Conversely, the model-based approach that leads to the CI-CUB proposal gives satisfactory performances and it easily lends itself to encompass more refined item-based analysis, as when overdispersion [4] or *shelter effect* [5] are suspected for certain items: this extension and further developments on the selection of optimal weights are left for future works.

## References

1. Arezzo, M.F., Guagnano, G. (2014), Il problema della valutazione mediante indicatori compositi in presenza di correlazione tra gli indicatori elementari, Pubblicazioni Dipartimento MEMOTEF, Sapienza Universit di Roma, - The Future of Europe, Patron Editori Bologna, 271–279.
2. D'Elia, A., Piccolo, D. (2005), A mixture model for preference data analysis, *Computational Statistics & Data Analysis*, 49, 917–934.
3. Giambona F., Vassallo, E. (2014), Composite indicator of social inclusion for European countries, *Social Indicators Research*, 116. 269–293.
4. Iannario, M. (2014), Modelling Uncertainty and Overdispersion in Ordinal Data, *Communications in Statistics. Theory and Methods*, 43: 771–786.
5. Iannario, M., Piccolo, D. (2016), A generalized framework for modelling ordinal data, *Statistical Methods and Applications*, 25, 163–189.
6. Maggino, F. (2009), *The state of the art on indicators construction in the perspective of a comprehensive approach in measuring well-being of societies*, Firenze University Press, Archivio E-Prints, .
7. Maggino, F. (2016), *Challenges, needs and risks in defining well-being indicators*, in: F.Maggino ed., *A Life devoted to Quality of Life. Festschrift in Honor of Alex. C. Michalos*, Springer, Switzerland, pp.209–233.
8. Mazziotta, M., Pareto, A. (2016), Methods for Constructing Non-compensatory Composite Indices: A Comparative Study, *Forum for Social Economics* 45:2-3, 213–229.
9. Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., Giovannini, E. (2005), *Handbook on Constructing Composite Indicators: Methodology and User guide*, OECD Statistics Working Paper. Revised edition, 2008.
10. Piccolo, D. 2003, On the moments of a mixture of uniform and shifted binomial random variables, *Quaderni di Statistica*, 5, 85–104.
11. Ravallion, M. (2012), Poor or just feeling poor? On subjective data in measuring poverty, *Policy Research Working Papers*, n.5968, The World Bank, Washington, D.C.
12. Saisana, M., Saltelli, A., Tarantola, S. (2015), Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators, *Journal of the Royal Statistical Society, Series A*, 168(2), 307–323.
13. Zanarotti, M.C., Pagani, L. (2015), Some considerations to carry out a composite indicator for ordinal data, *Electronic Journal of Applied Statistical Analysis*, 8(3), 384–397.

# The distribution of Net Promoter Score in socio-economic surveys

## *La distribuzione del Net Promoter Score nelle indagini socio-economiche*

Stefania Capecchi and Domenico Piccolo

**Abstract** In marketing studies devoted to customer satisfaction and consumers' loyalty analysis, Reichheld (2003, 2006) proposed the Net Promoter Score (NPS) to synthesize by means of an excess index the distribution of the sample responses to a question as: "How likely is that you would recommend our Company/Institution to a friend or colleague?", on an ordinal scale from 1 to 10. More specifically, this measure is obtained by the difference between the proportion of "enthusiastics" minus that of "passives". This index may be fruitfully exploited in different research fields, of course, where the whole set of information is meant to be summarized by comparing the relative frequencies of "supporters" and "detractors" with respect to products, services, items, etc. Although the literature remarks critical and positive aspects of such an index, only recently Rocks (2016) are faced with inferential procedures with regard to NPS. In this study, we search for the distribution of NPS based on a convenient structure of the response patterns. Indeed, we assume a parametric mixture for the responses and verify the behaviour of NPS over the parameter space.

**Abstract** *Nell'ambito degli studi di marketing, Reichheld (2003, 2006) ha introdotto il Net Promoter Score (NPS), una sorta di misura di eccesso per la distribuzione delle risposte alla domanda, su scala ordinale da 1 a 10: "Quanto raccomanderesti ad un tuo amico o collega la nostra Società/Istituzione?". L'indice è la differenza tra la proporzione dei rispondenti entusiasti e quella dei detrattori. Tale misura può essere proficuamente utilizzata anche in altri ambiti dove è opportuno confrontare la frequenza relativa dei "molto favorevoli" a servizi, prodotti, items, etc. con quella dei "critici". In letteratura si discutono aspetti positivi e critici di tale proposta e recentemente Rocks (2016) ha affrontato la questione da un punto di vista inferenziale. In questo lavoro, esaminiamo la distribuzione dell'indice NPS sotto l'ipotesi*

---

Stefania Capecchi and Domenico Piccolo

Department of Political Sciences, University of Naples Federico II, Via Leopoldo Rodinò, 22, I-80138 Napoli, Italy,

e-mail: stefania.capecchi@unina.it; domenico.piccolo@unina.it

*che le risposte siano generate da una mistura idonea per le indagini con risposte ordinali riguardanti giudizi e/o opinioni.*

**Keywords:** Ordinal data, Net Promoter Score, Mixture models

## 1 Introduction

Customer satisfaction is one of the most important concern of the companies since this variable summarizes reactions and sentiments of clients. More specifically, loyalty is indicated as a fundamental component to maintain success. Proposals have been introduced to alert companies in order to monitor and predict this key driver.

Among the several syntheses aimed at interpreting the mood of clients, a main question emerged as a signal of confidence and loyalty towards the Company: “*How likely is that you would recommend our Company/Institution to a friend or colleague?*”, with responses on an ordinal scale from 1 to 10. Thus, Reichheld [6, 7] introduced the Net Promoter Score (NPS) as the proportion of extremely favourable respondents minus the proportion of disaffected ones. Notice that NPS is a trademark of Stametrix Systems, Inc., Bain & Company, Inc. and Freid Reichheld.

Briefly, NPS is considered as a customer loyalty metric able to measure bond, endorsement and sponsor support between a provider and a consumer. Although some critical comments, this measure is now regularly applied by thousands of companies. In general, it has become a benchmark for the policy of the companies and its use may be easily extended to the fields of products, services, holidays, financial advisors, banks, diets, sanitary protocols, educational training, and so on.

From a statistical point of view, NPS is an estimate of the mean value of a discrete random variable whose probabilities are generated by a distribution expressing the graduated opinions of a sample of respondents on an ordinal scale, ranging from 1 to  $m$ , for a given  $m$ .

In this paper we show that a large collection of models generate the same NPS. In addition, the (underestimated) uncertainty always present in human decisions as well as the heterogeneity of the respondents may largely affect the NPS value. The framework of the analysis is a model-based approach for the data generating process by which respondents express their judgements about the selected question.

## 2 Notation and formal background

Let  $R$  be a discrete random variable defined for a given  $m$  on the support  $\{1, 2, \dots, m\}$  and able to describe the mechanism of the ordinal responses. Assume that  $R$  is fully characterized by the probability distribution  $p_r = p_r(\theta) = Pr(R = r | \theta)$ , for  $r = 1, 2, \dots, m$ , where  $\theta \in \Omega(\theta)$ . Then, the NPS is defined by:

$$NPS = \sum_{r=b}^m p_r - \sum_{r=1}^a p_r; \quad -1 \leq NPS \leq +1;$$

where  $1 \leq a < b \leq m$  and  $a$  and  $b$  are given integers.

People with scores in  $R \in [1, a]$ ,  $R \in [a + 1, b - 1]$  and  $R \in [b, m]$  are denoted as “detractors”, “passives” and “promoters”, respectively. Thus, we let:  $p_{det} = p_1 + \dots + p_a$ ,  $p_{pas} = p_{a+1} + \dots + p_{b-1}$ , and  $p_{pro} = p_b + \dots + p_m$ . In common analyses,  $m = 10$ ,  $a = 6$ ,  $b = 9$ ; sometimes, the Likert scale starts at  $r = 0$ .

It is immediate to show that NPS coincides with the expectation of a discrete random variable  $X$  defined on the support  $\{-1, 0, 1\}$  with probabilities  $\{p_{det}, p_{pas}, p_{pro}\}$ , respectively. Thus, all the characteristic of this index are specified in the ternary simplex. In particular, according to Huber[2],

$$\mathbb{E}(X) = NPS = p_{pro} - p_{det}; \quad Var(X) = p_{pro} + p_{det} - [p_{pro} - p_{det}]^2.$$

As a mean value, NPS may be effectively estimated by

$$\widehat{NPS} = \sum_{r=b}^m f_r - \sum_{r=1}^a f_r; \quad -1 \leq \widehat{NPS} \leq +1;$$

where  $f_r = n_r/n$  and  $n_r$ , for  $r = 1, 2, \dots, m$ , are the relative and absolute frequencies derived by the sampling distribution  $(n_1, n_2, \dots, n_m)$  of scores, with  $n = n_1 + \dots + n_m$ . Of course, such frequencies are realizations of a Multinomial distribution characterized by  $n$  and  $p_r = p_r(\theta)$ , for  $r = 1, 2, \dots, m$ .

These results imply that  $\widehat{NPS}$  is an unbiased and consistent estimator of NPS with variance  $n^{-1}Var(X)$ . Moreover, the standardized  $\widehat{NPS}$  is asymptotically Normal distributed; thus, tests and confidence intervals may be assessed. A survey of inference for the  $\widehat{NPS}$  estimator, with some improvements, is discussed by [8].

### 3 A model-based approach

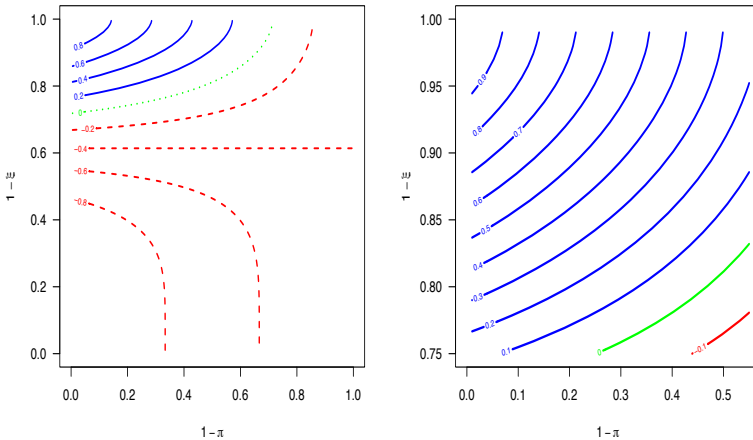
On the basis of experimental evidence and statistical reasoning, we assume that ordinal responses of the customer judgements/opinions are generated by a CUB model as in [5, 1]. More specifically,  $R \sim CUB(\pi, \xi)$  is a random variable defined over the support  $\{1, 2, \dots, m\}$ , for a given  $m$ , whose probability mass distribution is:

$$Pr(R = r | \theta) = \pi \binom{m-1}{r-1} \xi^{m-r} (1-\xi)^{r-1} + (1-\pi) \frac{1}{m}, \quad r = 1, 2, \dots, m.$$

The model is well defined over the parameter space:  $\Omega(\theta) = \Omega(\pi, \xi) = \{(\pi, \xi) : 0 < \pi \leq 1; 0 \leq \xi \leq 1\}$  and it is identifiable for any  $m > 3$ , whereas  $m = 3$  represents a saturated model [3].

Then,  $(1 - \pi)$  increases with the indecision/heterogeneity of the responses whereas  $(1 - \xi)$  increases with the confidence/loyalty of the client. Subjects' covariates are generally linked to parameters by a logistic function for simplicity; however, other mappings are legitimate.

The advantage of this parameterization is the ability to capture different patterns of the observed distributions (in terms of modal values, skewness and flatness) by means of only two parameters  $\theta = (\pi, \xi)'$  which are easily interpreted with respect to the components of the random process [4]. In addition, any CUB model admits a visual representation as a point in the parameter space  $\Omega(\theta)$ . Then, the introduction of subjects' covariates permits to investigate if and how the individual characteristics of the respondents affect the expressed opinions.



**Fig. 1** NPS contour plots over the parameter space of a CUB model (left panel). Subregion where NPS is positive (left panel)

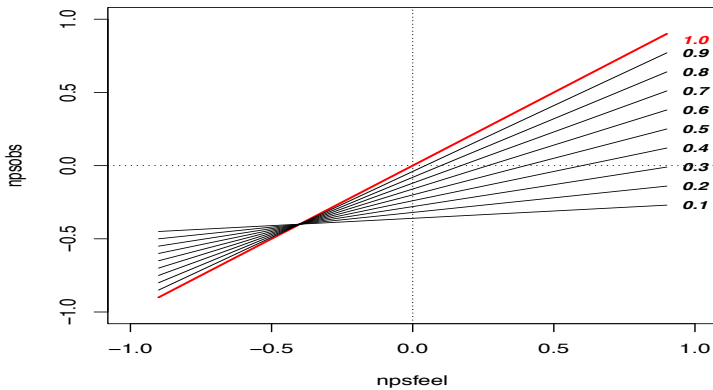
In this line of reasoning, the behaviour of NPS when responses follow a CUB distribution is investigated. The left panel of Figure 1 shows the contour lines for given NPS over  $\Omega(\theta)$  and distinguishes negative, null and positive values of this measure. Then, the right panel magnifies the top-left area of the parameter space where  $NPS > 0$ : that is the area of interest for companies searching for a growth. As it is expected, a rewarding NPS is the consequence of both a moderate uncertainty and a substantially positive endorsement; however, infinitely many CUB models refer to the same NPS.



### 4 The role of uncertainty

In CUB models framework it is important to emphasize that  $(1 - \pi)$  is the weight for the Uniform distribution and thus it measures both personal indecision of respondents and the presence of different reactions in the sample (heterogeneity). This is a twofold meaning which manifests in the same parameter: in fact, if people converge on a category this means a low level of indecision among respondents and it generates a low heterogeneity within the sample. On the contrary, when respondents are fuzzy and quirky they generate high heterogeneity. This component (which may be effectively estimated in sample data) affects also the interpretation of NPS.

As a matter of fact, Figure 1 emphasizes the role of uncertainty in the assessment of the NPS index. For instance, a feeling as high as  $1 - \xi = 0.85$  and a very low indecision/heterogeneity of respondents expressed by  $1 - \pi = 0.10$  generates an  $NPS = 0.50$  whereas an increase of uncertainty up to  $1 - \pi = 0.30$ , say, lowers NPS to 0.25. This implies that a modal value very high in the responses distribution is a necessary but not a sufficient condition to get an appreciable positive NPS.



**Fig. 2** Effect of uncertainty in the specification of NPS

Further insight may be derived from the consideration that the feeling of respondents is modified by uncertainty/heterogeneity. In a sense, we would like to obtain just NPS for the feeling ( $=NPS_{feel}$ , say) but we estimate NPS on the basis of the expressed responses (which are a mixture of both components). A simple algebra proves that:

$$NPS = \pi NPS_{feel} + (1 - \pi) const,$$

where  $const = (m + 1 - a - b)/m$  and, in common cases,  $const = -0.4$ . Except for  $\pi = 1$ , that is a model without uncertainty, the presence of indecision/heterogeneity in the data reduces NPS.

Figure 2 shows how the effect of uncertainty modify the correspondence between the desired  $NPS_{feel}$  and the observed NPS by systematically lowering the second one but for the bisector (where  $\pi = 1$ ).

This effect propagates in a different way when a CUB model with covariates is considered. In the simplest case of a dichotomous variable only affecting the feeling (for instance,  $D_i = 0, 1$  for men and women, respectively), the differential effect of NPS between women and men, with obvious notation, is:

$$NPS^{(1)} - NPS^{(0)} = \pi \left[ NPS_{feel}^{(1)} - NPS_{feel}^{(0)} \right].$$

As a consequence, a possible discrimination between the two clusters would appear attenuated by a factor of  $\pi$  due to the presence of uncertainty.

## 5 Concluding remarks

A widespread experience derived by more than a thousand of observed NPS [8] shows that this measure takes values mostly between 0 and 0.50 with a variance included in  $[0.50, 0.75]$ . As a consequence, effective studies should be concentrated on the NPS distributions putting a large mass on this sub-region of the parameter space.

**Acknowledgements.** This work has been implemented within CUBREMOT project financially supported by University of Naples Federico II.

## References

1. D'Elia, A., Piccolo, D. (2005), A mixture model for preference data analysis, *Computational Statistics & Data Analysis*, 49, 917–934.
2. Huber, W. (2011), How can I calculate margin of error in a NPS (Net Promoter Score) result? available at <http://stats.stackexchange.com/a/18609>.
3. Iannario, M. (2010), On the identifiability of a mixture model for ordinal data, *Metron*, LXVIII, 87–94.
4. Iannario, M., Piccolo, D. (2016), A comprehensive framework of regression models for ordinal data, *METRON*, 74, 233–252.
5. Piccolo, D. (2003), On the moments of a mixture of uniform and shifted binomial random variables, *Quaderni di Statistica*, 5, 85–104.
6. Reichheld, F.F. (2003). The One Number You Need to Grow, *Harvard Business Review*, 81, 46–54.
7. Reichheld, F.F. (2006). *The Ultimate Question: Driving Good Profits and True Growth*, Harvard Business School Press, Boston.
8. Rocks, B. (2016). Interval Estimation for the “Net Promoter Score”, *The American Statistician*, 70(4), 365–372.

# News, Volatility and Price Jumps

## *News, Volatilità e Salti nei Prezzi*

Massimiliano Caporin and Francesco Poli

**Abstract** From two professional news providers we retrieve news stories and earnings announcements of the S&P 100 constituents and 10 macro fundamentals, moreover we gather Google Trends of the assets. We create an extensive and innovative database, useful to analyze the link between news and asset price dynamics. We detect the sentiment of news stories using a dictionary of sentiment words and negations, and propose a set of more than 5K information-based variables that provide natural proxies of the information used by heterogeneous market players and of retail investors attention. We first shed light on the impact of information measures on daily realized volatility and select them by penalized regression; then, we use them to forecast volatility and obtain superior results with respect to models that omit them. Finally, we relate news with intraday jumps using penalized logistic regression.

**Abstract** *Ricaviamo da due news provider professionali le news e gli annunci sugli utili dei componenti dell'S&P 100 e 10 indicatori macroeconomici, inoltre raccogliamo i Google Trends associati ai titoli. Creiamo un database esteso ed innovativo, utile per analizzare il legame tra le news gli andamenti dei prezzi dei titoli. Rileviamo il sentiment delle news usando un dizionario di parole associate a un sentiment e delle negazioni, e proponiamo un insieme di più di 5K variabili che rappresentano l'informazione usata da agenti eterogenei e l'attenzione dei piccoli investitori. Facciamo luce sull'impatto delle misure di informazione sulla volatilità realizzata giornaliera e le selezioniamo con la regressione penalizzata; poi le usiamo per prevedere la volatilità, ottenendo risultati superiori rispetto a modelli che le omettono. Infine, mettiamo in relazione le news con i salti intragiornalieri usando la regressione logistica penalizzata.*

---

Massimiliano Caporin

University of Padova, Department of Statistical Sciences, Via Cesare Battisti, 245, 35121, Padova PD, e-mail: massimiliano.caporin@unipd.it

Francesco Poli

University of Padova, Department of Economics and Management, Via del Santo 33, 35123, Padova PD, e-mail: francesco.poli.2@studenti.unipd.it

**Key words:** news, Google Trends, sentiment, volatility, forecasting, jumps, regularization, big data

## 1 Introduction

According to the MDH mixture of distributions hypothesis: “A *serially correlated mixing variable measuring the rate at which information arrives to the market explains the GARCH effects in asset returns.*” We want to verify its validity and, more generally, to shed light on the link between news and volatility. In addition, we want to understand which news indicators are likely to provoke price jumps.

We create a database which contains information useful to face the previous questions. From two news providers we retrieve news stories and EPS earnings per share announcements of the S&P 100 constituents, and 10 macroeconomic announcements. We also collect Google Trends of the assets, and use them as a proxy for retail investors attention. We detect the sentiment of news stories using the sentiment-related word lists developed by [6] and introduce a set of negations, with the aim of extracting the sentiment of a financial text independently from its type, length and audience. We propose a set of news measures that provide natural proxies for the information used by heterogeneous market players. We end up with a large set of news measures, each representing a different type of information potentially causing a different market reaction. We test the MDH and shed light on the impact of news on volatility using the information-related variables we develop. We perform an application using the database to explain realized volatility and selecting the most important indicators with *LASSO*, then we improve volatility forecasting in an out-of-sample analysis. Finally, we relate news with intraday jumps with *Elastic Net*.

## 2 Database Construction

We collect news and indicators from two news providers, FactSet-StreetAccount and Thomson Reuters, and from Google Trends. We utilize the latter as a proxy for retail investors attention, while providers gather information more relevant for professional investors. Time range of the dataset corresponds to the period February 4th 2005 - February 25th 2015 and all data has minute-precision, except for Google Trends that are daily.

We get firm-specific news and Google Trends of the S&P 100 constituents, since they are highly capitalized and attention grabbing companies. We exclude from the database 11 stocks since news about them were not available from both providers for the whole sample. The information of the database can be classified in five types:

1. **StreetAccount news stories** (firm-specific). They are classified along 11 topics, and we use 7 of them. News are filtered from irrelevant ones and are not redundant, that is each news is released only once.

2. **Thomson Reuters news stories** (firm-specific). Each story is organized according to a topic and a level of significance. There are 36 topics, and we use 6 of them, and four levels of significance: *low, medium, high, top*. They are also filtered from irrelevant ones and are not redundant.
3. **EPS announcements**. They are released by StreetAccount and comprehend both the company's reported actual quarterly EPS and the consensus forecast figure.
4. **Macro announcements**. 10 macroeconomic indicators released by Reuters: *consumer confidence, CPI, FOMC rate decisions, GDP, industrial production, balance of payments, jobless claims, non-farm payrolls, PPI and retail sales*. They also comprehend both the reported indicator and its consensus forecast.
5. **Google Trends**. Relative indicators of internet search volume available from Google. They summarize the searches performed through Google and represent how many web searches have been done for a keyword in a period of time in a given geographical area relative to the total in the same period and area. For each stock, we look at the global volume of search queries for the name of the company.

### 3 Sentiment Detection

We detect the sentiment of news stories, that is an indicator of whether the content of a document is good, bad or neutral in relation to the issue it talks about.

We use the sentiment-related word lists developed by [6], which are tailored for financial texts. They account for negation but use only six words and only if one of them precedes a negative word, and apply the methodology to US companies 10-Ks. We deal, instead, with news created by news providers, that are less limited in the use of language. We introduce the following improvements: we invert the sentiment each time a word, irrespective of whether it is positive or negative, is preceded by a negation, and extend negations by employing 28 single words, 24 sequences of two words (e.g. "far from") and 6 sequences of three words (e.g. "by no means"). We believe that this modification allows to extract the sentiment of a financial text with more confidence and independently of its type, length and audience. The procedure we develop works as follows:

1. positive words are given a value of 1, negative words -1 and the value is inverted in case of negation
2. values of all words with a sentiment are summed up to get the sentiment sum:  

$$Sent\_Sum = \sum_{i=1}^N s_i$$
 where  $i$  is the word index,  $N$  is the number of words with a sentiment in a text and  $s_i$  is the sentiment of the word indexed by  $i$
3.  $Sent\_Sum$  is divided by the number of words with a sentiment, obtaining the relative sentiment  $Rel\_Sent$ , comprised between -1 and 1:  $Rel\_Sent = Sent\_Sum/N$
4. If  $Rel\_Sent$  is bigger than 0.05 or smaller than -0.05 we associate, respectively, a positive (1) or a negative sentiment (-1) to the news, otherwise neutral (0).

## 4 Measures Creation

We go beyond the standard techniques used to assign numbers to textual information: we identify a set of concepts/events which are based on how news are released over different time horizons, with the aim to reconstruct the different portions of information on which the different market players base their decisions. In total, for each asset we end up with 5,159 news-related variables for daily analysis and 878 news-related variables for high-frequency analysis.

### Concepts for News Stories Variables

The variables are built following a scheme of several concepts, each of which is peculiar in the reaction it potentially causes in the market. All concepts refer to a reference period and to previous periods of equal or longer length. We list the main ones.

1. **standard measures:** number of news, number of words, sentiment. The first two represent proxies for the quantity of information, sentiment was illustrated above.
2. **abnormal quantity:** quantity of news above a threshold. Investors' reaction could be triggered by the release of an unusual quantity of information.
3. **uncertainty:** occurrence of news with opposite sentiment within the reference period. When this event happens, information is released but it is likely that investors are unable to detect whether it is good or bad.
4. **quantity variation:** variation across periods of the quantity of news, or words. This concept takes into account the chance that investors' reactions are triggered not only by the release of information, but more generally by increases or decreases in the quantity of information.
5. **news persistence/interaction:** event in which the quantity of news is above a threshold in each of two consecutive periods. Reminding that providers do not supply redundant news, the occurrence of this event denotes persistence in the release of news that are related in each period to a different issue.
6. **sentiment inversion:** event in which the sentiment of the reference period equals the opposite of the sentiment of previous periods.

### Standardized Surprises of EPS and Macro Announcements

EPS and Macro Surprises are constructed using techniques widespread in the literature. With regard to EPS, from actual figure and consensus forecast we compute the *SUE* Standardized Unexpected Earnings score, which measures the number of standard deviations the reported actual EPS differ from the consensus forecast.

With regard to macro announcements, from actual and consensus forecast of the indicators we compute the standardized surprise as we do for earnings.

### Google Search Index

Google restricts the access to daily data for intervals longer than 10 months but allows to gather daily data (relative to the maximum) for shorter intervals. From the set of the daily series for each month and the monthly-aggregated series for the whole sample we reconstruct the daily Google Search Index for the whole sample.

### Proposed Measures Based on Different Time Horizons

We propose a set of news measures suitable to be linked to daily asset price dynamics, by aggregating the information released during the following time horizons:

1. **daily**: from market closing time of day  $t-1$  to market closing time of day  $t$ ;
2. **overnight**: from market closing time of day  $t-1$  to market opening time of day  $t$ ;
3. **weekly**: last 5 days;
4. **monthly**: last 22 days.

We then develop a set of news indicators which can be related to high-frequency asset price dynamics, by aggregating the information released during market opening times in three different lagged intervals:

1. **lag 0**: last 10 min;
2. **lag 1**: from -30 min to -10 min;
3. **lag 2**: from -60 min to -30 min.

## 5 Volatility Forecasting and Intraday Jumps

We want to verify the validity of the MDH and to shed light on the link between news and volatility. We also want to understand which news indicators cause jumps.

### News Impact on RV

We compute daily realized volatility from five-minute returns. Then, we decompose it into its continuous and jump components resorting to the jump test of [4], and we model daily realized volatility with the HAR-TCJ linear model of [4], based on the HAR-CJ model of [1] using their corrected threshold multipower variation measures.

**HAR-TCJ model:**

$$RV_t = \beta_0 + \beta_d \widehat{C}_d + \beta_w \widehat{C}_w + \beta_m \widehat{C}_m + \beta_j \widehat{J}_d + \varepsilon_t \quad (1)$$

$$(RV_{t_1:t_2} = \frac{1}{t_2-t_1+1} \sum_{t=t_1}^{t_2} RV_t, \widehat{C}_d = \widehat{C}_{t-1}, \widehat{C}_w = \widehat{C}_{t-5:t-1}, \widehat{C}_m = \widehat{C}_{t-22:t-1}, \widehat{J}_d = \widehat{J}_{t-1})$$

Adding the news measures as regressors we obtain the **HAR-TCJN model**:

$$RV_t = \beta_0 + \beta_d \widehat{C}_d + \beta_w \widehat{C}_w + \beta_m \widehat{C}_m + \beta_j \widehat{J}_d + \beta_{News}^T News_{t-1} + \varepsilon_t \quad (2)$$

where  $\beta_{News}$  is the  $k \times 1$  vector of coefficients and  $News_{t-1}$  is the  $k \times 1$  vector of news measures built on the basis of the information available before the market opening time of day  $t$ . We face a dimensionality problem in the HAR-TCJN model since the number of regressors is higher than the number of observations, and we resort to **LASSO** to select the most useful measures. LASSO [7] is an estimation method for linear models that performs variable selection and coefficients shrinkage, and was already used to model realized volatility by [3].

We implement an in-sample analysis using the logarithmic counterparts ([4]) of the models, and estimate the parameters of the HAR-TCJN model with LASSO. Ranking the indicators by the number of assets for which their estimated  $\beta$  is different from zero, it is possible to see that macro announcements and EPS are the most important drivers of volatility, but news stories and Google Trends also have a role. Macro announcements per se count, as well as surprises from expectations. Markets tend to react more strongly to negative surprises, and on the basis of the information released during several previous time horizons, from overnight to the last month. EPS announcements per se and surprises are both important as well, and there is no evident asymmetric effect between positive and negative surprises. Only EPS information released during the last day seems relevant. News stories from StreetAccount are slightly more useful to explain market reactions than Reuters news, and variables based on day-to-day variations of the rate of information arrival are the most useful. Earnings is the most important news topic. Retail investors attention during the last week, caught by Google Trends, is positively linked with volatility.

In order to test the MDH, we perform two different OLS regressions with HAC standard errors: one for the HAR-TCJ model and one for the HAR-TCJN model employing as news variables only the previously selected ones, and compare the estimated autoregressive coefficients between the two models. Table 1 presents the estimation results for the autoregressive coefficients (cross-sectional average)  $\beta_0$ ,  $\beta_d$ ,  $\beta_w$ ,  $\beta_m$  and  $\beta_j$  for both models and their variation after the inclusion of news. Coefficients are all positive and, with the exception of  $\beta_m$ , their value is lower for the model HAR-TCJN, while the intercept  $\beta_0$  is higher for the HAR-TCJN model. These variations highlight the relevance of news as a driver of additional information, which involves effects on the estimated autoregressive coefficients. Results are consistent with the MDH.

### Evaluating the Forecasting Performance Improvement

Using a rolling window long 1000 observations, we iteratively estimate the HAR-TCJ and the HAR-TCJN models and apply the estimated coefficients to the information available the day following the last day used for estimation, obtaining the one-step-ahead forecast of realized volatility. The forecasting performance of the two models is compared with five metrics: MAE mean absolute error, MSE mean square



**Table 1**

	log HAR-TCJ	log HAR-TCJN	$\Delta\beta$
$\beta_0$	0.34 (2.50)	0.65 (3.35)	0.31
$\beta_d$	0.26 (2.64)	0.23 (2.52)	-0.03
$\beta_w$	0.46 (2.88)	0.38 (2.86)	-0.08
$\beta_m$	0.20 (2.13)	0.22 (2.26)	0.02
$\beta_J$	0.18 (0.83)	0.14 (0.52)	-0.04

Estimated (cross-sectional average)  $\beta_0$ ,  $\beta_d$ ,  $\beta_w$ ,  $\beta_m$  and  $\beta_J$  and their t-statistics in brackets for the log HAR-TCJ and the log HAR-TCJN models, and variation of the coefficients between the two models. OLS regression with HAC standard errors, using as explanatory variables for the HAR-TCJN model the regressors of the log HAR-TCJ plus the news variables selected by LASSO.

error,  $R^2$  of Mincer-Zarnowitz forecasting regressions, HRMSE heteroskedasticity adjusted mean square error, QLIKE.

Table 2 reports the cross-sectional mean over all assets of the metrics, and includes in brackets, for all metrics except for the  $R^2$  MZ, the percentage of assets for which the Diebold-Mariano test [5] rejects with a 5% significance level the null hypothesis of equal predictive accuracy in favor of each model, and in brackets for the  $R^2$  MZ the percentage of assets for which the metric is higher (i.e. a superior predictive accuracy) for each model. The HAR-TCJN model yields on average lower MAE, HRMSE and QLIKE and a higher  $R^2$  MZ. The average MSE is instead lower for the HAR-TCJ model. The HAR-TCJN model imply a better forecasting power which is statistically significant for a percentage of stocks ranging, depending on the metrics, from 11.24% to 82.02%. The test never signals a statistically significant superior predictive accuracy of the HAR-TCJ model.

**Table 2**

	log HAR-TCJ	log HAR-TCJN
MAE	0.96 (0.00%)	0.95 (26.97%)
MSE	33.82 (0.00%)	34.30 (11.24%)
$R^2$ MZ	0.50 (26.97%)	0.51 (73.03)
HRMSE	0.92 (0.00%)	0.82 (59.55%)
QLIKE	1.45 (0.00%)	1.44 (82.02%)

One-step-ahead MAE, MSE,  $R^2$  MZ, HRMSE, and QLIKE of the log HAR-TCJ and the log HAR-TCJN models (cross-sectional average). In brackets, for each metric except for  $R^2$  MZ: percentage of assets for which the Diebold-Mariano test rejects with a 5% sign. level the null hypothesis of equal predictive accuracy in favor of that model; for  $R^2$  MZ: percentage of assets for which the metric is higher for that model.

## News Measures and Intraday Jumps

We identify the precise intraday intervals at which jumps occur, relying on the procedure of [2] using the corrected threshold multipower variation measures of [4]. Indicators are selected using the Elastic Net ([9]) in a logistic regression with the occurrence of jumps (1 for occurrence, 0 otherwise) as dependent variable. [8] point out that the logistic regression is often plagued with degeneracies when the number of covariates  $p$  is greater than the number of observations  $N$  and exhibits wild behavior even when  $N$  is close to  $p$ ; the Elastic Net penalty alleviates these issues, and regularizes and selects variables as well.

Results tell us that macro announcements, especially FOMC rate decisions, as well as news stories, independently of their topic, cause jumps, and that all lagged intervals used to aggregate information are relevant.

## 6 Concluding Remarks

Our empirical results validate the Mixture of Distributions Hypothesis, showing the relevance of news as an important driver of volatility. Macro news and EPS are the most influential, followed by news stories and Google Trends. Aggregating information over different time horizons is important. By including news-based information, we are able to improve volatility forecasting. Macro announcements, especially FOMC rate decisions, and news stories are related to intraday jumps, which can follow immediately or with a delay ranging from few minutes to one hour.

## References

1. Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P.: Modeling and forecasting realized volatility. *Econometrica* **71**, 579–625 (2003).
2. Andersen, T.G., Bollerslev, T., Dobrev, D.: No-arbitrage semi-martingale restrictions for continuous-time volatility models subject to leverage effects, jumps and i.i.d. noise: theory and testable distributional implications. *J Econom* **138**, 125–180 (2007).
3. Audrino, F., Knaus, S.D.: Lassoing the HAR model: a model selection perspective on realized volatility dynamics. *Econom Rev* **35**, 1485–1521 (2016).
4. Corsi, F., Pirino, D., Renò, R.: Threshold bipower variation and the impact of jumps on volatility forecasting. *J Econom* **159**, 276–288 (2010).
5. Diebold, F., Mariano, R.: Comparing predictive accuracy. *J Bus Econ Stat* **13**, 253–263 (1995).
6. Loughran, T., McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J Finance* **66**, 35–65 (2011).
7. Tibshirani, R.: Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol* **58**, 267–288 (1996).
8. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* **33**, 1–22 (2010).
9. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* **67**, 301–320 (2005).

# Growing happiness: a model-based tree

Carmela Cappelli, Rosaria Simone and Francesca di Iorio

**Abstract** Tree based methods in statistics are gaining a renewed interest in the Big Data era since they entail effective interpretation of results. In this setting, we apply a model-based technique to build trees for ordinal responses relying on a class of mixture models whose characteristic feature is the probabilistic specification of *uncertainty*. An application to the perception of happiness shows that the integration of tree methods with the chosen modelling boosts cluster analysis of respondents.

**Abstract** *Nell'era dei Big Data, i metodi statistici basati sugli alberi ottengono una grande rilevanza poichè permettono un'efficace interpretazione dei risultati. In questo contesto, consideriamo una tecnica per crescere alberi per risposte ordinali basata su una classe di modelli statistici il cui valore aggiunto è la specificazione probabilistica dell'incertezza. La validità dell'approccio e una sua applicazione al clustering vengono discusse sulla base di un survey sulla percezione della felicità.*

**Keywords:** Tree based methods; Uncertainty; Ordinal Responses

## 1 Motivations

Among the consolidated literature on ordinal data analysis [1], an alternative approach is based on CUB models [4, 5], whose rationale is that discrete choices arise from a psychological process that involves two components: a personal *feeling* and an inherent *uncertainty*. The effectiveness of this paradigm improves with the inclusion of explanatory covariates for parameters, leading to CUB regression models. Lately, tree based methods [2] have gained widespread popularity because they are a simple, yet powerful data analysis tool particularly useful to analyze large data sets characterized by both qualitative and quantitative covariates. A key advantage of trees methodology is the automatic selection of the most relevant covariates as

---

Department of Political Sciences, University of Naples Federico II, Via Leopoldo Rodinò, 22, 80138 Napoli, e-mail: carcappe@unina.it, e-mail: rosaria.simone@unina.it, e-mail: fdiiorio@unina.it

well as their interpretability.

In the streamline of [6], in [3] the authors proposed a method to grow model-based trees in which every node is associated with a CUB regression model. Then, the terminal nodes of the tree identify alternative profiles of respondents based on the covariates values and classified according to levels of *uncertainty* and *feeling*. Additionally, similarity of the clusters so determined can be investigated exploiting a graphical feature of the chosen modelling [3].

The paper is organized as follows: first we recall the basics of the chosen model-based procedure (section 2). Then, in section 3, we present an application to a data set investigating the perception of happiness. The whole analysis has been run within the free R environment: the code is available upon request from authors.

## 2 Background and Methodology

Trees have proven to be a useful tool for high dimensional data analysis, able to capture nonlinear structures and interactions. Growing trees [2] for a response variable (either continuous or categorical) relies on a top-down partitioning algorithm that is known as recursive binary splitting, as it is based on a splitting criterion that allows to choose at each tree node (subset of observations), the best split, i.e. binary division, of the current node, based on a set of explanatory variables. At each tree node  $t$ , given an *impurity measure*  $I(s, t)$  that assesses the homogeneity of node  $t$ , the algorithm chooses the split  $s^*$  that induces the highest decrease in impurity with respect to the child nodes of ( $t_l$  and  $t_r$  respectively):

$$s^* = \underset{s}{\operatorname{argmax}} \Delta I(s, t), \quad \Delta I(s, t) = i(t) - [i(t_l)p_l + i(t_r)p_r]$$

where  $p_l$  and  $p_r$  represent the node weights. Once a node is partitioned, the splitting process is recursively applied to each child node until either they reach a minimum size or no further reduction of impurity can be achieved.

The tree methodology based on CUB models (CUB REgression MOdel Trees - CUBREMOT for short) has been advanced in [3]. In a nutshell, CUB models paradigm [4] designs the data generating process yielding to an ordinal evaluation out of the latent perception as the combination of a *feeling* component (which drives substantial likes and agreement)-shaped by a shifted Binomial distribution- and an *uncertainty* component- which is assigned a discrete Uniform distribution. Denoting  $R_i = 1, \dots, m$  the score assigned by the  $i$ -th respondent to a given item of a questionnaire, we say that  $R_i$  is a CUB distributed random variable with uncertainty parameter  $\pi$  and feeling parameter  $\xi$  (for short  $R_i \sim \text{CUB}(\pi, \xi)$ ) if:

$$\Pr(R_i = r | \pi_i, \xi_i) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r} (1 - \xi_i)^{r-1} + (1 - \pi_i) \frac{1}{m}, \quad r = 1, \dots, m.$$

In particular, the mixing proportion  $\pi_i$  is an indirect measure of heterogeneity while  $1 - \xi_i$  indicates a positive tendency in the data w.r.t. the topic under investigation. Explanatory variables may be included in the model in order to relate feeling and/or uncertainty to respondents' profiles. Then, consider a CUB regression model with a *logit* link between parameters and a dichotomous factor  $D$ :

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 D_i, \quad \text{logit}(\xi_i) = \gamma_0 + \gamma_1 D_i. \quad (1)$$

If no covariate is considered neither for feeling nor for uncertainty, the  $\pi_i = \pi$  and  $\xi_i = \xi$  are constant among subjects. Estimation of CUB models relies on likelihood methods and on the implementation of the Expectation-Maximization (EM) algorithm for mixtures.

Since the process of descending son nodes from a father node is a binary splitting, the starting point to grow a CUBREMOT is the selection of a set of explanatory variables to be sequentially transformed and associated with a set of dichotomous factors. Then, for a given  $k \geq 1$  and a dichotomous variable  $D$ , a CUB regression fit to the  $k$ -th node provides it with a CUB  $(\hat{\pi}_k, \hat{\xi}_k)$  distribution whose log-likelihood at the final ML estimates is  $\mathcal{L}_{n_k}(\hat{\pi}_k, \hat{\xi}_k)$ . Then, the split induced by  $D$  associates the left son node  $2k$  (right son node  $2k + 1$ , resp.) with the conditional distribution  $R|D = 0$  ( $R|D = 1$ , respectively). The proposal in [3] relies on a splitting criterion based on the log-likelihood increment from the father to the sons level for each possible split, and at the given step chooses the one that maximizes the deviance:

$$\Delta \mathcal{L}_k = [\mathcal{L}_{n_0}(\hat{\pi}_{2k}, \hat{\xi}_{2k}) + \mathcal{L}_{n_1}(\hat{\pi}_{2k+1}, \hat{\xi}_{2k+1})] - \mathcal{L}_n(\hat{\pi}_k, \hat{\xi}_k) \quad (2)$$

(here,  $n_i$  denotes the size of the sub-sample conditional to  $D = i, i = 0, 1$ ). Finally, a node is declared terminal if none of the available covariates is significant (neither for feeling nor for uncertainty), or if the sample size is too small to allow a CUB model fit. Figure 1 shows the formal configuration of the split at node  $k$ .

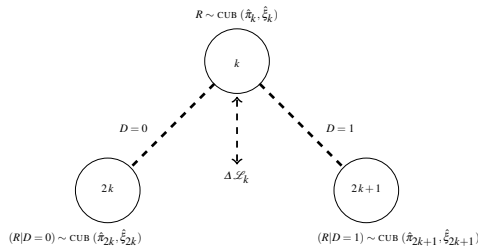


Fig. 1 CUBREMOT : split at node  $k$

### 3 Application

The focus of the present contribution is the derivation of a CUBREMOT for measurements on perceived happiness collected at University of Naples Federico II in December 2014 (the data set is available at: <http://www.labstat.it/home/wp-content/uploads/2015/09/relgoods.txt>). Every participant was asked to rate the quality and the importance attributed to selected relational goods on a  $m = 10$  point ordinal scale (1 = “Never”, “Not at all good”, to 10 = “Always”, “A lot”, “Absolutely good”), and to self-evaluate his/her happiness by marking a sign along a horizontal line (with the left-most bound standing for “extremely unhappy”, and the right-most one for the status “extremely happy”). This continuous measurement has been uniformly discretized into an ordinal variable *Happiness* over  $m = 10$  categories in order to allow direct comparisons with other questionnaire items. For illustrative purposes, the case study considers only few subjects’ dichotomous characteristics: *Gender* (1 for women), the marital status *Married* (1 for married), and the smoking habit *Smoke* (= 1 for smokers). Also the association between happiness and relationships with both *Friends* and *Parents* is considered (the latter measured by a proxy quantifying the time spent with them).

The final CUBREMOT is displayed in Figure 2 (terminal nodes are squared): at each split, the value of the deviance splitting criterion (2) and the sample sizes are reported. Noticing that in this context the estimated  $1 - \hat{\xi}$  is a direct indicator of happiness, Overall people are fairly happy ( $1 - \hat{\xi}_1 = 0.653$ ) and evaluations are affected by a modest level of indecision ( $1 - \hat{\pi}_1 = 0.398$ ). Some main comments about CUBREMOT classification can be summarized as follows:

- The happiest group of respondents corresponds to males giving a low evaluation for relationships with parents (*Parents*  $\leq 5$ ) and an extremely positive perception about friendship (*Friends*  $\geq 9$ ), with  $1 - \hat{\xi}_{53} = 0.823$ . However, these responses are affected by a not negligible uncertainty, indicating that there could be unobserved factors, as response styles effects. A comparable level of happiness is observed within married people that speak very often with their parents and that rate their relationships with friends of high quality (with an estimated happiness of  $1 - \hat{\xi}_{31} = 0.808$ ). In this case, this index can be considered also as a precise classification measure since the node is characterized by a low uncertainty ( $1 - \hat{\pi}_{31} = 0.258$ ). In addition, friendship is recognized a major role in the perception of happiness especially among married people (perceived happiness increases when descending from node 15). Instead, it is noticeable that the married status does not come into play among those not having a good relationships with their parents (from node 6 downward), whereas married people are happier among those evaluating positively their relationships with parents ( $1 - \hat{\xi}_{14} = 0.663$  against  $1 - \hat{\xi}_{15} = 0.744$ ).
- The unhappiest are those with a very poor quality of relationships both with parents and with friends ( $1 - \hat{\xi}_4 = 0.147$  and a fairly high level of indecision:  $1 - \hat{\pi}_4 = 0.641$ ). More specifically, among respondents assessing unsatisfactory the relationships with friends (*Friends*  $\leq 5$ ), one observes a sharp improvement

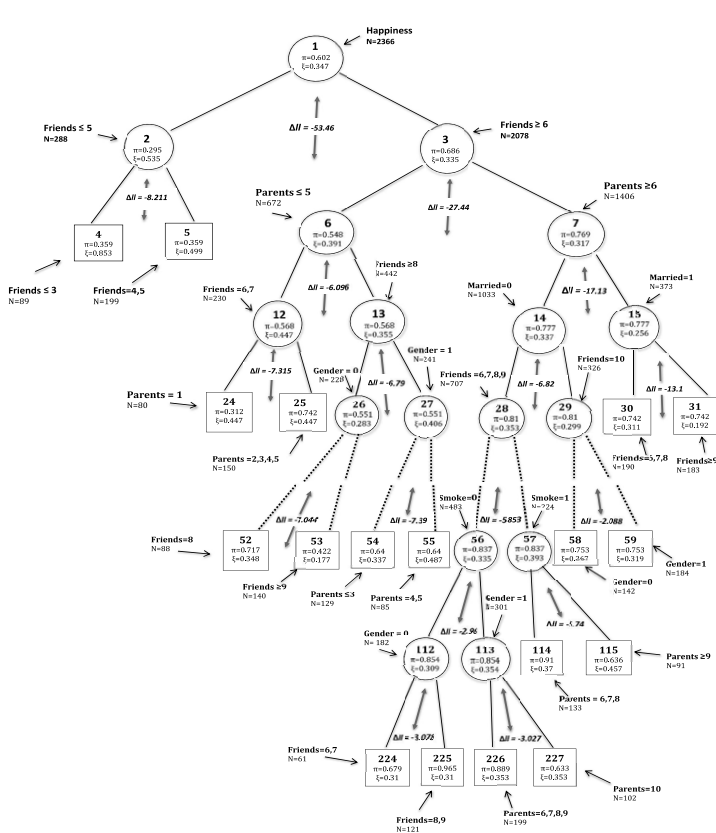


Fig. 2 CUBREMOT for *Happiness*

from  $1 - \hat{\xi}_4 = 0.147$  to  $1 - \hat{\xi}_5 = 0.501$  in perceived happiness as soon as one switches from giving a very low judgment ( $Friends \leq 3$ ) to a medium-low scoring ( $Friends = 4, 5$ ). In addition, the split of node 2 into nodes 4 and 5 has indeed identified more homogeneous groups ( $1 - \hat{\pi}_2 = 0.705$  against  $1 - \hat{\pi}_4 = 1 - \hat{\pi}_5 = 0.641$ ).

- By looking at nodes 26 and 27, 112 and 113, and nodes 58 and 59, it can be inferred that happiness hinges more on friendships for men than for women. For instance, among the unmarried respondents which are satisfied with their family

bonds ( $Parents \geq 6$ ) and giving the highest evaluation for friendships ( $Friends = 10$ ), females are slightly unhappier than males ( $1 - \hat{\xi}_{59} = 0.681$  against  $1 - \hat{\xi}_{58} = 0.733$ ).

As a by-product, a cluster analysis of CUBREMOT nodes can be performed once they are represented as points in the parameter space with coordinates given by corresponding uncertainty  $1 - \pi$  and feeling  $1 - \xi$ . Figure 3 shows the scatterplot of the fist 59 nodes for *Happiness*: three clusters, highlighted by different symbols, are identified using a simple  $k$ -means algorithm with  $k = 3$ .

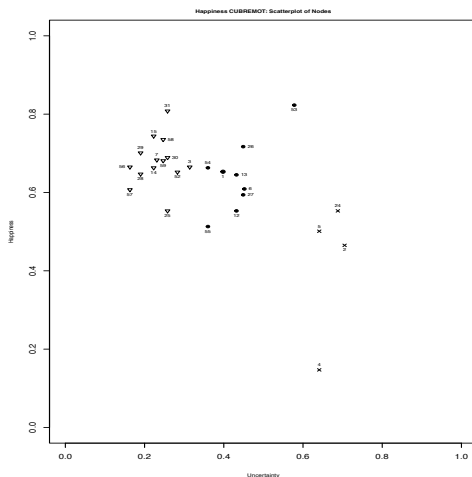


Fig. 3 CUBREMOT nodes in CUB parameter space

## References

1. Agresti A. (2010). *Analysis of Ordinal Categorical Data*, II edition. J.Wiley & Sons, Hoboken.
2. Breiman L., Friedman J.H., Olshen R.A, Stone C.J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks: Monterey (CA).
3. Cappelli, C., Simone, R., di Iorio, F. (2017). Model-based trees with uncertainty for ordinal responses. *Submitted*.
4. D'Elia A., Piccolo D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, **49**, 917–934.
5. Iannario M., Piccolo D. (2016). A Generalized Framework for Modelling Ordinal Data. *Statistical Methods and Applications*, **25**, 163-189.
6. Zeileis A., Hothorn T., Hornik K. (2008). Model-Based Recursive Partitioning, *Journal of Computational and Graphical Statistics*, **17**, 492–514.



# **Inequalities in access to job-related learning among workers in Italy: evidence from Adult Education Survey (AES)**

## *Differenze nella formazione dei lavoratori in età adulta nei risultati della Adult Education Survey (AES)*

Paolo Emilio Cardone<sup>1</sup>

**Abstract** Equitable access to adult learning for all is a goal for European education, training and employment policies. In particular, all workers should be able to acquire, update and develop their skills over their lifetime. How is it possible to improve access to learning for older workers? This report provides a statistical picture of older workers participation in job-related training in Italy, investigating its variability and relevant inequalities. The analysis is carried out using Italian AES, provided by Eurostat. It analyses adults' learning activities and distinguishes formal, non-formal and informal learning. Using logistic regression model it is possible to estimate the learning-age gap between those aged under and over 50 years more accurately. Overall the data confirm the existence of strong inequalities in access to job-related learning among workers.

**Abstract** L'allungamento delle aspettative di vita e i cambiamenti demografici rendono necessario lavorare più a lungo e sostenere una forza lavoro competente, adattabile al cambiamento e competitiva. Un equo accesso ai percorsi di apprendimento per tutti i lavoratori, in particolare per quelli più adulti, è uno dei principali obiettivi della Commissione Europea per le politiche di istruzione, formazione e occupazione. In particolare, tutti i lavoratori devono essere in grado di acquisire, aggiornare e sviluppare le proprie competenze nel corso della loro vita lavorativa. Come è possibile migliorare l'accesso alla formazione continua per i lavoratori più adulti? Il presente contributo fornisce un quadro statistico della partecipazione dei lavoratori adulti ai programmi di formazione legata al lavoro in Italia, indagando la sua variabilità e le disuguaglianze più rilevanti. L'analisi è stata

---

<sup>1</sup> Paolo Emilio Cardone, INAPP-Statistical Office/Sapienza University of Rome; email: [paoloemilio.cardone@uniroma1.it](mailto:paoloemilio.cardone@uniroma1.it)

effettuata utilizzando i dati italiani dell'indagine AES, condotta da Eurostat. Essa analizza le attività di formazione degli adulti, distinguendo tra formazione formale, non formale e apprendimento informale. Utilizzando un modello di regressione logistica è inoltre possibile stimare il divario tra gli occupati over 50 e under 50 con maggiore precisione. Nel complesso i dati confermano l'esistenza di forti disuguaglianze tra i lavoratori per quanto riguarda l'accesso ai programmi formativi legati al lavoro.

**Key words:** Age management; Adult education; Lifelong learning; Logistic regression model.

## 1 Introduction

Demographic ageing is an irreversible process. The direct effect of population ageing is the increasing share of elderly people, who are in retirement age, compared to the decreasing share of young people.

Furthermore, the European Commission 2012 Ageing Report suggests that population ageing has been also affecting the age structure of population working age. This is extremely important in the overall context of labour force in the EU (particularly in Italy). On the labour market, the proportion of jobs that require medium and high-level qualifications is expected to increase. However, there is still an extremely high number of those of working age in Europe who have either low or no qualifications.

The nature of jobs is changing, necessitating changes in the skills that are required of workers and adapting lifelong learning systems to the needs of an ageing workforce. The recent crisis has also highlighted the importance of education and training at all stages of life, in particular for older adults to avoid unemployment, vindicating the messages that "it is never too late to learn" and learning must be for all. This requires older people to maintain and update the skills they have, particularly in relation to new technologies. Continuous learning and development of an ageing workforce are important for employers' survival in competitive markets, as well as for maintaining older people's employability.

Equitable access to adult learning for all is a goal for European education, training and employment policies. In particular, all workers should be able to acquire, update and develop their skills over their lifetime. However, despite the increasing need for learning later in life, participation and access to learning decrease with age. How is it possible to improve access to learning for older workers? This report provides a statistical picture of older workers participation in job-related training in Italy, investigating its variability and relevant inequalities due to key factors such as the influence of individual characteristics, jobs and workplaces.

## 2 Data and methods

In order to achieve this goal, the analysis is carried out using microdata from the second and latest wave of Italian Adult Education Survey (AES-2011), provided by Eurostat. The survey analyses the learning activities of adults and distinguishes between formal, non-formal and informal learning, which takes place inside or outside the workplace. It investigates adult participation in training in depth and includes a sample of 11.500 individuals, 6.000 of which are workers (if weighted they become 22 million, exactly the workers' amount in Italy).

Regular participation in learning activities does not include taking part in formal training only, but also learning in non-formal and informal learning settings. In particular, informal learning plays a greater role for older employees than formal learning because it facilitates the transfer of knowledge and know-how between generations, allows practical skills to be gained quickly and ensures the inclusion, particularly for older workers, within the circles of relationships.

The organizing concept of the CLA (Classification of Learning Activities) is based on 3 broad categories: Formal Education (F), Non Formal Education (NF) and Informal Learning (INF). It is possible to classify all learning activities into these 3 categories using some general concepts and definitions:

Lifelong Learning (LLL) is defined as encompassing “all learning activity undertaken throughout life, with the aim of improving knowledge, skills and competences, within a personal, civic, social and or employment related perspective.”

Formal Education as “education provided in the system of schools, colleges, universities and other formal educational institutions that normally constitutes a continuous “ladder” of full-time education for children and young people, generally beginning at age of five to seven and continuing up to 20 or 25 years old. Formal education refers to institutionalised learning activities that lead to a learning achievement that can be positioned in the National Framework of Qualifications (NFQ).

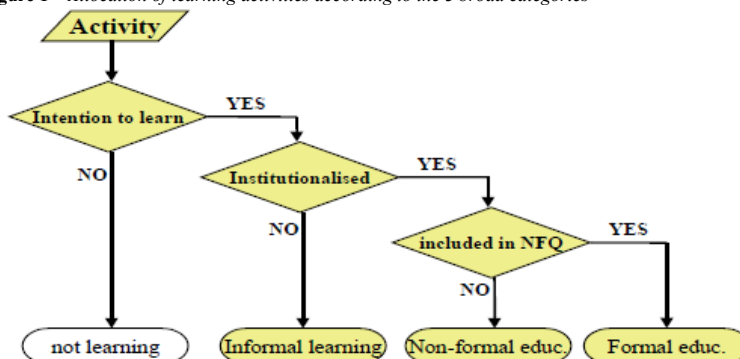
Non Formal Education is defined as “any organised and sustained educational activities that do not correspond exactly to the above definition of formal education. Non-formal education may therefore take place both within and outside educational institutions, and cater to persons of all ages. Non formal education programmes do not necessarily follow the “ladder” system, and may have a differing duration. Non-formal education refers to institutionalised learning activities, which are not part of the NFQ.

Informal Learning is defined as “...intentional, but it is less organised and less structured ...and may include for example learning events (activities) that occur in the family, in the work place, and in the daily life of every person, on a self-directed, family-directed or socially directed basis. Informal learning activities are not institutionalised.

The National Framework of Qualification (NFQ) is defined as “the single, nationally and internationally accepted entity<sup>1</sup>, through which all learning achievements may be measured and related to each other in a coherent way and which define the relationship between all education and training awards”.

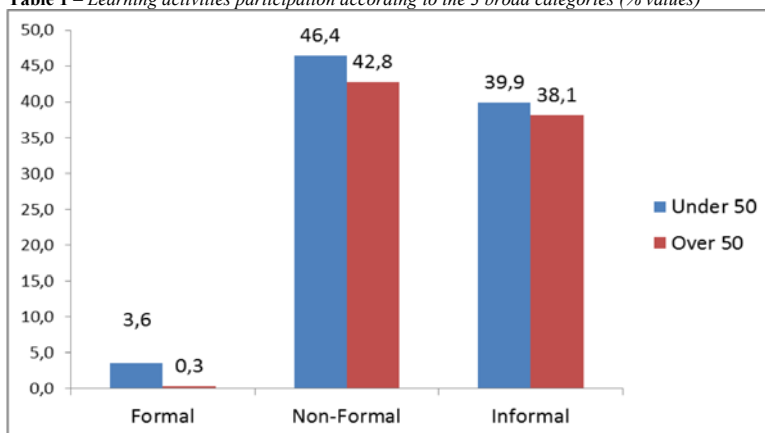
In synthesis, the process to allocate education and learning according to the broad categories is presented in the decision making flowchart shown in Figure 1:

**Figure 1** – Allocation of learning activities according to the 3 broad categories



As shown in table 1, descriptive analysis shows a strong inequality between under and over 50 workers for all broad categories.

**Table 1** – Learning activities participation according to the 3 broad categories (% values)



Source: own elaboration on AES data

<sup>1</sup> The entity can take the form of an organization/body, or regulatory document. It stipulates the qualifications and the bodies that provide or deliver the qualification (awarding bodies) that are part of the National Framework of Qualifications.

Using multivariate analysis (logistic regression models with Stata software) it is possible to estimate the learning-age gap between those aged under and over 50 years more accurately. The model has been developed for employed adults only and includes, first of all, adults' socio-demographic characteristics (age, gender and citizenship), secondly, job and size enterprise.

In order to achieve this goal, we have used "Learning" as the dependent variable (weighted model). Learning=1 if the worker has participated at least one training activity (formal, non formal or informal). Concretely, in our study the following variables are considered:

- **Gender.** Categorical. Dummy variable: Female, Male (reference cat.).
- **Citizen.** Categorical. Three values. Italian citizenship (reference cat.), Other citizenship UE, citizenship extra UE.
- **JobISCO.** Categorical. Nine levels. Elementary occupations (reference cat.), Managers, Professionals, Technicians and associate professionals, Clerical support workers, Service and sales workers, Skilled agricultural, forestry and fishery workers, Craft and related trades workers, Plant and machine operators, and assemblers.
- **Sizefirm.** Categorical. Four intervals. From 1 to 10 (micro, reference cat.), between 11 and 49 (small), between 50 and 249 (medium) and more than 250 (large).
- **Age.** Dummy variable: Over 50 (reference cat.), Under 50.

**Table 2** – Logistic regression models

Variables		Beta	ODDS	Sign.
• Gender				
Male (ref.)	Female	-0,10	0,91	0.174
• Citizen				
Italian (ref.)	EU	-0,58	0,56	0.025
	Extra EU	-0,56	0,57	0.001
• Size firm				
Micro (1-10) (ref.)	Small (11 - 49)	0,19	1,21	0.019
	Medium (50- 249)	0,48	1,62	0.000
	Large (250 +)	0,59	1,81	0.000
• Job ISCO				
Elementary occupations (ref.)	Managers	1,24	3,44	0.000
	Professionals	2,24	9,44	0.000
	Technicians and associate professionals	1,53	4,61	0.000
	Clerical support workers	0,96	2,61	0.000
	Service and sales workers	0,66	1,94	0.000
	Skilled agricultural, forestry and fishery workers	0,84	2,32	0.002
	Craft and related trades workers	0,34	1,40	0.018
	Plant and machine operators, and assemblers	0,55	1,74	0.000
• Age				
Over 50 (ref.)	Under 50	0,20	1,22	0.009
	Intercept	-0,65	0,52	0.000

Source: own elaboration on AES data

### 3 Conclusions

One principal finding of such an analysis is that people under 50 have a probability of 1.22 and higher of participating in training when compared to those aged 50 and more (table 2). Secondly, women are less likely to take part in training than men.

Overall the data confirm the existence of strong inequalities in access to job-related learning among workers: foreign individuals, in micro and small enterprises and in occupations with lower skills participate in job-related learning to a much lower extent.

This requires policy attention, to increase the focus on job-related training as part of active labour market policies, to prevent skills' obsolescence. In addition, it is important develop a "learning culture". It is a key factor for increasing the productivity of older workers increasing e.g. the capacity to deal with technological change ("it is never too late to learn").

However, it will be crucial to increase the level of continuous vocational training for all workers in future.

This is (or should be) the real challenge.

### References

1. Boeren, E. (2011). Gender differences in formal, non-formal and informal adult learning. *Studies in continuing education*, Vol. 33, No 3, pp. 333-346.
2. Cedefop (2015). *Job related adult learning and continuing vocational training in Europe: a statistical picture*. Luxembourg: Publication Office.
3. European Commission (2011). *Supporting vocational education and training in Europe: the Bruges communiqué*. Luxembourg: Publication Office.
4. [http://ec.europa.eu/education/library/publications/2011/bruges\\_en.pdf](http://ec.europa.eu/education/library/publications/2011/bruges_en.pdf)
5. Eurostat (2006). *Classification of learning activities: manual*. Luxembourg: Publication Office. <http://ec.europa.eu/eurostat/documents/3859598/5896961/KS-BF-06-002-EN.PDF/387706bc-ee7a-454e-98b6-744c4b8a7c64?version=1.0>
6. Liu, X. (2016). *Applied Ordinal Logistic Regression using Stata*. Sage Publications. <http://www.stata.com/bookstore/applied-ordinal-logistic-regression-using-stata>

# Signal detection in high energy physics via a semisupervised nonparametric approach\*

## *Individuazione di un segnale fisico mediante un approccio non parametrico semi-supervisionato*

Alessandro Casa and Giovanna Menardi

**Abstract** In particle physics, the task of identifying a new signal of interest, to be discriminated from the background process, shall be in principle formulated as a clustering problem. However, while the the signal is unknown, usually even missing, the background process is known and always present. Thus, available data have two different sources: an unlabelled sample which might include observations from both the processes, and an additional labelled, sample from the background only. In this context, semisupervised techniques are particularly suitable to discriminate the two class labels; they lies between unsupervised and supervised ones, sharing some characteristics of both the approaches. In this work we propose a procedure where additional information, available on the background, is integrated within a nonparametric clustering framework to detect deviations from known physics. Also, we propose a variable selection procedure that allows to work on a reduced subspace.

**Abstract** Nell'ambito della fisica delle particelle la ricerca di un segnale di interesse, che si manifesta come una deviazione dal processo di background, può essere formulata in termini di problema di raggruppamento. Tuttavia, mentre la presenza del segnale non è certa, lo è quella del background, che rappresenta un processo noto. Nelle analisi empiriche, si dispone non solo di dati non etichettati, che potrebbero contenere segnale, ma anche di un campione di dati etichettati, provenienti dal solo processo di background. Ha senso allora adottare un approccio semisupervisionato, che si colloca a metà strada tra i metodi supervisionati e non. In questo lavoro si propone una procedura che integra l'informazione aggiuntiva a disposizione a tecniche di clustering non parametrico per individuare deviazioni dalle teorie fisiche esistenti. Viene inoltre proposta una procedura di selezione delle variabili che permette di operare su un sotto-spazio ridotto.

---

Alessandro Casa, Giovanna Menardi  
Dipartimento di Scienze Statistiche, Università degli Studi di Padova  
via C. Battisti 241, 35121, Padova; e-mail: casa@stat.unipd.it, menardi@stat.unipd.it

\*This report is part of a project that has received funding from the EU's Horizon 2020 research and innovation programme under grant agreement 675440. The authors wish to thank Dr. T. Dorigo, from the National Institute of Nuclear Physics (INFN) for providing the data.

**Key words:** high energy physics, nonparametric clustering, semisupervised classification

## 1 Introduction

Since the early Sixties, the *Standard Model* has represented the state of the art in High Energy Physics. It describes how the fundamental particles interact with each others and with the forces between them, giving rise to the matter in the universe. Despite its empirical confirmations, there are indications that the Standard Model does itself not complete our understanding of the universe. Model independent searches aim to explain the shortcomings of this theory by empirically looking for any possible *signal* which behaves as a deviation from the *background* process, representing, in turn, the known physics.

The considered problem can be recasted to a classification framework, although of a very peculiar nature. While the background process is known and a sample of virtually infinite size can be drawn from it, the signal process is unknown, possibly even missing. Available data have, consequently, two different sources: a first, labelled, sample from the background class only, and a second, unlabelled sample which might include observations from the signal. A semisupervised perspective [2] shall be then adopted, either by relaxing assumptions of supervised methods, or by strenghtening unsupervised clustering structures through the inclusion of additional information available from the labelled data.

In [5], the problem has been faced by building on a suitable adaptation of parametric density-based clustering to the semisupervised framework, according to the same logic of anomaly detection tasks. In this work we follow a similar route, yet in a nonparametric guise. Such formulation appears consistent with the physical notion of signal, *i.e.* a new particle would manifest itself as a peak emerging from the background process. Nonparametric *-modal-* clustering, in turn, draws a correspondence between groups and the modal peaks of the density underlying the observed data. Thus, the one-to-one relationship between clusters and modes of the distribution would provide an immediate physical meaning to the detected clusters.

The main idea underlying this work is to semisupervise nonparametric clustering by exploiting information available from the background process. Specifically, we tune a nonparametric estimate of the unlabelled data by selecting the smoothing amount so that the induced modal partition will classify the labelled background data as accurately as possible. As a side contribution we propose a variable selection procedure, specifically conceived for this framework, linked to the concept of stability of the distribution underlying the data.

We adopt the following notation:  $\mathcal{X}_b = \{\mathbf{x}_i\}_{i=1,\dots,n_b}$  denotes the set of labelled data, supposed to be a sample of *iid* multidimensional observations from the background distribution  $f_b$ . Since the background is known and well explained by the existing physical theories, we may assume  $n_b$  to be as large as needed to estimate  $f_b$  arbitrarily well.  $\mathcal{X}_{bs} = \{\mathbf{x}_i\}_{i=1,\dots,n_{bs}}$  has the same structure as  $\mathcal{X}_b$  and denotes



the unlabelled set of data, assumed to be drawn from the distribution  $f_{bs}$  underlying the whole process. We assume that  $f_{bs}$  and  $f_b$  could be different just because of the presence of a signal which features as a new mode of  $f_{bs}$ , not arising from  $f_b$ .

## 2 The statistical framework

According to the nonparametric formulation of density-based clustering, the observed data  $\mathcal{X} = \{\mathbf{x}_i\}_{i=1,\dots,n}$ ,  $\mathbf{x}_i = (x_{i1}, \dots, x_{ij}, \dots, x_{id})' \in \mathbb{R}^d$  are supposed to be a sample from a random vector with unknown probability density function  $f$ , whose modes are regarded as the archetypes of the clusters, in turn represented by the surrounding regions. After building a nonparametric estimate  $\hat{f}$  of  $f$ , the identification of the modal regions may occur according to different directions. One strand of methods looks for an explicit representation of the modes of  $f$  and associates each cluster to the set of points along the steepest ascent path towards a mode, e.g. via the mean-shift algorithm. A second class of methods does not attempt explicitly the task of mode detection but associates the clusters to disconnected density level sets of the sample space, as the modes correspond to the innermost points of these sets. See [4] for a review of these approaches.

Whatever direction is followed, any estimate of  $f$  leaves defined the modal structure and hence the clustering. However, nonparametric density estimation is a critical task, at least with respect to two aspects. First, the shape and the number of modes of the density estimate depend on the regulation of some smoothing parameter, whatever estimator is chosen. While not binding, in the rest of the paper, we focus on the specific case of product kernel estimator:

$$\hat{f}(\mathbf{x}; \mathcal{X}, h) = \frac{1}{n \cdot h^d} \sum_{i=1}^n \prod_{j=1}^d K\left(\frac{x_j - x_{ij}}{h}\right), \quad (1)$$

where  $K$  is the kernel, usually a symmetric probability density function, and  $h > 0$  is the bandwidth. A large bandwidth tends to oversmooth the density, possibly pulling out its modal structure, while a small bandwidth favours the appearance of spurious modes. How to set the amount of smoothing is then an issue to be tailored.

A second critical aspect related to density estimation, and worth to be accounted for, is the dimensionality of the problem at hand. The curse of dimensionality is known to have a strong impact on nonparametric density estimators and this explains a worsened behaviour of modal clustering for increasing  $d$ . In high dimensions, much of the probability mass flows to the tails of the density, possibly giving rise to the birth of spurious clusters and averaging away features in the highest density regions. Since a typical aspect of high dimensional data is the tendency to fall into manifolds of lower dimension, dimension reduction methods are often advisable.

### 3 A nonparametric semisupervised approach

Our contribution, to include a source of supervision in nonparametric clustering, builds on the idea of exploiting available information on the known labelled process to aid the most critical aspects of the nonparametric framework, i.e. density estimation in high dimensions and selection of the smoothing amount.

To address the former issue, here we propose a variable selection approach specifically formulated for this context. The procedure is based on the idea that a possible different behavior between  $f_b$  and  $f_{bs}$  shall be only due to the presence of a signal of interest in  $f_{bs}$ ; hence, a variable will be considered to be relevant if it contains any trace of signal. The approach here adopted consists in comparing repeatedly the estimates of multivariate marginal distributions of  $f_b$  and  $f_{bs}$ , at each step on a different, randomly selected subset of variables. In this way we operate in lower dimensional spaces, with a gain in density estimation accuracy, while accounting for relations among variables. The comparison is based on the use of the non parametric statistic [3] to test equality of two distributions. If a different behavior is detected, the procedure updates a counter for the variables selected at that step; at the end of the procedure the counter will indicate the relevance of each single variable. The procedure allows for selecting a smaller subset of variables to work with, leading both to interpretative and computational advantages.

To address the second critical aspect discussed above, we propose a procedure whose rationale is the following. We identify the modal partition of the unlabelled data associated with the nonparametric estimate  $\hat{f}_{bs}$  which guarantees the most accurate classification of the labelled background observations. Given an estimate  $\hat{f}_b$  of  $f_b$ , supposed to be arbitrarily accurate due to our knowledge of the background process, a partition  $\mathcal{P}_b(\mathcal{X}_b)$  of the background data remains associated. Then, we get multiple estimates  $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, h_{bs})$  of  $f_{bs}$  for  $h_{bs}$  varying in a range of plausible values. Each of these estimates identifies a partition  $\mathcal{P}_{bs}(\mathcal{X}_{bs})$  and, eventually, also a partition  $\mathcal{P}_{bs}(\mathcal{X}_b)$  of the background data, both defined by the modal regions of  $\hat{f}_{bs}$ . The latter classification is obtained by assigning a background observation to the cluster of  $\hat{f}_{bs}$  for which its density is the highest.  $\mathcal{P}_{bs}(\mathcal{X}_b)$  is then compared with  $\mathcal{P}_b(\mathcal{X}_b)$  via the computation of some agreement index  $I$ . The bandwidth  $h_{bs}$  that maximizes  $I$  is then selected to estimate  $f_{bs}$  and identify the ultimate partition  $\mathcal{P}_{bs}(\mathcal{X}_{bs})$ . The main steps of the procedure are listed in the Pseudo-algorithm 1.

From an operational point of view we use, to obtain partitions, the clustering method [1] and, as agreement index, the *Adjusted Rand Index*. Furthermore,  $\mathcal{P}_{bs}(\mathcal{X}_b)$  and  $\mathcal{P}_b(\mathcal{X}_b)$  are not, in fact, compared on the whole background sample  $\mathcal{X}_b$  but on a number of different bootstrap samples from  $\mathcal{X}_b$ ; this allows us to get the empirical distribution of the agreement index  $I$  and obtain more reliable results.

Eventually we note that, besides the background process is known and  $\mathcal{X}_b$  is arbitrarily large, the procedure presented above requires an estimate of the background density  $f_b$ , i.e. the relative choice of the bandwidth. To this aim we rely on the concept of stability of the density estimate and select the bandwidth that minimizes the *integrated squared distance* among density estimates computed from different samples drawn from the background process.

**Pseudo-algorithm 1** *Semisupervised procedure for bandwidth selection*

Denote with:  $\mathcal{X}_b$  the *background* sample,  $\mathcal{X}_{bs}$  the *unlabelled* sample from the whole process; it is assumed that the dimensionality of both samples has been already reduced via variable selection. Let  $h_b$  be the *background* bandwidth;  $h_{bs}$  the whole process bandwidth (to be determined);  $h_{grid}$ : a grid of plausible values for  $h_{bs}$ . Finally let  $\mathcal{P}_k(\mathcal{X})$  be a partition of data  $\mathcal{X}$  identified by the modal structure of density  $f_k$  and  $I(\mathcal{A}, \mathcal{B})$  an agreement index between partitions  $\mathcal{A}$  and  $\mathcal{B}$

---

**Input**  $\mathcal{X}_b, \mathcal{X}_{bs}, h_b, h_{grid}$ .

- 1: compute  $\hat{f}_b(\cdot; \mathcal{X}_b, h_b)$ ;
- 2: obtain  $\mathcal{P}_b(\mathcal{X}_b)$ ;
- 3: **for**  $h$  in  $h_{grid}$  **do**
- 4:   compute  $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, h)$ ;
- 5:   obtain  $\mathcal{P}_{bs}(\mathcal{X}_{bs})$ ;
- 6:   compute  $I(\mathcal{P}_b(\mathcal{X}_b), \mathcal{P}_{bs}(\mathcal{X}_{bs}))$ .
- 7: **end for**
- 8:  $h_{bs} = \operatorname{argmax}_{h \in h_{grid}} I(\mathcal{P}_b(\mathcal{X}_b), \mathcal{P}_{bs}(\mathcal{X}_{bs}))$
- 9: compute  $\hat{f}_{bs}(\cdot; \mathcal{X}_{bs}, h_{bs})$ ;
- 10: obtain  $\mathcal{P}_{bs}(\mathcal{X}_{bs})$ ;

**Output:**  $\mathcal{P}_{bs}(\mathcal{X}_{bs})$ .

---

## 4 Empirical results

In this section, we show the results of the application of the proposed procedure on a *Monte-Carlo* physical process simulated within the CMS experiment; the experiment refers to high-energy proton-proton collisions where each observation corresponds to a single collision event and may produce particles from two different physical processes: the *QCD multijet background*, and a signal known as *top pair production*.  $\mathcal{X}_b$  includes  $n_b = 20000$  background observations, while  $\mathcal{X}_{bs}$  include  $n_{bs} = 10000$  observations, whose the 16% comes from the signal process. For each dataset we observe  $d = 30$  variables related to the kinematic characteristics of the particles produced by the proton collisions. While both  $\mathcal{X}_b$  and  $\mathcal{X}_{bs}$  are labelled, labels of  $\mathcal{X}_{bs}$  have been employed only for evaluating the quality of the results.

In Figure (1) the results of the variable selection procedure are displayed. Two features (*dp12* and *jsv1*) show a remarkably different behavior between the background and whole process densities. In the subsequent analyses we have worked with these two variables only.

After estimating  $f_b$  based on a bandwidth selected to guarantee the density stability as explained in the previous section, we applied the procedure reported in Pseudo-algorithm 1. The obtained bandwidth was used to estimate the density  $f_{bs}$  of the whole process and thus obtaining a partition of  $\mathcal{X}_{bs}$  via the subsequent application of nonparametric clustering. Results, reported in the right table of Figure 1, compare the obtained partition with the known actual labels. The procedure identifies four different groups: two of them clearly refer to the background process, while the other two mostly contain observations coming from the *top pair production* sig-

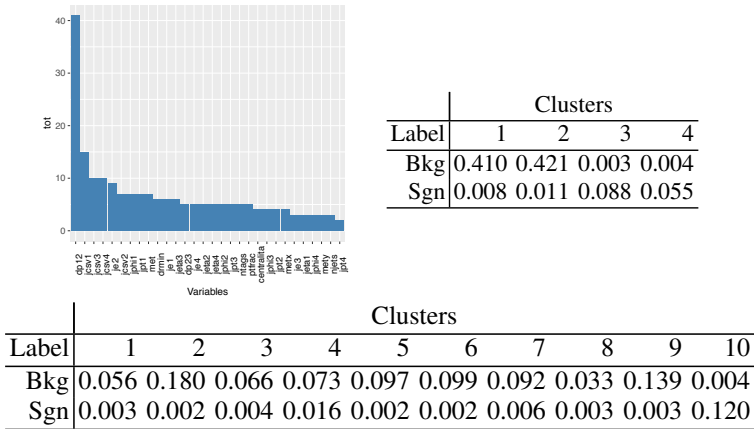


Fig. 1: Top left: results of variable selection procedure; variables are ordered decreasingly by importance (higher bar implies higher importance). Top right: true process labels vs clusters detected by the proposed semisupervised procedure. Bottom: true process labels vs clusters detected by the benchmark method [5].

nal. The overall misclassification error is equal to 2.6% with a true positive rate larger than 88%. For comparison purposes we also present results of the application of the competitive methodology proposed in [5]. Data dimensionality have been previously reduced by keeping four principal components, as proposed by the authors. Working in a parametric framework there is one-to-one relationship between mixture components and clusters; hence the method find 9 background clusters and an additional one capturing the signal. The overall error is equal to 4.5% with a true positive rate amounting to the 74.5%.

## References

1. Azzalini, A., Torelli N.: Clustering via nonparametric density estimation. *Stat Comput*, 17(1): 71-80 (2007).
2. Chandola, V., Banerjee A., Kumar V.: Anomaly detection: A survey. *ACM Comput Surv* 41(3): 1-58 (2009).
3. Duong, T., Goud, B., Schauer, K.: Closed-form density-based framework for automatic detection of cellular morphology changes. *P Natl Acad Sci Usa*, 109(22): 8382-8387 (2012).
4. Menardi, G.: A review on modal clustering. *Int Stat Rev*, 84(3): 413-433 (2016).
5. Vatanen, T., Kuusela, M., Malmi, E., Raiko T., Aaltonen T., Nagai, Y.: Semi-supervised detection of collective anomalies with an application in high energy particle physics. *Int Jt Conf Neural Netw*: 1-8 (2012).

# Employment study methodologies of Italian graduates through the data linkage of administrative archives and sample surveys

## *Metodologie per lo studio dell'occupazione dei laureati Italiani attraverso il data linkage di archivi amministrativi con quelli di indagini campionarie*

Claudio Ceccarelli, Silvia Montagna and Francesca Petrarca

**Abstract** We discuss the issues and the related study methodologies raised by the data linkage among different Istat archives to provide information on the employment status of Italian graduates. To this aim many different administrative archives are integrated with data from sample surveys on university graduates' vocational integration. From this integration, a very complex situation emerges which must be analysed and correctly interpreted. In this paper in order to show the feasibility of our method, we discuss the comparison among these sources and we present the strategy of constructing appropriate indicators.

**Abstract** *Si discutono le problematiche e le relative metodologie di studio necessarie messe in luce dall'integrazione di differenti archivi Istat per ottenere informazioni sullo stato occupazionale dei laureati Italiani. A questo scopo, sono stati integrati diversi archivi amministrativi con i dati provenienti dalle indagini campionarie sull'inserimento professionale dei laureati. Da questa integrazione è emersa una situazione molto complessa che deve essere analizzata e correttamente interpretata. In questo articolo, per mostrare l'applicabilità del nostro metodo, si discute il confronto tra le diverse fonti e si presenta la strategia per costruire indicatori appropriati.*

**Key words:** administrative data, data linkage, sample surveys

---

Claudio Ceccarelli

Istat, Via Cesare Balbo 16, Rome, Italy, e-mail: [clceccar@Istat.it](mailto:clceccar@Istat.it)

Silvia Montagna

Istat, Via Cesare Balbo 16, Rome, Italy, e-mail: [montagna@Istat.it](mailto:montagna@Istat.it)

Francesca Petrarca

Istat, Via Cesare Balbo 16, Rome, Italy, e-mail: [francesca.petrarca@uniroma1.it](mailto:francesca.petrarca@uniroma1.it)

## 1 Introduction

In the last few years, Italian universities have shown an increasing interest in studying the characteristics of the labour market demand for their alumni in line with the requests coming from the Italian Industry [4]. Periodic reports have been produced by public institutions, for example, the Italian National Institute of Statistics (Istat) [10]-[9] or private entities, such as [5] and AlmaLaurea [1].

The information on the contracts obtained by each graduate retrieved in the integrated database include data about the type and the job qualification, the sectors of economic activity, the location of businesses and the educational curriculum of the alumni in the period from upper secondary school diploma to university graduation.

A general plan of modernization which includes the Italian Statistical Institute [11]-[2] and other National Statistical offices has the purpose to produce the best possible estimates to meet user needs from multiple data sources, from surveys, administrative archives and new sources such as big data, and moreover to reduce burden and costs. Of course the problem of the integration of administrative archives to produce useful statistics is an issue also addressed and discussed in the rest of the world, (see, [6] and [8]).

In this paper, data are used to develop an explorative study on the transition of graduates into the Italian labour market taking advantage from data coming from administrative archives which guarantee sophisticated and robust statistical analyses [7]-[3]-[12]-[14]. Our database also contains data coming from the sample survey on university graduates' vocational integration.

## 2 Sources of data

For the first time, administrative data concerning Italian graduates in the year 2011 and their employment status information in the following four years are available. These data can be used for a benchmark among the answers given by the interviewee and the evidence from administrative sources. It must be anticipated that the study has a strong experimental character dictated by various reasons mainly related to criticalities of the available sources.

The main data sources are:

1. Survey sample on university graduates' vocational integration on graduates in 2011: the interview conducted in 2015, four years after graduation, detects the training paths and employment outcomes in relation to different moments or time intervals:
  - Before graduation;
  - at moment of graduation;
  - at one year from graduation;
  - at the moment of the interview.

Survey final data are stored in an archive called *Armida*, which contains data related to 58,400 individuals. For more details on the survey see [9] and [10].

In Tab. 1 we report the percentage of Italian graduates who were working at the moment of the degree, after one year and after four year for different level of Italian University degree coming from the survey.

**Table 1** Percentage of Italian graduates which work at the moment of the degree, after one year and after four year for different level of Italian University degree

	At the moment of grad.	1 year from grad.	4 years from grad.
First cycle degree (bachelor)	28.7	37.4	72.8
Second cycle degree (master)	34.7	55.7	84.5
Single-cycle master degree	27.0	40.3	80.3

Sample Survey Source

2. The National Register of students (called ANS), Ministry of Education University and Research (MIUR) source, provides, for each individual, the personal data and its university career from enrolment to graduation. In this paper, we considered all the graduates who got their university degree in the year 2011. Tab. 2 reports the number of graduates in 2011 for different levels of Italian University degrees.

**Table 2** Numbers of graduates in the 2011 for different levels of Italian University degrees

	Graduates
All	299,449
First cycle degree (bachelor)	169,232
Second cycle degree (master)	86,593
Single-cycle master degree	43,624

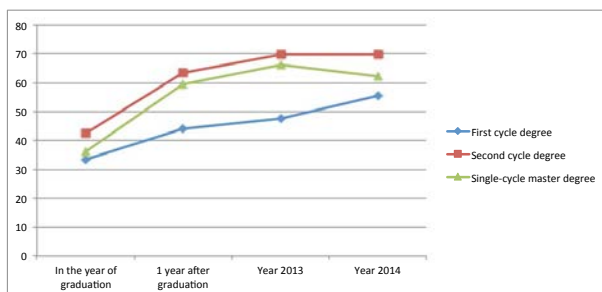
MIUR administrative Source

3. The *Integrated Base of Administrative Sources* of Istat contains the employment status of the Italian population. This administrative archive records all the *business relationships* of the Italian population. In order to identify a business relationship, it is necessary to have in the database an administrative record confirming a relationship with an employer (evidence of type LEED-Linked Employer Employee Dataset or Database<sup>1</sup>). The administrative record must be related to a contributory position (INPS source) or a social security position (INPDAP source), or other event associated with the worker and recorded in one of other archives on employment available at the time of the analysis. [13].

Therefore, in this case the employment status of the graduates is given through their contributory position or social security position. This information is recorded for each month of the year. The main issue with administrative data is that we

<sup>1</sup> LEED is the result of research linking every employer to its employees and vice versa every worker to his employer.

must define a proper definition of graduate employed. In this study, we define *worker* a graduate for whom at least a contributory position in a month of the considered year has been recorded in the Integrated Base of Administrative Sources archive.



**Fig. 1** Behaviors of employment in the subsequent four years after graduation for different levels of Italian University degrees coming from the administrative archive.

We report in Tab.3 data for the employment of the graduates in the year of graduation and after graduation. Looking at the year 2014, first cycle degrees have less chance of getting an employment respect to the other kind of degrees. The variation of employment, in the four years, can be seen in Fig. 1; this shows a similar behavior for the curve of the second cycles and single-cycle master degree. The two curves increase up to the year 2013 and then there is a saturation. In the case of the first cycle degrees, the curve grows almost linearly. In the last column of Tab. 3 are reported the percentage of graduates who have an employment for each year of the subsequent four years after graduation. Also for this type of graduates, the first cycle degrees are the less favored.

**Table 3** Percentage of Italian graduates who work in the four years after graduation for different levels of Italian University degrees

	in the year of graduation	1 year after graduation	2013	2014	ALL
First cycle degree (bachelor)	33.41	44.07	47.66	55.54	22.22
Second cycle degree (master)	42.33	63.58	69.58	69.56	31.77
Single-cycle master degree	36.13	59.25	66.17	62.16	28.11

Integrated Base of Administrative Sources of Istat



### 3 Comparison

The statistical sample of the graduates in the 2011 (58,400 units) is merged with the Integrated Base of Administrative Sources with the aim to compare the percentage of working graduates. We call this data *combined*. In first column of Tab. 4, we show the percentage of graduates which are recognized as working graduates at one year from graduation by analysing combined data. These percentages are obtained by analysing the administrative information contained in the combined data records. For the sake of comparison, in the second column of Tab. 4 are reported the percentages obtained by the sample survey information contained in the combined data records.

**Table 4** Comparison between the percentage of Italian graduates who work at one year from graduation.

	Combined data	
	Administrative information	Sample survey information
First cycle degree (bachelor)	37.86	37.40
Second cycle degree (master)	55.06	56.70
Single-cycle master degree	40.89	40.30

We observe small differences of the results: the administrative information of the combined data are slightly higher. This was expected because the use of administrative data reduces the possibility to loose units, in fact the results of the sample survey are subjective and over dependent on the memory of the interviewee. Moreover, sometime, during the interview, people do not declare employments which are not coherent with the educational path or of short duration. It is worth nothing that in the case of administrative data the indicator of working graduates is able to get a better classification of the working units.

We underline the importance of using administrative data for the study of the entrance of graduates into the Italian labour market. We have briefly shown that an administrative archive is flexible and rich enough to analyse the work paths of graduates in the years following graduation. Moreover, the administrative data allow us to study the evolution of graduates after the graduation and therefore to analyse changes in their job position. We plan to perform these studies through longitudinal analyses. The analyses on the Integrated database could be adopted as a permanent monitor of the entrance of graduates into the Italian labour market over the years which may be accompanied from the sample survey.

### References

1. AlmaLaurea (2015) XVII rapporto 2015 AlmaLaurea sulla condizione occupazionale dei laureati. Tech. rep., Consorzio Universitario AlmaLaurea

2. Barcaroli G, Falorsi PD, Fasano A, Mignolli N (2014) A Business Architecture Model to Foster Standardisation in Official Statistics. European Conference on Quality in Official Statistics- Quality 2014
3. Capecchi S, Iannario M, Piccolo D (2012) Modelling job satisfaction in AlmaLaurea surveys. AlmaLaurea Working paper n 50
4. Carpita M (2011) Laureati Stella: Rapporto statistico 2008-2011- Progetto CILEA. Grafiche Porpora
5. CENSIS (2012) Quarantaseiesimo Rapporto sulla situazione sociale del Paese 2012. Franco Angeli
6. CESS2014 (2014) CONFERENCE OF EUROPEAN STATISTICS STAKEHOLDERS-methodologists, producers and users of european statistics. <http://cdss.sta.uniroma1.it/files/site/Program.pdf>
7. Ciriaci D, Muscio A (2011) University choice, research quality and graduates' employability: Evidence from Italian national survey data. AlmaLaurea Working paper n 48
8. Citro CF (2014) From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology* 40 (2):137161
9. ISTAT (2015a) University graduates vocational integration. <http://www.istat.it/en/archive/82425>
10. ISTAT (2015b) I percorsi studio e lavoro dei diplomati e dei laureati. <https://www.istat.it/it/files/2016/09/I-percorsi-di-studio-e-lavoro-dei-diplomati-e-laureati.pdf?title=Percorsi+lavorativi+di+diplomati+e+laureati+-+29%2Fset%2F2016+-+I+percorsi+di+studio+e+lavoro+dei+diplomati+e+laureati.pdf>
11. ISTAT (2016a) Programma di modernizzazione. [http://www.istat.it/it/files/2010/12/Programma\\_modernizzazione\\_Istat2016.pdf](http://www.istat.it/it/files/2010/12/Programma_modernizzazione_Istat2016.pdf)
12. Petrarca F (2014) Assessing Sapienza university alumni job careers: Enhanced partial least squares latent variable path models for the analysis of the UNICO administrative archive. PhD thesis, University of Roma Tre, [On- line; accessed 3-Jun-2014]
13. Runci M C, Di Bella G and Galìè L. (2016), Il sistema di integrazione dei dati amministrativi in Istat, Istat working paper n. 18/2016. [http://www.istat.it/it/files/2016/11/IWP\\_18\\_20161.pdf](http://www.istat.it/it/files/2016/11/IWP_18_20161.pdf)
14. UNICO Group (2015) La Domanda di Lavoro per i laureati. I risultati dell'integrazione tra gli archivi amministrativi dell'Università Sapienza di Roma e del Ministero del Lavoro e delle Politiche Sociali. Edizioni Nuova Cultura

# Prediction of Firm's Creditworthiness Risk using Feature Selection and Support Vector Machine

Ikram Chairi, Amina El Gonnouni, Sarah Zouinina, Abdelouahid Lyhyaoui  
Laboratory of Innovative Technologies, Abdelmaklek Essaadi University  
{Chairikram, sarahzouinina1, lyhyaoui}@gmailcom, amina\_elgo@yahoo.fr

**Abstract.** A series of challenges have recently emerged in the data mining field, triggered by the rapid shift in status from academic to applied science and the resulting needs of real-life applications. The recourse to statistical learning models as support vector machines (SVM), or neural networks, is a common practice. However, the performances of those algorithms strongly depend on the quality of the data used. This constraint, oblige the data scientist to employ different statistical methods before using those algorithms. This paper aims to apply feature selection method on financial data of 20 firms in order to to set up our Support Vector Machine (SVM) Model through which we can predict firms' creditworthiness risk.

**Key words:** Feature Selection, Dimensionality Reduction, Factor Analysis Model, SVM, Intelligent Financial Solution, Financial Health.

## 1 Introduction

Nowadays, due to the development that sciences and technologies have known lastly, several data anomalies has appeared. The most common anomalies that we find in real world data are in general the incomplete records, irrelevant and/or redundant pieces of information, imbalanced class distribution and imbalanced error costs, but the most redundant problem stills the big size of data.

Indeed, many scientific studies are featured by the huge number of variables used. Because of these big numbers of variables that are into play, the study can become rather complicated. In these cases, dimensionality reduction techniques are highly used. In this paper, we perform a dimensionality reduction using correlation matrix and factor analysis. The reduced dataset will be used as an input of a Support Vector Machine algorithm in order to generate a prediction model. Experimental results on a set of financial data for 20 Moroccan firms will be presented and discussed.

This paper is structured as follows. Section 2 presents an overview on feature selection and dimensionality reduction, in section 3 we will present the SVM approach. In section 4 we present the data set and the experimental results. Finally, the last section provides some concluding comments.

## 2 Feature Selection And Dimensionality Reduction of Data

Feature selection techniques have become an apparent need in many different fields because of the constant growth of data. The objective of feature selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data [1]

Among the most used methods of feature selection we found the factor analysis which aims to reduce “the dimensionality of the original space and to give an interpretation to the new space, spanned by a reduced number of new dimensions which are supposed to underlie the old ones” [2]

The Factor Analysis Model is one of several techniques which seek to explain the correlation between a set of variables by a smaller set of random variables. It uses a small number of imaginary variables to express the basic data structure by studying the internal relationship between the variables, and reflects the main information and interdependence of these original data.

The following assumes that the  $p$  observed variables (The  $X_i$ ) that have been measured for each of the  $n$  subjects have been standardized.

$$\begin{aligned} X_1 &= a_{11}F_1 + \dots + a_{1m}F_m + e_1 \\ X_2 &= a_{21}F_1 + \dots + a_{2m}F_m + e_2 \\ &\vdots \\ X_p &= a_{p1}F_1 + \dots + a_{pm}F_m + e_p \end{aligned}$$

$F_j$  are the  $m$  common factors,  $e_i$  are the  $p$  specific errors, and the  $a_{ij}$  are the factor  $p \times m$  factor loadings.

In matrix form this can be written as:

$$X_{px1} = A_{pxm}F_{mx1} + e_{px1}$$

It doesn't make sense to use factor analysis if the different variables are unrelated; this is why the starting point of factor analysis is a correlation matrix.

The correlation Matrix provides the inter-correlations between the studied variables. The dimensionality of this matrix can be reduced by looking for variables that correlate highly with a group of other variables, but correlate very badly with variables outside of that group. These variables with high inter-correlations could well measure one underlying variable, which is called a factor.

### 3 Support Vector Machine Model

Support Vector Machine (SVM) is a powerful method for pattern recognition and classification introduced by Vapnik [3]. The SVM maps the input data into a higher dimensional feature space via a nonlinear map and construct a separating hyperplane with maximum margin. It has been proposed as a technique in times series prediction. The key characteristic of SVM is that a nonlinear function is learned by a linear learning machine in a kernel induced feature space while the capacity of the system is controlled by a parameter that does not depend on the dimensionality of the space. The following shows the SVM algorithm [4]:

Consider a given training set  $\{x_i, y_i : i=1, \dots, l\}$  randomly and independently generated from an unknown function, where  $x_i \in X \subseteq R^n, y \in Y \subseteq R$  and  $l$  is the total number of training data.

The SVM approximates the unknown function using the following form:

$$f(x) = \langle w, \Phi(x) \rangle + b \tag{1}$$

Where  $\langle \cdot \rangle$  is the dot product,  $w$  and  $b$  are the estimated parameters and  $\Phi$  is a nonlinear function used to map the original input space  $R^n$  to high dimensional feature space. So, the nonlinear function estimation in original space becomes linear in feature space.

The optimization goal of standard SVM is formulated as:

$$\text{Minimize } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \tag{2}$$

Subject to:

$$\begin{aligned} y_i - \langle w, \phi(x_i) \rangle - b &\leq \varepsilon + \xi_i, \\ \langle w, \phi(x_i) \rangle + b - y_i &\leq \varepsilon + \xi_i^*, \\ \xi_i^*, \xi_i &\geq 0, i=1, \dots, l. \end{aligned}$$

Where the constant  $C > 0$  determines the tradeoff between the flatness of  $f$  and the amount up to which deviations larger than are  $\varepsilon$  tolerated and  $\xi_i, \xi_i^*$  are slack variables and they are introduced to accommodate, respectively, the positive and the negative errors on the training data. The formulation above corresponds to dealing with the so called  $\varepsilon$ -insensitive cost function:

$$|\xi|_{\varepsilon} := \begin{cases} 0 & \text{if } |\xi| < \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases}$$

The nonlinear function is obtained as:

$$f(x) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) K(x, x_i) + b \quad (5)$$

Where  $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$  is defined as the kernel function. The elegance of using the kernel function is that one can deal with feature spaces of arbitrary dimensionality without having to compute the map  $\Phi$ . Any function that satisfies Mercer's condition can be used as the kernel function.

## 4 Proposed method and experimental results

In this paper, we propose to reduce the dimension of a real world data set using factorial analysis before using the reduced data on a SVM in order to generate a prediction model.

For our experiences, we choose a financial data over 3 years 2009-2011, which all come from 20 Moroccan companies that belong to different sectors and have different sizes. Firstly, we have selected 39 variables, over 3 years, which are impacting the financial health of companies:

**Table 1:** Selected variables

V1	Turnover	V14	Equity / Total assets	V27	Financial costs / Gross operating profit
V2	Net equity	V15	Working capital / Working capital requirement	V28	Financial costs / Operating cash surplus
V3	Net cash	V16	Leverage ratio	V29	Gross operating profit / Turnover
V4	Net profit	V17	Coverage ratio	V30	Net profit margin ratio
V5	Working capital	V18	Change in debts / Cash flow	V31	Net profit / Net equity
V6	Working capital requirement	V19	Rotation net cash	V32	Net profit / Permanent capital
V7	Value added	V20	Inventory turnover	V33	Long term debts / Cash flow
V8	Gross operating margin	V21	Delay of payment of customers	V34	Net profit / Equity
V9	Gross operating profit	V22	Delay of payment to vendors	V35	Staff costs
V10	Operating cash surplus	V23	Gross margin / Turnover	V36	Lenders
V11	Free cash flow	V24	Cash flow / Turnover	V37	State
V12	Cash flow	V25	Productivity ratio	V38	Current operating income
V13	Solvency	V26	Staff costs / Value added	V39	Non-operating income

When we assess a firm's creditworthiness risk, we want to know if a company is solvent, if it is profitable and if it is still productive; this is why we have computed the correlation coefficients, using SPSS 10, between our 39 variables involved and the three relevant components: Solvency, productivity and profitability, in order to detect the variables that influence the most

We next reduced the dimensionality of the input factors using Principal Factor Analysis (PFA) approach by SPSS10, in order to obtain factors that create a new dimension and that can be visualized as classification axes along which measurement variables can be plotted [5].

Table 2 shows the 8 factors that will be retained instead of 39 variables. The eight factors based on the data from 2009-2010 can explain 86% of the variance contribution, which means the model has a good measuring effect. So, we can reduce the input factors dimension to eight factors to predict the creditworthiness risk.

**Table 2:** Input Factors.

Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
V1	0.921	-0.082	0.077	0.042	-0.021	-0.050	-0.109	-0.061
V2	0.995	0.020	-0.022	-0.049	0.013	-0.008	-0.032	-0.015
V3	-0.299	0.072	-0.095	0.747	0.401	0.081	0.166	0.188
V4	0.986	-0.006	-0.023	0.073	0.061	-0.003	-0.004	0.001
V5	0.941	-0.005	-0.023	-0.135	-0.009	-0.005	-0.006	-0.012
V6	0.934	-0.016	-0.007	-0.242	-0.070	-0.017	-0.031	-0.040
V7	0.996	-0.031	-0.004	0.008	0.025	-0.013	-0.019	-0.013
V8	0.997	-0.032	0.000	-0.007	0.016	-0.015	-0.025	-0.017
V9	0.990	-0.031	-0.009	0.035	0.039	-0.010	-0.010	-0.009
V10	0.912	-0.035	-0.024	0.335	0.178	0.014	0.040	0.054
V11	0.873	-0.031	-0.030	0.404	0.201	0.026	0.063	0.077
V12	0.901	0.009	-0.045	0.362	0.211	0.019	0.045	0.070
V14	0.299	0.571	0.010	0.016	-0.076	0.194	0.263	-0.192
V15	-0.010	-0.014	0.000	0.144	-0.309	-0.179	-0.219	0.586
V16	-0.070	-0.320	-0.170	-0.266	0.521	0.488	-0.175	-0.086
V17	0.068	0.313	0.899	-0.105	0.176	0.062	0.002	0.088
V18	0.041	0.381	0.849	-0.001	0.101	-0.013	0.098	0.039
V19	-0.026	0.469	-0.066	0.351	-0.204	0.405	-0.290	-0.143
V20	-0.090	-0.091	-0.076	-0.220	0.163	-0.065	0.762	0.210
V21	0.215	0.226	-0.345	-0.067	0.139	-0.510	0.159	0.058
V22	-0.032	0.585	-0.400	-0.142	0.040	0.051	-0.133	0.273
V23	0.318	0.432	-0.384	-0.219	-0.154	0.320	0.403	0.256
V24	0.026	0.878	-0.282	-0.064	0.141	0.028	-0.150	0.001
V26	0.974	-0.030	0.016	-0.098	-0.026	-0.023	-0.056	-0.026
V27	0.920	-0.006	0.010	-0.209	-0.075	-0.021	-0.077	-0.036
V28	0.987	-0.029	-0.012	0.083	0.064	-0.009	-0.003	-0.001
V29	0.991	-0.028	-0.012	0.032	0.040	-0.009	-0.006	-0.007
V30	-0.911	0.040	-0.014	0.345	0.135	0.044	0.064	0.084
V31	-0.045	0.007	0.038	0.133	-0.020	-0.422	-0.142	-0.302
V32	-0.031	-0.736	-0.059	0.093	-0.061	0.172	-0.092	0.021
V33	0.036	-0.159	0.084	-0.130	-0.054	0.220	-0.365	0.580
V35	0.537	-0.341	0.077	-0.005	-0.309	0.351	0.242	0.175
V36	0.060	0.871	-0.277	-0.066	0.089	0.011	-0.175	-0.017
V37	0.209	0.249	0.209	0.329	-0.697	-0.159	0.115	0.066
V38	-0.024	0.085	0.023	0.165	-0.346	0.431	0.161	-0.371
V39	0.068	0.313	0.899	-0.105	0.176	0.062	0.002	0.088

As cited before, we use SVM to predict the model for the reduced data. To obtain a good performance, we have carefully chosen some parameters that include the

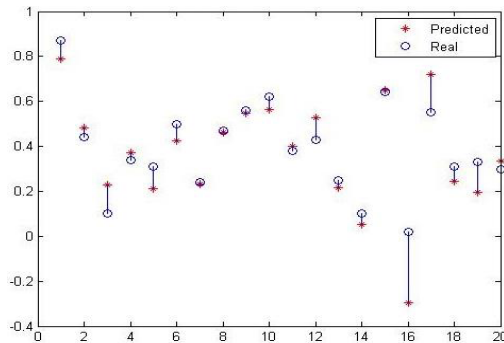
regularization parameter  $C$ , which determines the trade-off between minimizing the training error and minimizing model complexity, and parameters of the Kernel function.

In this simulation we test the classification using the kernel function RBF so two parameters need to be chosen; they are the  $\gamma$  width of the RBF function and the soft margin parameter  $C$  of SVM [6].

One method often used to select the parameters is grid search on the log ratio of the parameters associated with cross-validation. Value pairs  $(C, \gamma)$ , respectively was assessed using cross-validation and then we have chosen the pair with highest precision:  $(C, \gamma) = (100, 0.1)$ .

According to the architecture of the support vector machine, only the training data near the boundaries are necessary. In addition, because the training time becomes longer as the number of training data increases, the training time is shortened if the data far from the boundary are deleted. Therefore, we have implemented a sample of 40 Moroccan companies whose financial data is extracted over (2009-2011) and reduced on our 8 factors analysis. Then we have applied our SVM model over the training set on a new sample of 20 Moroccan companies whose financial data is selected over (2009-2011), with the purpose to measure the precision of creditworthiness risk prediction as compared to the actual data of 2011.

In order to test the effectiveness of the proposed method, a series of simulations were carried out to predict solvency, productivity and profitability, as follows:



**Figure 1:** Solvency risk prediction



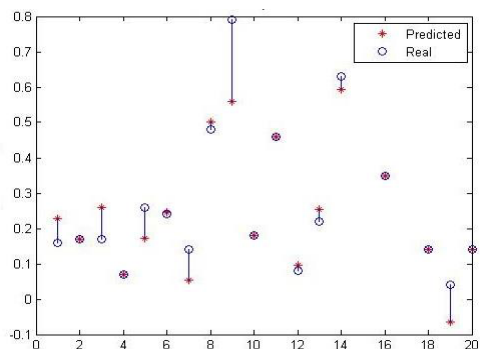


Figure 2: Productivity risk prediction

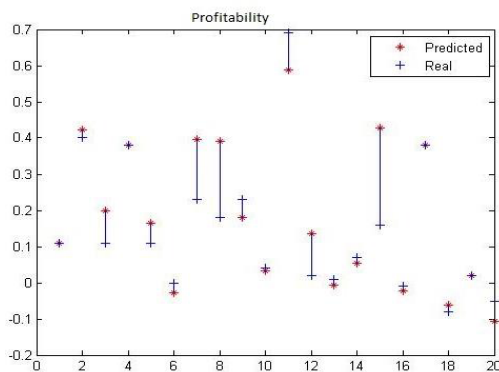


Figure 3: Profitability risk prediction

As proved by the results above, the fact that the precision of the creditworthiness risk prediction is about 90% means that the model has a good measuring effect.

## 5 Conclusion

We presented in this paper a simple and efficient way to improve the performance of the SVM predictor and this by reducing the dimensionality of the training set using factor analysis as feature selection technique. Experiences were generated on a sample of financial data of 20 companies over 2009-2011.

The simulation results show that our SVM model gives good precisions, and that we are able to forecast the companies' default and to give intelligent financial solutions to investors and financial institutions to help them in decision-making.

We consider this study as a beginning of a line of research in which we will explore more parameters that can improve the performance of prediction.

## References

- [1] G. Isabelle and E. Andre, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research* 3, 2003.
- [2] T. Rietveld and R. Van Hout, "Statistical Techniques for the Study of Language and Language Behaviour," Berlin – New York, 1993.
- [3] M. Sayan, O. Edgar and G. Federico, "Nonlinear Prediction of Chaotic Times Series Using Support Vector Machines," in *Proceedings of the IEEE NNSP'97*, Sep 1997.
- [4] C. Nello and s.-t. john, "An introduction to support vector machines, and other kernel based learning methods," in *Cambridge press*, 2000.
- [5] D. Bartholomew, "The Foundation of Factor Analysis," *Biometrika*, vol. 71, pp. 221-32., 1984.
- [6] N. Marcelo, R. Kapp, P. Sabourin and A. Maupin, " PSO-Based Framework for Dynamic SVM Model Selection," in *GECCO'09*, Montréal Québec, 2009.

# Contribution of extracting meaningful patterns from semantic trajectories

Sana Chakri, Said Raghay, Salah El Hadaj<sup>1</sup>

**Abstract** Explosive growth in geospatial and temporal data emphasizes the need for automated discovery of spatiotemporal knowledge. Different algorithms have been proposed in the last few years for discovering different types of behaviours in trajectory data, integrating trajectory sample points with geographical and contextual data before applying mining techniques can be more gainful for the application users. It contributes to produce significant knowledge about movements and provide applications with richer and more meaningful patterns. Trajectory Outliers are a sort of patterns that can be extracted from trajectories. We propose a new approach for trajectory outlier detection based on semantic data besides than geometric data.

**Key words:** moving objects analysis, spatial databases, data-mining, semantic clustering, semantic trajectories

## 1 Introduction

Researchers from spatial databases, GIS, data mining, and knowledge extraction communities have developed several techniques for mobility analysis. As consequence three research areas have been expended; The first one focuses on data modelling to provide definitions and extensions of trajectory related data types such as moving objects, points, lines, or regions. The second deals with data management to optimize the storage of mobility data with suitable indexing and querying techniques. And the last one that is the main topic of this research deals with the analysis of patterns that can be extracted from stored data like trajectories by using spatiotemporal data mining algorithms. Several data mining methods have been proposed for extracting patterns from trajectories. However, the majority of them use

---

<sup>1</sup> Sana Chakri, Cady Ayyad University; [chakri.sana@gmail.com](mailto:chakri.sana@gmail.com)  
Said Raghay, Cady Ayyad University; [s.raghay@uca.ma](mailto:s.raghay@uca.ma)  
Salah El Hadaj, Cady Ayyad University; [elhadajs@yahoo.fr](mailto:elhadajs@yahoo.fr)

trajectories without looking for any additional information, and yet by considering only the raw trajectory data, discovering why an object followed a different route become very complex since no additional information (called semantic) is given about the moving object. This additional information can hide behind a lot of meanings; in fact it can lead to a better understanding of the patterns extracted. This is can be achieved by combining the raw mobility tracks (e.g., the GPS records) with related contextual data in order to use semantic trajectories instead of focusing only on the geometric side of trajectories [1, 2]. Semantics refers essentially to additional contextual and geographical information available about the moving object, apart its position. Semantics contain both the geometric properties of the moving object as well as the geographic properties and any other additional information like the moving object's activity, mode of transportation, speed or any data that can help give more meaning to the behaviour extracted. The purpose of this research is to find spatial, spatiotemporal and temporal outliers among semantic trajectories.

In this paper we try to go further in semantic trajectory outlier detection, in order to deduce the possible reason why an object moves differently than its group. More specifically, we try to enrich trajectories with semantic data, and then extract outliers based on both geometrical and semantic information to give more meaning to the behavior extracted. The rest of the paper is organized as follows: section 2 the methodology pursued to detect semantic outliers. Section 3 presents the methodology to add meaning to patterns extracted, section 4 gives experimental results on real trajectory data. Finally, section 5 concludes the paper with discussion and comparisons.

## **2 Enriching trajectories with semantics**

The purpose is to find spatiotemporal and temporal outliers between regions of interest [3], Analysing them with semantic data to understand the meaning of the outliers detected. Spatiotemporal outliers refers to sub-trajectories that have spatial and temporal difference compared to common trajectories, while temporal outliers refers to moving objects that behave spatially like the majority of the other moving objects, but temporally they are different; for instance moving objects that took the same route but they accelerate or they mark an important number of stops which make them seen as suspicious moving objects. The analysis presented in this paper are made on sub-trajectories that rely regions of interest which are shapes that have different size and format, depending on the application, they can be regions ROI, lines LOI, or even points POI, they can be districts, dense areas, hotspots, important places, etc. generally a region of interest can be a pre-defined important place or computed by an algorithm that finds dense areas. In our case we consider a region as a point, line or polygon, which is a well-known concept in GIS community. The use of regions allows filtering from the whole dataset only the sub-trajectories that move between the same regions, outliers will be searched among these sets what significantly reduces the search space for outliers. Among the trajectories that cross

all regions, we are only interested in the part of trajectories (sub-trajectories) that move between specific regions, we call these sub-trajectories Nominees. After defining the set of nominees, we start looking for temporal outliers, and spatial outliers in which we extract from them spatiotemporal outliers. A nominee will be a spatial outlier when it follows a different path in relation to the majority of the sub-trajectories from its group, and it can be a temporal outlier if it follows the same path, but shows different behaviours compared to the other moving objects. In general, we have two types of path: Populated path that have many trajectories in its proximity. And depopulated Path, it has less trajectories around. The spatial and the spatiotemporal outliers will be extracted from depopulated paths, while the temporal outliers will be extracted from the populated paths.

### 3 Giving Meaning to patterns extracted

After extracting outliers from semantic trajectories, the main goal of the next step is to add meaning to the outliers extracted. The next step is about splitting the outliers extracted to several types according to their semantic interpretation;

#### A. Spatiotemporal outliers

##### 1) Stop outliers

It occurs when the moving object made a stop for some time during the deviation. We consider as a stop a sub-trajectory that its speed is close to zero for a minimal amount of time (MT).

##### 2) Emergency outliers

It occurred when the moving object took an alternative route and shows an important acceleration of its speed, the reasons can be almost about an emergency case like an ambulance transporting patient, or someone trying to escape from police, etc. We consider that there is an emergency outlier if the speed of the fast outlier is higher than the double of the average speed of the synchronized outliers detected in the same derived route.

##### 3) Regular outliers

It occurs when the moving object deviates from the populated route without an important change of speed, or with a degradation of speed. This may reveal that the populated route is temporarily busy or is under reconstructions, or there is an accident, or even there is an event that block the path, so the moving object is forced to deviate from the populated route, Which can cause a big traffic on the alternative ways, and as consequence, the speed of the moving object may decrease. Our algorithm assembles all these reasons in three types of outliers: the blocked route outlier, the avoided route outlier, and the traffic jam outlier:

- a) Blocked route outliers: Expresses any deviation because something happens close to the populated route which causes some blockage, for instance, an accident, and route reconstructions. To discover the reason, we start by analysing only the part of the closest populated route deviated by the outlier, then we look if there is an activity around the main

segments, if yes, we verify the time of this activity to be sure that the outlier was generated in the moment of the action. And finally we verify that at the time of the activity, there are no synchronized segments in the populated route, to prove that the path was blocked by the event.

- b) **Avoided route outliers:** This type of outliers is similar of the first type, the only difference is that there is an activity in the populated route, but this activity doesn't cause any blockage, an example could be a police checkpoint.
- c) **Traffic jam outliers:** Expresses deviations due to a heavy charge at the rush hour, it occurs if we found an outlier, but no activity is blocking the populated route, so we start looking if there is a traffic jam. For that we look for the slow traffic in the populated route at the time of the outlier.

### **B. Temporal outliers**

Temporal outliers are common trajectories that follow the populated route, but with an important difference of the speed compared to the other common trajectories. We extract two essential types; temporal emergency outliers, and temporal stop outliers.

#### **1) Temporal emergency outliers**

It occurred when the moving object stay in the populated route but shows an important acceleration of its speed, the reasons can be almost about an emergency case. We consider that there is a temporal emergency outlier if the speed of the fast common trajectory is higher than the double of the average speed of the synchronized common trajectories detected in the same populated route.

#### **2) Temporal stop outliers**

the temporal stop outliers are common trajectories that Travers the populated route with a very slow speed compared to the synchronized common trajectories in the same route, it occurs when the moving object made a stop for some time in the populated route. We consider as a temporal stop outlier a sub-trajectory that its speed is close to zero for a minimal amount of time (MT).

## **4 Experimental results and comparison**

For the experimental results we try to analyse real data sets to prove the efficiency of our method, these datasets are taken from [4, 5, 6, and 7]. It contains trajectories of Trucks dataset which consists of 276 trajectories of 50 trucks delivering concrete to several construction places around Athens metropolitan area in Greece for 33 distinct days. Notice that we analysed only trajectories from Monday to Friday The structure of each record is as follows: {obj-id, traj-id, date(dd/mm/yyyy), time(hh:mm:ss), lat, lon, x, y} where (lat, lon) is in WGS84 reference system and (x, y) is in GGRS87 reference system. These datasets are interesting for analysing outliers because this type of drivers, in general, knows different routes to reach the

Contribution of extracting meaningful patterns from semantic trajectories same place. Therefore, we can find the alternative routes (outliers) in relation to the standard path.



Figure 1: Trucks trajectories



Figure 2: Common Trajectories



Figure 3: Outliers extracted

Table 1: Trucks outliers extracted

Nominees	Expected outliers	Spatiotemporal outliers	Common trajectories	Temporal outliers
35750	9402	1157	26348	421

Table 2: Semantic spatiotemporal outliers trucks trajectories

Spatiotemporal outliers					
Stop	Emergency	Regular			
		Blocked route	Avoided route	Traffic Jam	Others
512	14	14	345	225	47

Table 3: Semantic temporal outliers from trucks trajectories

Temporal outliers	
p	Emergency
387	43

In this experiment we consider as interesting regions the districts around Athens metropolitan area. The application domain data are all about information about drivers, the number of students for the school buses, the type and the number of products for the trucks, the noun of the districts and the activities of the drivers and regions in this period. The closest algorithms to our approach are TRAOD algorithm [8] and Tra-SOD algorithm [9]. Until now we don't have experimental comparisons between the algorithms because the three algorithms provide different outputs, but

we tried to classify some characteristics of each of them in the table below to clarify the important difference between their characteristics. The algorithm TRAOD does not consider regions, the main route and it does not perform any further analysis over outliers. The algorithm TRA-SOD considers regions and searches for the main route, and gives some further analysis for the outliers extracted. While the algorithm SOA present two levels of classification of the outliers extracted; the first one is the speed to classify the emergency outliers, then use semantic to reveal the reasons for the regular outliers by extracting the reason of such type of outliers. The table below resume the characteristics of each one the three algorithms.

**Table 4:** comparison between algorithms

	<b>TRAOD</b>	<b>Tra-SOD</b>	<b>SOA</b>	
<b>Region of interests</b>	No	Yes	Yes	
<b>Populated/ Depopulated route</b>	No	Yes	Yes	
<b>Level of classification</b>				
<b>Spatial / temporal</b>	No	No	Yes	
<b>Speed</b>	No	No	Yes	
<b>Semantics</b>	No	Yes	Yes	
		Stops	<b>Spatiotemporal</b>	<b>Temporal</b>
<b>Types of outliers extracted</b>	Geometrical outliers	Avoidance Traffic Jam	Stops Emergency Regular	Stops Emergency

## 5 Conclusion

In this paper we have shown that trajectory knowledge discovery depends directly on the application domain, there is a need to integrate geographic information into trajectories in order to create semantic trajectories before extracting meaningful patterns. Several algorithms have been proposed for trajectory data mining, but only a few consider semantics, and very few of them deal with semantics on trajectory outlier detection. In this paper we gave importance to outliers extracted from semantic trajectories, the algorithm shown in this experiment finds the main route that the majority of trajectories followed, then detect all other deviations that trajectories can followed to reach the same place, after that the algorithm divided the results to spatial or spatiotemporal outliers according to their natures, the next step will be the interpretation of the outliers detected, since the semantic data allow more understanding to the behaviors detected.



## References

1. Sana Chakri, Said Raghay and Salah El Hadaj. Modeling, Mining, and Analyzing Semantic Trajectories: The Process to Extract Meaningful Behaviors of Moving Objects. *International Journal of Computer Applications* 124(8):15-21, August 2015. Published by Foundation of Computer Science (FCS), NY, USA.
  2. Sana chakri, Said Raghay, Salah El Hadaj, enriching trajectories with semantic data for a deeper analysis of patterns extracted. *Proceedings of the 16th International Conference on Hybrid Intelligent Systems (HIS 2016)*. The *Advances in Intelligent Systems and Computing* series. Volume 552.
  3. AR Aquino, Alvares L. O., Renso C. and Bogorny V. Towards Semantic Trajectory Outlier Detection. *GeoInfo*. (2013).
  4. C. Panagiotakis, N. Pelekis, I. Kopanakis, E. Ramasso, Y. Theodoridis, "Segmentation and Sampling of Moving Object Trajectories based on Representativeness", *IEEE Transactions on Knowledge and Data Engineering*, 07 Feb. 2011. IEEE computer Society Digital Library. IEEE Computer Society.
  5. N. Pelekis, I. Kopanakis, E. Kotsifakos, E. Frentzos and Y. Theodoridis. "Clustering Uncertain Trajectories", *Knowledge and Information Systems (KAIS)*, DOI 10.1007/s10115-010-0316-x, 2010.
  6. E. Frentzos, K. Gratsias, N. Pelekis and Y. Theodoridis. "Algorithms for Nearest Neighbor Search on Moving Object Trajectories". *Geoinformatica*, 11:159–193, 2007.
  7. N. Pelekis, I. Kopanakis, E. Kotsifakos, E. Frentzos and Y. Theodoridis. "Clustering Trajectories of Moving Objects in an Uncertain World", In the *Proceedings of the IEEE International Conference on Data Mining (ICDM'09)*, Miami, U.S.A., 2009. - O. Abul, F. Bonchi, M. Nanni, Never Walk Alone: uncertainty for anonymity in moving objects databases, in: *Proceedings of the 24nd IEEE International Conference on Data Engineering (ICDE'08)*, 2008.
  8. Lee, J.-G., Han, J., and Li, X. (2008). Trajectory outlier detection: A partition-and-detect framework. In *ICDE*, pages 140–149. IEEE.
- Fontes, V. C., de Alencar, L. A., Renso, C., and Bogorny, V. (2013). Discovering trajectory outliers between regions of interest. In *GeoInfo*.



# Towards The Register-Based Statistical System: A New Valuable Source for Population Studies

Chieppa A., Ferrara R., Gallo G., Tomeo V.<sup>1</sup>

## Abstract

The strong effort in micro-level integration among different statistical sources together with the availability of an increasing number of administrative archives is determining a big change in the processes that the National Institutes of Statistics adopt to produce population statistics. The Italian National Statistical Institute, Istat, is planning a new design for the next Census round, based on a convenient integration of administrative data and surveys.

A thematic database has been created to study how administrative sources could improve the quality and information of population registers: sources integrated are official municipal population registers and Istat statistical population together with administrative archives from labour market, education, data on income and taxation. The aim of this work is to point out how this integration of data in proper registers could allow discovery of new relevant information about population: clusters of individuals determined by patterns emerging when analyzing ‘signals of presence’ in different sources and their geographical distribution could be of great interest for population studies. Moreover, emerging patterns could be very useful to design population surveys and producing population counts: both for definition of statistical models to assess accuracy of population enumeration and for implement sampling frames, including Populations Census ones.

**Key words:** Permanent census, administrative data, Integrated microdata system, groups of population

---

<sup>1</sup> Chieppa A., Istat, [chieppa@istat.it](mailto:chieppa@istat.it);  
Ferrara R., Istat, [rferrara@istat.it](mailto:rferrara@istat.it);  
Gallo G., Istat, [gegallo@istat.it](mailto:gegallo@istat.it) ;  
Tomeo V., Istat, [tomeo@istat.it](mailto:tomeo@istat.it)

For years, main sources for population studies and statistics have been demographic surveys and Population Census on one hand, municipal population registers on the other. In the past, the integration among these sources was set up at aggregated level: results of Census were used to increase accuracy of municipal population register; current demographic survey results were used to update Census population counts to provide intercensal estimation or projection; and so on. Nowadays, the strong effort in micro-level integration among different statistical sources together with the availability of an increasing number of administrative data archives is determining a big change in the processes that the National Institutes of Statistics adopt to produce population statistics.

The population Census still remains the largest and most important statistical data collection to provide population figures at the smallest geographical units: while most statistical advanced countries still use the traditional scheme, with complete enumeration of population and housing units (i.e. Usa and Canada), an increasing number of countries base their Census production on statistical registers. Census register-based can use exclusively registers data, as in the case of Scandinavian countries (Netherlands, Sweden, Denmark, Finland and Norway), or can use a combination of both registers and sample surveys data within the frame of the so-called 'combined Census' (Spanish 2011 Population Census, for example).

The Italian National Statistical Institute, Istat, is planning a combined Census scheme for the next Census round, by conveniently carrying on the integration of administrative data and surveys and then exploiting this new informative richness.

During the last decade, Istat has been increasingly adopting administrative sources for statistical purposes. Municipal population registers were the first administrative microdata sources used in the field of populations statistics: from 2011, Istat has been yearly collecting individual and households archives from this source. These data were used to define the primary list of households respondents for 2011 Census. Currently, variables collected from population registers (i.e. citizenship, age, gender and place of residence) are used to fit the sampling social surveys (i.e., Labour Force, Living Conditions, EuSilc and Consumer Expenditure surveys).

Starting with Population Census microdata and adding vital events (births, deaths, internal and international migrations), Istat has been computing a statistical population register, the so-called "ANagrafe Virtuale Statistica" (ANVIS). This register ensures higher level of quality than municipal administrative population one and represents a solid component for a frame of register-based production for population statistics.

Since 2015 many experiments at Istat has been investigating the use of other administrative sources to increase quality and information derived from population registers. To manage the increasing number of administrative data sets and to maximize the benefit, Istat built an integrated system of available administrative sources, named SIM (Integrated System of Microdata). When a new administrative archive is loaded in this system, recognition processes identify any individual or economic unit present in data and assign it a permanent and unique identification number (ID): if the unit is already present in Istat databases, this ID is the same the unit was assigned in the past. ID assignment and the creation of proper links among archives coming from different sources is the 'core' of microdata integration. Then,

starting from this base, it is possible to construct specific data structures for statistical processes and to create thematic (di Bella and Ambroselli, 2014).

The SIM, among all the archives loaded, comprises data coming from municipal population registers and ANVIS, permits to stay, data referring to employees and self-employed workers, compulsory education students, university students, retired people, non-pension benefits records, and individual data on income and taxation: these integrated data have been used to create a thematic database to study how administrative sources could improve the quality and information of population registers (Chieppa et al, 2016).

The aim of this work is to point out how this integration of data in proper registers could allow knowledge discovery of new relevant information about population, not directly deriving from exploitation of singular archives: clusters of individuals determined by patterns emerging when analyzing 'signals of presence' in different sources and their geographical distribution could be of great interest for population studies. Moreover, emerging patterns could be very useful to design population surveys and producing population counts: both for definition of statistical models to assess accuracy of population enumeration and for implement sampling frames for population surveys, including Populations Census ones.

The process of knowledge discovery to extract useful information need expertise in domains of specific administrative sources, together with expertise in population studies and technical skills: so a multidisciplinary team and tasks are needed. This collaboration is essential, among other things, when selecting specific administrative sources from all those available and identifying their hierarchy: labor and education registers rank higher than the other sources given that they provide more detailed information on the territorial level and activity duration.

With the goal of discover useful information to increase population registers quality, signals of presence of individuals on the national territory are extracted from the sources integrated in the database. These signals were deeply investigated to identify more discriminant attributes and possible latent dimensions. Once relevant features of signals have been selected or derived (in case of latent ones), they are used to study specific clusters of individuals useful for the permanent Census purposes.

The duration of signals for each individual has result of a big importance, together with their geographical distribution. Signals of administrative data can represent a temporary or occasional presence, so it is necessary to carry out a characterization process, by the means of analysis of data and subsequently constructing calculated variables. Patterns of signals duration emerged: continuous and steady signals, seasonal pattern, new entrance, and so on. This goal was achieved thanks to the longitudinal perspective that integration of microdata allows.

So, signals can be used to derived new relevant variables for related individuals and their type of living conditions. More specifically with regard to population counts, this new information could identify cases of permanent presence that correspond to the usual residence definition and concept of the international regulations, that not always correspond to what is record into official administrative population registers. Demographic variables, especially gender, age and country of citizenship as well as the location of the signal on the territory have proved to be very significant variables in defining specific sub-population profiles

A more detailed analysis of the characteristics of sub-populations at risk of over- and under-coverage in populations registers has highlighted some important topics and specific clusters of individuals. The division into different clusters reveals the sub-populations that require further thematic analysis: for example, the typical foreign communities that elude the registers, but perform specific labor activities, or people who frequent certain territories. These same groups can form the basis for defining a census procedure formulated on the use of mixed techniques that combine specific surveys and appropriate statistical models.

The potential over-coverage of population register counts 3 million individuals, three out of four have Italian citizenship, and the foreigners are on average more than six years younger than the Italians. This sub-population mainly consists of people of working age (15-64 years) and it is linked to the geographical areas where unemployment is higher such as the municipalities in the South and in some central areas of Italy.

When considering under-coverage, the analysis shows the presence of several distinct clusters. The continuous signals showing stable presence on the territory are related to just over 400 thousand people who are mainly foreign nationals. Geographical location and specific citizenship are essential for identifying the cross-border workers, whose absence from the population register is admissible.

Analysis revealed that also some individuals with weak (not continuous) signals may be associated with stable presence on the territory, and for this reason an improved characterization of signals is needed.

Clusters of individuals emerged, relevant to assure the quality of populations counts, could be effectively used to measure improve quality of Census survey: contributing in the sampling design and also to properly assess quality of enumeration counts.

## References

1. di Bella G. and Ambroselli, S. (2014). Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat. Paper presented to the European Conference on Quality in Official Statistics Q2014. Vienna: 2-5, June:1-14.
2. Chieppa, A., Gallo, G., Tomeo, V., Borrelli, F., Di Domenico, S. (2016). Knowledge Discovery Process to Derive Usually-Resident Population from Administrative Registers. Mimeo.
3. Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*. 40(2): 137-161.
4. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds). AAAI Press, Menlo Park: 134.
5. ISTAT (2014). La misurazione della qualità del 15° Censimento generale della popolazione e delle abitazioni: i risultati dell'indagine di copertura (PES). Seminario del 27 giugno, Roma: <http://www.istat.it/it/archivio/126014>.

# Consulting, knowledge transfer and impact case studies of statistics in practice

## *Consulenza, trasferimento della conoscenza e casi reali di impatto pratico della statistica*

Shirley Coleman<sup>1</sup>

**Abstract** There is a growing interest by companies in learning more about how data science can benefit them by analysis of their company data. This has led to a flurry of requests for consultancy work and a corresponding bottleneck of people available to do the work. Many companies value the stamp of a university and highly qualified business development staff are being employed by universities to bring in more outside interest partly to achieve the impact case study requirements. This mismatch can be problematic but does not detract from the satisfaction of having so many enquiries. This talk will describe some of the interesting work currently undergoing at Newcastle University with small and medium enterprises.

**Abstract** Negli ultimi anni si osserva da parte delle aziende un crescente interesse per come il Data Science può dare dei vantaggi nell'analisi dei loro dati aziendali. Tutto ciò ha portato ad una imponente richiesta di consulenza, ma con una conseguente carenza di persone competenti disponibili ad effettuare il lavoro. Molte aziende attribuiscono valore al marchio di una università, mentre le stesse università stanno impiegando personale altamente qualificato nello sviluppo commerciale per suscitare interesse all'esterno e in parte qualificare l'impatto dei loro casi di studio. Tale discrepanza potrebbe essere problematica, ma questo non riduce la soddisfazione di avere così tante richieste. Questa presentazione sarà dedicata ad alcuni interessanti progetti con imprese medio-piccole che sono attualmente in corso presso l'Università di Newcastle (UK).

**Key words:** data science, data analytics, SMEs, business development

## 1 Introduction

The practical application of statistics is fundamental to the scientific approach to managing a business and running a successful industry. Although many companies have made use of statistics, many sectors are slow in the uptake and barriers in communication hinder a more complete realisation of the benefits. The extent of the successes and losses of opportunity varies across Europe and it has been a great

---

<sup>1</sup> School of Mathematics and Statistics, Newcastle University  
email: shirley.coleman@ncl.ac.uk

advantage to have a pan European network of practitioners who can share experiences and learn from each other.

Statistics in practice is the main focus of the European Network of Business and Industrial Statistics (ENBIS) which was set up in 2000 as a network of statistical practitioners. The Industrial Statistics Research Unit (ISRU) in the School of Mathematics and Statistics at Newcastle University, UK was one of the founder members of ENBIS and has supported its activities ever since its launch and throughout its 17 years of annual conferences, subject-focused Spring meetings and other activities.

ISRU is a self-financing consultancy unit dedicated to spreading the message of statistics in practice. The unit was set up by G. Barrie Wetherill in 1980 as a response to his growing involvement with process industries, in particular the massive ICI which employed over 20,000 people in nearby Teesside. ISRU evolved from providing consultancy in statistical process control (Wetherill and Brown 1992) and design of experiments to tackling the underlying issues of understanding the needs of industry and implementing the changes needed to realise the benefits of statistical interventions. The work involved methods from Total Quality Management and applying the Six Sigma approach.

ISRU functions in a deprived economic area where heavy industry such as shipping and mining have given way to light engineering and many start up SMEs. Typically these SMEs were slow to take advantage of innovation in terms of new management methods, partly because of their lack of finance. The importance of European funding was soon realised as a way to give SMEs an equal chance of improving the quality of their products and processes as that enjoyed by large organisations. Several million £ of support was obtained for helping companies make use of statistics and take a more scientific approach to problem solving. Amongst the EU projects that ISRU proposed was substantial funding to set up the thematic network called pro-ENBIS. This initiative led to the growth of ENBIS into the established and well-respected group of statistical practitioners that it is today.

In response to the increasing availability of data and awareness by companies of its importance, ISRU developed data mining offerings for business and industry and carried out several projects with SMEs part funded by the UK government. These included developing and implementing segmentation methods in retail, explorations of the Kansei Engineering approach (<http://www.kansei.eu>), statistical training in the healthcare sector and extensive data analytics in the gas industry.

The interest in statistics and data mining is now escalating with many SMEs making serious enquiries as to what support is available for them. ISRU has the capability to help companies analyse and benefit from their data. There is now an issue of how to service these requests for consultancy services. University staff are fully occupied with research and teaching and do not have time and motivation to carry out consultancy tasks. However, all UK universities are now required to demonstrate the impact of their academic activities although enthusiasm for taking on consultancy is lacking, the increasing importance of impact should help make it more appealing.



ISRU is in a pivotal position to help academics demonstrate impact. One major impact theme is around big data analytics for SMEs. The rest of this paper will focus on this theme. Section 2 describes SMEs, gives some background to big data and considers SME relationships with big data analytics, section 3 describes a case study from this impact work, section 4 discusses UK university impact requirements, section 5 considers the implications of the growth in data science for statisticians and for the practice of statistics, and section 6 is the conclusion.

## 2 SMEs and their relationship with big data analytics

Small to medium enterprises (SMEs) have fewer than 250 employees; a turnover of less than £40m or a balance sheet less than £34m. They are considered to be the driving force of the economy employing over half of EU employees. Eurostat shows that the % of total value added by SMEs in 2014 was around 60% on average in different EU countries (<http://ec.europa.eu/eurostat/>). However, it is noted that SMEs are behind in their commercial exploitation of big data:

“SMEs are lagging behind in the usage of business and big data analytics. In 2012, the adoption rate of big data analytics among UK SMEs was only 0.2 per cent, compared to 25 per cent for businesses with over 1,000 employees (e-skills UK 2013). Market studies expect an annual growth rate of the global SME big data market by 42 percent over the period of 2013 until 2018 (TechNavio 2014). “

Big data analytics here refers to the analysis of the masses of data regularly collected by businesses as part of their everyday activities. It includes data from sales, inventory, logistics, quality improvement, new product development and promotion. In particular where companies are concerned, the opportunities lie in the monetisation of company owned data in conjunction with secondary and open data (Coleman, 2016).

Compared with larger organisations, SMEs have

- Less money, available staff and access to consultants
- Greater concern with privacy, security and confidentiality
- And are more risk averse

The ENBIS community is keen to address this shortfall and are in process of composing a European co-operation in science and technology (COST) funding bid, <http://www.cost.eu/>. As background to the proposal, ENBIS members published a paper reviewing the state of the art and the needs (Coleman et al 2016). One of the findings reported in this paper is that SMEs have 14 main challenges which are concerned with skills, operational issues and practicalities. These are fully described in the paper and in summary are:

- *Skills*: Lack of understanding of big data; Dominance of domain specialists; Cultural barriers and intrinsic conservatism; Shortage of in-house data science expertise; Bottlenecks in the labour market.

- *Operational issues*: Lack of business cases; Shortage of affordable consulting; Confusing software market; Lack of intuitive software; Lack of management and organisational models.
- *Practicalities*: Concerns on data security; Concerns about data protection and privacy; Over-focus on venture concept; Financial barriers.

Of these challenges, the ENBIS COST proposal will in particular address the lack of business cases; shortage of affordable consulting and financial barriers. There are currently few examples of data analytics case studies in SMEs. Ahlemeyer-Stubbe et al (2014) give recipes for dealing with company data but the focus is mainly on marketing issues. This lack of business cases will be ameliorated by making exemplars and case studies available. Consulting companies are disinclined to offer their services to SMEs as they prefer larger more lucrative contracts; the COST project will tackle this problem initially by clarifying what services are available, helping SMEs find assistance and facilitating the communication between SMEs and consultants, and rationalising the kinds of intervention that will be beneficial to SMEs. As more and more SMEs take up the big data opportunity consultants will find them a safer, more appealing prospect and it is to be expected that their services will become more affordable as the consultants see the advantage of working with SMEs. The financial barriers arise because of the intrinsically low cash flow of most SMEs but also because there are fewer avenues for debt finance. There is often an asymmetry of information between the SME and lender because it is difficult for the SME to express the benefits they expect from investing resources in big data analytics. The COST project will help clarify the costs and benefits and make financial applications easier to compose and more likely to be successful.

### 3 Case study

An established approach to helping SMEs and large organisations to innovate using big data analytics is funded knowledge transfer from universities to businesses. In the UK, Innovate UK funds 67% of the costs for an SME undertaking a knowledge transfer partnership (KTP) or 50% of the costs for a large organisation. These partnerships have to address a substantial, specific research area of need in the company and are usually of around 2 years duration. A graduate research associate is appointed to work full time in the business whilst being employed by the university and being mentored and supervised by an academic for at least 2 days per month. KTPs are a 4 way partnership with the company standing to benefit from the work on their problem without the risks of employing a new worker, the research associate gaining valuable work-based experience, the academic gaining new material for teaching and publishing, and Innovate UK showing added value for the funds they have spent in terms of increased turnover and profit for a UK company.

Newcastle University currently has 17 such KTPs including projects in agriculture, energy and historic records. KTPs were initiated in 1976 and over the

years ISRU staff have supervised projects in retail, finance, shipping, energy and assistive technology. One such KTP specializing in assistive technology was completed in 2016 and provides a good example of the work. Assistive technology refers to equipment and services that assist older people to deal with losing their ability to undertake activities of daily living. Activities of Daily Living (ADLs) include dressing, toileting, bathing, feeding and moving around. It is not always clear which assistive technologies will be the most appropriate for each individual, and the current socio-political-medical environment means that there are frequently long waiting periods to access expert opinion. It is quite common for products to have been developed specifically for an individual who has a particular need or requirement, then the product is mass-produced and marketed. However, bespoke products are not necessarily suitable for all who present with similar problems.

The SME in this project had developed an expert system to enable people to interrogate information resources and find out which assistive technologies are available and are suitable for them in their physical environment. The company had 14 years' worth of assessment data resulting in several million cases with data that can be indexed by person, assessment question, ADL problem and ADL solution. The company was aware that their data is a rich source of insight and that more use could be made of it and so they were eager to explore what could be done.

The KTP was designed to explore the use of data analytics to reveal insight about the ageing sector and to be of benefit to the many stakeholders of the company. As the company was built around the needs of the public sector, their income was dependent upon Government sources, an additional important aim of the data analytics was to provide a new revenue stream for the company to make it more stable and robust to economic changes.

The project produced guidelines to help SMEs get started with analysing their data:

1. Review strategic objectives.
2. Review IT options for data manipulation, access, analysis and presentation.
3. Identify relevant dimensions of the data.
4. Carry out a stakeholder analysis.
5. Quantify the importance of each data dimension to each stakeholder
6. Determine suitable revenue strategies.
7. Use data analytical methods to create insight.
8. Pilot and revisit step 1.

Exemplars were prepared and the completed KTP was considered to be a great success by all partners. In particular, the company has completely changed its relationship with data and sees data analysis as a key company offering. In addition various stakeholders have committed to paying for insight and create the desired new revenue stream.

## 4 Impact

A considerable part of the funding of UK universities is based on impact. Universities need to show that their research has a broad and deep impact on society and the world at large. In the 2014 research excellence framework 25% of non-student related income depended upon impact. Each department had to show:

- Quantifiable reach and breadth of research impact in the last 5 years on a cohesive theme
- With several >2\* underpinning publications over the last 20 years
- With 1 impact case study per 7-10 academic staff

ISRU and academic staff who undertake KTPs and consulting are in a good position to provide impact case studies. ISRU contributed one of 3 case studies submitted by the School of Mathematics and Statistics (<http://www.ncl.ac.uk/research/ref/unit/uoal0>) and staff are currently involved with preparing cases and helping other academics for the next research excellence framework expected to be in 2020.

There is a growing mismatch between university administrative staff who want to maximise the university income and impact opportunities and academic staff who feel they are judged by their academic publications. The growing importance of impact may help to bridge this gap. Somehow the profile of staff involved with impact needs to be raised and the rewards made commensurate.

Consultancy does not have to be carried out by a university group, however, many businesses like to have the stamp of approval of a university as an independent voice on their methods and conclusions.

## 5 Consultancy

A European wide initiative to work with big data and SMEs is just one approach to introducing and eventually embedding statistics in companies. Universities and colleges have realised the need for more accessible training and there are an increasing number of Master's and other postgraduate courses available for study. These are labelled variously as data analytics, business intelligence, data science and others. Most such courses offer a combination of statistical training, business awareness and IT skills (Coleman and Kenett 2016).

Data science is a fast growing concept that is becoming accepted in many walks of life. It has attracted interest from the claims made on bill-boards and the fact that even Governments are giving it extensive attention, for example the UK Parliamentary Office of Science and Technology, <http://www.parliament.uk/mps-lords-and-offices/offices/bicameral/post/work-programme/big-data>. Citizen scientists are encouraged to collect data and contribute to major research programs. For

example, in the Zooniverse project, <https://www.zooniverse.org/>, citizens are invited to take part in research.

Data science has caused a stir amongst professions such as statistics and operations research because many data scientists do not feel the need to be part of these professional societies. The Royal Statistical Society debated whether to include data science in all of its special interest sections or to start a new section devoted to data science. After considerable discussion it was decided to start a new data science section. The president strongly expressed the opinion that data science needs a sound basis in mathematics and statistics and that the professions are an intrinsic part of the new field of data science.

In ISRU's consultancy work we are teaming up with computer science specialists and with the client domain specialists so that we are addressing the three aspects of data science in a cohesive way. Computer scientist input includes data access and manipulation, machine learning methods and bespoke visualisation. Together we can produce tailored client solutions which can be implemented and developed within the company in a highly satisfactory manner.

The data science consultancy needs of large organisations are often met by services provided as one small part of the offering from big consultancy firms who are more familiar with other business activities. There is a large body of statisticians working in government, health, finance and drug development but these statisticians and statistical groups have limited opportunities to help SMEs. The issue of where consultancy units are best placed was discussed in pro-ENBIS and a database of consultancy provision across Europe was one of the project deliverables. A comparison of academic based and commercial consultancy units was included in the project book (Coleman et al 2008) produced as part of the dissemination. This issue is still pertinent. Academic staff are judged by the quality and quantity of their research output. Research income is not always valued. We have found that it is difficult to attract academics to be involved in consultancy, but as stated above impact requirements may be the key to solving the issue.

## 6 Conclusions

Consultancy is a vital service getting added value from research and providing real-life examples for students. Mechanisms are needed to facilitate the exchange and knowledge transfer partnerships are a good method. SMEs are increasing motivated to explore data analytic options for their company data. The current focus on impact in UK universities is helpful in giving consultancy a higher profile and encouraging more staff to service the need. Academic based consultancies have an important role to play and need to be encouraged wherever possible. European funding is needed to help SMEs to take part in the data analytics revolution. ENBIS is an important peer group to co-ordinate such activities.

## References

1. Ahlemeyer-Stubbe, A., Coleman, S.Y.: *A Practical Guide to Data Mining in Business and Industry*. Wiley (2014)
2. Coleman, S.: Data mining Opportunities for Small to Medium Enterprises from Official Statistics, *Journal of Official Statistics*, 32(4), 849-866 (2016)
3. Coleman, S.Y., Gob, R., Manco, G., Pievatolo, A., Tort-Martorell, X., Reis, M.: How Can SMEs Benefit from Big Data? Challenges and a Path Forward, *Journal of Quality and Reliability Engineering International* (2016)  
<http://onlinelibrary.wiley.com/doi/10.1002/qre.2008/full>
4. Coleman, S.Y., Kenett, R.: The Information Quality Framework for Evaluating Data Science Programs in Data science world scientific encyclopedia (2016)  
<http://ssrn.com/abstract=2911557>
5. Coleman, S.Y., Greenfield, T., Stewardson, D.J., Montgomery, D.: (eds.) *Statistical practice in business and industry*, Wiley (2008)
7. E-skills UK 2013. *Big Data Analytics, Adoption and Employment Trends, 2012-2017*.  
[http://www.e-skills.com/Documents/Research/General/BigDataAnalytics\\_Report\\_Nov2013.pdf](http://www.e-skills.com/Documents/Research/General/BigDataAnalytics_Report_Nov2013.pdf) [accessed 21 December 2015]
8. TechNavio 2014. *Global SME Big Data Market 2014-2018*. TechNavio - Infiniti Research Ltd.
9. Wetherill, G.B., Brown, D.W.: *Statistical Process Control*, Chapman and Hall (1992)

# The evaluation of the inequality between population subgroups

## *La valutazione della disuguaglianza tra i sottogruppi di una popolazione*

Michele Costa

**Abstract** This paper illustrates the advantages to evaluate inequality between population subgroups with respect to a maximum compatible with the observed data, thus going beyond the traditional approach to the analysis of inequality between, where the maximum corresponds to total inequality. The new proposal improves both the measurement and the interpretation of the contribution of inequality between to total inequality.

**Abstract** *In questo lavoro vengono illustrati i vantaggi di valutare la disuguaglianza tra i gruppi di una popolazione rispetto a un massimo compatibile con i dati osservati, superando in questo modo l'approccio tradizionale alla misura della disuguaglianza tra, nel quale il massimo viene rappresentato dalla disuguaglianza complessiva. La nuova proposta consente un miglioramento sia nella misura, sia nell'interpretazione del contributo della disuguaglianza tra alla disuguaglianza totale.*

**Key words:** Inequality between, Inequality decomposition, Gini index, Inequality factors

## 1 Introduction

Inequality between population subgroups represents perhaps the most important component of total inequality. By means of inequality between, different sources of inequality are evaluated and compared, with the twofold goal to detect the main determinants of poverty condition and to implement socio-economic policies able to reduce poverty.

The measurement of inequality between can be achieved following different approaches, since inequality literature presents a wide collection of contributions on

---

Michele Costa

Department of Statistical Sciences, University of Bologna, e-mail: michele.costa@unibo.it

inequality decomposition. However, the size of inequality between is usually evaluated with respect to its theoretical maximum, which corresponds to total inequality, when the inequality within subgroups is equal to 0. The case of null inequality within is a quite unrealistic situation, which can be essentially considered as a theoretical reference, without a proper phenomenal correspondence. That is, we really do not expect to achieve a situation where each unit of each subgroup possesses the subgroup mean.

Furthermore, by comparing inequality between to total inequality, we can observe two unfortunate effects. First, the size of inequality between is frequently unreasonably small, thus suggesting a too low influence of the underlying inequality factor. Second, the measure of inequality between is strongly influenced by the number of subgroups used into the partition of the total population, thus preventing a direct comparison between different inequality factors when the number of subgroups is not the same.

In order to overcome these drawbacks, we propose a new framework for the evaluation of the inequality between, where the basis for comparison is not represented by total inequality, but by the maximum which can be obtained given the observed data. We build on [4] and develop new indicators for the evaluation of the inequality between. The new indexes allow to assess the importance of the different inequality factors into the observed data, thus improving our knowledge of inequality.

## 2 Methodology

In the following we will refer to the Gini index ([1],[5]) as our inequality measure and to the Dagum's decomposition to get the inequality between subgroups.

For the case of a population disaggregated into  $k$  subgroups of size  $n_j$ , with  $\sum_{j=0}^k n_j = n$ , the Gini index  $G$  can be expressed as follows

$$G = \frac{1}{n\bar{y}^2} \sum_{j=1}^k \sum_{h=1}^k \sum_{i=1}^{n_j} \sum_{r=1}^{n_h} |y_{ji} - y_{hr}| \quad (1)$$

where  $\bar{y}$  is the arithmetic mean of  $Y$  in the overall population,  $y_{ji}$  is the value of  $Y$  in the  $i$ -th unit of the  $j$ -th subgroup and, accordingly,  $y_{hr}$  is the value of  $Y$  in the  $r$ -th unit of the  $h$ -th subgroup. Among the many methods which allow to decompose the Gini index (see, e.g., [2],[6],[7]), we use the decomposition proposed by Dagum [3], where the differences  $|y_{ji} - y_{hr}|$  in (1) are assigned to  $G_w$ , the component of inequality within subgroups, when  $j = h$ , to  $G_b$ , the component of inequality between subgroups, when  $j \neq h$ ,  $\bar{y}_j \geq \bar{y}_h$ ,  $y_{ji} \geq y_{hr}$ , and to  $G_t$ , the component of overlapping, when  $j \neq h$ ,  $\bar{y}_j \geq \bar{y}_h$ ,  $y_{ji} < y_{hr}$ . Globally we have  $G = G_w + G_b + G_t$ .

In the framework of the Dagum's decomposition, as well as following any other approach, inequality between reaches its maximum when two conditions are verified. First, the  $k$  subgroups should not overlap, that is, the component  $G_t$  is equal to 0. Second, the variability within the subgroups should be equal to 0, that is



the component  $G_w$  is equal to 0 and each subgroup unit possesses the subgroup mean:  $y_{ji} = \bar{y}_j, j = 1, \dots, k; i = 1, \dots, n_j$ . On the basis of these two conditions, we have  $G_{bmax} = G$  and the evaluation of  $G_b$  is obtained by means of the ratio

$$I_{G_b} = G_b/G \quad (2)$$

We propose to relax the condition  $G_w = 0$  and to compare  $G_b$  not to ist theoretical maximum  $G$ , but to the maximum  $G_{bmax}$  which can be achieved given the observed data. That is, we propose to compare  $G_b$  not to the unrealistic case of equidistributed subgroups, but to a case more coherent and compatible with the data.

If we maintain the condition of no overlapping, we have that  $G_{bmax} = G - G_{wmin}$ , where  $G_{wmin}$  is the minimum inequality within, which can be obtained partitioning the observed data into  $k$  non overlapping subgroups. We have many ways to divide  $n$  units into  $k$  non overlapping subgroups: with the aim of preserving the structure of the original partition, we propose two possible solutions. First, we obtain the  $k$  subgroups by using the original  $p_i = n_i/n$  values, thus keeping the same population shares of the original partition. Second, we obtain the  $k$  subgroups by using the original  $s_i = (n_i \bar{y}_i)/(n \bar{y})$  values, thus keeping the original income shares.

The second step of our method refers to the calculus of  $G_{wmin}$ , the minimum inequality within compatible with the new  $k$  subgroups. We propose to permute the sequence of the  $p_i$  (or  $s_i$  for the second solution), to get a set of  $k$  subgroups for each permutation, to calculate the related  $G_w$  and to chose the minimum value among all disposable  $G_w$ . Let be  $G_{wmin(p)}$  the minimum inequality within, which can be obtained by permutating the values  $p_i$  and, correspondingly,  $G_{wmin(s)}$  the minimum inequality within, which can be obtained by permutating the values  $s_i$ .

In the last step we derive the new indexes for the evaluation of  $G_b$ , obtained as

$$I_{G_b(p)} = G_b/(G - G_{wmin(p)}) = G_b/G_{bmax(p)} \quad (3)$$

$$I_{G_b(s)} = G_b/(G - G_{wmin(s)}) = G_b/G_{bmax(s)} \quad (4)$$

The new indexes depend on the minimum inequality within compatible with the observed data and, therefore, are not strongly affected by  $k$  as for  $I_{G_b}$ .

### 3 Case study

In order to illustrate the advantages of our proposal, we present a case study related to the Italian households for the 2014. The data are from the Survey on Households Income and Wealth, a multidimensional survey on Italian households performed every two years by the Bank of Italy. The study analyzes the income inequality among the Italian households, divided into subgroups by means of two of the main determinants of inequality: the area of residence of the household and the educational level of the head of household. In order to evaluate the effect of the number of subgroups on inequality between, we consider the cases  $k = 2, 3, 5$ .

Table 1 illustrates the income mean, the population share and the income share for the two partitions. From Table 1 it is possible to observe some well known stylized facts of income inequality in Italy, clearly evident from the values  $\bar{y}_i$  and from the differences  $(p_i - s_i)$ .

**Table 1** Mean income, population share and income share for Italian households divided by area of residence and by educational level of the head of household, 2014

Area	mean	p	s	Education	mean	p	s
North West	33750	0.254	0.279	None	14676	0.03	0.02
North East	35150	0.221	0.229	Elementary	22329	0.20	0.16
Center	32636	0.202	0.226	Middle school	26753	0.37	0.31
South	23365	0.244	0.173	High school	35893	0.26	0.31
Islands	24095	0.081	0.093	University	46641	0.13	0.20

Our focus is on the effects of the differences between the subgroups on total inequality. Table 2 illustrates the Dagum's Gini index decomposition by area of residence. By increasing  $k$ , we can observe the usual pattern in inequality decomposition: the decrease of inequality within  $G_w$  and the consequent greater importance of inequality between  $G_b$  and of overlapping component  $G_t$ . The evaluation of  $G_b$  on the basis of  $I_{G_b}$  strictly depends on  $k$ : for  $k = 2$  we have that the area of residence contributes for the 31% to total inequality, while for  $k = 5$  its importance rises to the 48%. From Table 2 we can also observe how  $I_{G_{b(p)}}$  and  $I_{G_{b(s)}}$  are not a monotone function of  $k$ , since they depend on the minimum inequality within. The new indexes show quite similar results, with the contribution of the geographical dimension ranging from the 36% for  $k = 2$  to the 50-57% for  $k = 5$ .

**Table 2** Income inequality decomposition by area of residence<sup>a</sup>, Italian households 2014

	k	Gw	Gb	Gt	$I_{G_b}$	$I_{G_{b(p)}}$	$I_{G_{b(s)}}$
NC, SI	2	0.194	0.107	0.049	0.306	0.355	0.359
N, C, SI	3	0.125	0.139	0.086	0.397	0.562	0.574
NW, NE, C, S, I	5	0.073	0.168	0.109	0.479	0.508	0.566

<sup>a</sup> N north, NW north-west, NE north-east, C center, S south, I islands.

The results related to the decomposition by educational level of the head of household are reported on Table 3. The components  $G_w, G_b, G_t$  show a behavior similar to the previous case, however we can observe how  $G_b$  has a greater importance, while  $G_t$  is smaller: two signals of a stronger relevance of the educational dimension.  $I_{G_b}$  confirms this indication, showing higher levels with respect to Table 2. Also the new indexes are higher, but their increase with respect to the results of Table 2 is less accentuated. Moreover,  $I_{G_{b(p)}}$  and  $I_{G_{b(s)}}$  show only slight variations to changes of  $k$ , thus indicating an important degree of robustness into the evaluation

**Table 3** Income inequality decomposition by educational level<sup>a</sup> of the head of household, Italian households 2014

	k	Gw	Gb	Gt	$I_{G_b}$	$I_{G_{b(p)}}$	$I_{G_{b(s)}}$
NEM, HU	2	0.162	0.149	0.038	0.426	0.590	0.626
NEM, H, U	3	0.130	0.171	0.049	0.487	0.613	0.568
N, E, M, H, U	5	0.081	0.200	0.069	0.570	0.615	0.613

<sup>a</sup> N none, E elementary, M middle school, H high school, U university.

of  $G_b$ . The educational dimension is considered an inequality factor more important than the geographical dimension by all indexes, however, within the new proposals, the difference between the two factors is not so high as on the basis of  $I_{G_b}$ . In both cases the new indexes attribute to the inequality factors a stronger role, overcoming the usual underestimation and truly reflecting the effective importance of these determinants of total inequality.

## 4 Conclusions

We propose to modify the traditional evaluation of the inequality between population subgroups by introducing a maximum compatible with the observed data. Our purpose is to assess the determinants of inequality with respect to the observed data, and not by referring to the unrealistic case of equidistributed subgroups. Our proposal also allow to strongly reduce the effect of the number of subgroups on the evaluation of inequality between. Two new indexes are illustrated and we believe that their foundation on observed data represents an improvement for our knowledge of the inequality structure.

## References

1. Dagum, C.: Gini ratio. In: The New Palgrave Dictionary of Economics. Mac Millian Press, London (1987)
2. Dagum, C., Zenga M.: Income and Wealth Distribution, Inequality and Poverty. Springer, Berlin (1990)
3. Dagum, C.: A new decomposition of the Gini income inequality ratio. *Empirical Economics*, **22**, 515–531 (1997)
4. Elbers, C., Lanjouw, P., Mistiaen J.A., Ozler, B.: Reinterpreting Between-Group Inequality. *Journal of Economic Inequality*, **6**, 231–245 (2008)
5. Giorgi, G.M.: Gini's scientific work: an evergreen. *Metron*, **63**, 299–315 (2005)
6. Giorgi, G.M.: The Gini inequality index decomposition. An evolutionary study. In Deutsch, J., Silber, J.: The measurement of individual well-being and group inequalities. Routledge, London (2011)
7. Yitzhaki, S., Lerman, R.: Income stratification and income inequality. *Review of Income and Wealth*, **37**, 313–329 (1991)



# Bayesian Non–Negative $\ell_1$ –Regularised Regression

## *Regressione LASSO Bayesiana non negativa*

Costola Michele

**Abstract** This paper proposes a novel Bayesian approach to the problem of variable selection and shrinkage in high dimensional sparse non–negative linear regression models. The regularisation method is an extension of the LASSO which has been recently cast in a Bayesian framework by Park and Casella (2008). Moreover, to deal with the additional problem of variable selection we propose a Stochastic Search Variable Selection (SSVS) method that relies on a dirac spik–and–slab prior where the slab component induces the sparse non–negative regularisation. The methodology is then applied to the problem of passive index tracking of large dimensional index in stock markets without short sales.

**Abstract** *In questo lavoro introduciamo un nuovo metodo per la selezione sparsa delle variabili in un modello di regressione lineare quando i parametri di regressione sono vincolati ad essere positivi. Il metodo di regolarizzazione impiegato è una estensione della metodologia LASSO recentemente estesa all'ambito bayesiano da Park and Casella (2008). L'obiettivo della selezione dei regressori viene raggiunto attraverso l'estensione del metodo Stochastic Search Variable Selection (SSVS) al caso di regressori non negativi con introducendo una distribuzione a priori di tipo spike–and–slab. La metodologia sviluppata è applicata al problema della replicazione passiva di un indice finanziario.*

**Key words:** Bayesian inference, Non–Negative Lasso, Sparsity, Spike and Slab prior, Index Tracking.

## 1 Introduction

High–dimensional data analysis dealing with models where the number of parameters is larger than the sample size, is becoming one of the most important and active

---

Costola Michele

SAFE, Goethe University Frankfurt e-mail: costola@safe.uni-frankfurt.de

research area of statistics. Since the seminal paper of Tibshirani (1996) that introduced the least absolute shrinkage and selection operator (LASSO), i.e., the first and most popular method that can simultaneously perform parameters estimation and selection in regression models, a number of relevant contributions have been proposed with the same purpose of delivering sparse estimators in high–dimensions. The least angle regression (LARS) of Efron et al. (2004), the adaptive LASSO of Zou (2006) and the group LASSO of Yuan and Lin (2006) are among the most important shrinkage methods proposed in the last 20 years.

In this paper, we propose a novel Bayesian approach to the problem of variable selection and shrinkage in high dimensional sparse linear regression models where regression coefficients are also subject to non–negative constraints. The regularisation method is an extension of the LASSO which has been recently cast in a Bayesian framework by Park and Casella (2008), Carvalho et al. (2010) and Hans (2009). Moreover, since as realised by Tibshirani (2011) the Bayesian LASSO of Park and Casella (2008) based on the Laplace prior does not deliver sparse estimates, we deal with the additional problem of variable selection using a Stochastic Search Variable Selection (SSVS) method that relies on a dirac spike–and–slab prior where the slab component induces the sparse non–negative regularisation. The methodology is then applied to solve the practical issues of passive index tracking of large dimensional index in stock markets without short sales.

## 2 The Bayesian non–negative $\ell_1$ –regularised regression

Let  $\mathbf{y} = (y_1, y_2, \dots, y_T)'$  be the vector of observations on the scalar response variable  $Y$ ,  $\mathbf{X} = (\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_T)'$  is the  $(n \times p)$  matrix of observations on the  $p$  covariates, i.e.,  $\mathbf{x}_{j,t} = (x_{j,1}, x_{j,2}, \dots, x_{j,p})$  and consider the following regression model

$$\pi(\mathbf{y} \mid \mathbf{X}, \alpha, \beta, \sigma_\varepsilon^2) = \mathcal{N}(\mathbf{y} \mid \iota_T \alpha + \mathbf{X}\beta, \sigma_\varepsilon^2) \quad (1)$$

$$\pi(\alpha \mid \tau, \sigma_\varepsilon) = \mathcal{L}_+(\alpha \mid \tau, \sigma_\varepsilon) \quad (2)$$

$$\pi(\beta \mid \tau, \sigma_\varepsilon) = \prod_{j=1}^p \mathcal{L}_+(\beta_j \mid \tau, \sigma_\varepsilon), \quad (3)$$

where  $\iota_T$  is the  $T \times 1$  vector of unit elements,  $\alpha \in \mathbb{R}$  denotes the parameter related to the intercept of the model,  $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$  is the  $p \times 1$  vector of regression parameters and  $\sigma_\varepsilon^2 \in \mathbb{R}^+$  is the scale parameter. We assume equations (2)–(3) as the prior distributions on the constant and regression parameters, respectively, which are assumed to be exponential distributed

$$\mathcal{L}_+(x \mid \tau, \sigma_\varepsilon) = \frac{\tau}{\sigma_\varepsilon} \exp\left\{-\frac{\tau x}{\sigma_\varepsilon}\right\} \mathbb{1}_{[0, \infty)}(x), \quad (4)$$

where  $\tau \in \mathbb{R}^+$  acts as the shrinkage parameter in the Lasso framework and  $\sigma_\varepsilon$  is the scale parameter. We want to project the regression parameters  $\beta_* = (\alpha, \beta')'$  onto

the  $\ell_1^+$  space, i.e.,  $\mathbb{1}_{\mathcal{O}_{p+1}^+}(\beta_*)$ , where  $\mathcal{O}_{p+1}^+$  denotes the positive orthant of dimension  $p+1$ . From a Bayesian probabilistic point of view, this projection is equivalent to introduce the truncated Laplace prior on  $\beta_*$  defined in equation (3). The next proposition introduces the  $\ell_1^+$ -regularised version of the static regression parameters.

**Proposition 1.** *Applying Bayes' theorem to the lasso regression model, the posterior distribution in orthant-wise Normal*

$$\pi(\beta_* | \mathbf{y}, \mathbf{X}, \sigma_\varepsilon, \tau) = \varpi(\mathbf{y}, \mathbf{X}, \tau, \sigma_\varepsilon) \frac{\mathcal{N}(\beta_* | \hat{\beta}_*^+, \Sigma)}{\Phi_{p+1}^+(\hat{\beta}_*^+, \Sigma)} \mathbb{1}_{\mathcal{O}_{p+1}^+}(\beta_*), \quad (5)$$

where  $\Phi_{p+1}^+(\mathbf{m}, \mathbf{S}) = \int_{\mathcal{O}_{p+1}^+} \mathcal{N}(\mathbf{t} | \mathbf{m}, \mathbf{S}) d\mathbf{t}$ , and

$$\hat{\beta}_*^+ = \hat{\beta}_* - \tau \sigma_\varepsilon^{-1} \Sigma \mathbf{1} \quad (6)$$

$$\Sigma = \sigma_\varepsilon^2 (\mathbf{X}'_* \mathbf{X}_*)^{-1} \quad (7)$$

$$\hat{\beta}_* = (\mathbf{X}'_* \mathbf{X}_*)^{-1} \mathbf{X}'_* \mathbf{y} \quad (8)$$

$$\mathbf{X}_* = [\mathbf{1}_T \ \mathbf{X}] \quad (9)$$

$$\varpi(\mathbf{y}, \mathbf{X}, \tau, \sigma_\varepsilon) = \int \pi(\mathbf{y} | \mathbf{X}, \alpha, \beta, \sigma_\varepsilon^2) \pi(\alpha, \beta | \tau, \sigma_\varepsilon) \mathbb{1}_{\mathcal{O}_{p+1}^+}(\beta_*) d\beta_* \quad (10)$$

$$= \left( \frac{\tau}{\sigma_\varepsilon} \right)^{p+1} \frac{\Phi_{p+1}^+(\hat{\beta}_*^+, \Sigma) \prod_{t=1}^T \phi(y_t | 0, \sigma_\varepsilon^2)}{\phi_{p+1}(\mathbf{0} | \hat{\beta}_*^+, \Sigma)}, \quad (11)$$

and  $\mathbf{1}_{p+1}$  is the  $(p+1) \times 1$  vector of unit elements.

## 2.1 Dirac Spike-and-slab prior

Using standard notation, let  $\gamma$  be the  $p$ -vector where  $\gamma_j = 1$  if the  $j$ -th covariate is included as explanatory variable in the regression model and  $\gamma_j = 0$ , otherwise. Assuming that  $\gamma_j \sim \mathcal{B}er(\omega)$ , the prior distribution for  $\beta_j$ ,  $j = 1, 2, \dots, p$  can be written as the mixture

$$\pi(\beta_j | \tau, \sigma_\varepsilon, \omega) = (1 - \omega) \delta_0(\beta_j) + \omega \mathcal{L}_+(\beta_j | \tau, \sigma_\varepsilon), \quad (12)$$

where  $\delta_0(\beta_j)$  is a point mass at zero. The regression model defined in equations (1)–(3) with the spike and slab  $L_1^+$  prior defined in equation (12) becomes

$$\pi(\mathbf{y} | \mathbf{X}, \alpha, \beta, \sigma_\varepsilon^2) = \mathcal{N}(\mathbf{y} | \mathbf{1}_T \alpha + \mathbf{X} \beta, \sigma_\varepsilon^2) \quad (13)$$

$$\pi(\beta | \tau, \sigma_\varepsilon, \omega) = \prod_{j=1}^p \left[ (1 - \omega) \delta_0(\beta_j) + \omega \mathcal{L}_+(\beta_j | \tau, \sigma_\varepsilon) \right], \quad (14)$$

while we retain the same prior defined in equation (2) for the parameter  $\alpha$ . Under the spike and slap prior in equation (14), an iteration of the Gibbs sampling algorithm cycles through the full conditional distribution  $\beta_j | \mathbf{y}, \mathbf{X}, \alpha, \beta_{-j}, \tau, \sigma_\varepsilon^2, \omega$ , where  $\beta_{-j}$  denotes the vector of regression parameters without the  $j$ -th element, for  $j = 1, 2, \dots, p$ . The next proposition provides the analytical expression for the full conditional distribution of  $\beta_j$ , for  $j = 1, 2, \dots, p$ .

**Proposition 2.** *Applying Bayes' theorem to the lasso regression model, the full conditional distributions of the parameters  $(\alpha, \beta)$  is*

$$\pi(\alpha | \mathbf{y}, \mathbf{X}, \beta_{-j}, \sigma_\varepsilon) = \mathcal{N}\left(\alpha | \hat{\alpha}^+, \frac{\sigma_\varepsilon^2}{T}\right) \mathbb{1}_{(0, \infty)}(\beta_j), \quad (15)$$

with  $\hat{\alpha}^+ = \frac{1}{T} \sum_{t=1}^T (y_t - \mathbf{x}'_t \beta) - \frac{\sigma_\varepsilon \tau}{T}$ , and, for  $j = 1, 2, \dots, p$

$$\begin{aligned} \pi(\beta_j | \mathbf{y}, \mathbf{X}, \alpha, \beta_{-j}, \sigma_\varepsilon, \tau, \omega) &= \tilde{\omega}_j^0(\mathbf{y}, \mathbf{X}, \alpha, \beta_{-j}, \tau, \sigma_\varepsilon, \omega) \delta_0(\beta_j) \\ &+ \left(1 - \tilde{\omega}_j^0(\mathbf{y}, \mathbf{X}, \alpha, \beta_{-j}, \tau, \sigma_\varepsilon, \omega)\right) \\ &\times \frac{1}{\Phi_1\left(\frac{\hat{\beta}_j^+}{\sigma_j^{2+}}\right)} \mathcal{N}\left(\beta_j | \hat{\beta}_j^+, \sigma_j^{2+}\right) \mathbb{1}_{(0, \infty)}(\beta_j), \end{aligned} \quad (16)$$

where  $\Phi_1(x) = \int_{-\infty}^x \phi(z) dz$  and

$$\hat{\beta}_j^+ = (\mathbf{x}'_j \mathbf{x}_j)^{-1} \left[ \mathbf{x}'_j \left( \mathbf{y} - \iota_T \alpha - \mathbf{X}_{-j} \beta_{-j} \right) - \sigma_\varepsilon \tau \right] \quad (17)$$

$$\sigma_j^{2+} = \sigma_\varepsilon^2 (\mathbf{x}'_j \mathbf{x}_j)^{-1}, \quad (18)$$

with

$$\begin{aligned} \omega_j^+ (\mathbf{y}, \mathbf{X}, \alpha, \beta_{-j}, \tau, \sigma_\varepsilon, \omega) &= \int \pi(\mathbf{y} | \mathbf{X}, \alpha, \beta, \sigma_\varepsilon^2) \pi(\beta_j | \tau, \sigma) \mathbb{1}_{(0, \infty)}(\beta_j) d\beta_j \\ &= \frac{\omega \tau}{(\sigma_\varepsilon^2)^{\frac{1-p}{2}}} \frac{\Phi_1\left(\frac{\hat{\beta}_j^+}{\sigma_j^{2+}}\right) \prod_{t=1}^T \phi(y_t | 0, \sigma_\varepsilon^2)}{\exp\left\{-\frac{\mathbf{y}' \mathbf{X}_{*, -j} \beta_{*, -j}}{\sigma_\varepsilon^2}\right\}} \phi\left(0 | \hat{\beta}_j^+, \sigma_j^{2+}\right) \\ &\times \frac{\phi_p\left(\mathbf{0} | \beta_{*, -j}, \sigma_\varepsilon^2 \left(\mathbf{X}'_{*, -j} \mathbf{X}_{*, -j}\right)^{-1}\right)}{|\mathbf{X}'_{*, -j} \mathbf{X}_{*, -j}|^{\frac{1}{2}}}, \end{aligned} \quad (19)$$

and



$$\begin{aligned}
\varpi_j^0(\mathbf{y}, \mathbf{X}, \alpha, \beta_{-j}, \tau, \sigma_\varepsilon, \omega) &= \int \pi(\mathbf{y} | \mathbf{X}, \alpha, \beta, \sigma_\varepsilon^2) \pi(\beta_j | \tau, \sigma) \mathbb{1}_{(0, \infty)}(\beta_j) d\beta_j \\
&= \frac{(1-\omega) \prod_{t=1}^T \phi(y_t | 0, \sigma_\varepsilon^2)}{(\sigma_\varepsilon^2)^{-\frac{p}{2}} \exp\left\{-\frac{\mathbf{y}' \mathbf{X}_{*, -j} \beta_{*, -j}}{\sigma_\varepsilon^2}\right\}} \\
&\quad \times \frac{\phi_p(\mathbf{0} | \beta_{*, -j}, \sigma_\varepsilon^2 (\mathbf{X}'_{*, -j} \mathbf{X}_{*, -j})^{-1})}{|\mathbf{X}'_{*, -j} \mathbf{X}_{*, -j}|^{\frac{1}{2}}}, \quad (20)
\end{aligned}$$

and

$$\bar{\omega}_j^0(\mathbf{y}, \mathbf{X}, \alpha, \beta_{-j}, \tau, \sigma_\varepsilon, \omega) = \left[ 1 + \frac{\omega \tau}{(1-\omega) \sigma_\varepsilon} \frac{\Phi\left(\frac{\hat{\beta}_j^+}{\hat{\sigma}_j^+}\right)}{\phi\left(0 | \hat{\beta}_j^+, \hat{\sigma}_j^{2+}\right)} \right]^{-1}, \quad (21)$$

where  $\mathbf{X}_{*, -j} = [1_T \ \mathbf{X}_{-j}]$  and  $\beta_{*, -j} = [\alpha \ \beta_{-j}]$ , for  $j = 1, 2, \dots, p$ .

## 2.2 The Gibbs sampler

The scale parameter  $\sigma_\varepsilon$  and the shrinkage parameter  $\tau$ , as well as the prior inclusion probability  $\omega$  are parameters that have to be estimated. Common choices for the prior on those parameters are  $\sigma_\varepsilon^2 \sim \mathcal{IG}(\sigma_\varepsilon^2 | \lambda_\sigma, \eta_\sigma)$ ,  $\tau \sim \mathcal{G}(\tau | \lambda_\tau, \eta_\tau)$  and  $\omega \sim \mathcal{Be}(\omega | \lambda_\omega, \eta_\omega)$ , where  $(\lambda_\sigma, \eta_\sigma, \lambda_\tau, \eta_\tau, \lambda_\omega, \eta_\omega)$  are prior hyperparameters. Under this prior the full conditional distribution of the scale parameter  $\sigma_\varepsilon^2$  will be equal to

$$\pi(\sigma_\varepsilon^2 | \mathbf{y}, \mathbf{X}_\gamma, \alpha, \beta_\gamma, \tau, \omega) \propto \mathcal{IG}(\sigma_\varepsilon^2 | \tilde{\lambda}_\sigma, \tilde{\eta}_\sigma) \prod_{j=1}^{n_\gamma} \exp\left\{-\frac{\tau \beta_j}{\sigma_\varepsilon}\right\} \mathbb{1}_{(0, \infty)}(\sigma_\varepsilon^2), \quad (22)$$

where  $n_\gamma$  is the number of nonzero elements of  $\beta$ ,  $\mathbf{X}_\gamma$  is the  $(T \times n_\gamma)$  matrix collecting the observations on the variables included in the regression, i.e., with  $\gamma_j = 1$ ,  $\beta_\gamma$  is the  $(n_\gamma \times 1)$  vector of regressors included, and  $\tilde{\lambda}_\sigma = \lambda_\sigma + \frac{T+n_\gamma}{2}$ ,  $\tilde{\eta}_\sigma = \eta_\sigma + \frac{1}{2} \mathbf{S}_\gamma$  and  $\mathbf{S}_\gamma = (\mathbf{y} - \mathbf{X}_\gamma \beta_\gamma)' (\mathbf{y} - \mathbf{X}_\gamma \beta_\gamma)$ . The full conditional distribution of the penalty parameter  $\tau$  is

$$\pi(\tau | \mathbf{y}, \mathbf{X}_\gamma, \alpha, \beta_\gamma, \sigma_\varepsilon, \omega) \propto \mathcal{G}(\tau | \tilde{\lambda}_\tau, \tilde{\eta}_\tau), \quad (23)$$

with parameters  $\tilde{\lambda}_\tau = \lambda_\tau + \frac{n_\gamma}{2}$  and  $\tilde{\eta}_\tau = \eta_\tau + \frac{\sum_{j=1}^{n_\gamma} \beta_j}{\sigma_\varepsilon}$ . The full conditional distribution of the sparsity parameter  $\omega$  is

$$\pi\left(\omega \mid \mathbf{y}, \mathbf{X}_\gamma, \alpha, \beta_\gamma, \sigma_\varepsilon, \tau\right) \propto \mathcal{B}\left(\omega \mid \tilde{\lambda}_\omega, \tilde{\eta}_\omega\right), \quad (24)$$

with parameters  $\tilde{\lambda}_\omega = \lambda_\omega + n_\gamma$  and  $\tilde{\eta}_\omega = \eta_\omega + p - n_\gamma$ .

The final algorithm for the linear regression model consists of choosing the initial parameters values  $\left(\beta^{(0)}, \sigma_\varepsilon^{2(0)}, \tau^{(0)}, \omega^{(0)}\right)$  and iteratively sampling  $\left(\beta^{(k)}, \sigma_\varepsilon^{2(k)}, \tau^{(k)}, \omega^{(k)}\right)$ , for  $k = 1, 2, \dots$  from

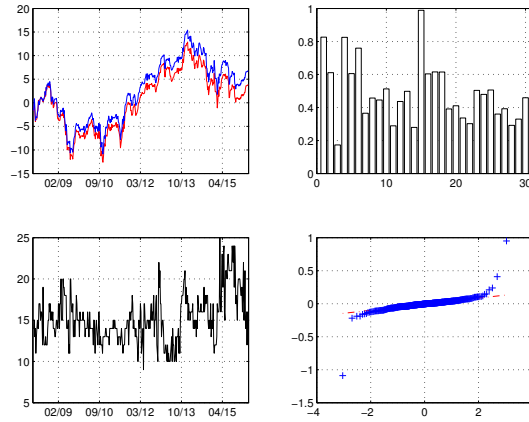
- (i)  $\alpha^{(k)} \sim \pi\left(\alpha \mid \mathbf{y}, \mathbf{X}_\gamma, \beta^{(k-1)}, \sigma_\varepsilon^{2(k-1)}, \tau^{(k-1)}, \omega^{(k-1)}\right)$  defined in equation (15), for  $j = 1, 2, \dots, p$ ;
- (ii)  $\beta_j^{(k)} \sim \pi\left(\beta_j \mid \mathbf{y}, \mathbf{X}_\gamma, \alpha^{(k)}, \beta_{-j}^{(k-1)}, \sigma_\varepsilon^{2(k-1)}, \tau^{(k-1)}, \omega^{(k-1)}\right)$  defined in equation (16), for  $j = 1, 2, \dots, p$ ;
- (iii)  $\sigma_\varepsilon^{2(k)} \sim \pi\left(\sigma_\varepsilon^2 \mid \mathbf{y}, \mathbf{X}_\gamma, \alpha^{(k)}, \beta^{(k)}, \tau^{(k-1)}, \omega^{(k-1)}\right)$  defined in equation (22);
- (iv)  $\tau^{(k)} \sim \pi\left(\tau \mid \mathbf{y}, \mathbf{X}_\gamma, \alpha^{(k)}, \beta^{(k)}, \sigma_\varepsilon^{2(k)}, \omega^{(k-1)}\right)$  defined in equation (23);
- (v)  $\omega^{(k)} \sim \pi\left(\omega \mid \mathbf{y}, \mathbf{X}_\gamma, \alpha^{(k)}, \beta^{(k)}, \sigma_\varepsilon^{2(k)}, \tau^{(k)}\right)$  defined in equation (24).

### 3 Application to index tracking

In this section, we focus on the application of non-negative Bayesian regularised regression in financial modelling. The performance of the variable selection of non-negative regularised regression is tested when the method is applied to tracking the index. We use genuine data from the DJIA index. Index tracking is a quantitative passive trading scheme which aims to replicate the returns of a given portfolio of assets over a certain time horizon by peaking portfolios' constituents which are most correlated with the portfolio returns. Index tracking, which attempts to match the performance of index as closely as possible, is one of the most popular topic in statistical finance. Two main reasons justify the use of non-negative Bayesian regularised regression for index tracking. First, statistical modelling of large dimensional indexes is a typical high-dimensional problem where the number of regressors  $p$  is usually larger than the number of observations. Second, for the cost concern, the optimal replication index should match the the performance of the entire index by choosing a the smallest subset target stocks. Bayesian non-negative regression with spike-and-slab prior successfully leads to sparse estimates of the regression parameters leading to tracking solutions where only few regressors are nonzero.

#### 3.1 Data and results

Our data set consists of the end-of-week weekly prices of stocks in DJIA 30 Index, from 20 March 2008 to 10 March 2017 (the data come from the database). Weekly prices are subsequently converted to weekly returns. For a price  $P_t$ , weekly returns



**Fig. 1** Replication results of the DJIA 30 index using a window of dimension  $W = 24$  weekly observations (about 6 months). (*Top left panel*) plots the true value of the DJIA index (red line), along with the replicating portfolio, (blue line). (*Top right panel*) posterior inclusion probabilities of each regressor averaged over all the subsampling periods. (*Bottom left panel*) number of non-zero regressors for each rolling window estimate. (*Bottom right panel*) Normal qq-plot of the replicating portfolio.

are defined as  $r_t = \frac{P_t}{P_{t-1}} - 1$ , for  $t = 1, 2, \dots, T$ . Let  $x_{i,t} = r_{i,t}$ , for  $i = 1, 2, \dots, 30$  represent the returns of the  $i$ -th constituent stock and  $y_t = r_t^M$  represent the return of the index. Then we can describe the relationship between  $x_{i,t}$  and  $y_t$  by the linear regression model defined in (1), where  $\beta$  is sparse since partial replication for index tracking which only selects a small subset. The Bayesian non-negative regularisation is then repeatedly applied to get the estimation of the regression parameters  $\beta$  using a rolling windows estimation of  $W = 24$  observations over the entire sample of observations while retaining the last observation of each subsample for tracking the index. Estimation results are reported in Figure 1. The top-left panel of Figure 1 plots compares the cumulative weekly log-returns over the whole sampling period of the true index (red line) and the replicating index (blue line). It is evident from the figure that the replicating index is quite close to the true index denoting that our Bayesian non-negative regularised regression method provide satisfactory results. The bottom-left panel plots the number of asset used to replicate the index for each estimating windows. The average number of non-zero coefficients is less than half of the number of components denoting that our method replicates well the index using a quite small subset of constituents.

## 4 Conclusion

We propose Bayesian non-negative regularised regression that performs parameters estimation and variable selection under spike-and-slab- $\ell_1$  prior via the Stochastic Search Variable Selection algorithm in high-dimensional linear regression models where regression coefficients are also constrained to be non negative. We propose an efficient Gibbs sampling algorithm for posterior simulation from the augmented space of parameters and inclusion indexes. In the following empirical application, we use the Bayesian non-negative regularised regression to track the DJIA 30 index return by choosing a subset of its constituent stocks. We demonstrate that our algorithm provides satisfactory results in replicating the index using a quite small subsets of constituents.

## Acknowledgement

The author acknowledges financial support from the Marie Skłodowska-Curie Actions, European Union, Seventh Framework Program HORIZON 2020 under REA grant agreement n.707070. He also gratefully acknowledges research support from the Research Center SAFE, funded by the State of Hessen initiative for research LOEWE.

## References

- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2):407–499.
- Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.
- Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.
- Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.

# Industrial Production Index and the Web: an explorative cointegration analysis.

## *L'indice della produzione industriale e il web: un'analisi esplorativa di cointegrazione.*

Lisa Crosato, Caterina Liberati, Paolo Mariani and Biancamaria Zavanella

**Abstract** In this paper we explore the relationship between the Industrial Production Index (IPI), the confidence index for the manufacturing sector and its sub-indexes and Google searches for several words linked to the economic situation, for the period January 2004 - September 2016 on Italian data. Significant correlations between the selected indicators point to probable comovements of same. Adding one observation at a time since the first forewarning signs of the 2008 crisis, we find that a few Google searches and the IPI cointegrate, particularly during the strong downward trend leading to January 2009, while no confidence indicators cointegrate with the IPI. These findings suggest that concern about economic conditions expressed through searches in google and the IPI or the confidence indexes are influenced by common circumstances. Recursive forecasts of the IPI through VECM models suggest that the evolution of the IPI can be well mimicked using the real time Gtrends selected variables.

**Abstract** *In questo articolo vengono esplorate le interrelazioni fra l'indice della produzione industriale (IPI), l'indice di fiducia del settore manifatturiero e suoi sub-indici e le ricerche su google per tre parole legate alla situazione economica del Paese durante il periodo gennaio 2004 - settembre 2016. Le relazioni fra variabili significativamente correlate vengono approfondite tramite un'analisi della cointegrazione sulle serie storiche degli indicatori a partire dai primi segnali della crisi del 2008. I risultati suggeriscono che la preoccupazione riguardo alle condizioni economiche del paese espressa tramite le ricerche online e l'indice della produzione o gli indicatori di fiducia delle imprese sono influenzati da circostanze comuni. Una previsione sequenziale dell'IPI conclude il lavoro.*

**Key words:** Industrial Production Index, Big Data, Google Trends, Confidence Indicators, Cointegration

---

Università di Milano-Bicocca - DEMS - Via Bicocca degli Arcimboldi, 8, 20126 Milano. e-mail: [lisa.crosato@unimib.it](mailto:lisa.crosato@unimib.it); [caterina.liberati@unimib.it](mailto:caterina.liberati@unimib.it), [paolo.mariani@unimib.it](mailto:paolo.mariani@unimib.it); [biancamaria.zavanella@unimib.it](mailto:biancamaria.zavanella@unimib.it). We thank Chiara Zannier for the use of her Google trends data.

## 1 Introduction

The Industrial Production Index is probably the main monthly indicator attesting the current health of a country's economy. Accordingly, several contributions in the literature proposed simple to complex models to forecast it usually imputing hard data as regressors, from macroeconomic variables to business-specific indicators (Bodo and Signorini, 1987; Bruno and Lupi, 2004; Hassani et al., 2013). Soft data, such as text analysis in media and other sentiment indicators were introduced instead by Ulbricht et al. (2016) to predict the German IPI with more than 17,000 models.

In this paper we intend to pursue much a simpler goal. The idea is to explore the comovements of official statistics on industrial production and non-official indicators built on google searches for words related to the general and personal economic situation. The main goal of the paper is to understand whether web based soft index numbers together with confidence indicators may help in predicting the hard IPI. Our empirical strategy is to proceed by subsequent selection of variables, firstly by simple visual inspection on the range of variability and secondly by analysing their correlation with the IPI. Were the correlations between the IPI and one or more soft indicators significant, one could try and see whether the relationship may be represented also through time series modeling. So our third step for selection of indicators is to leave behind the stationary ones in order to proceed to the final cointegration analysis. Finally we test for more than one cointegration relationship among confidence indicators, google searches and the IPI, to end up with VECM based short term forecasts of the IPI. Our work is on the lines of Daas et al. (2014), however this paper differs at least in two aspects, besides the objective variable: to begin with, we analyse integration and cointegration between and among indicators in a recursive fashion, moreover we also forecast the IPI.

## 2 Data description

This paper makes use of three data sources, two of which official and a third one non-official. The first is the Industrial Production Index (IPI hereafter) monthly released by ISTAT (Italian Institute of Statistics) with two months of delay with the reference period. The IPI is a 2010 fixed base Laspeyres index and is the main conjunctural indicator measuring real output for all facilities located in Italy.

The second data source is the Italian confidence index for manufacturing, monthly released by ISTAT with about 15 days of delay with respect to the interviews. We have selected in particular opinions on current level of orders, current economic situation, future level of orders and future economic situation and the composite confidence indicator.

The third data source we use is Google Trends, a free tool by Google that allows to download the number of times a word or a sentence has been searched in the Google and YouTube websites. The idea in using Google trends is to build statistical variables for measuring the interest of the people of a country in spe-

cific ambits over time. The economic literature has been using Google trends since its appearance in 2004 (see Hassani and Silva (2015) for a recent review on forecasting using Big Data). Google trends data are released as monthly frequencies of searches starting from January 2004, therefore this is the initial date for all our time series. Since in this paper we are interested in understanding whether Google searches can be considered and used as proxies of the IPI, the words we have searched for in Google Trends are related to the economic situation, especially regarding general concerns about the economic situation. We searched for *economic crisis*, *recovery*, *GDP*, *gross domestic product*, *public debt*, *spread*, *recession*, *unemployment*, *employment*, *job*. We also construct naive composite Gtrends indicators by summing up frequencies associated to related words so obtaining four more variables: *Total cycle=economic crisis+recession+recovery*, *Total occupation=unemployment+employment+job*, *Total Debt=public debt+spread* plus a mixed-up variable *three words=economic crisis+unemployment+public debt*. The actual Italian words searched for are: crisi economica, ripresa, PIL, prodotto interno lordo, spread, recessione, disoccupazione, occupazione, lavoro.

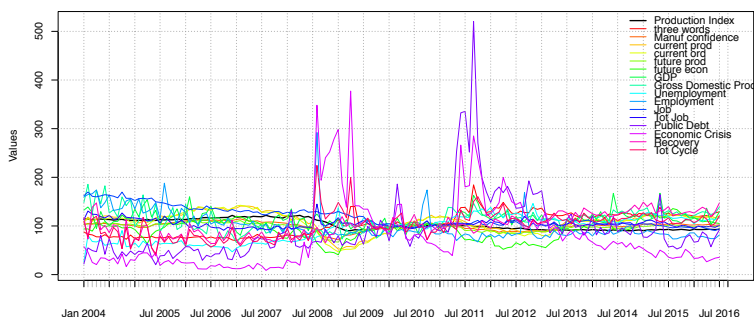
The official statistics we use in the paper are all expressed as index numbers in base 2010, so in order to have a fair comparison we have indexed to 2010 the Gtrends data. To this end, the single and composite words monthly frequencies were divided by the mean of 2010 respective frequencies.

### 3 Methodology and main results

The time series from the three data sources we used differ at least in two aspects. First, the IPI and the confidence indicators (when needed) are published already deseasonalized, while the Gtrends variables must be treated for seasonality. Therefore, we apply the R-interface to X13ARIMA-SEATS method by the United States Census Bureau. Second, they are released with different lags with respect to the date of the information they are referred to. At the end of each month we dispose of the IPI of two months earlier, while confidence indicators and Gtrends variables refer to the current month. Accordingly, we shape the data matrix anticipating all confidence and Gtrends indicators by two months. All the time series thus obtained are represented in figure 1. A quick glance to the series reveals different degrees of variability among the time series, highlighting the structural difference among the indicators. The flatter series is for sure the IPI, followed by the confidence indicators and the Gtrends variables. Gtrends variables are clearly more volatile and subject to sudden jumps in correspondance of particular events (for instance, see in figure 1 the spikes in *economic crisis* from spring 2008 onwards and of *three words* at the end of the Berlusconi Government in summer-fall 2011).

The final aim of the paper is to explore whether Gtrends variables and confidence indicators may show some predictive power on IPI. We intend to do this by a multivariate time series model (VAR or VECM if any cointegration relationship appears). In particular, we adopt a forward approach adding one observation at a

**Fig. 1** Time series of the selected indicators. Sources: ISTAT official statistics and our own elaborations on google trends data.



time from April 2008 onwards to monitor changes in the cointegration relationship during the observed period. Thus, the length of the series increases by one from 52 to 151 observations.

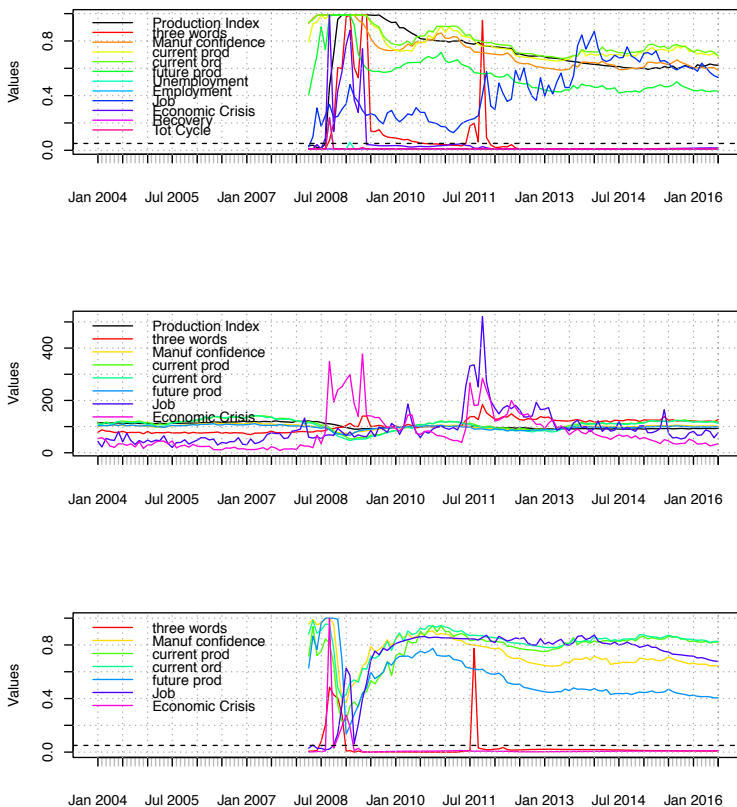
We proceed by subsequent selection of the initial variables by first eliminating those indicators showing too wide a range of variation (*spread*, *recession* and *total debt*). Secondly, we restrict the choice to variables which correlate with the IPI. We thus decided to eliminate all variables showing no significant correlation relationship with the IPI and, among the remaining, those presenting a correlation coefficient lower than 0.3. This way, we have discarded *GDP*, *Gross domestic product*, *total job*, *public debt*.

The third step in variable selection was to test for the presence of Unit root in the series, as a preliminary information for the cointegration analysis. We performed the Phillips Perron test for unit root on the whole set of 100X12 series. The p-values in figure 2, top panel, show that not all variables are integrated but most importantly that the IPI is integrated at least from July 2008 together with all the confidence indexes. Among Gtrends variables, *Job* is never stationary while *three words* and *economic crisis* are not stationary only over slightly different subperiods. Accordingly, we have left behind also *employment*, *unemployment*, *recovery* and *tot cycle* (see the remaining series in figure 2, central panel). Now we move to the cointegration analysis of the IPI index with one of the remaining variables in turn (i.e. all the confidence indicators and *three words*, *job*, *economic crisis*). Results of the Engle and Granger test for cointegration, reported in figure 2 (p-values, bottom panel) point to no cointegration neither between the IPI and the confidence indexes, nor between the IPI and *job*. On the contrary, the IPI and *three words* do cointegrate and so do IPI and *economic crisis*, although there are some spikes when the turbulence in the two Gtrends variables is higher. The cointegration analysis between confidence indicators and *threewords* reveals a similar outcome.

We think this can be viewed as a first result of the paper contributing to define a selection strategy for Gtrends variables to augment forecasting models with, although at present restricted to this particular case. If variables cointegrate when



**Fig. 2** Unit root tests (top panel, p.values of the Phillips Perron test), I(1) series (central panel) and Engle and Granger cointegration test (bottom panel, p-values of the Phillips Perron test on residuals). Both tests are applied adding one observation at a time from April 2008 onwards. Sources: ISTAT official statistics and our own elaborations.

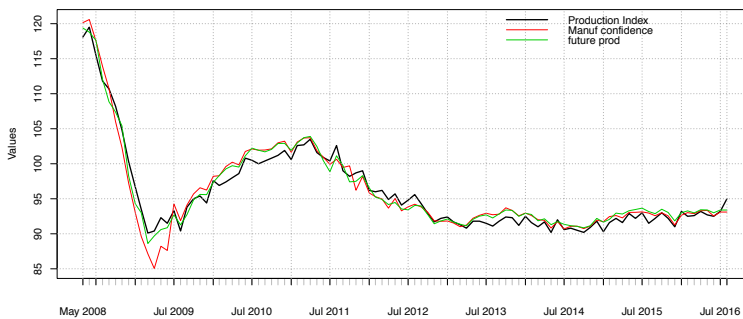


shaped by a common factor or by a combination of common factors, we may tentatively say that a few of the Gtrends variables and the IPI share some pattern drivers.

We conclude this exercise with a simple prediction based on a VECM model estimated on the IPI, the *threewords* index and one confidence indicator in turn. This is a way to measure the possible contribution to prediction of IPI by one or more confidence indicators and to better exploit pieces of information shared by Gtrends variables and the confidence indicators, although none of the latter cointegrate with the IPI. Again the VECM model was estimated 100 times on the month by month augmented time series, resulting in 100 forecasted values of the IPI, from May 2008 to September 2016 (see figure 3). The preliminary Johansen test confirms one rank of cointegration almost always for the confidence indicator on future orders and to a minor extent for the composite manufacturing confidence index,

while the introduction of current orders or current production indicators seems to weaken the cointegration relationship between *three words* and the IPI. Therefore, the 100 IPI forecasted values in figure 3 are obtained through VECM models based on IPI, *three words* and the manufacturing confidence index or the confidence in future production. As can be seen predictions closely follow the actual values of the production index, in downward as well as in upward changes. The median percentage absolute error is smaller for the confidence in future production (0.9%) with respect to the manufacturing composite confidence (1.1%) mainly due to the protracted fall in the forecast for April 2009, when the IPI had already turned up. Note that these predictions are available two months earlier than the official IPI.

**Fig. 3** Recursive forecast of the IPI by VECM models using one of the listed variables together with the IPI and the *three words* Gtrends variable. Sources: ISTAT official statistics and our own elaborations.



## References

- Bodo, G. and L. F. Signorini (1987). Short-term forecasting of the industrial production index. *International Journal of Forecasting* 3(2), 245–259.
- Bruno, G. and C. Lupi (2004). Forecasting industrial production and the early detection of turning points. *Empirical economics* 29(3), 647–671.
- Daas, P. J., M. J. Puts, et al. (2014). Social media sentiment and consumer confidence. Technical report, European Central Bank.
- Hassani, H., S. Heravi, and A. Zhigljavsky (2013). Forecasting UK industrial production with multivariate singular spectrum analysis. *Journal of Forecasting* 32(5), 395–408.
- Hassani, H. and E. S. Silva (2015). Forecasting with big data: A review. *Annals of Data Science* 2(1), 5–19.
- Ulbricht, D., K. A. Kholodilin, and T. Thomas (2016). Do media data help to predict german industrial production? *Journal of Forecasting*.

# Comparison of conditional tests on Poisson data

## *Un confronto di test condizionati su dati di Poisson*

Francesca Romana Crucinio and Roberto Fontana

**Abstract** We compare four conditional tests for Poisson data through a simulation study: the exact binomial test, its asymptotic approximation, a Markov Chain Monte Carlo test and the standard permutation test. Despite being non-parametric, we observe that permutation tests are as effective as the others. From a theoretical point of view we justify this result by observing that the orbits of permutations form a *good* partition of the conditional space.

**Abstract** *Si confrontano quattro test condizionati per dati di Poisson: il test binomiale esatto, la sua approssimazione asintotica, un test Markov Chain Monte Carlo e un test di permutazione standard. Si osserva che il test di permutazione, pur non parametrico, ha un comportamento simile agli altri. Una giustificazione teorica di questo risultato sta nell'osservare che le orbite di permutazione costituiscono una buona partizione dello spazio condizionato.*

**Key words:** Algebraic statistics, Conditional test, Permutation test, Poisson data

## 1 Introduction

We address the problem of comparing the means of two Poisson distributions with unknown parameter  $\lambda_i$ ,  $i = 1, 2$ . We consider two independent samples,  $\mathbf{Y}_1^{(n_1)} = (Y_1, \dots, Y_{n_1})$  of size  $n_1$  from Poisson( $\lambda_1$ ) and  $\mathbf{Y}_2^{(n_2)} = (Y_{n_1+1}, \dots, Y_{n_1+n_2})$  of size  $n_2$

---

Francesca Romana Crucinio  
Politecnico di Torino, Dipartimento di Scienze Matematiche  
e-mail: francesca.crucinio@gmail.com

Roberto Fontana  
Politecnico di Torino, Dipartimento di Scienze Matematiche  
e-mail: roberto.fontana@polito.it

from Poisson( $\lambda_2$ ). Then we use the joint sample  $\mathbf{Y} = (\mathbf{Y}_1^{(n_1)}, \mathbf{Y}_2^{(n_2)})$  to perform the test  $H_0 : \lambda_1 = \lambda_2$  against  $H_1 : \lambda_1 \neq \lambda_2$ .

The problem has been extensively studied in the literature. Among the several testing procedures available to researchers, we consider *conditional* tests, i.e. tests that are performed considering only samples  $\mathbf{Y}$  such that the sum  $\mathbf{Y}_+$  of their elements is equal to the sum  $\mathbf{y}_{obs,+}$  of the elements of the observed sample  $\mathbf{y}_{obs}$

$$\mathbf{Y}_+ = \sum_{i=1}^{n_1+n_2} Y_i = \sum_{i=1}^{n_1+n_2} y_{i,obs} = \mathbf{y}_{obs,+}. \quad (1)$$

A justification for this choice is that, if we assume that the model for the means of the two distributions is the standard one-way ANOVA model, which according to [6] is  $\log(\lambda_i) = \beta_0 + \beta_1 x_i$  with  $x_i = 1$  if  $1 \leq i \leq n_1$  and  $x_i = -1$  if  $n_1 + 1 \leq i \leq n_1 + n_2$ , the statistic  $T = \mathbf{Y}_+ = \sum_{i=1}^{n_1+n_2} Y_i$  is sufficient for the population constant  $\beta_0$ , which is the nuisance parameter of the test.

For the sake of simplicity we denote the sum of the observed sample  $\mathbf{y}_{obs,+}$  by  $t$  and the set of the samples  $\mathbf{Y}$  which satisfy (1) by  $\mathcal{F}_t$ . We refer to  $\mathcal{F}_t$  as the *fiber* corresponding to  $t$ . We focus on four conditional tests:

1. the exact binomial test by Przyborowski and Wilenski [8];
2. an asymptotic version of the exact binomial test [8], which is based on the normal approximation of the binomial distribution [4];
3. a Markov Chain Monte Carlo testing procedure which exploits Markov basis [3] and the Metropolis-Hastings algorithm [9];
4. a standard permutation test [7].

In Section 2 we briefly describe the structure of the tests under study. In Section 3 we compare the effectiveness of the tests through a simulation study and in Section 4 we analyse the link between fibers and permutations from a theoretical perspective. Conclusions are in Section 5.

## 2 Conditional Tests

### *Exact and Asymptotic Conditional Binomial Test*

It is well-known that the distribution of the sum of  $n$  independent Poisson variables of mean  $\lambda$  is a Poisson variable with mean  $n\lambda$ . Then it can be shown that the distribution of the variable  $T_1|T = t$ , i.e. of the variable  $T_1 = \sum_{i=1}^{n_1} Y_i$  conditioned to  $T = \sum_{i=1}^{n_1+n_2} Y_i = t$ , is a Binomial distribution with probability of success  $\theta = (n_1\lambda_1)/(n_1\lambda_1 + n_2\lambda_2)$  and  $t$  trials. It follows that under  $H_0 : \lambda_1 = \lambda_2$  the variable  $T_1|T = t$  follows a binomial distribution with probability of success  $\theta_0 = n_1/(n_1 + n_2)$  and  $t$  trials. If  $t_1$  is the observed value of  $T_1$  the p-value is computed as

$$\min\{2 \min\{p(T_1 \leq t_1), p(T_1 \geq t_1)\}, 1\} \quad (2)$$

where  $p(T_1 \leq t_1) = \sum_{k=0}^{t_1} \binom{t_1}{k} \theta_0^k (1 - \theta_0)^{t_1 - k}$  and  $p(T_1 \geq t_1) = \sum_{k=t_1}^t \binom{t}{k} \theta_0^k (1 - \theta_0)^{t - k}$ .

The asymptotic version of the conditional binomial test uses the asymptotic test statistic

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1 - \theta_0)/n}} \sim N(0, 1) \quad \text{where } \hat{\theta} = T_1/n_1.$$

The p-value is computed as  $2 * (1 - \Phi(|z_{obs}|))$  where  $\Phi$  is the cumulative distribution of the standard normal variable and  $z_{obs} = (t_1/n_1 - \theta_0) / \sqrt{\theta_0(1 - \theta_0)/n}$ .

*The Markov Chain Monte Carlo Test*

As mentioned above we condition on the sum  $t$  of the elements of the observed sample  $\mathbf{y}_{obs}$  and we explore the fiber

$$\mathcal{F}_t = \{(Y_1, \dots, Y_{n_1+n_2}) \in \mathbb{N}^{n_1+n_2} : \sum_{i=1}^{n_1+n_2} Y_i = t\}. \tag{3}$$

To explore the fiber  $\mathcal{F}_t$  as defined in (3) we set up a connected Markov chain by means of a Markov basis, i.e. a set  $\mathcal{B}$  of moves which have to be added/subtracted to the vectors in  $\mathcal{F}_t$  in order to move on the fiber (see [3] for a formal definition of Markov Basis). This basis can be found using the 4t i 2 software [10] or, in this specific case, simply by induction on the sample size  $N = n_1 + n_2$ . We get that  $\mathcal{B}$  is made of  $N - 1$  moves  $\mathbf{m}_U = (1, \delta_{1,U}, \dots, \delta_{N-1,U}), U = 1, \dots, N - 1$  where  $\delta_{a,b} = -1$  if  $a = b$  and 0 otherwise.  $\mathcal{B}$  allows us to build a graph over the fiber, where each pair of vectors  $\mathbf{y}, \mathbf{x} \in \mathcal{F}_t$  is linked by an edge if a move  $\mathbf{m} \in \mathcal{B}$  exists such that  $\mathbf{y} = \mathbf{x} \pm \mathbf{m}$ . An example when  $t = 6$  and  $N = 3$  is shown in Figure 1.

Under  $H_0 : \lambda_1 = \lambda_2 = \lambda$  we exploit the Metropolis Hastings algorithm (an accelerated version as in [1], [2]) to modify the transition probabilities and grant convergence to

$$p(\mathbf{y}) = e^{-\lambda} \frac{\lambda^{y_1}}{y_1!} \dots e^{-\lambda} \frac{\lambda^{y_N}}{y_N!} = e^{-N\lambda} \frac{\lambda^t}{\prod_{i=1}^N y_i!} = C \prod_{i=1}^N \frac{1}{y_i!} \propto \prod_{i=1}^N \frac{1}{y_i!} \tag{4}$$

where  $C = e^{-N\lambda} \lambda^t$ . At each step if we are in state  $\mathbf{y}$  we select a random move  $\mathbf{m}_U \in \mathcal{B}$  and we consider every possible transition  $\mathbf{y} + \gamma \cdot \mathbf{m}_U$  with  $\gamma \in \Gamma = \{\gamma \in \mathbb{Z} : \mathbf{y} + \gamma \cdot \mathbf{m}_U \in \mathcal{F}_t\} = [-y_1, y_{U+1}] \cap \mathbb{Z}$ . We move to  $\mathbf{y} + \gamma^* \cdot \mathbf{m}_U$  with  $\gamma^*$  randomly drawn from the set above with probability

$$q_{\gamma^*} = \frac{p(\mathbf{y} + \gamma^* \cdot \mathbf{m}_U)}{\sum_{\gamma \in \Gamma} p(\mathbf{y} + \gamma \cdot \mathbf{m}_U)} \propto \frac{1}{(y_1 + \gamma^*)! \cdot (y_{U+1} - \gamma^*)!}.$$

This walk on  $\mathcal{F}_t$  allows us to build an approximation of the distribution, under  $H_0$ , of the test statistic  $W = \bar{Y}_1 - \bar{Y}_2 = T_1/n_1 - T_2/n_2$ . Finally the p-value is computed as

$$\frac{\#(|W| \geq |w_{obs}|)}{M} \tag{5}$$

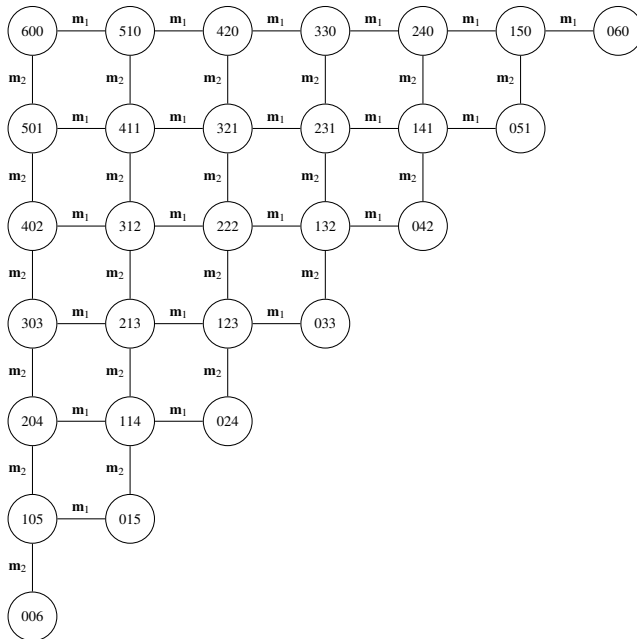


Fig. 1: Graph on the fiber  $\mathcal{F}_t$  with  $t = 6$  and  $N = 3$

where  $M$  is the number of transitions and  $w_{obs}$  is the observed value of  $W$ .

#### Permutation Test

We perform a standard permutation test [7], randomly selecting  $M$  permutations of  $\mathbf{y}_{obs}$  ( $M$  is at least 1,000), computing the corresponding values of  $W$  and the p-value as in (5).

### 3 Simulation Study

We consider 27 scenarios that have been built taking three different sample sizes  $(n_1, n_2)$  (Table 1a) and, for each sample size, nine different population means  $(\lambda_1, \lambda_2)$  (Table 1b).

For each scenario 1,000 samples have been randomly generated. For each sample the corresponding p-values for the four testing procedures under study have been

	1	2	3
$n_1$	3	8	35
$n_2$	17	12	15

(a) Sample sizes

	1	2	3	4	5	6	7	8	9
$\lambda_1$	0.5	0.5	0.5	1	1	1	5	5	5
$\lambda_2$	0.5	0.75	1	1	1.5	2	5	7.5	10

(b) Population means

computed. Specifically for the MCMC test 10,000 moves after the 1,000 used for the burn-in step have been used. For the permutation test 2,000 permutations have been used .

We summarise the most important results:

- the behaviour of the binomial tests (exact and asymptotic) looks different from the behaviour of the Monte Carlo tests (MCMC and permutation). This difference is due to the non-equivalent definitions of p-value ((2) and (5)) and, possibly, to the sampling of the fiber;
- the significance values achieved by the permutation test are almost equivalent to the ones achieved by the MCMC test although this test explores a much wider sample space. We discuss this point in Section 4.

### 4 Fiber and Permutation Sample Space

The permutation operator does not alter the sum of entries. Hence the *orbits* of permutations  $\pi_{\mathbf{y}}$ , where  $\mathbf{y}$  is the generating vector, are subsets of the fiber. The orbits do not intersect and then we can create a partition of  $\mathcal{F}_t$  made up of  $\text{part}(t, N)$  orbits  $\pi_{\mathbf{y}}$ , where  $\text{part}(t, N)$  is the partition function defined in [5].

In the same orbit,  $p(\mathbf{y})$  is constant and then the probability of taking  $\mathbf{y} \in \pi_{\mathbf{y}}$  is  $p(\pi_{\mathbf{y}}) = \sum_{\mathbf{y}^* \in \pi_{\mathbf{y}}} p(\mathbf{y}^*) = \#\pi_{\mathbf{y}} \cdot p(\mathbf{y}) = \#\pi_{\mathbf{y}} \cdot C \prod_{i=1}^N \frac{1}{y_i!}$ , where  $\#\pi_{\mathbf{y}}$  is the cardinality of  $\pi_{\mathbf{y}}$ . It can be proved that  $C$ , the normalizing constant defined in (4), can be computed as  $C = (\sum_{\pi_{\mathbf{y}} \subseteq \mathcal{F}_t} \#\pi_{\mathbf{y}} \prod_{i=1}^N \frac{1}{y_i!})^{-1}$ , an expression that does not contain the unknown parameter  $\lambda = \lambda_1 = \lambda_2$ .

As an example let us consider the fiber in Figure 1. It can be partitioned into  $\text{part}(6, 3) = 7$  orbits. We get  $C = 80/81$  and we can compute the probability of each orbit

$\mathbf{y}$	$p(\mathbf{y})$	$\#\pi_{\mathbf{y}}$	$p(\pi_{\mathbf{y}})$
(6, 0, 0)	$80/(81 \cdot 6!0!0!)$	3	$3/729$
(5, 1, 0)	$80/(81 \cdot 5!1!0!)$	6	$36/729$
(4, 2, 0)	$80/(81 \cdot 4!2!0!)$	6	$90/729$
(3, 3, 0)	$80/(81 \cdot 3!3!0!)$	3	$60/729$
(3, 2, 1)	$80/(81 \cdot 3!2!1!)$	6	$360/729$
(4, 1, 1)	$80/(81 \cdot 4!1!1!)$	3	$90/729$
(2, 2, 2)	$80/(81 \cdot 2!2!2!)$	1	$90/729$

The partition of  $\mathcal{F}_i$  into permutation orbits looks somehow *optimal*, because we can approximate well the fiber with one orbit if its probability  $p(\pi_{\mathbf{y}})$  is large enough. This result is confirmed in Figure 1. If we select  $n_1 = 2$  and  $n_2 = 1$  and we compute the exact null cumulative distribution of  $W$  over  $\mathcal{F}_6$  and its approximation using the orbit  $\pi_{(1,2,3)}$  (which has the highest probability), we obtain two distributions which are considerably close, even if the cardinality of the selected orbit is low ( $\#\pi_{(1,2,3)} = 6$ ) compared to the the cardinality of  $\mathcal{F}_6$ , which is 28.

Table 1: Cumulative distribution of  $W$  on  $\mathcal{F}_6$  and  $\pi_{(1,2,3)}$

$w$	-6	-4.5	-3	-1.5	0	1.5	3
$\mathcal{F}_6$	0.001	0.018	0.100	0.320	0.649	0.912	1
$\pi_{(1,2,3)}$	0	0	0	0.333	0.667	1	1

## 5 Conclusion

This study can easily be extended to the non-negative discrete distributions of the exponential family. The convergence of the MCMC to the exact binomial and a mathematical statement on the *optimality* of the partition of the fiber into orbits of permutations are part of our ongoing research.

## References

1. Aoki, S., Hara, H., Takemura, A.: Markov Bases in Algebraic Statistics. Springer Series in Statistics. Springer New York (2012)
2. Aoki, S., Takemura, A.: Markov chain monte carlo tests for designed experiments. Journal of Statistical Planning and Inference **140**(3), 817 – 830 (2010)
3. Diaconis, P., Sturmfels, B.: Algebraic algorithms for sampling from conditional distributions. Ann. Statist. **26**(1), 363–397 (1998). DOI 10.1214/aos/1030563990
4. Fleiss, J.L., Levin, B., Paik, M.C.: Statistical methods for rates and proportions. John Wiley & Sons (2013)
5. Kunz, M.: Partitions and their lattices. ArXiv Mathematics e-prints (2006)
6. McCullagh, P., Nelder, J.: Generalized Linear Models, Second Edition. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis (1989)
7. Pesarin, F., Salmaso, L.: Permutation tests for complex data: theory, applications and software. John Wiley & Sons (2010)
8. Przyborowski, J., Wilenski, H.: Homogeneity of results in testing samples from poisson series: With an application to testing clover seed for dodder. Biometrika **31**(3/4), 313–323 (1940)
9. Robert, C., Casella, G.: Monte Carlo Statistical Methods. Springer Texts in Statistics. Springer New York (2013). URL <https://books.google.it/books?id=lrvfBwAAQBAJ>
10. 4ti2 team: 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces. Available at [www.4ti2.de](http://www.4ti2.de)



# Non-parametric micro Statistical Matching techniques: some developments

## *Tecniche micro non-parametriche per Statistical Matching: alcuni sviluppi*

Riccardo D'Alberto and Meri Raggi

**Abstract** Sometimes, the integration of different data sources is the only suitable solution to microdata shortage. Among the several data integration methodologies, Statistical Matching (SM) imputation allows to integrate different datasets when the same records are not uniquely identifiable through the observed variables and/or beyond a modelled rescaling procedure from an observed sample. Particularly, non-parametric micro SM imputation (“hot deck”) techniques allow researchers both to work always with observed (real) data and to avoid model misspecification bias. Nevertheless, non-parametric methods still lack a proper theoretical formalization and a sound methodology to evaluate the imputation quality. Therefore, we propose new combinations of distance functions and “hot deck” techniques, analysing how they perform in different donor-recipient datasets scenarios and elaborating a robust, recursive strategy for the imputation validation.

**Abstract** *L'integrazione di diverse fonti di dati può risultare a volte la sola soluzione percorribile alla mancanza di microdati. Tra le molteplici metodologie, l'imputazione tramite Statistical Matching (SM) permette di integrare dataset differenti quando per gli stessi record non sono disponibili variabili identificative e/o al di là di un modello di re-scaling del campione osservato. Nello specifico, le tecniche (“hot deck”) micro non-parametriche, permettono sia di lavorare sempre con dati osservati (reali) sia di evitare l'errore da misspecificazione del modello. La loro incompleta formalizzazione teorica e la mancanza di una strategia per la validazione della bontà dell'imputazione sono l'oggetto del presente lavoro. Infatti, proponiamo nuove combinazioni di tecniche “hot deck” e funzioni di distanza, analizzando le loro performance in differenti scenari ricevente-donatore ed elaborando una strategia per la validazione dell'imputazione.*

**Key words:** statistical matching, imputation, hot deck techniques

---

Riccardo D'Alberto, e-mail: [riccardo.dalberto@unibo.it](mailto:riccardo.dalberto@unibo.it)

Meri Raggi, e-mail: [meri.raggi@unibo.it](mailto:meri.raggi@unibo.it)

Department of Statistical Sciences “P. Fortunati” - Alma Mater Studiorum University of Bologna, Via Delle Belle Arti, 41 - 40126 Bologna (BO)

## 1 Introduction

Nowadays researchers are experiencing both a relevant increase of privately collected data sources (e.g. big data, ad hoc project surveys, etc.) and a simultaneous reduction of official administrative data sources. Nevertheless, despite the wide range of possibilities these data offer, researchers cannot always neglect microdata which are sometimes essential for the research. Their shortage is due to several issues; whole official administrative data sources are often unavailable for research purposes and/or hardly accessible whereas their release is ruled by strict procedures which reduce variables' informative power due to privacy claims restrictions. Moreover, the collection of the same amount of information through ad hoc surveys is extremely expensive and long lasting.

A reliable solution then, is to resort to data integration methodologies among which, Statistical Matching (SM) imputation techniques (both parametric and non-parametric) have gained a relevant attention in the most recent years. They allow different datasets integration when the same records are not identifiable through a unique observed variable and/or beyond a modelled rescaling procedure from an observed sample (as it is instead applying, respectively, Record Linkage and the Statistical Up/Downscaling methodologies).

SM techniques have been properly formalized through a rigorous theoretical framework by Rässler [4] and D'Orazio [2]. At the best of our knowledge, these two works constitute the most relevant and complete references for the whole state of the art in SM imputation. Nevertheless, if the parametric framework has been carefully studied and developed, non-parametric methods still lack a proper theoretical formalization and a sound methodology to evaluate the imputation quality.

In this paper, we focus on non-parametric micro SM imputation ("hot deck") techniques since they allow researchers both to work always with observed (real) data and to avoid model misspecification bias. Indeed, if the parametric approach requires the specification of the variables' distribution family and of an imputation model, "hot deck" techniques "allow researchers to handle the missing data issues by replacement" [3] from the most similar observed unit.

We explore new combinations of distance functions within the "hot deck" techniques matching algorithms, analysing how these combinations perform in different donor-recipient datasets scenarios and elaborating a robust, recursive strategy for the imputation validation. In Sect. 2 we describe the state of the art in SM imputation w.r.t. the non-parametric techniques and the proposed distance functions whereas, in Sect. 3, we propose our developments and the results achieved through a simulation study.

## 2 State of the art in SM imputation

There are four non-parametric micro SM imputation techniques, i.e. the Nearest Neighbour Distance Hot Deck (nnd), the Constrained Nearest Neighbour Hot Deck (nndc), the Random hot deck (rnd) and the Rank hot deck (rnk) [2].

For sake of simplicity, we define a basic imputation context of two different datasets, the recipient (R) and the donor (D) ones. Let be  $i$  and  $j$  two different units with  $i = 1, \dots, n_R$  and  $j = 1, \dots, n_D$ . Defining  $\mathbf{X} = \{X_l, l = 1, \dots, L\}$  the set of common variables between R and D, we have that  $\mathbf{X}_l^R$  is a vector of dimension  $(n_R \times 1)$  and  $\mathbf{X}_l^D$  is a vector of dimension  $(n_D \times 1)$ .

Assuming that  $L = 1$  so that  $X$  is a single (continuous) variable, defining  $i$  the recipient unit in R and  $j^*$  the donor unit in D chosen to be matched (i.e. to constitute a matching unit pair) among all the units  $j$ , from [2], we know that the nnd technique associates units pairs in the way that the equation  $d_{ij^*} = \left| x_i^R - x_{j^*}^D \right| = \min_{j=1, \dots, n_D} \left| x_i^R - x_j^D \right|$  holds, where  $d$  is the difference in absolute value between the two units  $i$  and  $j$  ( $j^*$ ), always computed such that  $1 \leq j \leq n_D$ .

This technique can also be sharpened in the nndc technique by imposing a constraint by minimizing the function  $\sum_{i=1}^{n_R} \sum_{j=1}^{n_D} (d_{ij} \omega_{ij})$ , where  $\omega_{ij} \in \{0, 1\}$  represents the matching unit pair of  $i$  and  $j$  such that  $\omega_{ij}$  is equal to 0 if they are matched, equal to 1 otherwise. Two conditions have to hold in order to use just once each donor unit  $j$  in the setting up of a matching unit pair with a recipient unit  $i$ , i.e.:  $\sum_{j=1}^{n_D} \omega_{ij} = 1$  and  $\sum_{i=1}^{n_R} \omega_{ij} \leq 1$ .

The rnd technique constitutes a matching unit pair by picking at random the donor units. It can be sharpened in several ways but, for sake of brevity, we present only the easiest one, i.e. by building donation classes. Indeed, if the usual possible set of donor and recipient units pairs is defined by  $n_D^{n_R}$ , it is possible to define within the chosen matching variables some homogeneous subsets. Let be  $X_1$  and  $X_2$  two existing common variables between R and D upon which we can constitute a donation class, the possible set of units pairs is reduced to  $(n_{X_1}^D)^{n_{X_1}^R} + (n_{X_2}^D)^{n_{X_2}^R}$ .

We stress that it is possible to build donation classes also w.r.t. the nnd and nndc techniques, improving the imputation quality but decreasing the computational speed.

Finally, the rnk technique works in two recursive steps. Firstly, it ranks recipient and donor units w.r.t. their empirical cumulative distribution functions  $F_{X^R}(x^R) = \frac{1}{n_R} \sum_{i=1}^{n_R} I(x_i \leq x)$  and  $F_{X^D}(x^D) = \frac{1}{n_D} \sum_{j=1}^{n_D} I(x_j \leq x)$ , being  $I$  the set of indices of  $x_i \leq x$  and  $x_j \leq x$ , respectively.

Secondly, rnk associates to each recipient unit a donor unit in the way that the following equation  $\left| F_{X^R}(x_i^R) - F_{X^D}(x_{j^*}^D) \right| = \min_{j=1, \dots, n_D} \left| F_{X^R}(x_i^R) - F_{X^D}(x_j^D) \right|$  holds, where the minimum of the distance between  $F_{X^R}(x^R)$  and  $F_{X^D}(x^D)$  is computed such as  $1 \leq j \leq n_D$ .

At the best of our knowledge, nnd, nndc and rnd techniques work by applying to their matching algorithms a default distance function (the Manhattan one) whereas not so much is known both about their performances in different recipient-donor

datasets scenarios and w.r.t. the quality of the imputation. We stress that what is mainly known, is derived from the parametric framework as a sort of “prescription” we summarize in the following sentences: *i.* being equal the dimensionality ratio between R and D, their variability is crucial, i.e. when the variance of the matching variable(s) in R is lower than the variance of the matching variable(s) in D, this condition is always preferable; *ii.* if, instead, the variance of the matching variable(s) in R is higher than the variance of the matching variable(s) in D, the condition of the widest dimensionality ratio between R and D is always preferable; *iii.* being different the dimensionality ratio between R and D, the key “assumption” is “the biggest, the best”, i.e. the choice of the recipient and the donor datasets has always to respect the condition  $n_R < n_D$ ; *iv.* donation classes always benefit the imputation quality.

Studying new combinations of “hot deck” techniques and the Manhattan (mn), Mahalanobis (ms) and Exact (e) distance functions (for all the details we refer to [1]) we formalize and validate the above-mentioned expectations, also proposing an imputation validation strategy.

### 3 Our proposal: a simulation study

The simulation study is based on two steps, a previous R and D simulation and a consequent imputation. R and D are characterised in several scenarios w.r.t. the different dimensionality ratio, the different variability of the matching variables and the SM imputation running both with and without donation classes.

For both R and D we simulated two sets of common variables:  $\mathbf{X}^A = \{X_1^A, X_2^A, X_3^A\}$  (used as matching variables) and  $\mathbf{X}^B = \{X_1^B, X_2^B\}$  (used as imputation variables). These latter are simulated as the realization of a log-Normal( $\mu$ ,  $\sigma^2$ ) multiplied for a Bernoulli( $\theta$ ).  $X_1^A$  is simulated as the realization of a Bernoulli( $\theta$ );  $X_2^A$  is a categorical variable indicating the main variable value between  $X_1^B$  and  $X_2^B$ ;  $X_3^A$  is simulated as the sum of the values of the variables  $X_1^B$  and  $X_2^B$ . The simulation scheme (for more details we refer to [1]) is shown in Table 1.

**Table 1** Simulation study and imputation scheme

Simulation Nr.	1		2		3		4	
Ratio	1 to 10		1 to 10		1 to 3		1 to 3	
Variability	$\text{var}(R) > \text{var}(D)$		$\text{var}(R) < \text{var}(D)$		$\text{var}(R) > \text{var}(D)$		$\text{var}(R) < \text{var}(D)$	
Imputation Nr.	1	2	3	4	5	6	7	8
Donation classes	with	without	with	without	with	without	with	without

We propose an imputation quality validation strategy upon three tools: *i.* the distributions checking of both the variables originally present in R and imputed from D; *ii.* the distributions checking of the differences between the values of the original variables in R and the values of the imputed variables from D (defining these differences “z” variables); *iii.* the valuation of the MSE of the “z” variables.

We refer to [1] for the detailed description of the whole simulation results referred to each R-D scenario, applying each one of the proposed tool. Here, we discuss the main simulation results only w.r.t. the valuation of the MSE of the “z” variables.

Firstly, a wider dimensionality ratio between the R and D is determinant when the variance of the matching variables in R is higher than the variance of the matching variables in D, as Table 2 shows.

**Table 2** MSE values of differences z (imputations 1, 2, 5, 6)

	donation classes				no donation classes			
	1 to 10		1 to 3		1 to 10		1 to 3	
	var(R) > var(D)				var(R) > var(D)			
	Imputation 1		Imputation 5		Imputation 2		Imputation 6	
	$\bar{z}_{X_1^B}$	$\bar{z}_{X_2^B}$	$\bar{z}_{X_1^B}$	$\bar{z}_{X_2^B}$	$\bar{z}_{X_1^B}$	$\bar{z}_{X_2^B}$	$\bar{z}_{X_1^B}$	$\bar{z}_{X_2^B}$
nnd.mn	101.536	9.617	102.534	10.017	176.171	83.896	182.890	90.273
nnd.ms	101.536	9.617	102.534	10.017	176.171	83.896	182.890	90.273
nnd.e	1,972.411	136.508	2,113.379	121.772	1,850.420	180.590	2,047.865	187.587
nndc.mn	101.527	9.608	102.679	10.293	175.903	83.628	183.459	90.858
nndc.ms	101.526	9.606	102.815	10.368	176.010	83.734	183.573	90.964
nndc.e	2,688.750	139.780	2,728.813	131.305	108.465	14.920	108.465	14.920
rnd.mn	1,000.011	15.570	1,186.610	19.674	1,253.199	85.351	1,192.059	73.047
rnd.ms	1,005.479	17.575	1,121.168	16.839	1,257.923	90.165	1,465.474	105.852
rnd.e	1,794.635	127.224	1,756.882	137.068	1,798.596	182.784	1,883.323	164.871
rnk	165.375	45.464	133.446	23.293	281.824	167.775	203.317	99.555

Secondly, being the dimensionality ratio between R and D equal, the lower variance of the matching variables in R w.r.t. the variance of the matching variables in D, is always determinant, as Table 3 shows.

**Table 3** MSE values of differences z (imputations 1, 2, 3, 4)

	donation classes				no donation classes			
	1 to 10				1 to 10			
	var(R) > var(D)		var(R) < var(D)		var(R) > var(D)		var(R) < var(D)	
	Imputation 1		Imputation 3		Imputation 2		Imputation 4	
	$\bar{z}_{X_1^B}$	$\bar{z}_{X_2^B}$	$\bar{z}_{X_1^B}$	$\bar{z}_{X_2^B}$	$\bar{z}_{X_1^B}$	$\bar{z}_{X_2^B}$	$\bar{z}_{X_1^B}$	$\bar{z}_{X_2^B}$
nnd.mn	101.536	9.617	9.532	9.528	176.171	83.896	77.918	77.904
nnd.ms	101.536	9.617	9.532	9.528	176.171	83.896	77.918	77.904
nnd.e	1,972.411	136.508	444.579	157.936	1,850.420	180.590	786.865	208.549
nndc.mn	101.527	9.608	9.466	9.465	175.903	83.628	84.813	84.770
nndc.ms	101.526	9.606	9.494	9.492	176.010	83.734	84.515	84.474
nndc.e	2,688.750	139.780	343.698	163.905	108.465	14.920	46.965	37.842
rnd.mn	1,000.011	15.570	8.273	7.295	1,253.199	85.351	78.321	81.351
rnd.ms	1,005.479	17.575	9.421	9.767	1,257.923	90.165	92.751	88.203
rnd.e	1,794.635	127.224	407.317	94.668	1,798.596	182.784	583.777	121.647
rnk	165.375	45.464	2,943.404	98.975	281.824	167.775	2,963.817	160.906

Thirdly, we found evidence that a narrower dimensionality ratio between R and D, being the variance of the matching variables in R lower than the variance of the matching variables in D, can produce the best imputation results if the matching variables in D have a proper variability, as Table 4 shows. In other words, oppositely to “the biggest, the best” common prescription, the bond between R and D can be relaxed if the variance of the matching variable(s) in R is lower than the variance of the matching variable(s) in D, and if the variance of the matching variable(s) in the smaller of the two donor datasets is the widest one.

**Table 4** MSE values of differences  $z$  (imputations 3, 4, 7, 8)

	donation classes				no donation classes			
	1 to 10		1 to 3		1 to 10		1 to 3	
	$\text{var}(R) < \text{var}(D)$				$\text{var}(R) < \text{var}(D)$			
	Imputation 3		Imputation 7		Imputation 4		Imputation 8	
	$z_{X_1^B}$	$z_{X_2^B}$	$z_{X_1^B}$	$z_{X_2^B}$	$z_{X_1^B}$	$z_{X_2^B}$	$z_{X_1^B}$	$z_{X_2^B}$
nnd.mn	9.532	9.528	7.872	7.945	77.918	77.904	87.838	88.045
nnd.ms	9.532	9.528	7.872	7.945	77.918	77.904	87.838	88.045
nnd.e	444.579	157.936	477.174	158.138	786.865	208.549	666.437	205.484
nndc.mn	9.466	9.465	7.867	7.976	84.813	84.770	95.708	95.738
nndc.ms	9.494	9.492	7.913	8.022	84.515	84.474	77.219	77.183
nndc.e	343.698	163.905	420.386	169.801	46.965	37.842	46.965	37.842
rnd.mn	8.273	7.295	12.321	16.484	78.321	81.351	104.761	99.260
rnd.ms	9.421	9.767	9.950	16.915	92.751	88.203	85.926	87.745
rnd.e	407.317	94.668	573.707	106.418	583.777	121.647	334.443	76.499
rnk	2,943.404	98.975	2,834.001	86.592	2,963.817	160.906	2,953.937	143.025

Therefore, all the “prescriptions” from the literature were tested and validated with the remarkable exception of the “the biggest, the best” one. Rather than it is commonly thought and prescribed, this condition is not mandatory and can be relaxed whenever either the variance of the matching variable(s) in R is lower than the variance of the matching variable(s) in D or, comparing two potential donor datasets, the variance of the matching variable(s) in the smaller of the two ones, is the widest. Further developments are already under studying, w.r.t. both a deepest theoretical formalization of the proposed combinations and in order to structure and elaborate more the imputation quality validation strategy.

## References

1. D'Alberto, R.: Statistical Matching Imputation among different farm data sources, [Dissertation thesis] Alma Mater Studiorum Università degli Studi di Bologna, (2017)
2. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical matching: Theory and practice. John Wiley & Sons, Chichester (2006)
3. Little, R., Rubin, D.: Statistical analysis with missing data. John Wiley & Sons, Hoboken (2002)
4. Rässler, S.: Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches. Springer, New York (2002)

# Measuring tourism from demand side

## *La misura del turismo dal lato della domanda*

Stefano De Cantis, Mauro Ferrante, Anna Maria Parroco

**Abstract** This Paper proposes an analysis of tourism from the demand side, taking into account for both the total level of tourism demand produced by some European countries (domestic and outgoing) and its general tendency, and for the seasonal fluctuations which characterize many tourism-related aggregates. Tourist flows from the demand side at the European level are analyzed in the last decade, and a special focus on Italian tourism demand is provided, jointly with an analysis of its seasonal fluctuations. The analysis of general tendency of tourism demand and of the impacts of seasonality is a fundamental pre-requisite for the implementation of tourism policies.

**Abstract** *Il presente articolo propone un'analisi dei flussi turistici dal lato della domanda, prendendo in esame sia il livello complessivo e la tendenza generale della domanda turistica, che il suo andamento stagionale. A tal fine vengono analizzati dati sulla domanda turistica a livello Europeo nell'ultimo decennio e viene proposto un focus specifico sulla domanda turistica in Italia, ponendo particolare enfasi sull'andamento stagionale. Una maggiore comprensione delle dinamiche della domanda turistica nonché degli impatti della stagionalità rappresentano un pre-requisito essenziale per l'implementazione delle politiche turistiche.*

**Key words:** Tourism Statistics, Tourist behavior, Seasonal pattern, Seasonal amplitude

---

Stefano De Cantis

Dipartimento di Scienze Economiche Aziendali e Statistiche, University of Palermo, Viale delle Scienze, ed.13 90128, Palermo, e-mail: stefano.decantis@unipa.it

Mauro Ferrante

Dipartimento di Culture e Società, University of Palermo, Viale delle Scienze, ed.15 90128, Palermo, e-mail: mauro.ferrante@unipa.it

Anna Maria Parroco

Dipartimento di Psicologia, University of Palermo, Viale delle Scienze, ed.15 90128, Palermo, e-mail: annamaria.parroco@unipa.it

## 1 Introduction

Country-specific measurement and the analysis of tourism activity of residents (both in terms of physical and economic volumes) allow us to describe the determinants of consumer behavior, to investigate motivations regarding the choices of making or not making tourism, and finally to estimate new tendencies in tourism related behaviors. In order to describe the main features of tourism demand at the country level, it is important to consider: a) the total level of tourism demand produced by each country (domestic and outgoing); b) the general tendency which, in the medium run, characterizes the demand; c) the seasonal fluctuations of tourism demand. In particular, seasonality plays an important role, since the same level of tourism demand could determine very different impacts according to its distribution over time. Starting from these premises, the present work aims at analyzing tourist flows from the demand side at the European level. In the European context, the European Regulation (EU) No 692/2011 [3] concerning European statistics on tourism aims at establishing a common framework in the European Union, concerning the collection of statistical information on tourism. The Regulation states that data to be transmitted by the Member States concerns: a) the participation in tourism and the characteristics of tourism trips and visitors, and b) the characteristics of same-day visits. Collected and integrated data is provided by Eurostat in the tourism section of the database available at Eurostat's website. Moreover, a special focus is given for the Italian case; thanks to the availability of micro-data on trips made by residents, the seasonal component of tourism demand is analyzed in detail and some synthetic measures of seasonal concentration are provided.

## 2 Analysis of tourism demand in Europe

In the European context, the level of tourism demand of residents in the European Union is largely determined by a relatively limited number of countries. Germany, France, UK, Italy, and Spain represent a relevant share of the EU's population, and, beyond Scandinavian countries, also have among the highest rates of tourism propensity. Germany presents the highest tourism demand with more than 300 millions trips made by residents in 2011, followed by France, the UK, Italy and Spain.

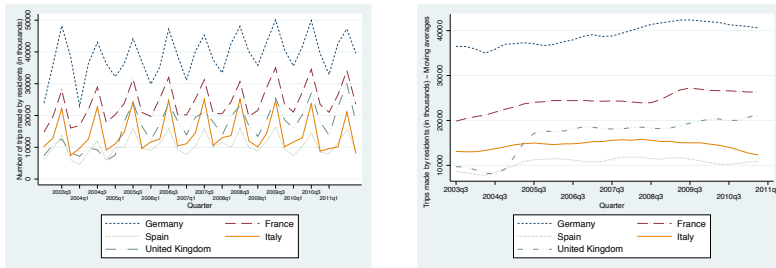
In order to offer a wider perspective of tourism demand at the European Level, in Tab. 2, Net Travel Propensity Index is reported for a set of European Countries for 2011, along with their number of trips and resident population values. Scandinavian countries (Norway, Sweden, and Finland) are those which present the highest Net Travel Propensity, but also Germany, with a value of about 60% in the third quarter, demonstrate a high travel propensity. On the contrary, countries such as Bulgaria, Poland, Portugal, and Romania are those in which the highest value of net travel propensity index does not exceed 25%. Moreover, although almost similar in terms of population, the five selected countries (Germany, France, UK, Italy, and Spain) are very different in terms of travel propensity (Tab. 1). First, European Countries



**Table 1** Quarterly Net Travel Propensity Index for selected European countries (*data in thousands*), 2011

Country	Population	Quarterly number of resident travelers				Net Travel Propensity Index			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Bulgaria	7,369	369	668	1,170	624	5.0%	9.1%	15.9%	8.5%
Czech Republic	10,487	3,627	4,487	4,596	2,802	34.6%	42.8%	43.8%	26.7%
Denmark	5,561	2,560	2,736	3,181	2,726	46.0%	49.2%	57.2%	49.0%
Germany	81,752	32,897	42,750	47,208	39,569	40.2%	52.3%	57.7%	48.4%
Greece	11,123	1,019	1,824	3,253	1,287	9.2%	16.4%	29.2%	11.6%
Spain	46,667	7,902	11,478	15,624	9,035	16.9%	24.6%	33.5%	19.4%
France	64,979	21,012	26,550	34,169	23,560	32.3%	40.9%	52.6%	36.3%
Croatia	4,290	755	1,013	1,521	955	17.6%	23.6%	35.5%	22.3%
Italy	59,365	9,575	10,140	21,095	8,113	16.1%	17.1%	35.5%	13.7%
Latvia	2,075	317	484	623	353	15.3%	23.3%	30.0%	17.0%
Lithuania	3,053	573	911	1,210	761	18.8%	29.8%	39.6%	24.9%
Luxembourg	512	214	267	318	228	41.9%	52.2%	62.1%	44.5%
Hungary	9,986	1,584	2,067	2,787	1,886	15.9%	20.7%	27.9%	18.9%
Malta	415	72	78	117	56	17.4%	18.8%	28.3%	13.5%
Austria	8,375	2,528	3,342	4,652	2,854	30.2%	39.9%	55.5%	34.1%
Poland	38,530	4,610	5,830	9,560	4,807	12.0%	15.1%	24.8%	12.5%
Portugal	10,573	1,017	1,423	2,508	1,641	9.6%	13.5%	23.7%	15.5%
Romania	20,199	2,123	2,918	3,348	3,212	10.5%	14.4%	16.6%	15.9%
Slovenia	2,050	447	650	1,079	476	21.8%	31.7%	52.6%	23.2%
Slovakia	5,392	1,374	1,450	2,375	1,423	25.5%	26.9%	44.0%	26.4%
Finland	5,375	3,026	3,202	3,583	3,090	56.3%	59.6%	66.7%	57.5%
Sweden	9,416	5,411	6,360	6,871	5,883	57.5%	67.6%	73.0%	62.5%
United Kingdom	63,023	13,791	22,868	29,910	18,152	21.9%	36.3%	47.5%	28.8%
Norway	4,920	2,672	3,016	3,257	2,562	54.3%	61.3%	66.2%	52.1%

exhibit different levels of tourism demand, both in absolute and in relative terms. The causes of these differences in tourism propensity should be more deeply investigated, and can be related to economic and sociocultural differences, and with the country specific uses and habits related to tourist behaviors, which only partially have been discussed in academic literature [1, 2]. Second, the way in which tourism demand by residents is distributed during the year is also very different from one country to another with different seasonal variations both in terms of pattern and amplitude [4].



**Fig. 1** Quarterly number of trips made by Residents in top five travel generating European countries, actual and moving averages series (data in thousands), 2003-2011.

### 3 Tourism demand of Italian residents from Istat 'Trips and holidays' survey

Istat, until 2014, annually presented the estimates of the main aggregates of tourism demand in Italy, based on CATI survey 'Trips and Holidays' which had been conducted on a quarterly basis from 1997 until 2013. Starting from 2014 a new survey on 'Consumption of Italian Families' has been introduced, which replaces and updates the survey on 'Trips and Holidays'. In Tab. 2, annual data on trips (by purpose and destination) made by Italian residents are reported, from 2005 to 2013. Also the number of nights spent, average length of trip, and the gross travel propensity index are reported. This data highlights the decrease in the number of trips and nights registered in the last years: from about 123 million trips in 2008 to less than 65 million in 2013, with a loss of about 60 million trips (-48.6%). Also in terms of nights, a considerable decrease can be observed: from more than 706 million nights in 2008 to about 417 million in 2013, with a loss of about 289 million nights (-41.0%). In other words, if in 2008 there were about 210 trips per 100 residents, in 2013 there were only 106. From this preliminary analysis, the strong effect of the economic crisis is evident on tourism, representing a relatively less-investigated phenomenon.

In order to isolate the seasonal component from other potential sources of variability in the series several methods could be used. For our application, we implemented the TRAMO-SEATS procedure, which allows seasonal factors to be derived. A summary of SARIMA models estimated for each series is reported in Tab. 3. Additive seasonal factors were produced for the series related to propensity indices related to both holiday and business trips in Italy, whereas multiplicative seasonal factors were derived for the series related to holiday and business trips abroad.

Once seasonal factors are derived, the cycle plot is a useful tool to synthesize the seasonal behavior of the series over all the considered periods. The series of seasonal factors related to travel propensity in Italy for holiday purposes seems to be one that presents a very high degree of stability of pattern of seasonality. Both the series related to holiday trips, in Italy and abroad, present almost the same pattern with a peak in August, and values of seasonal factors above the trend-cycle component only in summer months (from June to September). Several measures can be used in order to summarize the amplitude of seasonal fluctuations, some of which are reported in Tab. 4. All the amplitude measures indicate a relative stability of seasonal amplitude with a slight decrease of inequality in the last considered years.

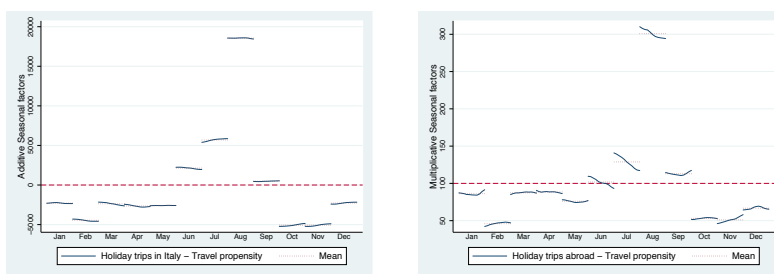
Tab. 4 reveals that the overall level of seasonality in each series is relatively stable during all the years considered, with a slight decrease in seasonal amplitude for the series related to holiday trips in Italy, holiday trips abroad, and business trips in Italy, and a strong irregular behavior for the series related to holiday trips abroad.

**Table 2** Number of trips, nights and related index of trips characteristics made by Italian residents in Italy and abroad, 2005-2013

Variable	2005	2006	2007	2008	2009	2010	2011	2012	2013
<i>Data in thousands</i>									
Holiday trips in Italy	77,860	78,606	80,972	90,463	82,265	71,926	59,807	54,733	46,062
Holiday trips abroad	14,268	15,284	16,201	16,347	16,412	15,524	12,751	13,967	11,389
Total trips ( <i>all purposes</i> )	107,100	107,895	112,369	122,938	114,099	100,040	83,417	78,703	63,154
Nights spent in Italy ( <i>holiday purposes</i> )	493,775	534,672	480,724	508,722	497,230	467,796	393,015	357,772	300,709
Nights spent abroad ( <i>holiday purposes</i> )	123,003	133,119	146,267	135,374	125,351	118,250	101,757	113,828	100,946
Total nights ( <i>all purposes</i> )	676,243	719,763	689,313	706,650	680,215	626,990	527,811	501,059	417,126
Population (1 <sup>st</sup> Jan)	57,875	58,064	58,224	58,653	59,001	59,190	59,365	59,394	59,685
Avg. duration of trips in Italy	6.34	6.80	5.94	5.62	6.04	6.50	6.57	6.54	6.53
Avg. duration of trips abroad	8.62	8.71	9.03	8.28	7.64	7.62	7.98	8.15	8.86
Avg. duration of trips	6.31	6.67	6.13	5.75	5.96	6.27	6.33	6.37	6.60
Gross travel propensity index in Italy ( <i>holiday purposes</i> )	1.35	1.35	1.39	1.54	1.39	1.22	1.01	0.92	0.77
Gross travel propensity index abroad ( <i>holiday purposes</i> )	0.25	0.26	0.28	0.28	0.28	0.26	0.21	0.24	0.19
Travel propensity index ( <i>all purposes</i> )	1.85	1.86	1.93	2.10	1.93	1.69	1.41	1.33	1.06

**Table 3** Summary of TRAMO-SEATS procedure on monthly travel propensity index series in Italy, by purposes and destination. Years 2002-2013

	Holiday trips in Italy	Business trips in Italy	Holiday trips abroad	Business trips abroad
Transformation	none	none	Log-transformation	Log-transformation
SARIMA Model	(0,1,1)(0,1,1)	(1,0,0)(0,1,1)	(0,1,3)(0,1,1)	(1,1,1)(1,0,0)
Parameter estimates	$\theta_1 = -0.7240$ $\theta_{12} = -0.8283$	$\phi_1 = -0.3629$ $\theta_{12} = -0.8435$	$\theta_1 = -0.8648$ $\theta_2 = 0.1459$ $\theta_3 = -0.1378$ $\theta_{12} = -0.7447$	$\phi_1 = -0.4643$ $\theta_1 = -0.9900$ $\phi_{12} = -0.4883$



**Fig. 2** Cycle plots of seasonal factors for travel propensity index, holiday trips in Italy and abroad, 2002-2013.

**Table 4** Measures of amplitude derived from seasonal factors of Travel Propensity Index, by purpose and destination of trips, Italian residents 2002-2013

Year	Holiday trips in Italy	Business trips in Italy	Holiday trips abroad		Business trips abroad	
	Seasonal Range	Seasonal Range	Coefficient of Seasonal Variation	Gini Index	Coefficient of Seasonal Variation	Gini Index
2002	23,787.7	1,269.9	0.68	0.12	0.21	0.34
2003	23,792.1	1,268.3	0.67	0.11	0.20	0.33
2004	23,773.5	1,262.2	0.67	0.11	0.19	0.33
2005	23,741.6	1,249.6	0.67	0.16	0.26	0.32
2006	23,725.9	1,233.9	0.66	0.20	0.32	0.32
2007	23,699.0	1,221.4	0.65	0.19	0.33	0.31
2008	23,643.4	1,207.0	0.64	0.13	0.20	0.31
2009	23,584.8	1,187.5	0.64	0.10	0.17	0.30
2010	23,495.8	1,176.3	0.64	0.15	0.25	0.30
2011	23,397.0	1,178.0	0.64	0.35	0.61	0.30
2012	23,335.6	1,181.2	0.63	0.20	0.32	0.30
2013	20,708.7	958.8	0.71	0.30	0.38	0.24

## 4 Conclusion

European residents have very different tourism behaviour in terms of overall travel propensity and very different seasonal travel demands. Thanks to microdata obtained from the survey on tourism made by Italian residents, a detailed analysis of seasonality has been possible. Despite of the relevant decrease in tourism trips following the recent economic crisis, a strong and persistent seasonal behaviour characterizes the pattern of tourism trips in Italy. In social phenomena it is common to observe changes in human behaviours and attitudes, but, quite surprisingly, this is not the case. From the methodological perspective, this work used the Gini index, as well as other indices for synthesizing the seasonal burden, which, as pointed out by De Cantis et al. [4], do not take into account for the natural ordering of months. Subsequently, a deep analysis on the causes that determine the persistent seasonal behaviour, as well as critical reflections on the appropriate measurement of seasonal fluctuations merits further investigations from a variety of perspectives.

## References

1. Alegre J, Mateo S, Pou L (2009) Participation in tourism consumption and the intensity of participation: an analysis of their socio-demographic and economic determinants. *Tourism Economics*, 15(3): 531–546.
2. Bernini C, Cracolici MF (2015) Demographic change, tourism expenditure and life cycle behaviour. *Tourism Management*, 47: 191–205.
3. European Parliament (2011) Regulation (EU) No 692/2011. *Official Journal of the European Union*, L192(54), 17–32.
4. De Cantis S, Ferrante M, Vaccina F (2011) Seasonal Pattern and amplitude – a logical framework to analyse seasonality in tourism: an application to bed occupancy in Sicilian hotels. *Tourism Economics*, 17(3): 655–675.

# Optimal Ethical Balance for Phase III Trials Planning

## *Bilancio Etico Ottimale nella Pianificazione della Fase III delle Prove Cliniche*

Lucio De Capitani and Daniele De Martini

**Abstract** The need of an ethical evaluation is mandatory for every clinical trial, Ethics Committees have been built for that. The distinction between individual and collective ethics has been introduced in a seminal work by Lellouch and Schwartz in 1971, where individual ethics regard concerns related to the patients enrolled in the trial, and collective ethics those of the patients not enrolled who would benefit of a positive trial result. In this paper, a metrization of individual and collective ethics is proposed in order to evaluate their balance in a confirmatory clinical trial. The ethical balance evaluation, among these two aspects of ethics, can be performed before trial starting in order to address sample size determination. The metrization is based, among other parameters, on the drug effect size, on the quality of life of patients under therapy or placebo, and of that induced by adverse reactions. Some numerical examples show that the optimal ethical balance can be provided by sample sizes far from those computed by adopting the usual paradigm based on the prefixed power of 80-90%.

**Abstract** *La valutazione etica nelle prove cliniche è assolutamente necessaria, e infatti ogni studio clinico viene sottoposto ad adeguato comitato etico. La distinzione tra etica individuale e collettiva è stata introdotta in un lavoro pionieristico da Lellouch e Schwartz nel 1971, in cui i possibili danni subiti dai pazienti coinvolti nello studio vengono valutati dall'etica individuale, mentre l'etica collettiva considera quelli dei pazienti non coinvolti, che beneficerebbero di un positivo risultato dello studio clinico. In questo lavoro, si propone una quantificazione dell'etica individuale e collettiva al fine di valutare il bilancio etico in uno studio clinico confermativo. La valutazione del bilancio tra questi due aspetti di etica può essere eseguita prima dell'inizio dello studio clinico al fine di calcolare la dimensione campionaria*

---

Lucio De Capitani

Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca, via Bicocca degli Arcimboldi 8, 20126 Milano, e-mail: lucio.decapitani1@unimib.it

Daniele De Martini

Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca, via Bicocca degli Arcimboldi 8, 20126 Milano, e-mail: daniele.demartini@unimib.it

*ottimale. La proposta quantificazione dell'etica dipende dall'entità dell'effetto del farmaco, dalla qualità della vita dei pazienti in terapia o placebo, e da quella indotta dai possibili effetti indesiderati. Attraverso degli esempi numerici si mostra che il bilancio etico ottimale pu essere raggiunto in corrispondenza di dimensioni del campione lontane da quelle calcolate adottando il paradigma usuale, in cui la numerosità campionaria viene scelta al fine di assicurare il raggiungimento di valori di potenza del test pari all'80-90%.*

**Key words:** Individual Ethics, Collective Ethics, Global Ethics, Assurance

## 1 Introduction

It is a common habit to prefix the power of phase III clinical trials at 80-90%. These power settings are suggested by the most relevant books on clinical trials methodology (see, for example, [8] or [2]). Unfortunately, the practical choice of power is seldom motivated as to concern its clinical, and therefore ethical, impact.

In fact, on the one hand 80% looks high enough to guarantee that if the new treatment is effective the trial will succeed - and the drug will be available for the ill population, with high probability. On the other hand, the power is often set not higher than 90% in order to minimize, for the enrolled sample of patients, risks and wastings due to potential harm and lack/or of efficacy of the new treatment. In other words, the power threshold of 80-90% seems accomplishing the need to affect on medical practices, being the drug effective, and that of preserve the safety of enrolled patients.

These two concerns remind to the concept of individual and collective ethics, introduced by [7]. Originally, collective ethics (CE) concerned maximizing total group benefit, and individual ethics (IE) concerned maximizing the benefit of each person to be treated.

However, it is not clear when adopting either 80% or 90%, and why. FDA and [12] encourage to set the power at 90%, but not for explicit ethical reasons: they argue to adopt this power to decrease the rate of unsuccessful trials. In fact, in the literature there are not precise indications about which power threshold should be adopted in different ethical situations.

In the seminal paper [7] the concepts of IE and CE were introduced, together with some mathematical formulations of them under both fixed and sequential designs. [6] provide a critical history of IE and CE, and remark that “very little follow-up research in the lineage of the 1971 paper considers mathematical models”.

The aim of this work is that of providing a model to quantify the ethical balance in fixed designs, which are the most widely adopted in phase II and phase III trials (see [www.clinicaltrials.gov](http://www.clinicaltrials.gov)). Our perspective is in agreement with the view on the ethic of the trial proposed in [11], which “is dictated by the type of evidence sought and by balancing various costs of aggregate harm and benefit” (see [6]).

In [4] it is recalled that Nuremberg Codes point 2 offers strong support of the existing connection between power and ethics, and they add: “the underlying principle is that for any given outlay of human risk or resources, there is an obligation to maximize the power and efficiency of the experimental design”. Here, we invert the point of view, by modeling ethics as a function of the sample size, and so of the statistical power. Then, we suggest to adopt the sample size that maximizes experimental ethics.

## 2 Theoretical framework

A two-arm parallel design with balanced sampling is considered for the phase III trial. A sample of size  $n$  is collected for each arm (i.e. new drug and standard treatment/placebo). The true, and unknown, proportions of healing are  $p_t$  and  $p_c$ . The statistical hypotheses are  $H_0 : p_t = p_c$  vs  $H_1 : p_t > p_c$ , and  $\alpha$  is the type I error probability. We assume that the proportions represent the responder rate under the two arms.

Given that  $\hat{p}_{t,n}$  and  $\hat{p}_{c,n}$  are the sample proportions, the test statistic is  $T_n = (\hat{p}_{t,n} - \hat{p}_{c,n}) / \sqrt{(\hat{p}_{t,n}(1 - \hat{p}_{t,n}) + \hat{p}_{c,n}(1 - \hat{p}_{c,n})) / 2}$ . The power function  $\pi(n) = P(T_n > z_{1-\alpha})$  is approximated by  $\Phi(ES\sqrt{n/2} - z_{1-\alpha})$ , where ES is the standardized effect size:  $ES = (p_t - p_c) / \sqrt{(p_t(1 - p_t) + p_c(1 - p_c)) / 2}$ .

## 3 Modeling ethics

Ethics is the sum of individual and collective ethical contributions. Usually, individual ethics concern the sample of patients involved in the experiment, where collective ethics consider the remaining population.

Basically, ethics are computed through Benefit/risk indicators times the duration of periods of interest. Quality of Life measures (QoL) are adopted together with the probabilities of responders and of harms, which are related to the “Number Needed to Treat” (NNT) and “Number Needed to Harm” (NNH), all being classical benefit/risk indexes (see, for example, [5]). In particular, the ethical contribution of a group is given by the group size times the quality of life indicator times the duration of the period where such a QoL persists. Since the new treatment is available to the population if the trial succeeds, collective ethics are also multiplied by the power of the experiment. Thus, in general, Ethics are:

$$E = \text{group size} \times \text{probability} \times \text{quality of life} \times \text{duration}$$

Individual ethics ( $IE$ ) concerns ethics of the population sample enrolled in the trial, and considers the placebo group just during the trial.  $IE$  depends on: the size of each sample ( $n$ ), the rate of responder under new therapy ( $p_t$ ), the harm probability under new therapy ( $hp_t$ ), the rate of responder under control treatment ( $p_c$ ), the quality of life during the disease and before treatments ( $QoL_d$ ), the benefit (quality of life)

after the new treatment ( $QoL_t$ ), the harm (risk) during the new treatment ( $QoL_r$ ), the benefit after control treatment ( $QoL_c$ ), life expectancy ( $D_L$ ), duration of the therapy ( $D_{th}$ ) and accrual rate ( $A_r$ , which is assumed to be uniform during enrollment). The duration of phase III trial is:  $D_{P3} = 2n/A_r + D_{th}$ .

Under the new treatment, the ethical contribution of (eventual) QoL improvement of responders and non responders is the sum of the two, resulting:

$$IE_T(n) = n \times \left( p_t \times (QoL_t \times (D_L - (D_{P3}(n) - D_{th}))/2) + \right. \\ \left. QoL_d \times (D_{P3}(n) - D_{th})/2 + (1 - p_t) \times QoL_d \times D_L \right) .$$

Note that the duration of the benefit of responders is given by life expectancy minus the average of the time elapsed from the beginning of the trial and the end of the therapy. For non responders, the quality of life remains that of the disease during all life.

The ethical contribution of harm due to the new drug is:

$$IE_{TH}(n) = n \times h p_t \times QoL_r \times D_{th} .$$

Under the placebo control there is no harm. The (eventual) QoL improvement of responder and non responder is evaluated just during the trial (at the end of phase III this group could be treated with the new therapy if the trial succeeds) and it results:

$$IE_C(n) = n \times (p_c \times (QoL_c \times D_{th} + QoL_d \times (D_{P3}(n) - D_{th})) + \\ + (1 - p_c) \times QoL_d \times D_{P3}(n)) .$$

Finally, IE results:  $IE(n) = IE_T(n) + IE_{TH}(n) + IE_C(n)$  .

Collective ethics (CE) concerns ethics of the population not involved in the trial, and that enrolled in the trial under the control treatment once the trial has been completed. Besides the quantities already introduced to define IE, CE depends on: the population size ( $N$ ), the incidence in the illness ( $Prev_i$ ), the power of the experiment  $\pi(n)$ . The size of the ill population is  $N_{ill} = N \times Prev_i - n$ , that is the ill population minus the group that tested the new treatment in the trial.

First, the “during trial” ethical balance of the population not involved in the experiment is:

$$CE_{DT}(n) = (N \times Prev_i - 2n) \times QoL_d \times D_{P3} .$$

When the trial succeeds, the ethical contribution related to the benefit of the population due to treatment, for responder and non responder, is:

$$CE_{ST}(n) = N_{ill} \times (p_t \times QoL_t + (1 - p_t) \times QoL_d) \times (D_L - D_{P3}(n) - D_{th}) \times \pi(n) .$$

Note that the duration of the quality of life (benefit or not) is given by life expectancy minus the duration of the trial minus that of the therapy. In other words, it is assumed that the ill population adopts the new treatment whenever it is available, that is, just after the end of the successful trial.



The ethical contribution of harm due to the new drug in the ill population also accounts for the power of the experiment, resulting:

$$CE_{TH}(n) = N_{ill} \times hp_t \times QoL_r \times D_{th} \times \pi(n) .$$

In case the trial is unsuccessful the harm is due to the loss of benefit, and the quality of life of ill population remains the same:

$$CE_{UT}(n) = N_{ill} \times QoL_d \times (D_L - DP_3(n)) \times (1 - \pi(n)) .$$

The total collective ethical contribution results:

$$CE(n) = CE_{DT}(n) + CE_{ST}(n) + CE_{TH}(n) + CE_{UT}(n) .$$

The global ethical contribution of the experiment is given by the sum of individual and collective ethics:

$$GE(n) = IE(n) + CE(n) . \quad (1)$$

It is of interest to compute the sample size providing the best ethical balance, and then to account for the power this optimal sample size provides, which could not be in the range of the classical range of power adopted for planning phase III trials, i.e. [80%, 90%].

## 4 Examples

Two numerical examples are reported here, based on the ethical model in (1).

Let us consider first a situation where the new drug works well, and where side effects are low. We expect that the power at the phase III sample size giving the best ethical balance is high.

We set the parameters as follows:  $\alpha = 2.5\%$ ,  $p_t = 0.5$ ,  $p_c = 0.1$ ,  $Prev_i = 10\%$ ,  $N = 1M$ ,  $hp_t = 0.05$ ,  $QoL_d = -2$ ,  $QoL_t = 5$ ,  $QoL_r = -5$ ,  $QoL_c = 0.5$ ,  $D_L = 20$ ,  $D_{th} = 0.5$ ,  $A_r = 200$ . In this situation, if  $\alpha$ ,  $p_t$  and  $p_c$  were considered only, standard sample size computation would give group samples of size 17 to achieve a power of 80%, and of size 23 with power 90%. However, the sample size giving the optimal Ethical Balance, that is providing the maximum of  $GE(n)$ , is  $argmax(GE(n)) = n_{opt} = 55$ , per group. The subsequent power is:  $\pi(55) = 0.9956$ , meaning that the optimal power is quite higher than those usually adopted, viz. 80-90%.

Now, consider a situation where the effect of the new drug is just moderate and the side effects are quite remarkable. Some of the above parameters are modified as follows:  $p_t = 0.3$ ,  $Prev_i = 0.1\%$ ,  $hp_t = 0.5$ ,  $QoL_t = 3$ ,  $QoL_r = -10$ ,  $A_r = 50$ . In this second situation, standard sample sizes per group would be of 59 and 79, with prefixed power of 80% and 90%, respectively. The optimal Ethical Balance is obtained here with samples of size  $argmax(GE(n)) = n_{opt} = 50$ . The power function has now changed, since  $p_t = 0.3$ . Consequently, the optimal power under the ethical

perspective is  $\pi(50) = 0.7054$ . This means that when the effect size is low and there are considerable side effects, the optimal power can be quite lower than 80-90%.

Often, the information on the parameters involved in the Ethical Balance is weak. To account for the possible deviation between the prefixed values of the parameters and their true value, a statistical distributions on parameters is usually introduced: this technique is called assurance (see [9]). We performed a sensitivity analysis of model (1) with assurance obtaining that, also in this case, the optimal power can be quite lower than 80-90%. These results are not reported here for the sake of brevity.

## 5 Conclusions

We have shown, through the model on individual and collective ethics we introduced, and a couples of appropriate examples, that defining the power within the range 80-90% can lead to poor choices in terms of the impact a new drug might have on the ill population.

To conclude, we would like to merge ethical models, like the one here introduced, and economical models, such as those developed in [10], [1], and in [3], since we believe that even profit and cost have an ethical impact on our society.

## References

1. Antonijevic Z., Kimber M., Manner D., Burman C-F., Pinheiro J., and Bergenheim K.: Optimizing Drug Development Programs: Type 2 Diabetes Case Study. *Therapeutic Innovation & Reg.Sci.* **47**(3), 363–374 (2013)
2. Chow S.C., Shao J., Wang H.: *Sample Size Calculations in Clinical Research*. Chapman and Hall/CRC, 2nd ed. Boca Raton (2008)
3. De Martini D.: Profit Evaluations when Adaptation by Design is Applied. *Therapeutic Innovation and Regulatory Science*, **50**(2), 213–220 (2016)
4. Gelfond J.A., Heitman E., Pollock B.H., Klugman C.H.: Power, ethics, and obligation. *Stat. Med.* **31**(29), 4140–4141 (2012)
5. Guo J.J., Pandey S., Doyle J., Bian B., Lis Y., Raisch D.W.: A Review of Quantitative RiskBenefit Methodologies for Assessing Drug Safety and Efficacy Report of the ISPOR RiskBenefit Management Working Group. *Value in Health* **13**(5), 657–666 (2010)
6. Heilig C.M., Weijer C.: A critical history of individual and collective ethics in the lineage of Lellouch and Schwartz. *Clinical Trials* **2**, 244–253 (2005)
7. Lellouch J., Schwartz D.: L'essai thérapeutique: éthique individuelle ou éthique collective? *Revue de l'Institut International de Statistique* **39** 127–136 (1971)
8. Meinert C.L.: *Clinical Trials - Design, Conduct and Analysis*. Oxford University Press, New York, 1986.
9. O'Hagan A., Stevens J.W., Campbell M.J.: Assurance in clinical trial design. *Pharmaceutical Statistics*, **4**, 187–201 (2005)
10. Patel N., Bolognese J., Chuang-Stein C., Hewitt D., Gammaitoni A., Pinheiro J.: Designing Phase II Trials Based on Program-Level Considerations: A Case Study for Neuropathic Pain. *Drug Inf.J.* **46**(4), 439–454 (2012)
11. Schwartz D., Flamant R., Lellouch J.: *Clinical Trials*. Academic Press, New York, 1980.
12. Wang S.J., Hung H.M.J., O'Neill R.T.: Adapting the sample size planning of a phase III trial based on phase II data. *Pharmaceutical Statistics* **5**, 85–97 (2006)

# Sampling schemes using scanner data for the consumer price index

## *Schemi di campionamento per la stima dell'indice dei prezzi al consumo usando gli scanner data*

Claudia De Vitiis, Alessio Guandalini, Francesca Inglese and Marco D. Terribili

**Abstract** The Italian National Institute of Statistics (ISTAT) is carrying out a redesign of Consumer Price Survey (CPS). The availability of Scanner Data (SD) from retail modern distribution, provided to ISTAT by Nielsen for a large number of stores selling food and grocery, is the starting point of this transformation. Indeed, SD represents a big opportunity for introducing improvements in the computation of Consumer Price Index (CPI). This work aims to study the properties of alternative aggregation formulas of the elementary price index in different sampling schemes implemented on SD. Bias and efficiency of the estimated indices are evaluated in a Monte Carlo simulation context. Finally, a comparison between a fixed and a dynamic approach in the compilation of the elementary price indices was performed.

**Abstract** *La disponibilità di dati scanner (SD) provenienti dalla grande distribuzione, che l'ISTAT acquisisce dalla Nielsen per un cospicuo numero di punti vendita (prodotti alimentari e per la casa), costituisce il punto di partenza per il ridisegno dell'indagine sui Prezzi al Consumo. Infatti, gli SD rappresentano una grande opportunità per l'introduzione di miglioramenti nel calcolo dell'indice dei prezzi al consumo (CPI). Questo lavoro si propone di studiare le proprietà di diversi schemi di campionamento implementati sugli SD con differenti formule di calcolo dell'indice elementare dei prezzi, valutando distorsione ed efficienza degli indici con una simulazione Monte Carlo. Infine, si presenta un confronto tra l'approccio fisso e dinamico per il calcolo dell'indice.*

**Key words:** Consumer Price index, scanner data, sampling, fixed and dynamic approaches

---

<sup>1</sup> Claudia De Vitiis, ISTAT; devitiis@istat.it

Alessio Guandalini, ISTAT; alessio.guandalini@istat.it

Francesca Inglese, ISTAT; fringles@istat.it

Marco D. Terribili, ISTAT; terribili@istat.it

## 1 Introduction

The Italian National Institute of Statistics (ISTAT) is carrying out a redesign of the Consumer Price Survey (CPS). The main aim of the project is to modernise the survey, improving and unburdening the data collection phase, together with the progressive introduction of more rigorous sampling procedures, probabilistic where possible, for the selection of outlets and products for the sectors where this is feasible (De Vitiis *et al.*, 2015; Bernardini *et al.*, 2016).

The availability of Scanner Data (SD) from retail modern distribution, provided through an agreement with Nielsen, are the starting point of this transformation. The SD represents a big opportunity for introducing improvements in terms of both data collection and sampling perspective.

Through a contract with Nielsen and an agreement with the six main retail chains operating in Italy, ISTAT started receiving, since the end of 2014, SD referred to food and grocery markets and treating them with the objective of experimenting the computation of the consumer price index (CPI). Scanner data files contain elementary information referred to single EAN codes<sup>1</sup> (European Article Number, GTIN) for specific outlets consisting of turnover and quantities sold during a week. This information does not provide the “shelf price” of the product individuated by the EAN code and outlet (reference or series), but allows to define a unit value or average weekly price. For reasons deriving from operational constraints of the productive process, a restriction is introduced regarding the observable weeks: only the relevant weeks are considered, defined as the first three full weeks (composed of seven days) in each month. Furthermore, usually SD do not include information about discounts or special sales.

For an accurate computation of the price index over time, the use of high-frequency data, as are SD, requires considerable efforts in both data collection and estimation phases. Completeness and correctness of the data are two important pillars for a correct use of SD (Vermeulen and Herren, 2006; Van der Grient *et al.*, 2010); formal and quality checks on the data flow must be implemented. In the estimation phase some important drawbacks associated to SD, as a high attrition rate of products, the temporary missing products, the entry of new products and volatility of the prices and quantities due mainly to sales, need to be addressed from both a theoretical and a practical point of view.

An important issue, out of the scope of this paper but crucial for the ISTAT CPI, is the necessity to combine estimates deriving from scanner data with the estimates that will continue to be produced by the current on field survey for the traditional retail distribution.

The aim of this paper is to present the SD experimental framework in which, first, probability and nonprobability selection schemes of series and, further, different probability sampling designs are compared. In particular, some important results on the properties of the price index at the elementary aggregate level, calculated according to different formulas and various probability sampling designs,

---

<sup>1</sup> Nielsen provided also the dictionary for the classification of EAN codes to GSI-ECR-Indicod product classification, while ISTAT ensures internally the translation from ECR to COICOP, the classification of products used for the CPI.

will be presented. A further experiment is also summarized to highlight the differences between a fixed and a dynamic population approach in the construction of the price indices.

The paper is organized as follows: section 2 describes the context and the methodological approach of experiments used to compare different sampling designs from SD; section 3 shows the most important results regarding accuracy of price indices estimates; in section 4 some conclusions and future developments are exposed.

## 2 Context and methods for sampling scanner data series

The experiments carried out so far on the scanner data aimed at evaluating the properties of the weighted and unweighted elementary price indices in different selection schemes of *series* (EAN and outlet codes). A series is a reference for which prices are observed during a certain period. In this phase of the experiments, the implications of life-cycle of series, seasonality issues and missing data have been not taken into account and a simplification have been used: only permanent series<sup>1</sup> are considered as universe for sampling and price index evaluation. A sample of series is selected at the beginning of the reference period and followed during all the year without considering either new entries nor discontinuities.

For each selection scheme, starting from the monthly price ratios with fixed base (December 2013) available for 2014, the elementary price indices are calculated using three classic aggregation formulas: Jevons (unweighted), Fisher (ideal) and Lowe (weights from quantities of previous year). The choice of these indices has been made on the basis of theoretical and empirical considerations: Fisher ideal index is thus preferred by economic theory, it uses quantities in different times and allows for substitution effects.

The experimental study has been developed in two phases: first, probability and nonprobability selection schemes of series have been compared; further, several sampling designs characterized by the use of different criteria of sample allocation, both for outlets and elementary items (EANs), and different selection methods of the sampling units, were considered. The comparison among the alternative selection schemes is made, for each price index, taking the corresponding true value of the index computed on the whole universe as a benchmark. Indices performance are evaluated in terms of bias for all selection schemes. For probability selection schemes, accuracy (bias and sampling variance) of the price indices have been studied with a Monte Carlo simulation: 500 samples are selected, according to different sampling designs. Indices variability and bias are computed on the estimated indices in the replicated samples. The sample selection and weighting of price indices is based on the total annual turnover of 2013. The analyses were conducted on SD relative to six retail chains (Conad, Coop, Esselunga, Auchan,

---

<sup>1</sup> Permanent series are referred to those references with not-null turnover for at least one relevant week (the first three full weeks) in each month of the considered year, starting from the December of previous year.

Carrefour, Selex) available in 2014 for Turin province and considering some consumption segments.

**Sampling designs** – In the first phase, nonprobability sampling is carried out by selecting series on the basis of cut-off thresholds of covered turnover in previous year, 2013: two samples are formed with all the series covering respectively the 60 and 80 percent of the total turnover in each of the considered consumption segment (coffee, pasta, mineral water) in the selected outlets. Moreover, considering the currently used fixed basket approach, a second selection scheme is defined selecting the most sold EANs for each representative product in the selected outlets. Nonprobability selection schemes are compared with two-stage probability sampling design, where primary stage units (PSU) and secondary stage units (SSU) are respectively outlets and EANs. The size of outlets sample has been fixed at a number of 30 out of 121 outlets available in SD. The sample size for SSU is fixed by a sampling rate of 5 percent of the number of EANs in each consumption segment in the sampled outlets. Outlets are stratified by chain and outlet type (hypermarket and supermarket). In each stratum, the sample has been allocated proportionally to the turnover. The selection of outlets is carried out in each stratum by simple random sampling (SRS), while the EANs are selected with probability proportional to size (PPS), in terms of total turnover of previous year, by adopting Sampford sampling (Sampford, 1967).

In the second phase of the experiments the following sampling designs have been compared: 1) one stage stratified sample of EANs; 2) cluster sample of outlets; 3) two-stage sampling with stratification of PSU (outlet) and SSU (EAN). For each sampling design the size of the final sample of EANs has been fixed in average at 7,400 to compare the different sampling strategies on equal computational effort. Moreover, different criteria of sample allocation, both for outlets and EANs, and different selection methods of the units were considered.

The first sampling design is carried out stratifying the EANs by market (ECR group) in each consumption segment (considering coffee, pasta, mineral water, olive oil, spumante and ice cream). Sample size is allocated among the strata through a Neyman formula, taking into account the variability of prices relatives in the markets observed in the reference year 2013. Two selection schemes have been considered, SRS and PPS.

In the second design, cluster sampling, a sample of outlets (14 out of 121 outlets) is selected. Outlets are stratified by chain and type. In each stratum, two different allocation of outlets are tested: proportional to the strata turnover and optimal allocation (Neyman). Outlets are selected with both SRS and PPS methods. All the EANs in the selected outlets are included in the sample.

Finally, two-stage sampling design is characterized by a stratification of both PSU and SSU. The stratifications adopted for the PSU and the SSU are the same of the two schemes described above. The size of the outlets sample has been fixed at a number of 30 out of 121 outlets. For both outlets and EANs, sample allocation in the strata is proportional to the strata turnover. PSU are selected with a PPS method, while SSU are selected both with SRS and PPS methods.

**Unbiased estimators** - The parameters of interest are monthly Jevons, Fisher and Lowe indices. Jevons index is an unweighted CPI that uses price information only (it assumes that expenditure shares remain constant), while Fisher and Lowe use also quantity information. Fisher and Lowe indices consider turnover shares at different time periods as weights (Gábor and Vermeulen, 2014). Indicating by the subscript  $t$  the current month (12 months in year 2014),  $t_0$  the reference month (December 2013),  $l$  the previous year (2013),  $c$  ( $c=1, \dots, C$ ) the generic homogeneous products group and  $m$  ( $m=1, \dots, M_c$ ) the series, unbiased sampling estimators of population parameters (elementary price indices aggregation) can be expressed as follows:

$$\begin{aligned}
 JEVONS_{ct}^{\bullet} &= \prod_m \left( \frac{p_{cmt}}{p_{cmt_0}} \right)^{1/\sum_m w_{cmt}} \\
 LASP_{ct}^{\bullet} &= \sum_m \left( \frac{p_{cmt}}{p_{cmt_0}} \right) * \left( \frac{p_{cmt_0} * q_{cmt_0} * w_{cmt}}{\sum_m p_{cmt_0} * q_{cmt_0} * w_{cmt}} \right) \\
 PAAS_{ct}^{\bullet} &= \sum_m \left( \frac{p_{cmt_0}}{p_{cmt}} \right) * \left( \frac{p_{cmt} * q_{cmt} * w_{cmt}}{\sum_m p_{cmt} * q_{cmt} * w_{cmt}} \right) \\
 FISH_{ct}^{\bullet} &= \sqrt{LASP_{ct}^{\bullet} * PAAS_{ct}^{\bullet}} \\
 LOWE_{ct}^{\bullet} &= \sum_m \left( \frac{p_{cmt}}{p_{cmt_0}} \right) * \left( \frac{p_{cmt_0} * q_{cmt}^z * w_{cmt}}{\sum_m p_{cmt_0} * q_{cmt}^z * w_{cmt}} \right)
 \end{aligned}$$

$$\text{with } q_{cmt}^z = \sum_{a=0}^{11} q_{cm(t_0-a)}$$

The  $q_{cmt}^z$  measure refers to the  $m$ -th quantity series in the previous year  $l$  (2013). The weight  $w_{cmt}$  is obtained as the inverse of the inclusion probability of the sampling unit deriving from the sampling design.

### 3 Main results of the experimental phase

The most meaningful results of the two experimental phases are shown in the following figures. Figure 1, from the first experimental phase, shows the level estimates of the monthly Jevons, Lowe and Fisher indices computed on probability (two stage sampling) and nonprobability samples and the true value (universe panel series SD, U) of the corresponding index for two consumption segments (coffee and pasta in Turin province).

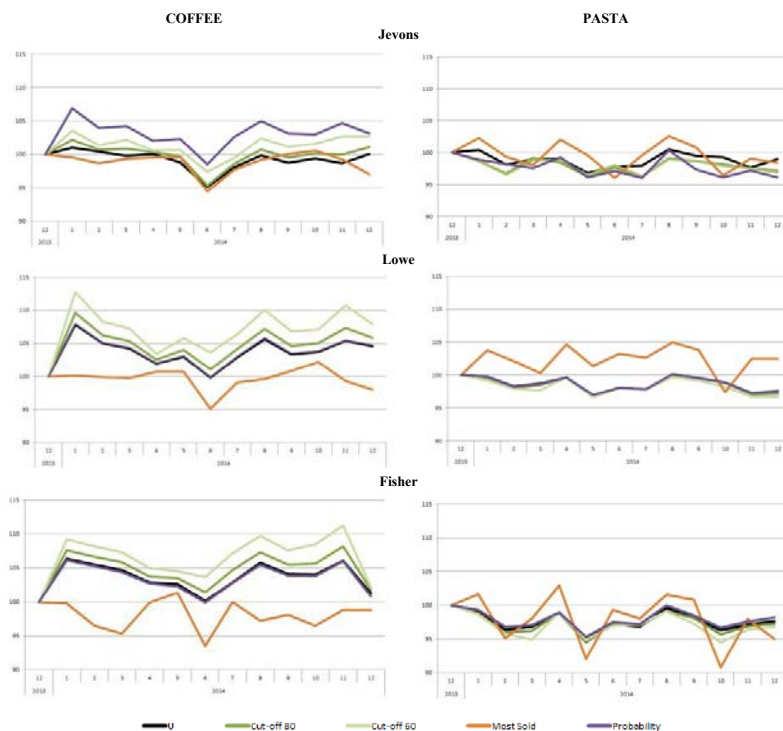


Figure 1 – Jevons, Lowe and Fisher indices computed with different selection schemes of series for coffee and pasta segments, Turin province, year 2014

The comparison between probability and nonprobability selection schemes shows a common evidence for both products: pps sample estimates of weighted index, Fisher and Lowe, results quite overlapped to the “true” value  $U$ ; cut-off estimates over-estimate, but follow the trend for coffee, while for pasta are quite overlapped to true value  $U$ . Most sold item estimates under-estimate and alter trend for coffee with weighted indices but not for Jevons, while for pasta they show different trends for the three indices. The mean of sample estimates of Jevons index strongly over-estimates the “true” value  $U$  for coffee but not for pasta. These opposite performance for the two product can be explained by the different number of items and turnover distributions. In general, also from other evidences not shown for sake of brevity, (i) probability sampling always produce more accurate estimates than nonprobability selection scheme; (ii) sampling scheme is not neutral with respect to the choice of aggregation formulas; (iii) sampling error varies among consumption segments.

Figure 2, from the second experimental phase, illustrates the difference among



the three indices estimated under two different sampling designs: cluster sample of outlets (with proportional allocation and PPS selection) versus two stage sample (proportional allocation and PPS selection of outlets and Neyman allocation and PPS selection of EANs).

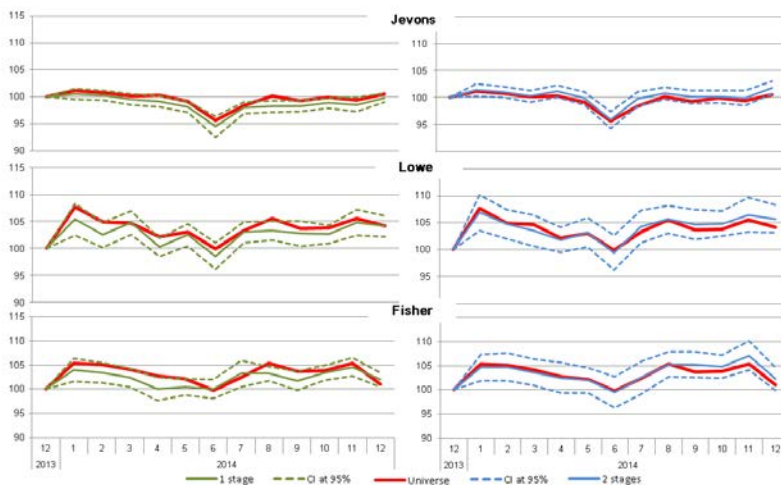


Figure 2 – Jevons, Lowe and Fisher indices for coffee segment estimated on one sample, confidence interval (CI) of estimates at 95% and true value (computed on the universe of SD). Turin, year 2014

The comparison between two probability selection schemes highlights that all the estimates seems to catch properly the level and the trend of the related true index. The estimator of Lowe and Fisher indices have in both cases wider confidence intervals (CI) with respect to the Jevons index, due to the variability of quantities involved in the weights. In general the width of CIs are greater under the two stage sampling than under cluster sampling design (one stage), even if the difference does not seem so large. In fact, the intra-group correlation is negative, even if close to 0. SO, variability within the outlets seems slightly higher that between the outlets. However, this aspect needs of further studies taking into account also all the consumption segments.

#### 4 Concluding remarks and future developments

The two experimental phases produced interesting results regarding the performance of sampling schemes and index formulas in a closed population context and fixed approach. They lead to the conclusion that probability sampling is the better choice in this context.

The successive phase, currently in progress, regards the comparison between a fixed and a dynamic approach, the latter consisting in considering all series of an

open population (Ivancic *et al.*, 2011). The elementary price indices are computed considering both closed and open population. Assuming a closed population, direct indices are built on a fixed basket of products defined at reference time, ignoring new products (fixed approach). In this context the indices are affected by shrinkage over time due to the attrition of products during the year. However, in reality many products disappear and new products enter continuously. By using chain indices the life cycle of products is taken into account as the basket of products changes months by months: the flexible basket is constituted by the matching products sold during two months in a row (dynamic approach).

In order to evaluate the impact of the life cycle of products, direct and chain price indices are compared. For this purpose, an artificial population has been generated, with products appearing and disappearing (momentarily and permanently). Starting from a panel of products, new products have been introduced considering the monthly birth rates and old products have been removed in accordance to a survival function. Both monthly birth and survival rates have been estimated on the real open population.

The construction of this artificial population is a trick enabling to evaluate the whole error, both sampling and non-sampling errors, due to appearing and disappearing products.

## References

1. Bernardini, A., De Vitiis, C., Guandalini, A., Inglese F. and Terribili M. D. Measuring inflation through different sampling designs implemented on scanner data. Paper presented at the UNECE meeting of the group of experts, Geneva 2-4 May (2016)
2. de Haan, J., Opperdoes, E., Schut, C.M. Item selection in the Consumer Price Index: Cut-off versus probability sampling. *Survey Methodology*, 25(1), 31-41. (1999)
3. de Haan, J., van der Grient, H.A. Eliminating chain drift in price indexes based on scanner data. *Journal of Econometrics*, 161, 36-46. (2011)
4. De Vitiis C., Casciano, M. C., Guandalini, A., Inglese, F., Seri, G., Terribili, M. D. and Tiero F. Sampling design issues in the first Italian experience on scanner data. Paper presented at the Scanner Data Workshop. Rome 1-2 October (2015)
5. Feldmann B. (2015). Scanner-data-current-practice. [http://www.istat.it/en/files/2015/09/5-WS-Scanner-data-Rome-1-2-Oct\\_Feldmann-Scanner-data-current-practice.pdf](http://www.istat.it/en/files/2015/09/5-WS-Scanner-data-Rome-1-2-Oct_Feldmann-Scanner-data-current-practice.pdf).
6. Gábor, E. and Vermeulen, P. New evidence in elementary index bias. (2014)
7. ILO, IMF, OECD, Eurostat, United Nations, World Bank Consumer Price Index Manual: Theory and Practice, Geneva: ILO Publications (2004)
8. Ivancic, L., Diewert, W.E., Fox, K.J. Scanner Data, Time Aggregation and the Construction of Price Indexes. *Journal of Econometrics*, 161(1), 24-35. (2011)
9. Nygaard, R. Chain drift in a monthly chained superlative price index. Workshop on scanner data, Ceneva, 10 may (2010)
10. Rosén, B. On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*. 62(2):159-191. (1997)
11. Sampford, M.R. On sampling without replacement with unequal probabilities of selection. *Biometrika*. 54(3-4):499-513. (1967)
12. Van der Grient, H. A. and de Haan J. The Use of Supermarket Scanner Data in the Dutch CPI. Paper presented at the Joint ECE/ILO Workshop on Scanner Data, Eurostat (2015)
13. Vermeulen, B. C. and Herren, H. M. Rents in Switzerland: sampling and quality adjustment. 11th Meeting - Ottawa Group - Neuchâtel 27-29 May (2006)

# Interactive machine learning prediction for budget allocation in digital marketing scenarios

## *Machine learning per l'allocazione del budget attraverso la previsione interattiva di scenari digitali di marketing*

Della Valle Ermelinda, Scardovi Elena, Iacobucci Andrea, Tignone Edoardo

**Abstract** Scenario Analysis estimates the relationship between budget allocation and some typical digital analytics metrics quantifying the performances of marketing campaigns. The actual values of the monitored Key Performance Indicators, deriving from different big data sources, are compared to their estimated value after a variation in the investments. Our ensemble approach combines multivariate generalized linear models with machine learning models. R implementation is embedded into an interactive dashboard for providing real time score predictions. The visualization simplifies the business fruition of the analysis and encourages new and deeper experimentations with data.

**Abstract** *L'analisi di scenario stima la relazione tra l'allocazione del budget e alcune metriche tipiche della digital analytics che quantificano le performance delle campagne di marketing. Alcuni indicatori chiave di performance (KPI), derivati da grandi quantità di dati provenienti da fonti differenti, sono confrontati con la stima del valore che assumerebbero se fossero modificati gli investimenti. L'ensemble predittivo da noi realizzato combina modelli lineari generalizzati e machine learning. L'implementazione in R è inserita in una dashboard interattiva per realizzare previsioni in tempo reale. La visualizzazione scelta rende semplice la fruizione dell'analisi da parte dei top manager e incoraggia una nuova e approfondita sperimentazione attraverso i dati.*

**Key words:** Marketing Campaign, Optimization, Google AdWords Auctions, Machine Learning, Big Data, Key Performance

---

<sup>1</sup> Della Valle Ermelinda, BitBang Srl; email: edellavalle@bitbang.com  
Scardovi Elena, BitBang Srl; email: escardovi@bitbang.com  
Iacobucci Andrea, BitBang Srl; email: aiacobucci@bitbang.com  
Tignone Edoardo, BitBang Srl; email: etignone@bitbang.com

## 1 Motivating Example

The Scenario Analysis project here described was realized for an international company active in the retail distribution that needed to understand the effectiveness of its digital allocation of investments on advertising, highlighting strategic opportunities otherwise ignored. We were required to build an interactive dashboard, addressed to top managers, showing the relationship between budget allocation and some typical digital analytics metrics coming from Google AdWords (GAW) and Adobe Analytics (AA).

In the last decades, strategic market management has often been addressed from a theoretical point of view [1]. Our aim is to drive business decisions through data and modelling, allowing non-statistical entrepreneurs to exploit and interpret evidences that are not usually examined together due to their different sources. The main goal of Scenario Analysis is to show what would happen if something changed in the business decision making process. Responses can be influenced by a large number of factors, many of which can not be controlled by the experimenters or even quantified; for this reason, the “Scenario” concept treats such factors as fixed.

In our application, we predicted Key Performance Indicators (KPIs) on the basis of differently allocated budget amounts. We addressed this Big Data problem by developing synthetic scores and working with machine learning models for quickly processing large amounts of data.

## 2 Data

We considered a temporal period ranging from December 2014 to February 2017 with weekly data updates. The large amount of digital data (~20GB), separately measuring different marketing actions, had to be joined and uniformed in both format and granularity. We addressed Clicks, i.e. the number of clicks on the advertisements (GAW); Impressions, i.e. the number of times the advertisement was shown (GAW); Impression Share, i.e. the percentage of impressions on the potential (available) impression amount (GAW); Page Views Per Visit, i.e. the average number of page views for each visit; Entries, i.e. the number of visits directly deriving from advertising (AA); Bounces, i.e. the number of entries with only one page view (AA); Orders (AA); Conversion Rate, i.e. the ratio between Orders and Entries cleansed from Bounces.

As requested by our customer, we treated these quantities and the allocated budget on a daily basis for each marketing campaign. The variable Cost was also available; in fact, advertisers bid on certain keywords in Google AdWords auction [2] in order for their clickable ads to appear in search results, and pay Google a certain amount of money (Cost) according to the clicks received by their ads, with a Cost per Click that depends on the behaviour of internet surfers and competitors (Fig. 1). Optimal allocation of budget allows to achieve the highest possible profits with the smallest effective expense.

Campaigns were a-priori classified into three distinct macro-areas of interest: “Brand”, dedicated to empower the firm name; “Cart”, focused on item selling; “Promo”, regarding temporal offers. We were requested to consider budget modifications in the Scenario at a macro-area level rather than on single campaigns.

### 3 Modelling and prediction

Objective of the model was to predict the values of the metrics after budget leverages under the assumption that all other factors were fixed.

The relationship between indicators changed according to the campaign macro-area (Fig. 2). The different origin, scale and variability of the KPIs under study required smart grouping on the basis of homogeneous behaviours. Clicks, Bounces and Entries were similarly influenced by investments, with a nearly linear tendency; on the other hand, Impressions and Potential Impressions depended on discrete budget jumps in a smooth nonlinear way, while Orders, Page Views and Visits required a model specification allowing different speeds of growth, while Orders required an ad-hoc approach able to address the non-negligible presence of zero values on the least relevant products. Finally, Page Views Per Visits and Conversion Rate consisted in calculated fields that were addressed by tuning the model on Page Views, Visits, Orders, Entries and Bounces.

Because of the strong influence of discontinuous budget investments and to Google AdWords instantaneous dynamics, no significant temporal tendencies encouraging a time-series approach were found; the temporal dimension is only considered to combine contingent data.

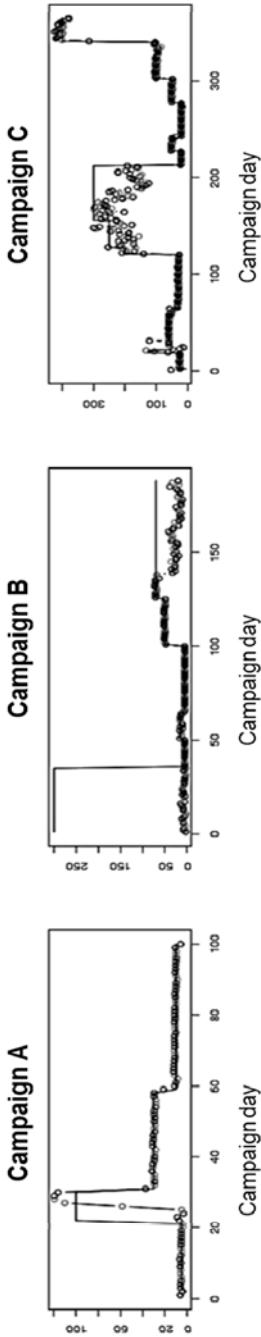
#### 3.1 Model specification

We adopted an ensemble approach [3] that combined a multivariate generalized linear model with Gamma or Poisson specification with a dynamic implementation of Random Forest [4]. The specific business requirement imposed the budget as the only actionable lever. In order to ensure model flexibility, we included campaign-level metadata and a campaign-group classification. The value for the  $n$ -th KPI reached by the  $i$ -th campaign on day  $t$  was modelled as

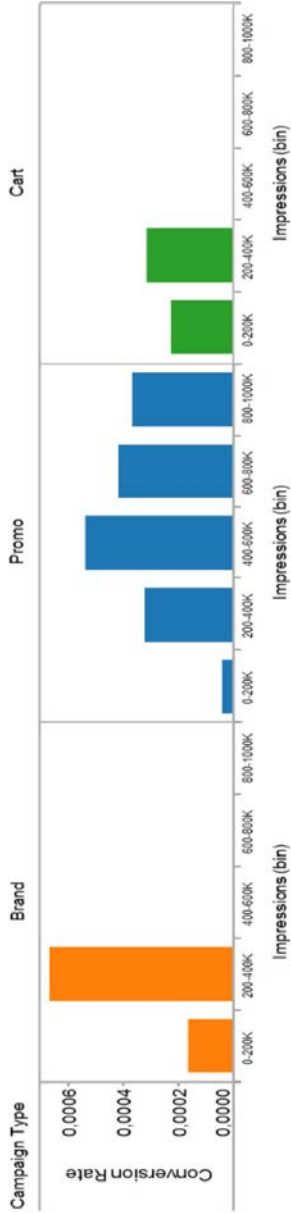
$$K_{n,i,t} = f_n(a_i, C_{i,t})$$

where  $f_n$  is a stochastic function defined for the  $n$ -th KPI,  $n=1, \dots, N$ , and  $C_{i,t}$  is the amount of money (Cost) spent on the category  $l_i$  of the  $i$ -th campaign,  $i=1, \dots, I$ :

$$C_{i,t} = B_{i,t} \cdot CoB_{l_i,t} \cdot (1 + p_{l_i,t})$$



**Figure 1:** Example of Budget (line) and Cost (circles) behaviour for three campaigns. Cost can be much lower than Budget when investments do not carry clicks (see Campaign B), or a bit higher in case of clicks abundance (e.g. peak on Campaign A)



**Figure 2:** Relation between (binned) Impressions and Conversion Rate for the three macro-areas

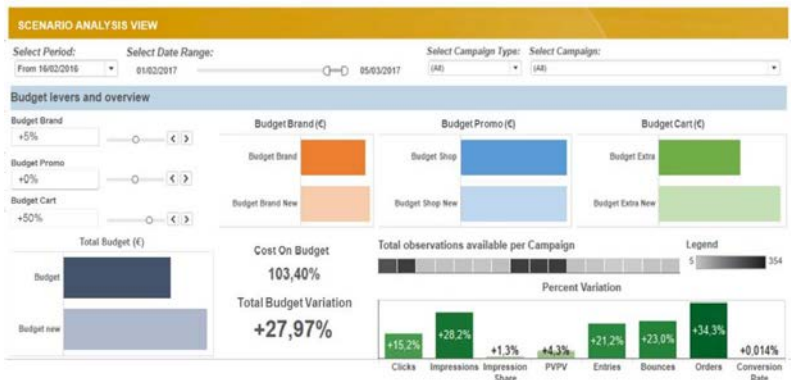
with  $CoB_{i,t}$  being the ratio between Cost and Budget  $B_{i,t}$ , that in our Scenario is assumed not to change after investment percent modification  $p_{i,t}$ . The number  $I$  of the campaigns was 77 on the initial six months of the project, and changed at each data update.  $N$  is equal to 8, since this approach addresses Clicks, Impressions, Potential Impressions, Entries, Bounces, Orders, Page Views and Visits. Notice that some of these KPIs needed to be combined for calculating the quantities listed in Sec. 2, that were shown to the end user<sup>1</sup>.

The inclusion of individual dynamics (with macro-area parameters  $a_{i,t}$  instead of  $a_i$ ) turned out to be crucial. In the nonparametric machine learning ensemble member, campaign information is included as an explanatory variable. A grid-based search found 300 trees as the best compromise between model efficiency and computational effort for Random Forest. Regression and machine learning predictions were combined in an ensemble favouring the model with higher fitting performances on a 10% test set.

Daily KPIs were obtained as  $K_{n,t} = \sum_{i=1,\dots,I} \bar{K}_{n,i,t}$  for  $n=1,\dots,N$ , with  $t$  denoting the day and  $\bar{K}_{n,i,t}$  being the ensemble result. Whereas, when a multi-day period  $\tau$  is desired,  $K_{n,\tau} = \sum_{t \in \tau} K_{n,t}$ . Composite KPIs were calculated from these basic KPIs.

## 4 Dashboard visualization and Conclusions

We realized an interactive dashboard [5] consisting in a simple visual interface relying on a computational engine integrating an R instance computing real time prediction for scenario comparison.



**Figure 3:** Dashboard Layout. Left: budget modifications. Top-right: filters. Bottom-right: KPIs variations.

<sup>1</sup> Direct modelling of composite KPIs (such as Conversion Rate) would have been possible, too; however, it would have required conservation of the ratio among its constituents as a constraint for preserving consistency of the scenario.

A crucial feature of our approach was auto-tuning on the base of filtering: campaign and period selection provided the user with the best fit referred to his choices. A dynamic choice of the training period was provided in order to preserve stability and prevent overfitting.

Model performances were evaluated through the improvement of the effectiveness of the data-driven marketing strategies on Brand, Cart and Promo, in terms of increase of Visits, Orders and Conversion Rate. As an example, in the period Sep 2016 - Feb 2017 w.r.t. Mar 2016 - Aug 2016 such KPIs registered +89%, +51% and +10% respectively, with a 16% increase in Budget and an effective 12% saving in Cost.

## References

1. Bood, R. P., Postma, T. J. B. M.: Scenario Analysis as a Strategic Management Tool. Dep. of Manag. and Organization, Fac. of Econ., Univ. of Groningen. SOM theme B: marketing and networks (1998)
2. Edelman, B., Ostrovsky, M., Schwarz, M.: Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords. *Am. Econ. Rev.* 97(1), 242–259 (2007)
3. Lanham, M. A., Badinelli, R. D.: Merging Business Kpis With Predictive Model Kpis For Binary Classification Model Selection. Proc. of the 2015 INFORMS Workshop on Data Mining and Analytics M. G. Baydogan, S. Huang, A. Oztekin, eds. (2015)
4. Shi, L., Li, B.: Predict the Click-Through Rate and Average Cost Per Click for Keywords Using Machine Learning Methodologies. Proc. of the 2016 Int. Conf. on Ind. Eng. and Oper. Manag. Detroit, Michigan, USA, Sept. 23-25 (2016)
5. Wankhade, R. S., Ingle, D. R., Meshram, B. B.: Web analytics dashboard and analysis system. *Adv. in Comput. Res.*, Vol 4, Issue 1, pp. 83-86 (2012)



# Nonparametric classification for directional data

Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor

**Abstract** We discuss nonparametric methods to address the problem of classification for directional data. We focus on local regression and kernel density estimation when the domain is the unit circle. We provide asymptotic theory for the proposed methods along with simulation results.

**Abstract** *Discutiamo metodi non parametrici per problemi di classificazione di osservazioni direzionali. In particolare consideriamo metodi di regressione locale e stima kernel di funzioni di densità nel caso in cui il dominio è il cerchio unitario, presentando alcune proprietà asintotiche e risultati simulativi.*

**Key words:** Density estimation, Discriminant analysis, Local weights, Logistic regression, von Mises density

## 1 Introduction

Circular data occur when the sample space is the unit circle. The peculiarity of a circular measurement scale is that its beginning and its end coincide. After both an origin and an orientation have been chosen, a circular observation can be measured, in radians, by an angle  $\theta \in [-\pi, \pi)$ . Circular data often arise in biology (migration

---

Marco Di Marzio  
DMQTE, Università di Chieti-Pescara. e-mail: mdimarzio@unich.it

Stefania Fensore  
DMQTE, Università di Chieti-Pescara. e-mail: stefania.fensore@unich.it

Agnese Panzera  
DiSIA, Università di Firenze. e-mail: a.panzera@disia.unifi.it

Charles C. Taylor  
Department of Statistics, University of Leeds. e-mail: charles@maths.leeds.ac.uk

paths, flight directions of animals), meteorology (wind and marine current directions), and geology (orientations of joints and faults, landforms, oriented stones).

We propose nonparametric classification methods for circular data based both on kernel estimation of the population densities and local logistic regression. In particular we consider the case when there are two sub-populations. This research filed seems unexplored, although the need for flexible methods is evident as is seen even in our very simple motivating example.

The paper is organized as follows. Section 2 collects some basic results on kernel estimation of circular densities. Section 3 deals with two different approaches to nonparametric regression with a circular predictor and a binary response, while Section 4 discusses kernel density estimation for discrimination. Finally, Section 5 presents some simulation examples.

## 2 Kernel circular density estimation

Given a random sample of angles  $\Theta_1, \dots, \Theta_n$  from an unknown circular density  $f$ , the kernel estimator of  $f$  at  $\theta \in [-\pi, \pi]$  can be then defined as

$$\hat{f}(\theta; \kappa) := \frac{1}{n} \sum_{i=1}^n K_\kappa(\Theta_i - \theta),$$

where the weight  $K_\kappa$  is a *circular kernel* with zero mean direction and concentration parameter  $\kappa > 0$ , see Definition 1 given by [2]. The weight function  $K_\kappa$  is usually chosen to be a continuous density function whose support is the circle with the property that as  $\kappa \rightarrow \infty$  the density tends to concentrate at the mode.

Now, for  $j \in \mathbb{N}$  and a circular kernel  $K_\kappa$ , we set

$$\eta_j(K_\kappa) := \int_{-\pi}^\pi K_\kappa(\alpha) \sin^j(\alpha) d\alpha \quad \text{and} \quad \nu(K_\kappa) := \int_{-\pi}^\pi K_\kappa^2(\alpha) d\alpha,$$

where  $K_\kappa$  is a  $r$ -th sin-order kernel if  $\eta_0(K_\kappa) = 1$ ,  $\eta_j(K_\kappa) = 0$  for  $j < r$  and  $\eta_r(K_\kappa) \neq 0$ , see Definition 2 in [2].

Assuming that:  $f''$  is continuous at  $\theta \in [-\pi, \pi]$ ;  $K_\kappa$  is a second sin-order kernel; as  $n \rightarrow \infty$ ,  $\eta_2(K_\kappa)$  and  $\nu(K_\kappa)/n$  both go to 0; then it results

$$E[\hat{f}(\theta; \kappa)] = f(\theta) + \frac{\eta_2(K_\kappa)}{2} f''(\theta) + o(\eta_2(K_\kappa)), \tag{1}$$

and

$$\text{Var}[\hat{f}(\theta; \kappa)] = \frac{\nu(K_\kappa)}{n} f(\theta) + o\left(\frac{\nu(K_\kappa)}{n}\right). \tag{2}$$

### 3 Nonparametric circular logistic regression

One of the possible regression models for dealing with dichotomous data is logistic regression. The goal is to find the best fitting to describe the relationship between a binary outcome and a set of independent predictors. Logistic regression determines the membership degree of each individual to one of the two groups by fitting a continuous function taking values in the interval  $[0, 1]$ .

Let  $Y$  and  $\Theta$  be a binary response and a circular predictor, respectively, and set  $\lambda(\theta) := P(Y = 1 \mid \Theta = \theta)$ . Denote the density functions in the circular covariate space for the successes ( $Y = 1$ ) and for the failures ( $Y = 0$ ) by  $f_1$  and  $f_2$ , respectively, and let  $\pi_1$  be the proportion of successes in the population, and  $\pi_2 = 1 - \pi_1$ . Then, for  $\theta \in [-\pi, \pi]$ ,

$$\lambda(\theta) = \frac{\pi_1 f_1(\theta)}{\pi_1 f_1(\theta) + \pi_2 f_2(\theta)}. \tag{3}$$

#### 3.1 Kernel estimator for binary regression

Given  $n$  independent copies of  $(\Theta, Y)$ ,  $(\Theta_1, Y_1), \dots, (\Theta_n, Y_n)$ , assume that the sample has been ordered in such a way that the first  $n_1$  pairs are successes and the last  $n_2 = n - n_1$  ones are failures. Replacing  $\pi_j$  in (3) with  $n_j/n$ ,  $j \in (1, 2)$ , a kernel estimator of  $\lambda(\theta)$ ,  $\theta \in [-\pi, \pi]$ , can be defined as

$$\hat{\lambda}(\theta; \gamma, \omega, \mu) = \frac{n_1/n \hat{f}_1(\theta; \gamma)}{n_1/n \hat{f}_1(\theta; \omega) + n_2/n \hat{f}_2(\theta; \mu)}, \tag{4}$$

where  $\hat{f}_j(\theta; \kappa)$ ,  $j \in (1, 2)$  and  $\kappa \in \{\gamma, \omega, \mu\}$ , stands for the kernel estimator of  $f_j(\theta)$  with a circular kernel  $K_\kappa$ . Estimators like the above one have been studied in the Euclidean setting by [4]. When the concentration parameters in (4) are  $\gamma = \omega = \mu$  the resulting estimator is the Nadaraya-Watson estimator with circular predictor, see [1]. When  $\gamma = \omega$ , assuming that: both  $f_1$  and  $f_2$  admit continuous derivatives up to order two; both  $K_\gamma$  and  $K_\mu$  are second sin-order circular kernels, with  $\eta_2(K_\kappa)$  and  $v(K_\kappa)/n$ ,  $\kappa \in \{\gamma, \mu\}$ , both going to 0 as  $n \rightarrow \infty$ ; and that  $\eta_2(K_\gamma) \sim \eta_2(K_\mu)$ , and  $v(K_\gamma) \sim v(K_\mu)$ ; using results (1) and (2), we obtain

$$E[\hat{\lambda}(\theta; \gamma, \mu)] - \lambda(\theta) = \frac{\pi_1 \pi_2 (\eta_2(K_\gamma) f_2(\theta) f_1''(\theta) - \eta_2(K_\mu) f_1(\theta) f_2''(\theta))}{2(\pi_1 f_1(\theta) + \pi_2 f_2(\theta))^2} + o(\eta_2(K_\gamma)),$$

and

$$\text{Var}[\hat{\lambda}(\theta; \gamma, \mu)] = \frac{\lambda(\theta)(1 - \lambda(\theta))}{n(\pi_1 f_1(\theta) + \pi_2 f_2(\theta))} [(1 - \lambda(\theta))v(K_\gamma) + \lambda(\theta)v(K_\mu)] + o\left(\frac{v(K_\gamma)}{n}\right).$$

For the case of a von Mises kernel, i.e.  $K_\kappa(\theta) := \{2\pi \mathcal{J}_0(\kappa)\}^{-1} \exp(\kappa \cos(\theta))$ , where  $\mathcal{J}_u(\cdot)$  stands for the modified Bessel function of the first kind and order  $u$ , it holds that for  $\kappa$  big enough

$$\eta_2(K_\kappa) \sim \frac{1}{\kappa}, \quad \text{and} \quad v(K_\kappa) \sim \frac{\kappa^{1/2}}{2\pi^{1/2}}. \tag{5}$$

As a consequence, using von Mises kernels for both  $K_\gamma$  and  $K_\mu$ , with  $\gamma \sim \mu$ , we have

$$\mathbb{E}[\hat{\lambda}(\theta; \gamma, \mu)] - \lambda(\theta) = O\left(\frac{1}{\gamma}\right), \quad \text{and} \quad \text{Var}[\hat{\lambda}(\theta; \gamma, \mu)] = O\left(\frac{\gamma^{1/2}}{n}\right).$$

Notice that in the special case where  $\gamma = \mu$ , it is easily seen that asymptotic bias and variance of the resulting estimator are the same of the local constant estimator with circular predictor, see [1].

Concerning optimal smoothing the standard approach in the Euclidean setting is to consider a weighted version of the mean squared error. For practical implementation the smoothing parameters are selected by minimizing an empirical version of the weighted mean squared error (for more details see, [4]).

### 3.2 Local polynomial binary regression

A different way to address the nonparametric binary regression estimation is based on the local likelihood approach. We start by defining the logit as a generic periodic function  $g$

$$\log\left(\frac{\lambda(\Theta_i)}{1 - \lambda(\Theta_i)}\right) := g(\Theta_i, \beta),$$

which depends on the observations and a vector of parameters  $\beta$ . The inverse transformation goes back from log-odds to probabilities, yielding the following circular logistic regression function

$$\lambda(\Theta_i) = \frac{\exp(g(\Theta_i, \beta))}{1 + \exp(g(\Theta_i, \beta))}. \tag{6}$$

The associated log-likelihood function at  $\theta \in [-\pi, \pi)$ , localized using kernel weights, is

$$\sum_{i=1}^n \left\{ Y_i \log\left(\frac{\lambda(\Theta_i)}{1 - \lambda(\Theta_i)}\right) + \log(1 - \lambda(\Theta_i)) \right\} K_\kappa(\Theta_i - \theta),$$

that can be reformulated in terms of  $g$  as

$$\log \mathcal{L}(\beta) = \sum_{i=1}^n \{g(\Theta_i, \beta) Y_i - \log(1 + \exp(g(\Theta_i, \beta)))\} K_\kappa(\Theta_i - \theta).$$

By modeling the log-odds ratio as a sin-series expansion yields a nonparametric method that we define Circular Local Polynomial Logistic Regression (CLP). In particular, define

$$g(\Theta_i, \beta) := \sum_{j=0}^p \frac{\beta_j \sin(\Theta_i - \theta)^j}{j!}, \tag{7}$$

and, letting  $\hat{\beta}_0$  be the solution for  $\beta_0$  of the maximization of  $\log \mathcal{L}(\beta)$  with respect to  $\beta = \{\beta_0, \dots, \beta_p\}$ , we get

$$\hat{\lambda}(\theta; \kappa) = \frac{\exp(\hat{\beta}_0)}{1 + \exp(\hat{\beta}_0)}.$$

Note that when  $p = 0$ ,  $\hat{\lambda}(\theta; \kappa)$  coincides with the estimator (4) with  $\omega = \gamma = \mu$ . Moreover, re-writing the log-likelihood function using different weights for the successes and failures as follows

$$\sum_{i=1}^n Y_i g(\Theta_i, \beta) K_\gamma(\Theta_i - \theta) - \log(1 + \exp(g(\Theta_i, \beta))) K_\mu(\Theta_i - \theta),$$

and using  $p = 0$  in approximation (7) we obtain the estimator (4) with  $\omega = \gamma$ .

### 4 Circular KDE discrimination

Kernel density estimation is commonly used for classification. Following the approach proposed by [3] we consider two groups of observations and estimate the difference between the two densities at the observation point allocating the label according to the highest density.

In particular, we consider the problem of estimating the difference between two circular densities,  $h(\theta) = f_2(\theta) - f_1(\theta)$ , using random samples of sizes  $n_2$  and  $n_1$  respectively. We are interested in solution  $h(\theta) = 0$  given by  $\theta_0$  such that  $f_1(\theta_0) = f_2(\theta_0) = f(\theta_0)$ . Letting  $\hat{h}(\theta; \kappa_1, \kappa_2) := \hat{f}_1(\theta; \kappa_1) - \hat{f}_2(\theta; \kappa_2)$ , under suitable assumptions on  $\eta_2(K_{\kappa_j})$ ,  $v(K_{\kappa_j})$  and  $f_j''(\theta)$ ,  $j \in (1, 2)$ , in virtue of results (1) and (2) we have that

$$E[\hat{h}(\theta; \kappa_1, \kappa_2)] = f_2(\theta) - f_1(\theta) + \frac{1}{2} \{ \eta_2(K_{\kappa_2}) f_2''(\theta) - \eta_2(K_{\kappa_1}) f_1''(\theta) \} + o(\eta_2(K_{\kappa_1}) + \eta_2(K_{\kappa_2})),$$

and

$$\text{Var}[\hat{h}(\theta; \kappa_1, \kappa_2)] = \frac{v(K_{\kappa_1})}{n_1} f_1(\theta) + \frac{v(K_{\kappa_2})}{n_2} f_2(\theta) + o\left(\frac{v(K_{\kappa_1})}{n_1} + \frac{v(K_{\kappa_2})}{n_2}\right).$$

When  $K_{\kappa_1}$  and  $K_{\kappa_2}$  are both von Mises kernels, the asymptotic mean squared error of  $\hat{h}$  at a point  $\theta_0$ , such that  $h(\theta_0) = 0$ , is

$$\text{AMSE}[\hat{h}(\theta_0; \kappa_1, \kappa_2)] = \frac{1}{4} \left\{ \frac{f_2''(\theta_0)}{\kappa_2} - \frac{f_1''(\theta_0)}{\kappa_1} \right\}^2 + \frac{1}{2\pi^{1/2}} \left\{ \frac{f_1(\theta_0)\kappa_1^{1/2}}{n_1} + \frac{f_2(\theta_0)\kappa_2^{1/2}}{n_2} \right\},$$

which is minimised by

$$\hat{\kappa}_1 = \{f(\theta_0)/[2\sqrt{\pi n_1} f_1''(\theta_0)^2 - (2\sqrt{\pi n_1} f_1''(\theta_0))^{5/3} (2\sqrt{\pi n_2})^{-2/3} f_2''(\theta_0)^{1/3}]\}^{-2/5},$$

and

$$\hat{\kappa}_2 = \{f(\theta_0)/[2\sqrt{\pi n_2} f_2''(\theta_0)^2 - (2\sqrt{\pi n_2} f_2''(\theta_0))^{5/3} (2\sqrt{\pi n_1})^{-2/3} f_1''(\theta_0)^{1/3}]\}^{-2/5}.$$

## 5 Numerical examples

In Figure 1 we observe examples of the behaviour of the parametric logistic regression. It is obtained using  $g(\theta_i, \beta) := \beta_0 + \beta_1 \cos(\theta_i) + \beta_2 \sin(\theta_i)$  as link function and a uniform circular density as a weight. We can see that it works perfectly for simple patterns, i.e. when the one-labelled group and the zero-labelled one are well separated. It works a bit worse when the data are disturbed. Finally the logistic regression is very poor when the patterns are more complex. A way to avoid this problem is to *localize* the logistic regression introducing a non-uniform spatial weight. Here the weight function used is the von Mises kernel with a concentration parameter equal to 3. An example of the behaviour of the estimator obtained using expression (7) with  $p = 1$  is depicted in Figure 2.

On the KDE discrimination side, we consider two samples of  $n_1 = n_2 = 100$  observations (black and red rugs) labelled by 1 and 0 respectively. They have different means and equal variances: the first sample is drawn from a  $\nu M(1.5, 3)$  population, the second one is drawn from a  $\nu M(4.5, 3)$  population. We obtain two kernel density estimates selecting the concentration parameters by least squared cross-validation (LSCV). We assign to each observation the label of the population exhibiting the highest density at it. In this example we obtain a misclassification rate equal to 0.5%. See Figure 3.

## References

1. Di Marzio, M., Panzera, A., Taylor, C.C.: Local polynomial regression for circular predictors. *Statistics & Probability Letters*, **79**, 2066–2075 (2009)
2. Di Marzio, M., Panzera, A., Taylor, C.C.: Kernel density estimation on the torus. *Journal of Statistical Planning and Inference*, **141**, 2156–2173 (2011)
3. Di Marzio, M., Taylor, C.C.: Kernel density classification and boosting: an  $L_2$  analysis. *Statistics and Computing*, **15**, 113–123 (2005)
4. Signorini, D.F., Jones, M.C.: Kernel Estimators for Univariate Binary Regression. *Journal of the American Statistical Association*, **99**, 119–126 (2004)

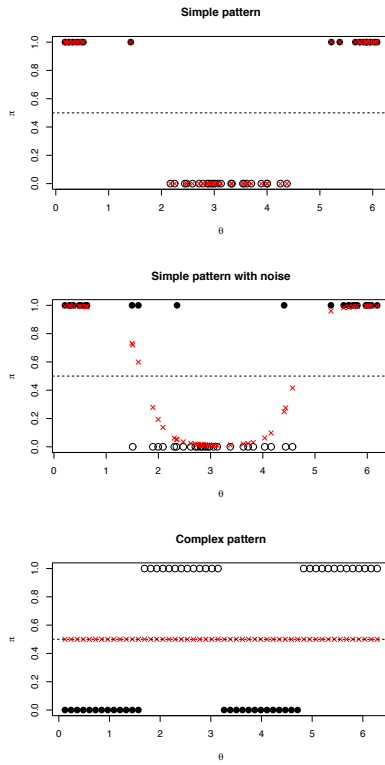
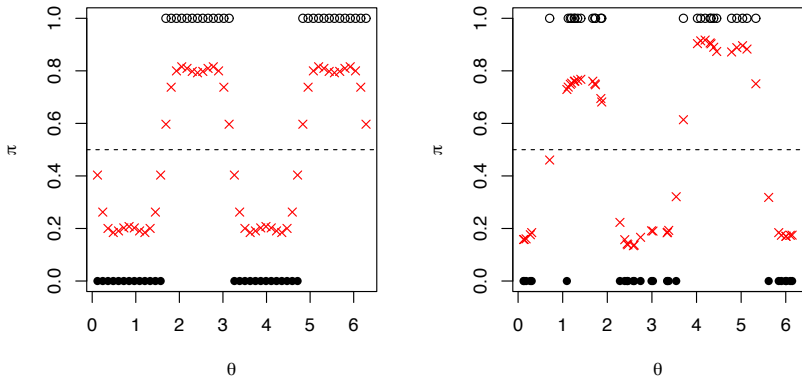
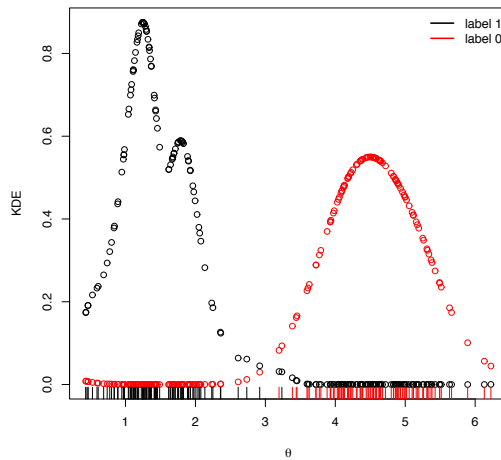


Fig. 1 Examples of Parametric circular logistic regression



**Fig. 2** Examples of circular local linear polynomial logistic regression



**Fig. 3** Example of circular discrimination using KDE between samples drawn from  $vM$  populations with different means and equal variances. The two bandwidths have been automatically selected using LSCV.



# Introduction to Symbolic Data Analysis and application to post clustering for comparing and improving clustering methods by the Symbolic Data Table that they induce

Edwin Diday

**Abstract** First we recall that Symbolic Data Analysis (SDA) is a way of thinking by classes in Data Science. We recall that classes of standard units are in SDA the new statistical units of higher level than the initial standard statistical units. In SDA classes are considered as objects to be described in all their facets by “symbolic data” taking care on their internal variability by staying close of the user language. Then we focus on different strategies of building a Symbolic data table from a standard data table by using: clustering (k-means, dynamic clustering), Fuzzy clustering (by EM, others), mixture decomposition of Copulas (by a “copula-EM” or a “copula-dynamic clustering”). Few words will be said also on how building classes at the second level (where the units are classes), by using Dirichlet models. Then, we give tools in order to measure the quality of the obtained symbolic data tables. By this way we can compare the different associated clustering methods and improve them. Finally, we show how to summarize the obtained symbolic data tables by symbolic data and also show how to visualize and compare them and their associated clustering methods, for example by an extension of PCA to symbolic data tables directly (Diday 2015) or by using a Wassenstein distance (Irpino, Verde 2015, 2016).

**Key words:** Symbolic Data Analysis, classes, objects, clustering, symbolic data table, Principal component analysis of symbolic data.

---

<sup>1</sup> E. Diday, CEREMADE Paris-Dauphine University (France), email: diday@ceremade.dauphine.fr

## 1 Introduction

Standard data tables are defined by a set of statistical units described by standard (numerical or categorical) variables. Complex data are data not reduced to just a standard data table but at the contrary they are defined by several data tables with different statistical units and different variables. Classes of statistical units are obtained in unsupervised learning by a clustering process allowing a concise and structured view on the data, in supervised learning, classes are used in order to extract efficient rules from the data.

A third way of thinking by classes consist in describing the classes (as « objects ») in their different facets by « symbolic data » in order to take care on their internal variability by intervals, bar charts, histograms, quantile functions, etc.. allowing the fusion of complex and big data in an explanatory framework. These symbolic descriptions leads to an extension of standard data tables (of numerical or categorical variables) to symbolic data tables (of symbolic value variables) and therefore to an extension of standard Data Analysis to Symbolic Data Analysis. One of the advantage of this approach is that unstructured data and unpaired samples at the level of row units, become structured and paired at the level of classes. Another advantage is that the symbolic class description has a high explanatory power as it is expressed in terms of the initial variables so in term of the user language. The study of such new type of data, built in order to describe classes in an explanatory way, has led to a new domain called “Symbolic Data Analysis” (SDA). Several international Journals as ADAC (first journal in Classification) and *Man And Cybernetics* has recently (2015, 2016), provided special issues on this topic. We can also mention several books (Bock, Diday (2000), Billard, Diday (2006), Diday, Noirhomme (2008)) and review papers (Billard, Diday (2003), Brito (2014), Diday (2016).

From the standard datable of the players described by their weight, nationality, number of goals, it is easy to build the Table 1 where each cell express the variability inside each class of the players of each team.

In the ground data table 1, the variability (of weight, nationality, number of goals) concerns the players of each team or of the team itself for the number of goals in a match. Therefore, by a fusion process (which transform the ground data table in a “symbolic data table”) each cell can contain: an interval (min-max, interquartile intervals or else of the weights of each team players), a sequence of categorical values (the nationality list of each team players), a sequence of weighted values as a bar chart (expressing the frequency of the number of goals in the match of a season), a histogram, a quantile function or simply a number (as the mean age of the team players, or the correlation between two variables as the correlation between the age and the weight inside each team), etc. This new kind of variables is called “symbolic” as they cannot be treated as numbers.

Any data set which cannot be considered as a unique data table of standard statistical units described by standard (numerical or categorical) variables is considered as a “complex data set”. This situation happen often and in many domains. For example, in the NSI (National Statistical Institutes), often the units to be studied are “regions”

characterized by different data tables of different units (hospitals, schools, inhabitants) and of different variables. A unique symbolic data table where the units are the regions described by symbolic data can be obtained by aggregation of the variables describing hospitals, schools and inhabitants.

**Table 1:** Teams are described by variables of symbolic values taking care on the variability of players inside each team for their weight and nationality and on the variability of the team itself for the number of goals in the match of a season.

TEAM	WEIGHT	NATIONALITY	NB OF GOALS
<b>BARSA</b>	[75 , 89 ]	{ Spane, Arg, ..}	{0.8 (0), 0.2 (1)}
<b>MANCHESTER</b>	[80, 95]	{Fr, Alg, Arg }	{0.1 (0), 0.3 (1), ...}
<b>PARIS-ST G.</b>	[76, 95]	{Fr, Tunisia }	{0.4 (0), 0.2 (1), ...}
<b>DORTMUND</b>	[70, 85]	{Fr, Engl, Arg }	{0.2 (0), 0.5 (1), ...}

## 2 Building symbolic data tables from clustering methods

By partitioning, the dynamic clustering method (DCM) and k-means is based on a “representation function” of any partition  $P = (P_1, \dots, P_k)$  called  $g$  such that  $g(P) = L$  with  $L = (L_1, \dots, L_k)$  and an “allocation function”  $f$  which associates to  $k$  representation  $L = (L_1, \dots, L_k)$  a partition  $P = (P_1, \dots, P_k)$ . The representation can associate to any class a distribution (Diday, Schroeder (1975)), a regression, factorial axis, points of the population (see Diday (1973)). For an overview see Diday, Simon (1979). In the special case where the representation function associates a mean to each class we get the K-means method as a case of DCM..

By using alternatively these two functions, this method converges towards a partition (i.e. a set of classes which covers the population and a representation associated to each class). The quality criterion can be written in the following way:

$W(P,L) = \sum_{i=1, k} w(P_i, L_i)$  where  $w$  measures the positive “fit” between each class and its representation and decreases when the fit increases. (For example, if  $L_i$  is a distribution, then  $w(C_i, L_i)$  can be the inverse of the likelihood of the class  $C_i$  for the distribution  $L_i$ ).

In the following we settle:  $P^{(n)} = (P^{(n)}_1, \dots, P^{(n)}_k)$  and  $L^{(n)} = (L^{(n)}_1, \dots, L^{(n)}_k)$ .

Starting from a partition  $P^{(0)}$ , the value of the sequence  $u_n = W(P^{(n)}, L^{(n)})$  decreases at each step  $n$  of the method. This can be proved in the following way.

First, by the reallocation of each individual  $x$  belonging in a class  $P^{(n)}_i$  in a new class  $f(L^{(n)}_i) = P^{(n+1)}_i$  such that the new classes  $P^{(n+1)}_i$  obtained by this reallocation improves the fit with  $L^{(n)}_i$  in order that we get  $w(P^{(n+1)}_i, L^{(n)}_i) \leq w(P^{(n)}_i, L^{(n)}_i)$  for  $i = 1, \dots, k$  which implies:

$\sum_{i=1, k} w(P^{(n+1)}_i, L^{(n)}_i) \leq \sum_{i=1, k} w(P^{(n)}_i, L^{(n)}_i)$  and therefore:

$W(P^{(n+1)}, L^{(n)}) \leq W(P^{(n)}, L^{(n)}) = u_n$

Second, by the representation process of each class  $P^{(n+1)}_i$ . In that way, we can define a new representation:  $L^{(n+1)} = (L^{(n+1)}_1, \dots, L^{(n+1)}_k)$  where  $L^{(n+1)}_i = g(P^{(n+1)}_i)$  fit better

$P^{(n+1)}_i$ , then  $L^{(n)}_i$ . This means that  $w(P^{(n+1)}_i, L^{(n+1)}_i) \leq w(P^{(n+1)}_i, L^{(n)}_i)$  for  $i = 1, \dots, k$  and therefore:

$u_{n+1} = W(P^{(n+1)}, L^{(n+1)}) \leq W(P^{(n+1)}, L^{(n)})$ . Hence, at this step, we get:

$u_{n+1} = W(P^{(n+1)}, L^{(n+1)}) \leq W(P^{(n+1)}, L^{(n)}) \leq W(P^{(n)}, L^{(n)}) = u_n$ . Therefore, as the positive sequence  $u_n$  decreases at each step, it converges.

Notice that in the reallocation process the allocation of an individual to a new class can be done at best, which means that it is allocated to the class which decreases the most the criterion. Notice also, that a simple condition in order that the sequence  $u_n$  decreases is that  $w(C, g(C)) \leq w(C, L), \forall L$  for any class  $C$  and any representation  $L$  of this class.

In partitioning by mixture decomposition, we can apply the DCM in the case where  $g(C)$  is the law which fit the best the class  $C$  and following a given model (Gaussian, Poisson, Gamma etc., see Diday, Schroeder (1975)). Its parameters are induced by a likelihood estimation. It is then easy to show that the sequence  $u_n$  converges until a partition and a vector of  $k$  laws such that each law fit at best its associated class.

The EM method (Dempster et al. (1977)) produces a fuzzy partitioning where each fuzzy class  $C_j$  (defined by a membership weight  $t_j(x_i)$ ) associated to each individual  $x_i$  is associated to a law of maximum likelihood of the mixture decomposition of the probability density of the population decomposed in  $k$  weighted density functions expressed in the following way:

$$f(x) = \sum_{j=1}^k p_j f(x, a_j),$$

with  $j = 1, k, \sum_{j=1}^k p_j = 1$  and  $0 < p_j < 1$  and  $f(x, a_j)$  is the probability density of parameter  $a_j$  and  $p_j$  is the likelihood estimator of the proportion of individuals following the density  $f(x, a_j)$  in the mixture.

The method aims to maximize alternatively the following likelihood equation:

$$L(x_1, \dots, x_n, a_1, \dots, a_k, p_1, \dots, p_k) = \sum_{i=1}^n \text{Log} \sum_{j=1}^k p_j f(x_i, a_j) \quad (1)$$

The method can be decomposed in two steps: "Estimation" and "Maximization". In the "estimation" step, for  $j = 1, k$  and  $i = 1, N$ , the  $t_j^n(x_i)$  which are the a posteriori probabilities that the individual  $x_i$  belongs to the class  $j$  at step  $n$  are determined by:

$$t_j^n(x_i) = p_j^n f(x_i, a_j^n) / \sum_{j=1}^k p_j^n f(x_i, a_j^n) \quad (2)$$

In other words  $t_j^n(x_i)$  is the fuzzy membership weight of the individual  $x_i$  to the fuzzy class  $C_j$  induced by the law  $f(\cdot, a_j^n)$ .

In the "maximization" step, first the weight associated to each law  $f(\cdot, a_j^n)$  is determined by:

$$p_j^{n+1} = \frac{1}{N} \sum_{i=1}^N t_j^n(x_i) \quad (3)$$

Then, the parameters  $a_j^{n+1}$  are obtained by the likelihood maximization of the equation (1), where  $l$  is the number of coordinates of the parameter  $a_j^{n+1}$ .

$$\forall j = 1, k, m = 1, l \sum_{i=1}^N t_j^n(x_i) \frac{\partial \text{Log} f(x_i, a_j^{n+1})}{\partial a_{jm}} \quad (4)$$

Notice that at the contrary of the DCM the clusters induced by the EM method by attributing to each individual, its best density function (i.e. the one of higher density value) does not follow this density function.

How to build the symbolic data table from a mixture decomposition?

By the analytical approach:

It is easy to build the marginal of the joints associated to each class by DCM or EM and then to obtain several kinds of symbolic variables describing each class. These symbolic variables can be built in order that their value be a density or a quantile or a distribution function, a histograms or a bar chart (depending on the kind of initial variable: numerical or categorical), a min-max or an inter-quartile intervals, a percentiles list etc. The graphical representation of the obtained densities functions can be obtained by a kernel method by using bandwidth which can allow a nice graphical representation inside the cells of the SDT (see Silverman (1986), Jones et al (2012). Histograms are another way to represent the obtained marginal of the joints with a good explanatory power. In both cases, the bandwidth for the density function and the intervals for the histograms has to follow two conditions. The first condition is that for a given variable they must be the same for all the classes (in order that they become comparable), the second condition is that the induced density function or histogram graphical representation must discriminate at best the classes. An efficient way in order to satisfy these two conditions in the case of histograms is given in Diday et al (2013).

By using the membership weight of the individuals to the classes

Another way to get the symbolic data table and the graphical representation of the marginal associated to each obtained joint, is to use the weights associated to each individual in each class. Then, by addition we can build the same kinds of symbolic variables. In case of a class  $C_j$  and an initial numerical variable  $Y_r$ , we obtain, a histogram denoted  $H_{jr}$ , such that:

$$H_{jr} = \left( \frac{\sum_{i=1}^N t_j(x_i) \delta_{I_1}(x_i^r)}{\sum_{v=1}^V \sum_{i=1}^N t_j(x_i) \delta_{I_v}(x_i^r)}, \dots, \frac{\sum_{i=1}^N t_j(x_i) \delta_{I_V}(x_i^r)}{\sum_{v=1}^V \sum_{i=1}^N t_j(x_i) \delta_{I_v}(x_i^r)} \right) \quad (5)$$

Where  $x_i^r = Y(x_i)$  and  $\delta(x_i^r)$  is a Dirac based vector, defined by a partition  $(I_1, \dots, I_V)$  of the numerical variable  $Y_r$  domain  $D_r$  in  $V$  intervals such that:  $\delta(x_i^r) = (\delta_{I_1}(x_i^r), \dots, \delta_{I_V}(x_i^r))$  where  $\delta_{I_v}(x_i^r)$  takes the value 1 if  $x_i^r \in I_v$  and 0 elsewhere.

In case of categorical variable with categorical values denoted  $I_1, \dots, I_V$ , we can obtain, in an analogous way, a symbolic variable with bar chart value. In this case,  $\delta_{I_v}(x_i^r)$  takes the value 1 if  $x_i^r$  takes the category  $I_v$  and 0 elsewhere.

Notice that the fuzzy partition produced by EM induces an exact partition  $(C'_1, \dots, C'_K)$  such that

$C'_k = \{x_i / f(x_i, a_k) \geq f(x_i, a_l)\}$ . In case of an exact partition obtained from EM or from DCM we have  $t_j(x_i) = 1$ , for any  $j = 1, k$ , and  $i = 1, N$  in the formula (5), from which we can built a histogram (resp. bar chart valued variables). from numerical (resp. categorical variables). In this case the result is biased for EM (resp. not biased for DCM) as it does not correspond (resp. correspond) to the EM (resp. DCM) obtained densities. At the contrary, we can use the formulas (5) for EM and for DCM, we can replace  $t_j(x_i)$  by  $w(\{x_i\}, L_j)$  in formulas (5) (where  $w(\{x_i\}, L_j)$  measure the fit between the individual  $x_i$  and the representation  $L_j$ , see section

4.1.1) of the class  $C_j$ . In this case the result is biased for DCM (resp. not biased for EM) as it does not correspond (resp. correspond) to the DCM (resp. EM) obtained densities.

Copulas have already been used in SDA at the level of a symbolic data table as input (see Vrac and al. (2011)). Here, our aim is to use copulas on the initial standard data table in order to produce a SDT. The advantage of this copulas approach is that we obtain directly the marginal law associated to each class which then induce directly the symbolic data table (and not in two steps as in the preceding clustering methods). By partitioning, we can apply the DCM in the case where  $g(C)$  is the copulas which fit the best the class  $C$  and following a given copulas model (Gaussian, Frank, etc.)

By fuzzy partitioning, the model is the following:

$f(x_i) = \sum_{j=1}^k p_j \text{Cop}_{ac_j}(f(x_{i1}, a_{1j}), \dots, f(x_{iq}, a_{qj}))$ , where here  $q$  is the number of initial variables and  $a = ((ac_1, a_{11}, \dots, a_{q1}), \dots, (ac_k, a_{1k}, \dots, a_{qk}))$  defines the parameters of the mixture where  $ac_j$  is the parameter of the copulas model associated to the fuzzy class  $j$ .

The EM algorithm can then be extended to a Copulas-EM algorithm by using in the formulas (1) to (4),  $\text{Cop}_{ac_j}(f(x_{i1}, a_{1j}), \dots, f(x_{iq}, a_{qj}))$  instead of  $f(x_i, a_j)$  for  $i = 1, N$  and  $j = 1, k$ .

Having obtained the marginal  $f(x_{is}, a_{sj})$  for  $i = 1, N$ ,  $s = 1, q$  and  $j = 1, k$ , we can build (as in the preceding clustering methods) a symbolic data table by many kinds of symbolic variables (density functions, histograms, min-max or inter-quartile intervals, percentiles list, etc.).

In the case where the initial data are defined by a symbolic data table, we start from a SDT where each row describe a class of individuals of the ground data table. By using random variables with symbolic values following for example a Dirichlet model, all the clustering methods presented in this section can be extended to this more general situation by transforming the description of each individual of the ground data table in a Dirac mass vector as in the preceding section. Notice that as we star from a SDT, the obtained marginal are laws of laws, therefore in order to get an explanatory SDT in the case of a Dirichlet-EM clustering method which leads to  $k'$  clusters of classes, it is better to use the sum  $\sum_{j'=1}^{k'} t_{j'}(C_j) h_r^{j'}$  where  $h_r^{j'}$  is the symbolic value (a histogram, for example) of the  $r$  th symbolic variable for the class  $C_j$  and  $t_{j'}(C_j)$  is the fuzzy membership weight of the class  $C_j$  to the fuzzy cluster of classes denoted  $C_{j'}$  for  $j' = 1, k'$ . Hence, in the obtained SDT,  $\sum_{j'=1}^{k'} t_{j'}(C_j) h_r^{j'}$  is the value of the  $r$  th symbolic variable for the row associated to the cluster  $C_{j'}$  of classes. Another kind of fuzzy clustering in case of initial symbolic distributional variables can be found in Irpino et al (2017).

### **3 Comparing and improving the clustering methods by using quality criteria of the Symbolic Data Table (SDT) that they induce.**

How to measure the explanatory power of a SDT?

The explanatory power of a clustering method can be measured from the explanatory power of the SDT that they induce. Basically, the more the rows of a SDT are different the higher is its explanatory power. Many other kinds of quality criterion of a SDT can be used as the entropy of the symbolic values and the correlation between the symbolic variables (see for example in Diday (2013, 2016)). A general way is to use the dissimilarities two by two between the symbolic descriptions of the classes. These dissimilarities can be adapted to each kind of variable (Wassenstein between quantiles functions, Hausdorff between intervals, etc., see Diday, Noirhomme (2008) for other dissimilarities between symbolic data). In that way we can associate to any SDT a set of dissimilarities values two by two between its rows. This set of numbers, which are between 0 and 1 is denoted  $D$ .

By this way, we can describe each clustering method by a set of symbolic variables (induced from  $D$ ), which values can be a probability density, a quantile function, a histogram, an inter-quartile interval, a set of percentile values, etc.. Finally, we get a SDT where each row is associated to a clustering method (and its induced SDT) and each column is associated to a symbolic variable induced by its set  $D$ . We can then compare the efficiency of the different clustering methods on a given ground data table, in term of the explanatory power or other qualities of the obtained SDT. Several SDA methods can be used, for example, a visualization allowing a positioning of the different clustering methods by an extension of PCA to symbolic data directly (Diday 2013) or by using a Wassenstein distance (Irpino, Verde 2016).

How improving the explanatory power of clustering methods by using the SDT?

Often clusters are overlapping and it is difficult to say in which cluster to allocate an individual which is at the bridge between two clusters. We can improve the explanatory power (or other quality criteria) of the SDT associated to obtained clusters by reallocating iteratively the individuals by improving at each step the explanatory of their associated SDT, until convergence. More precisely, having a partitioning as starting point, each individual can be allocated to the class which symbolic description improve the chosen quality criterion of the symbolic data table. The new classes can be described by new symbolic data, we can then reallocate each individual and so on until convergence.

## **4 Conclusion and perspectives**

In practice most of the given data contain a class variable from which an SDT can be build and analyzed by SDA. In this paper, we are interested in the case where the

classes are not given. Therefore, several kinds of clustering methods have been recalled or settled and we have shown how they yield to a SDT which enhance their interpretation, allows their comparison and improve them by several quality criterion (as their explanatory power). This paper opens several direction of research which need to be deeply studied and experimented.

## References

1. Bock H., Diday E. (2000). Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data. Springer Verlag, Heidelberg, 425 pages. L. Billard, E. Diday (2003). From the statistics of data to the statistic of knowledge: Symbolic Data Analysis. JASA . Journal of the American Statistical Association. Juin, Vol. 98, N° 462.
2. Billard L., Diday E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining". Wiley Series in Computational Statistics. Chichester: Wiley; 2006, 321. ISBN: 0-470-09016-2.
3. Brito P. (2014): Symbolic data analysis: another look at the interaction of data mining and statistics. Wiley Rev Data Mining Knowl Discov 2014, 4:281–295. doi:10.1002/widm.1133.
4. Dempster A., Laird N., Rubin D. (1977): Maximum likelihood from incomplete data via the EM algorithm. JRSS, B39.
5. Diday E, Noirhomme-Fraiture (2008). Symbolic Data Analysis and the SODAS software. Chichester: Wiley; 2008. Doi: 978-0-470-01883-5.
6. Diday E. (1973) "The Dynamic clusters method in non-hierarchical clustering". International journal of Computer and information Science. Vol 2, n° 1, 1973. DOI: 10.1007/BF00987153.
7. Diday E., Schroeder A. (1975): A new approach in mixed distributions detection. RAIRO v. 10 n°6.
8. Diday E., Simon J.C., 1979: "Clustering Analysis" Chapter in "Communication and Cybernetics Digital Pattern Recognition", K.S. FU edit. Springer Verlag.
9. Diday, E. (2013): Principal component analysis for bar charts and metabins tables. SAM (Statistical Analysis and Data Mining), Volume 6 Issue 5, October 2013 . Pages 403-430 . John Wiley & Son.
10. Diday E. (2016): Thinking by classes in Data Science: the symbolic data analysis paradigm. WIREs Comput. Stat 2016, 8: 172–205. Doi: 10.1002/wics.1384.
11. Jones M.C., Marron, J.S., Sheater (2012). A brief survey of Bandwidth selection for density estimation. Pages 401-407 JASA, Volume 91, 1996-issue 433. Published on line 27 Febr 2012.
12. Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC. p. 48. ISBN 0-412-24620-1.
13. Lebre R., Iovleff S., Langrognet F., Biernacki C., Celeux G., Govaert G. (2015). Rmixmod: The R Package Supervised Classification Mixmod Library. Journal of Statistical Software, 67(1), 1-29.
14. Vrac M., Billard L., Diday E., Chédin A. (2011): Copulas Analysis of mixture model. Computer Statistics. DOI 10.1007/s00180-011-0266-0
15. Diday E., Afonso F., Haddad R. (2013): The symbolic data analysis paradigm, discriminant discretization and financial application. HSDSA 2013, vol. RNTI-E-25, pp.1-14.
16. Irpino A., Verde R., De Carvalho F. A.T. (2015): Fuzzy clustering of distributional data with automatic weighting of variable components. Information Sciences 406–407 (2017) 24 8–26 8.
17. Verde R., Irpino A., Balzanella A. (2016): Dimension Reduction Techniques for Distributional Symbolic Data. IEEE Man and Cybernetics, Page(s): 340-355.



# Identifying Meta Communities on Large Networks

## *L'identificazione di meta comunità in reti di grandi dimensioni*

Carlo Drago

**Abstract** On large networks there is a specific need to consider specific patterns which can be related to structured groups of nodes which could be also defined as communities. In this sense we will propose an approach to cluster the different communities using interval data. This approach is relevant in the context of the analysis of large networks and in particular on discovering the different functionalities of the communities inside a network

**Abstract** *In networks di larghe dimensioni esiste la necessita di considerare strutture dati relativi a gruppi di nodi definite comunità. In questo lavoro proponiamo un approccio di clustering su rappresentazioni delle comunità basate su dati ad intervallo. Questo approccio e' rilevante nel contesto di networks di larghe dimensioni al fine di identificare le diverse funzionalità nella rete stessa.*

**Key words:** Social Network Analysis, Community Detection, Symbolic Data, Clustering

There are important cases in which it could be very important to cluster the communities of a network. For example an important case is explained by Fortunato 2010 [6]: different communities are associated with different behaviors on the network. The different nodes on the community can be associated to a specific function or behavior inside the network. In order to predict the future behavior of the different nodes it could be crucial to determine the different communities and to understand the different patterns of similarity it is possible to observe. In this sense understanding the concept of community on a network leads to a better understanding of the network behavior as a whole (see Coscia et al. 2011 [3]). So it is necessary to represent adequately the community to understand the entire network. The problem of the adequate representation is particularly relevant on big data. In this sense we have to decide the best representation to use. Clustering a community means taking into account on the clustering process all the structural features and the communities and the attributes of the different nodes considered as a whole. In this sense we can

---

University of Rome "Niccolò Cusano", via Don Carlo Gnocchi 3, e-mail: carlo.drago@unicusano.it

consider the structural features and the attributes of the nodes (which characterize the community). In order to cluster communities it is necessary to adequately represent the problem of representing the community structure of a network (Drago 2015 [5]). In particular it is relevant to find the nodes that show similar characteristics to each other and the nodes loosely connected with other nodes belonging to other communities (see Girvan and Newman 2002 [8]). At the same time it is particularly important to propose an approach which could be based on interval data because we want to consider the entire community (on the use of symbolic data in network analysis see Giordano and Brito 2014 [7]). Communities are a very relevant object to consider. In fact, on a specific network, the different vertices tend to react as a whole and so it could be relevant to cluster them as a whole.

## 1 Community Identification

The first step on the analysis is based on the need to determine the different communities inside the network. In order to determine the different communities we need to consider an appropriate community detection algorithm (Zhao et al. 2011 [14] and Blondel et al. 2008 [2]). We start from a network  $G$  defined as:

$$G = (V, E) \quad (1)$$

Typically the community detection methods tend to focus on the connections among the different nodes that are part of the same community. There are cases in which the different nodes tend not to fit with the communities identified. The general assumption of these methodologies is that there is a direct emphasis on considering the ties inside the communities, rather than the ties which connect members of different communities (Zhao et al. 2011 [14]). Usually the relevant requirement for detecting a community is connectedness. In particular we can expect a strong connection among the nodes that are part of the community. For detecting communities we need to take into account the modularity which can measure explicitly the capacity of a network to be divided into different modules (Blondel et al. 2008 [2]). At the same time the modularity allows the identification of the different communities. The modularity (see Newman 2006 [10]) needs to be computed by considering a null model not considering the community structure of the structure (i.e. a random graph). So following Fortunato 2010 [6] we can define the modularity in this way:

$$Q = \frac{1}{2m} \sum_{i,j} (A_{i,j} - K_{i,j}) \gamma(C_i, C_j) \quad (2)$$

where  $m$  is the number of the edges on the network,  $A$  is the adjacency matrix considered and  $K_{i,j}$  is the number of edges which can be considered between the vertices  $i$  and  $j$  on the null model. Finally it is possible to consider the  $\gamma$  function which returns two possible values: 1 where the two vertices  $i$  and  $j$  belong to the same community and 0 if they belong to different communities. However in order

to take into account also the degree of the vertices  $i$  and  $j$  (it is important to consider the degree distributions) we can write the modularity as follows (Fortunato 2010 [6] and Newman 2006 [10]) :

$$Q = \frac{1}{2m} \sum_{i,j} (A_{i,j} - \frac{k_i k_j}{2m}) \gamma(C_i, C_j) \quad (3)$$

where  $k_i$  and  $k_j$  are degree values for different vertices. In this way we obtain the different community inside the different network. A different approach in this sense is the one followed by Reichardt and Bornholdt [12] which introduces a different approach on null models and a general one (see Newman 2006-2 [11]). These communities are important because they are a stylized way to represent the different structure of the network. In this sense we need to take in to account the entire groups of nodes as a whole in order to cluster the communities considered entirely. In this sense we consider all the nodes singularly by considering their statistical characteristics. From the different communities we can start to cluster them by building an adequate data matrix. In particular we have to consider for each community the measurements for the taking into account of the different intervals.

## 2 K-Means Clustering of the Communities

The different communities are characterized by a vector with the different  $n$  observations related to the vertice for the same variable or attribute. So we can have:

$$X^b = (x_1, x_2, \dots, x_n) \quad (4)$$

We can write the interval data (the measurement for the network community) in this way:

$$X^{I,b} = [\bar{x}, \underline{x}] \quad (5)$$

Each interval of the considered variable represents a measurement for the single community. In this way we obtain for each different community the measurements relating to the single vertices, but at the same time observations related to the different communities (represented by the intervals). It is important to note that the intervals are at the same time characterized by their radii and the midpoints. In this case each community can be also represented by a midpoint value. In particular it is possible to obtain a value of the interval midpoint for the generic variable  $b$ :

$$X_{center}^{I,b} = \frac{1}{2}(\underline{x} + \bar{x}) \quad (6)$$

And the radius of the interval:

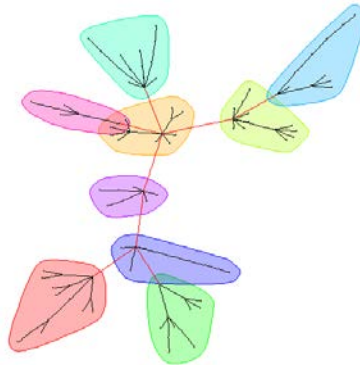
$$X_{radius}^{I,b} = \frac{1}{2}(\bar{x} - \underline{x}) \quad (7)$$

In order to cluster the different communities we depart from the measurement of the community characteristics by using interval data. Many different clustering algorithms were proposed in order to cluster interval data. In particular interval clustering was firstly considered.

### 3 Simulation Study

We can start from considering different networks simulated using different characteristics. In this sense we consider different networks obtained by using the R package *igraph* (see Csardi Nepusz 2006 [4]). In order to perform the data analysis at community level we have also used the package *RSDA* (Rodriguez (2004) [13]). In this sense we consider different groups of networks (for example: Barabasi Albert graph models Barabasi Albert 1999 [1])

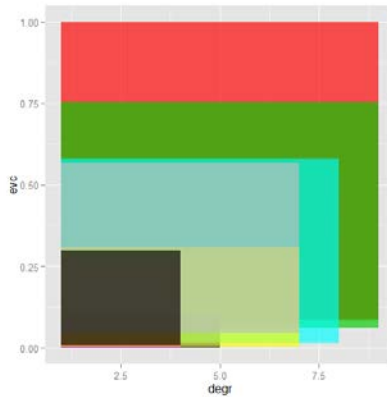
We have considered many different networks in order to test the approach on different structures. Here we present some different examples we have obtained from the simulation study. The results obtained on the examples are important in order to derive some interpretation rules which can be considered on the results of the proposed approach. In the first case we consider an example from the network based on Barabasi game. We consider on the example a network based on 100 nodes. We obtain 6 communities as part of the community structure.



**Fig. 1** Barabasi Model Simulation using 100 nodes: Community Structure

From the clustering analysis we can observe that there is a group or cluster of communities at the center which shows similar structural characteristics. This could be also observed by considering the different interval scatterplot diagrams. That

means we are able to identify some groups of communities which tend to show similar characteristics.



**Fig. 2** Barabasi Model Simulation using 100 nodes: Community Structure visualized. On the x-axis the degree, on the y-axis the eigenvector centrality scores

Then we consider the 9 different communities obtained on the clustering process using the K-Means. In this case we consider different data matrices using different specifications of the data matrices in order to evaluate the results using different data matrices and using different relevant structural measures. Two interpretative results need to be noted: on the simulations it is possible to observe that the lower bound seems not so relevant. In particular the upper bound is relevant to discriminate the different communities. In this sense by the differences which can be observed by the different intervals can be determined specifically the differences on the upper bounds. The visualization shows an overlapped structure because there is a centralized structure of the network. This structure tends to cluster specifically the communities in a central position. In this case the betweenness is related to the higher degree. It is also possible to note that the different nodes can be characterized to different groups of similar nodes when they are considered specifically on the communities. In fact it is possible to note it is different to consider the nodes as part of a community or the nodes singularly.

## 4 Conclusions

The results we have obtained confirm the usefulness of the approach on large networks. In particular the result is useful to determine specifically the community structure and some different meta-communities which can be identified on a specific

network. By starting from the meta communities (along the definition of Kalinka and Tomancak, (2011) [9]) it is possible to obtain the different prototypes. In particular a relevant insight related to the results is that in the case of the clustering communities as groups of nodes we can obtain different results from those considering the clustering of the single node. In this sense the analysis can be enriched by the fact that in some cases the nodes have on their communities relevant dissimilarities which need not be taken into account when the analysis is performed by considering the communities as a whole. At the same time clusters of communities (or meta-communities) can be characterized as behavior by nodes which participate in the community on different levels. The approach considered in this work allows the exploration of these levels of interaction between the different nodes.

## References

1. Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439), 509-512.
2. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
3. Coscia, M., Giannotti, F., & Pedreschi, D. (2011). A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(5), 512-546.
4. Csardi G, Nepusz T (2006) The igraph software package for complex network research, *InterJournal, Complex Systems* 1695. 2006. <http://igraph.org>
5. Drago C. (2015) Exploring the Community Structure of Complex Networks. *Annali del MEMOTEF - Note e Discussioni* 10/2015; 2(forthcoming).
6. Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3), 75-174.
7. Giordano G., Brito P. (2014) Social Networks as Symbolic Data, in Vicari D., Okada A., Ragozini G., Weihs C. Eds., *Analysis and Modeling of Complex Data in Behavioral and Social Science*, Springer: Heidelberg, pp. 133-142;
8. Girvan, M., & Newman, M.E.. (2002). Community Structure in Social and Biological Networks.
9. Kalinka, A. T., & Tomancak, P. (2011). linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics*, 27(14).
10. Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577-8582.
11. Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3), 036104.
12. Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(1), 016110.
13. Rodriguez O.R. with contributions from Olger Calderon and Roberto Zuniga (2014). RSDA: RSDA- R to Symbolic Data Analysis. R package version 1.2. <http://CRAN.R-project.org/package=RSDA>
14. Zhao, Y., Levina, E., & Zhu, J. (2011). Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18), 7321-7326.

# Random Forest-Based Approach for Physiological Functional Variable Selection for Driver's Stress Level Classification

Neska El Haouij, Jean-Michel Poggi, Raja Ghozi, Sylvie Sevestre Ghalila, and Mériem Jaïdane

**Abstract** With the increasing urbanization and technological advances, urban driving is bound to be a complex task that requires higher levels of alertness. Thus, the drivers mental workload should be optimal in order to manage critical situations in such challenging driving conditions. Past studies relied on drivers performances used subjective measures. The new wearable and non-intrusive sensor technology, is not only providing real-time physiological monitoring, but also is enriching the tools for human affective and cognitive states monitoring. This study focuses on a drivers physiological changes using portable sensors in different urban routes. Specifically, the Electrodermal Activity (EDA) measured on two different locations: hand and foot, Electromyogram (EMG), Heart Rate (HR) and Respiration (RESP) of ten driving experiments in three types of routes are considered: rest area, city, and highway driving issued from physiological database, labelled *drivedb*, available online on the PHYSIONET website. Several studies have been done on driver's stress level recognition using physiological signals. Classically, researchers extract expert-based features from physiological signals and select the most relevant fea-

---

Neska El Haouij  
CEA-LinkLab, Tunisia & Tunis El Manar University, Tunisia & Paris Sud University, France  
e-mail: elhaouij.nsk@gmail.com

Jean-Michel Poggi  
Paris Descartes University, France & Paris Sud University, France  
e-mail: Jean-Michel.Poggi@math.u-psud.fr

Raja Ghozi  
Tunis El Manar University, Tunisia & CEA-LinkLab, Tunisia  
e-mail: rjghozi@yahoo.com

Sylvie Sevestre Ghalila  
CEA-LinkLab, Tunisia  
e-mail: Sylvie.SEVESTRE-GHALILA@cea.fr

Mériem Jaïdane  
Tunis El Manar University, Tunisia & CEA-LinkLab, Tunisia  
e-mail: meriem.jaidane@planet.tn

tures in stress level recognition. This work aims to apply a random forest-based method for the selection of physiological functional variables in order to classify the stress level during real-world driving experience. The contribution of this study is twofold: on the methodological side, it considers physiological signals as functional variables and adapts a procedure of data processing and variable selection. On the applied side, the proposed method provides a "blind" procedure of driver's stress level classification that do not depend on the expert-based studies of physiological signals.

**Key words:** Random Forests, Variable Selection, Functional Data, Physiological Signals

## 1 Introduction

This paper aims to provide a random forests-based method for the selection of physiological functional variables in order to classify the stress level experienced during real-world driving. For that, we present first the context of our work which concerns the affective computing aspects with a summary of the study introducing the physiological database *drivedb*. Then, methods on functional data, variable selection using random forests and grouped variables importance are addressed. The contribution of this study is twofold: on the methodological side, it adapts the scheme proposed by [6] to take advantage of the functional nature of the physiological data and offers a procedure of data processing and variable selection. On the applied side, the proposed method provides a blind (i.e. without prior information) procedure of driver's stress level classification that does not depend on the extraction of expert-based features of physiological signals. This allows automatic exploration of promising signals to be included in statistical models for driver's state recognition.

## 2 Stress level recognition while driving

Many research groups tried to provide solutions and tools to vehicles and roadway users in order to improve safety, efficiency and quality in the sector of transport. [14] points out that according to the American Highway Traffic Safety Administration, high stress levels impact negatively drivers reactions especially in critical situations. It is one of the most prominent causes of vehicle accidents such as intoxication, fatigue and aggressive driving. In real world driving, human affective state monitoring can offer useful information to avoid traffic incidents and provide safe and comfortable driving.

With the increasing urbanization and technological advances, the new wearable and non-intrusive sensor technology, is not only providing real-time physiological monitoring, but also is enriching the tools for human affective and cognitive states



monitoring. In particular, several studies have been reported the last years in the field of driver's stress monitoring. In this paper, base our analysis on the study of [10] where they presented a protocol of physiological data collection in real-world driving conditions in order to detect stress levels. Specifically, physiological signals such as Electrodermal Activity (EDA), Electrocardiogram (ECG), Electromyogram (EMG) and Respiration (RESP) were captured for 24 driving experiences.

Features derived from non-overlapping segments of physiological signals taken from rest, highway and city of the driving experiences. The first analysis aiming to classify the stress levels allows to distinguish between the three levels of driver stress with an accuracy of 97%. The second analysis concerns the study of the correlation between extracted features from physiological signals and a stress levels metric created from the video tape. In this study, [10] reported that there is a correlation between driver's affective state quantified by the stress levels metric and the physiological signals, the highest correlation is with the EDA and HR. They have partially released their physiological database, labeled "*drivedb*", on-line on the PhysioNet website<sup>1</sup>. The data used in our work were extracted from the *drivedb* database which has a clear annotation of the several driving periods for each experience, allowing an easy exploitation of the information. Apart its availability on-line, various studies were based on this database which constitutes a main reference on stress level recognition in highway and city driving.

### 3 Functional Variable Selection

The main issue of variable selection methods is their instability where a set of selected variables may change when perturbing the training sample. The most widely used solution to solve this instability consists in using bootstrap samples where a stable solution is obtained by aggregating selections achieved on several bootstrap subsets of the training data. Random forests algorithm, introduced by [1], is one of these methods based on aggregating a large collection of tree-based estimators. These methods have good predictive performances in practice and they work well for high dimensional problems. Their power is shown in several studies summarized in [15]. Moreover, random forests provide several measures of the importance of the variables with respect to the prediction of the outcome variable. It has been shown that the permutation importance measure introduced by Breiman, is an efficient tool for selecting variables ([2, 5, 7]).

The standard approach in Functional Data Analysis (FDA) (see for example [13, 3]) consists in projecting the functional variables into a space spanned by a functional basis such as splines, wavelets, Fourier. Several regression and classification methods were the focus of studies in two situations: with one functional predictor and recently for several functional variables.

---

<sup>1</sup> <http://physionet.org/>

Classification based on several possibly functional variables has also been considered using the CART algorithm for similar driving experiences in the study of [12], using SVM in [16] work. Variable selection using random forests was achieved in the study of [4]. In our study, multiple FDA using random forests and the grouped variable importance measure proposed by [6] are used.

### ***3.1 Variable Selection using Random Forest-based Recursive Feature Elimination***

In this study, Random Forests-based Recursive Feature Elimination (RF-RFE) is used. The RF-RFE algorithm, proposed by [6], was inspired from [9] introducing Recursive Feature Elimination algorithm for SVM (SVM-RFE). At the first step, the dataset is randomly split into a training set containing two thirds of the data and a validation set containing the remaining one third. The procedure fits the model to all explanatory variables using Random Forests. Then, the variables are ranked using their importance measure. The grouped VI is computed only on the training set. The less important predictor is eliminated, the model is refit and the performance is assessed by a prediction error computed on the validation set. The variable ranking and elimination is repeated until no variable remains. The final model is chosen by minimizing the prediction error. It should be noted that at each iteration, the predictors importance is recomputed on the model composed by the reduced set of explanatory variables.

In the case of functional variables, the selection is performed using the algorithm on two different type of groups, thanks to the definition of importance of groups of variables. This allows to consider a group of variables as a whole, for example the group of the wavelet coefficients of a given signal, and to quantify its relative importance with respect to the other functional variables.

### ***3.2 Our procedure: Variable selection using iterative RF-RFE***

The proposed approach in this work aims to first eliminate the irrelevant physiological variables in the stress level classification task and then select among each kept variable the most relevant wavelet levels. In this study, the number of variables is very large (20480), compared to the number of the observations (68), thus the procedure is not stable. In order to reduce the variability of the selection, the procedure is repeated 10 times.

### 3.3 Variable selection results

The objective of variable selection is first to eliminate physiological signals that do not contribute significantly in the stress level classification, then for the retained physiological variables, the most relevant wavelet levels will be selected.

When applying our procedure to the *drivedb* database, we perform at a first stage functional variables decomposition using the Haar wavelet which is considered as the simplest one. We pick 12 as the decomposition level which corresponds to the maximum level compatible with the  $4096 = 2^{12}$  samples.

To achieve this work, we use the R software, with the `randomForest` package proposed by [11] and `RFgroove` packages developed by [8].

The proposed “blind” approach performs as the expert-based approach in terms of misclassification rate. This procedure offers moreover, additional information such as the physiological variables ranking according to their importance and the list of the relevant variables in stress level classification. The obtained results suggest that *EMG* and the *HR* are not very relevant when compared to the *EDA* and the respiration signals. This may help to investigate the list of physiological sensors that can be proposed to the smart vehicles designers, in order to determine the stress level.

**Acknowledgements** Jean-Michel Poggi, and all the authors, thank the organizers for the invitation to present this paper. The authors mention that the major part of the results of this presentation have been submitted for publication.

## References

1. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
2. R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):1–13, 2006.
3. F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
4. R. Genuer, J.-M. Poggi, and Tuleau-Malot. *Vsurf*: An r package for variable selection using random forests. *The R Journal*, 7(2):19–33, 2015.
5. R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225 – 2236, 2010.
6. B. Gregorutti, B. Michel, and P. Saint-Pierre. Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics and Data Analysis*, 90:15–35, 2015.
7. B. Gregorutti, B. Michel, and P. Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, pages 1–20, 2016.
8. Gregorutti, B. *RFgroove*: Importance measure and selection for groups of variables with random forests. <https://CRAN.R-project.org/package=RFgroove>, R package version 1.1, (2016), 2016.
9. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, March 2002.

10. J.-A. Healey and R.-W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, June 2005.
11. Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.
12. J.-M. Poggi and C. Tuleau. Classification of objectivization data using cart and wavelets. *Proceedings of the IASC 07, Aveiro, Portugal*, pages 1–8, 2007.
13. J.-O. Ramsay and B.-W. Silverman. *Functional Data Analysis*. Springer-Verlag New York, 2005.
14. R.-G. Smart, E. Cannon, A. Howard, P. Frise, and R.-E. Mann. Can we design cars to prevent road rage? *International Journal of Vehicle Information and Communication Systems*, 1(1-2):44–55, 2005.
15. A. Verikas, A. Gelzinis, and M. Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330–349, 2011.
16. K. Yang, H. Yoon, and C. Shahabi. A supervised feature subset selection technique for multivariate time series. *Proceedings of the Workshop on Feature Selection for Data Mining: Interfacing Machine Learning with Statistics*, pages 92–101, 2005.

# A risk index to evaluate the criticality of a product defectiveness

## *Un indice di rischio per variabili ordinali: un'applicazione nell'ambito del controllo della qualità*

Silvia Facchinetti and Silvia A. Osmetti

**Abstract** We propose a risk index naturally suitable in quality-control framework characterized by data often collected on ordinal scale, to measure the risk of failure of a product. A so-called *Severity Index* is defined on the basis of the relative frequencies of the ordinal variables. We examine the distribution and the statistical properties of its estimator. We apply the index to real data concerning the severity and the occurrence of defectiveness of the products of a multinational corporation manufacturer. Our index may be employed to communicate the level of risk, to compare among different risks and to identify interventions in the production system.

**Abstract** *Nel presente lavoro proponiamo una misura sintetica di rischio per variabili ordinali basata su frequenze relative. Tale indice risulta particolarmente adatto nell'ambito del controllo della qualità al fine di valutare il rischio di difettosità di un prodotto. In tale ambito infatti le informazioni disponibili sono spesso di natura ordinale. Noi analizziamo le caratteristiche dell'indice, proponiamo un suo stimatore corretto e consistente e studiamo la sua distribuzione asintotica. La nostra proposta viene applicata ai dati di una compagnia multinazionale, riguardanti la gravità delle diverse tipologie di difetti rilevati sulle componenti di un prodotto assemblato. L'indice calcolato consente di definire il livello di rischiosità dei componenti utile per programmare opportuni interventi sul sistema produttivo.*

**Key words:** risk index; categorical variables; failure modes and effects analysis.

---

Silvia Facchinetti

Department of Statistical Science, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Milano,  
e-mail: [silvia.facchinetti@unicatt.it](mailto:silvia.facchinetti@unicatt.it)

Silvia A. Osmetti

Department of Statistical Science, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Milano,  
e-mail: [silvia.osmetti@unicatt.it](mailto:silvia.osmetti@unicatt.it)

## 1 Introduction

One of the ways to measure and communicate the level of risk is through risk indices, that are based on a synthesis of quantitative or qualitative information. For a discussion on risk measures see e.g. [5].

For calculating the risk, often the companies employ approaches based on categorical data expressed on ordinal scale that are improperly considered as quantitative. In this context we discuss on a synthetic measure of risk available for data expressed on ordinal scale, called *Severity Index* ( $S$ ). It is defined on the basis of the relative frequencies of the ordinal variables. We study the distribution and the statistical properties of its estimator.

This index appears naturally suitable to provide a measure of risk of failure of a product in quality-control framework in testing and recalling phases of products or in similar situations where the quality is expressed on ordinal scale. More precisely, our aim is to propose a synthetic priority of intervention indicator for a product, based on the frequencies of specific ordinal variables used to measure the quality and the reliability of such product.

Other authors have proposed measures of risk for ordinal data. Figini and Giudici [2] propose a non parametric measure of operational risk for ordinal variables in a Bayesian framework. Figini *et al.* [3] use optimal scaling techniques to reduce the dimensionality of ordinal variables describing the service quality to a continuous score interpretable as a measure of operational risk. Cerchiello *et al.* [1] propose rank based models to asses perceived quality of academic teaching.

## 2 The Severity Index

Let  $X \sim \{x_j, p_j\}$  for  $j = 1, 2, \dots, K$  be a categorical random variable (r.v.) representing the level of quality/defectiveness of a product with ordered categories  $x_j$  and probabilities  $p_j = P(x_j)$ . We denote with  $\mathcal{P}_{K-1} \equiv (p_1, p_2, \dots, p_j, \dots, p_{K-1})$ ,  $\sum_{j=1}^{K-1} p_j \leq 1$  the parametric space of  $X$ . Let  $U \sim \{u_j = j, p_j\}$  for  $j = 1, 2, \dots, K$  be a discrete stochastic variable corresponding to  $X$  with parametric space  $\mathcal{P}_{K-1}$ , whose expected value

$$\mu_U = \sum_{j=1}^K j p_j = K - \sum_{j=1}^{K-1} (K-j) p_j \quad (1)$$

is usually adopted as a measure of risk.

If  $X$  represents growing levels of faulty then we are dealing with losses. Therefore  $\mu_U$  may be considered as a naive indicator of defectiveness of a product.

Sometimes it may be necessary to swap from one approach to the opposite; in this case the expected value of the reverse discrete r.v.  $U^* \sim \{u_j^* = (K+1) - u_j, p_j\}$  for  $j = 1, 2, \dots, K$  should be adopted. We denote such expected value *Severity Index* ( $S$ )

of the categorical r.v.  $X$ :

$$S = (K + 1) - \mu_U = 1 + \sum_{j=1}^{K-1} (K - j)p_j = \sum_{j=1}^K F_j, \tag{2}$$

where  $F_j = \sum_{l=1}^j p_l$  are the values of the cumulative distribution function of  $U$  for  $j = 1, 2, \dots, K$ .  $S$  is based only on the cumulative probabilities of the ordinal variable  $X$ , and is expressed as function of  $K$  and the parametric space  $\mathcal{P}_{K-1}$ . It assumes values in  $[1, K]$ ; the lower and the upper bounds occurs in the two situations of minimum heterogeneity.

The *Severity Index* proposed in (2) can be estimated by its empirical counterpart by using the empirical cumulative distribution function of  $X$ . Let  $(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$  be a simple random sample of size  $n$  from the categorical variable  $X$ , and let  $(\tilde{u}_1, \tilde{u}_2, \dots, \tilde{u}_n)$  be the corresponding sample of discrete values from the stochastic variable  $U$ .

The *Severity Index* estimator is defined as follows:

$$\hat{S} = \sum_{j=1}^K \tilde{F}_j = 1 + \sum_{j=1}^{K-1} (K - j) \frac{r_j}{n}. \tag{3}$$

$\tilde{F}_j = \sum_{l=1}^j \frac{r_l}{n}$  for  $j = 1, 2, \dots, K$ , is the empirical cumulative distribution function, where  $r_j$  is the number of the observations in the sample equal to the category  $x_j$ , with  $r_l \in \mathbf{N}$  and  $\sum_{l=1}^K r_l = n$ .

It is possible to show that the exact distribution of  $\hat{S}$  depends on the unknown values  $p_1, p_2, \dots, p_K$ . Consequently, in order to perform inferential procedures on  $S$ , that are robust with respect to the choice of  $p_1, p_2, \dots, p_K$ , it is possible to demonstrate that the *Severity Index* estimator is asymptotically normally distributed. Moreover,  $\hat{S}$  is an unbiased and consistent estimator for  $S$ .

### 3 Application in quality-control framework

We apply the proposed index to real data by a sales company of multinational corporation manufacturer of motion and control technology and systems providing precision-engineering solution for mobile, industrial and aerospace markets. The data concern information on severity and occurrence observed on potential failure of three components of a hose assembly (stripes, guard, hose). Severity is a measure of the gravity of a particular type of defect and occurrence is the frequency of the defects on  $n$  products. These information are typically available in companies that apply FMEA (failure modes and effects analysis)<sup>1</sup> to identify potential failures that could affect the customer’s expectations of product quality or process performance.

---

<sup>1</sup> FMEA is a reliability tool of product or process analysis that is conducted to identify potential failures that could affect the customer’s expectations of product quality or process performance.

For recent discussions and studies on the risk measures in FMEA see, among other, [7, 4, 6].

For each component of the hose assembly, the operator observe the type of defect and its frequency. Then he classifies every defect by the level of severity described on a 3-point scale (H=High severity, M=Medium severity, L=Low severity<sup>2</sup>). The observed levels of severity and the corresponding frequencies are summarized in Table 1. We call the ordinal variable in Table 1 SEVERITY.

**Table 1** Frequencies for levels of Severity for the components of the hose assembly

SEVERITY	stripes	guard	hose
H	0	0	0.58501
M	0.34286	0	0.29971
L	0.65714	1	0.11527

Our aim is to measure the risk associated to each component by an index that summarizes the SEVERITY. We calculate the sample *Severity Index*  $\hat{S}$ , according to (3), and its normalized version  $\hat{I} \in [0, 1]$ :

$$\hat{I} = \frac{\hat{S} - 1}{K - 1}, \tag{4}$$

and we provide the asymptotic confidence intervals for  $I$ . The results are reported in Table 2.

**Table 2** Point and 95% interval estimation for  $\hat{S}$  and  $\hat{I}$

INDEX	stripes	guard	hose
$\hat{S}$	1.34286 (1.18560; 1.50011)	1	2.46974 (2.43331; 2.50618)
$\hat{I}$	0.17143 (0.01417; 0.32869)	0	0.73487 (0.69844; 0.77131)

We observe that "hose" is the component with the highest level of risk. The other components have low level of risk. For "guard" a situation of minimum heterogeneity occurs: the index assumes its minimum value and  $Var(\hat{S}) = 0$ .

These results may be very useful for the company to prioritize intervention on the business line of the hose assembly, also in terms of improvement of the related process control.

Summarizing, the proposed index has been employed to communicate the level of risk, to compare among different risks in order to identify interventions on the pro-

<sup>2</sup> "H" indicates that the gravity of the defect is very serious and "L" not significant.



duction system or support decision making. Moreover, the index could be employed to understand how the risk change by monitoring it over time.

## References

1. Cerchiello, P., Dequarti, E., Giudici, P. and Magni, C. (2010). Scorecard models to evaluate perceived quality of academic teaching, *Statistica & Applicazioni*, 8, 145-155.
2. Figini, S. and Giudici, P. (2013). Measuring risk with ordinal variables, *Journal of Operational Risk*, 8, 35-43.
3. Figini, S., Kenett, R.S. and Salini, S. (2010). Optimal Scaling for Risk Assessment: Merging of Operational and Financial Data, *Quality and Reliability Engineering International*, 26, 887-897.
4. Lipol, L.S. and Haq, J. (2011). Risk analysis method: FMEA/FMECA in the organizations, *International Journal of Basic & Applied Sciences*, 11, 74-82.
5. MacKenzie, C.A. (2014). Summarizing Risk Using Risk Measures and Risk Indices, *Risk Analysis*, 34, 2143-2162.
6. Sellappan, N. and Palanikumar, K. (2013). Modified Prioritization Methodology for Risk Priority Number in Failure Mode and Effects Analysis, *International Journal of Applied Science and Technology*, 3, 27-36.
7. Wu, D.D., Kefan, X., Gang, C. and Ping, G. (2010). A Risk Analysis Model in Concurrent Engineering Product Development, *Risk Analysis*, 30, 1440-1453.



# Exponential family graphical models and penalizations

## *Modelli grafici basati su famiglie esponenziali e relative penalizzazioni*

Federico Ferraccioli, Livio Finos

**Abstract** In this paper we focus on the semantics of undirected graphical model. We present a general specification based on exponential family distributions that allows great model flexibility and leads to consistent inferential procedures. The model is extended to include prior distributions on the parameters, that reduce the variance of the estimates and permit to avoid over parametrization. Particular attention is devoted to non-differentiable  $l_1$  penalization, that leads to non-explicit gradient, for which we propose a new differentiable approximation. Experimental results and applications to large scale data are provided to demonstrate the increase in the rate of convergence and the variance reduction for different type of penalization priors.

**Abstract** Questo lavoro si concentra su modelli probabilistici basati su grafici indiretti. Viene presentata una specificazione generale, basata su distribuzioni appartenenti alla famiglia esponenziale, che permette grande flessibilità e conduce a stime consistenti. Il modello è dunque esteso introducendo vari tipi di distribuzioni *a priori* sui parametri, allo scopo di ridurre sia la varianza delle stime sia il rischio di sovra-parametrizzazione. Particolare attenzione è dedicata alla penalizzazione  $l_1$ , per la quale viene proposta una nuova approssimazione differenziabile. Infine vengono presentati risultati e applicazioni a dati di larga scala, con l'intento di dimostrare l'aumento del tasso di convergenza e la riduzione della varianza delle stime per i vari tipi di penalizzazioni.

**Key words:** graphical models, exponential family, non-differentiable penalization, contrastive divergence, hierarchical priors, topic modeling

---

Dipartimento di Scienze Statistiche, Università degli Studi di Padova  
via Cesare Battisti 241, 35100 Padova, e-mail: ferraccioli@stat.unipd.it  
Dipartimento di Psicologia, Università degli Studi di Padova  
via Venezia 8, 35131 Padova, e-mail: livio.finos@unipd.it

## 1 Introduction

Probabilistic graphical models are becoming a key part of the statistical modelling, particularly useful to deal with latent variables. Two major categories are Bayesian network and Markov random field, characterized by directed and undirected graph, respectively. In this paper we focus on this particular class of models, defined by undirected dependencies between observed and latent variables. From this perspective we can think of the model as a tool for dimensionality reduction, factor analysis and clustering. The major advantage consists in specification of both observed and latent variables distribution as elements of exponential family [4]. The most important disadvantage is the intractability of the partition function of the complete model, which complicates inference procedures using the likelihood. Here we focus our attention on a statistical specification of a method proposed by Hinton [2], the Contrastive Divergence, that greatly improve the efficiency of inference and opens the way for large scale problems. The properties of this method are still studied but the convergence in the case of exponential family distribution has been demonstrated in [5]. Using the properties of this set of probability distributions, we extend the general model introducing prior on the parameters: this extension permits the use of prior informations about the data and reduce the risk of over parametrization, a major issue in large scale problems. Moreover it reduces the variance of the estimates and increase the rate of convergence. We concentrate our discussion on the choice of the prior distribution studying the problems related to the optimization procedures for both differentiable and non-differentiable cases. The optimization in the latter case is not trivial. The loss function is non-convex and proximal gradient methods are not applicable. Furthermore in the case of non-differentiable function the gradient is not explicit. Here we give a possible strategy that consist in replacing the non-differentiable penalization with a differentiable approximations. Experiment results and applications to large scale data are provided to demonstrate the increase in the rate of convergence and the variance reduction for different type of penalization prior. The objective is not to claim superiority of the models proposed in this paper, but to give a general and viable alternative to directed probabilistic model discussing the possible advantages and disadvantages.

## 2 Exponential Family Model

Let  $\mathbf{X} = (X_1, \dots, X_N)$  a vector of observed random variables and  $\mathbf{Z} = (Z_1, \dots, Z_K)$  a vector of latent variables. We can choose from the exponential family  $N$  independent distributions for the observed variables and  $K$  independent distributions for the latent variables. The joint probability for the vectors  $\mathbf{X}$  and  $\mathbf{Z}$  are the following:

$$p_X(\mathbf{x} | \theta) = \prod_{i=1}^N p_0(x_i) \exp\left(\theta^\top \mathbf{T}(x_i) - A(\theta)\right) \quad (1)$$

$$p_Z(\mathbf{z} | \lambda) = \prod_{j=1}^K q_0(z_j) \exp\left(\lambda^\top \mathbf{U}(z_j) - B(\lambda)\right) \quad (2)$$

where  $\{\mathbf{T}(x), \mathbf{U}(z)\}$  are the vectors of sufficient statistics for the models,  $\{\theta, \lambda\}$  are the natural parameters of the models and  $\{A(\theta), B(\lambda)\}$  the log-partition functions.

We are also interested on dependencies between observed and latent variables, so we introduce a quadratic interaction term. The joint probability of the complete model is the following:

$$p(\mathbf{x}, \mathbf{z}) \propto \exp\left(\theta^\top \mathbf{T}(x) + \lambda^\top \mathbf{U}(z) + \mathbf{T}(x)^\top \Omega \mathbf{U}(z)\right) \quad (3)$$

### 3 Estimate procedures with Contrastive Divergence

The parameters estimation is done using an efficient method called Contrastive Divergence, that has the potential to greatly improve on the efficiency and reduce the variance of the estimates needed in the learning rule. The convergence of the method has been also proved for exponential family model in [5]. The idea is that instead of running the Gibbs sampler to its equilibrium distribution, we initialize Gibbs samplers on each data-vector and run them for only one (or a few) steps in parallel.

In order to obtain the parameter estimates, let  $\mathbf{X} = (X_1, \dots, X_N)$  an i.i.d sample generated from certain underlying distribution  $p_\theta$ .

Maximum likelihood estimation can be done by gradient ascent:

$$\theta^{new} = \theta + \alpha g(\theta) = \theta + \alpha \left( \frac{1}{N} \sum_{i=1}^N T(x_i) - \mu(\theta) \right) \quad (4)$$

where the learning rate satisfy  $\alpha > 0$ . The first term  $\frac{1}{N} \sum_{i=1}^N T(x_i)$  is easy to compute, but the second term  $\mu(\theta)$  is usually difficult to compute, since involves a complicated integral over  $X$ . To address this problem, Hinton [2] proposed the Contrastive Divergence (CD) method. The idea of CD is to replace the second term with  $\mu_{CD}(\theta) = \frac{1}{N} \sum_{i=1}^N T(x_i^{(m)})$ , where  $x_i^{(m)}$  is obtained by a small number ( $m$ ) of steps of an MCMC run starting from the observed sample  $X_i$ . We can now derive the parameters update for our bipartite model as follow:

$$\begin{aligned}
\theta^{new} &= \theta + \alpha \left( \frac{1}{N} \sum_{i=1}^N T(x_i) - \frac{1}{N} \sum_{i=1}^N T(x_i^{(m)}) \right) \\
\lambda^{new} &= \lambda + \alpha \left( \frac{1}{N} \sum_{i=1}^N \frac{\partial B(\tilde{\lambda})}{\partial \tilde{\lambda}} - \frac{1}{N} \sum_{i=1}^N \frac{\partial B(\tilde{\lambda}^{(m)})}{\partial \tilde{\lambda}} \right) \\
\omega^{new} &= \omega + \alpha \left( \frac{1}{N} \sum_{i=1}^N \frac{\partial B(\tilde{\lambda})}{\partial \tilde{\lambda}} T(x_i) - \frac{1}{N} \sum_{i=1}^N \frac{\partial B(\tilde{\lambda}^{(m)})}{\partial \tilde{\lambda}} T(x_i^{(m)}) \right)
\end{aligned}$$

#### 4 Hierarchical prior and penalization

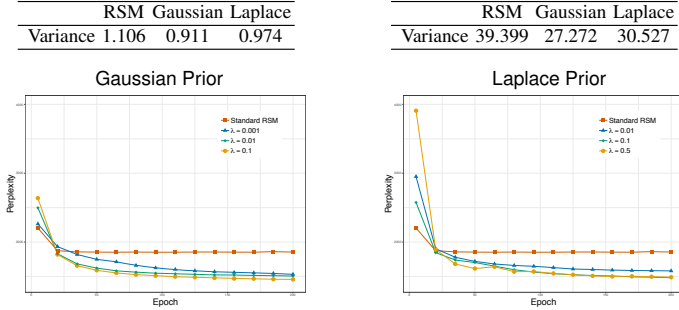
The main issue of this type of model, especially in the case of large scale problems, is the extremely high number of parameters. This leads to instability of the estimates and slows the rate of convergence. Given the likelihood of the model, we can think to add a prior over the parameters. A careful choice of the prior distribution permits to avoid over-parametrization, in particular due to the weights  $\omega$ ; the prior reduces also the variance of the estimates, increasing the rate of convergence. Two natural choices for the parameters are the Laplace and the Gaussian prior distributions, that lead to  $l_1$  and  $l_2$  regularization method, respectively. In the case of  $l_2$ , the gradient is explicit and we can easily derive the parameter updates as follows:

$$\omega^{new} = \omega + \alpha \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial B(\tilde{\lambda})}{\partial \tilde{\lambda}} T(x_i) - \frac{1}{n} \sum_{i=1}^n \frac{\partial B(\tilde{\lambda}^{(m)})}{\partial \tilde{\lambda}} T(x_i^{(m)}) - \gamma \sum_{\omega} \omega \right) \quad (5)$$

with  $\gamma$  the regularization parameter. Conversely, in the case of Laplace prior the non differentiability of absolute value leads to non explicit gradient. Furthermore the loss function is non-convex and proximal gradient methods are not applicable. An interesting way to solve this problem is given by a recent work on convolution based smooth approximations [3]. The smooth approximation to the non-differentiable absolute value function is computed via convolution with a Gaussian function as below:

$$\phi_{\sigma}(t) = t \operatorname{erf} \left( \frac{t}{\sqrt{2\sigma^2}} \right) + \sqrt{\frac{2\sigma^2}{\pi}} \exp \left( -\frac{t^2}{2\sigma^2} \right) \quad (6)$$

where  $\operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t \exp(-u^2) du$  is the standard error function and  $\sigma^2$  an hyperparameter. The advantage of using this approximation is that  $\phi_{\sigma}(t)$ , unlike the absolute value function, is a smooth function and we can compute its gradient. Furthermore,  $\phi_{\sigma}(t)$  converges uniformly to the absolute value function as  $\sigma \rightarrow 0$  with a higher rate of convergence than other approximations like square root. We can now add the gradient of the penalization term to the parameters update:



**Fig. 1:** Convergence plot for different value of the hyperparameter  $\lambda$ . The graph on the left corresponds to the Gaussian prior, the graph on the right to the Laplace prior with convolution based approximation. In both cases the penalization lead to faster convergence rate for all the hyperparameter values, reducing the variance of  $w$ , the estimated weights.

$$\omega^{new} = \omega + \alpha \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial B(\tilde{\lambda})}{\partial \tilde{\lambda}} T(x_i) - \frac{1}{n} \sum_{i=1}^n \frac{\partial B(\tilde{\lambda}^{(m)})}{\partial \tilde{\lambda}} T(x_i^{(m)}) - \gamma \sum_{\mathcal{W}} \text{erf} \left( \frac{\omega}{\sqrt{2\sigma^2}} \right) \right) \quad (7)$$

Without losing the asymptotic consistency, it is possible to use an annealed version of the learning rate  $\alpha_t = \frac{1}{t}$ , that increases the rate of convergence to  $\sqrt[3]{n}$ .

## 5 Application

In this section we present experimental results of the proposed prior penalization method applied to the Replicated Softmax (RSM)[1]. This is a particular case of the exponential family model proposed. The model can be view as the undirected counterpart of the Latent Dirichlet Allocation, one of the most used method in topic modeling. The basic idea is that documents can be represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. We use the NIPS proceedings papers dataset, that contains about 400 documents with a dictionary of more than 57.000 words. After several tests, we decided to use 10 variables in the hidden layer. To speed up the learning we used stochastic mini-batch: instead of computing the updates for all the observations, we permuted and divided the data in subsets of dimension  $m = 40$ . In presence of large scale problems the mini-batches help the parallel computation and the parameters convergence. Learning was carried out using Contrastive Divergence with ten full Gibbs step with loss function defined as:

$$\text{Perplexity} = \exp\left(\frac{1}{N} \sum_{n=1}^N \frac{1}{D_n} \log p(x_n)\right) \quad (8)$$

with  $D_n$  the number of words in the  $n$ -th document and log-likelihood  $\log p(x_n)$  estimated as described in Section 3. We compare the standard model without penalization (Replicated Softmax), with three values of the hyperparameter  $\lambda$  both for Gaussian and Laplace priors (figure 1). The rate of convergence is improved for both type of prior and in particular higher values of the hyperparameter  $\lambda$  lead to lower values of the loss function.

## 6 Conclusion

We develop a general framework for probabilistic graphical models, focusing our attention on formal definition of the Contrastive Divergence method. This greatly improves the efficiency of inference and opens the way for large scale problems. We also introduce hierarchical prior distributions on parameters to reduce the variance of the estimates and to increase the rate of convergence. We concentrate our discussion on the choice of the prior distribution studying the problems related to the optimization procedures for both differentiable and non-differentiable cases. Experiment results and applications to large scale data confirm the increase in the rate of convergence and the variance reduction for different type of penalization prior presented. Future work will concentrate on the model specification for an arbitrary number of layers, in order to capture hierarchical dependencies.

## References

1. Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Replicated softmax: an undirected topic model." *Advances in neural information processing systems*. 2009.
2. Hinton, Geoffrey E. "Training products of experts by minimizing contrastive divergence." *Neural computation* 14.8 (2002): 1771-1800.
3. Voronin, Sergey, Gorkem Ozkaya, and Davis Yoshida. "Convolution based smooth approximations to the absolute value function with application to non-smooth regularization." *arXiv preprint arXiv:1408.6795* (2014).
4. Welling, Max, Michal Rosen-Zvi, and Geoffrey E. Hinton. "Exponential Family Harmoniums with an Application to Information Retrieval." *Nips*. Vol. 4. 2004.
5. Wu, Tung-Yu, et al. "Convergence of Contrastive Divergence Algorithm in Exponential Family." *arXiv preprint arXiv:1603.05729* (2016).



# Key-indicators for maternity hospitals and newborn readmission in Sicily

## *Un sistema di indicatori per la classificazione dei punti nascita e re-ricoveri neonatali in Sicilia*

Mauro Ferrante, Giovanna Fantaci, Anna Maria Parroco, Anna Maria Milito, Salvatore Scondotto

**Abstract** This paper proposes a composite indicator for the classification of maternity units, which takes into account for the different dimensions of service delivery, as potential predictors of health outcomes. As a measure of outcome, infant readmissions is considered, being a proxy of morbidity. The results highlight that after controlling for risk factors of the newborn, and for the presence of neonatal intensive unit, infants born in lower level hospitals show readmission rates higher than infants born in higher level hospitals.

**Riassunto** *Il presente articolo propone un indicatore composito per la classificazione dei punti nascita che tenga conto delle diverse dimensioni coinvolte nella qualità del servizio erogato, quali potenziali predittori di esiti di salute. Come misura di esito vengono presi in esame i re-ricoveri neonatali, quale spia di possibili complicanze. I risultati, controllando per la presenza di unità di terapia intensive neonatale ed altri fattori di rischio, mostrano tassi di riammissione più elevate in strutture di basso livello, sulla base della classificazione proposta.*

**Key words:** *Composite indicator, birth-at-risk, healthcare evaluation, regionalization.*

---

Mauro Ferrante, Dipartimento Culture e Società, Università degli Studi di Palermo; email: [mauro.ferrante@unipa.it](mailto:mauro.ferrante@unipa.it)

Giovanna Fantaci, Dipartimento Attività Sanitarie e Osservatorio Epidemiologico della Regione Sicilia; email: [giovanna.fantaci@regione.sicilia.it](mailto:giovanna.fantaci@regione.sicilia.it)

Anna Maria Parroco, Dipartimento di Scienze Psicologiche, Pedagogiche e della Formazione, Università degli Studi di Palermo; email: [annamaria.parroco@unipa.it](mailto:annamaria.parroco@unipa.it)

Anna Maria Milito, Dipartimento Culture e Società, Università degli Studi di Palermo; email: [annamaria.milito@unipa.it](mailto:annamaria.milito@unipa.it)

Salvatore Scondotto, Dipartimento Attività Sanitarie e Osservatorio Epidemiologico della Regione Sicilia; email: [salvatore.scondotto@regione.sicilia.it](mailto:salvatore.scondotto@regione.sicilia.it)

## 1 Introduction

The relevance and quality of hospitals depend on various factors, among which the delivery volume, the geographic location, and the public or private ownership represent only a small component [2]. In regionalization programs the most used index to evaluate the quality of service delivery is given by the volumes of activity [4,7]. These programs generally aim at creating centers of excellence at which patients decide to receive care. Examples of this kind may be found in cancer care or complex survey procedures [6]. Nonetheless, there are situations, for example those related to acute situations for acute myocardial infarction [3] or in perinatal care [10], in which delivery volume alone is not able to provide an adequate picture of the complexity of factors which need to be evaluated in orienting regional healthcare programs.

Within the frame of maternity hospitals regionalization programs, the National Health and Medical Research Council recommend that pregnancies less than 33 week gestation be delivered at hospitals with neonatal intensive unit, in order to reduce the risk of mortality and morbidity. However, despite several studies have demonstrated better outcomes in high-level hospitals, with neonatal intensive unit and with high levels of delivery volume, for birth-at-risk (e.g. pre-term births) [1,9], less marked seem to be the differences in the case of low- or no risk childbirths [5,11]. Moreover, regionalization programs in perinatal care in Europe and North America, have determined a decrease in maternity hospitals with direct consequences also in terms of travel times required to reach the hospital [8]. This phenomenon determined a greater attention on the geographic location of maternity hospitals and, more in general, on issues related to maternity hospital accessibility.

Starting from these considerations, this Paper proposes a composite indicator for monitoring maternity hospitals in order to assist regionalization programs and to perform a classification of maternity hospitals in Sicily based on the proposed indicator. In order to evaluate the relationship between the proposed indicator and newborn readmission within 30 days from the childbirth event and hospital's category is performed.

## 2 Materials and methods

For the purposes of the present study several information sources were considered. The main information source is represented by Birth Certificates Records (CeDAP), which represent a unique information source for obtaining information on childbirth-related characteristics of both the mother and the infant. As a second information sources, Hospital Discharge Records have been used which, through an integration with CeDAP, allowed for the reconstruction of some characteristics of the childbirth event, e.g. transfers, complications, and readmissions. Finally, distance matrix among Sicilian municipalities has been considered in order to determine travel time

among the mother's Municipality of residence and the Hospital Municipality. In order to construct a composite indicator for maternity hospitals classification, several dimensions have been considered and for each dimension a set of variables have been selected, as reported in Table 1.

For the delivery *volumes* dimension, both the total number of childbirths and the caesarian section rates have been considered, given the relevance that both these aspects represent within the actual national evaluation system of quality of care. However, in order to take into account also for the degree of *complexity* of the childbirth event, also caesarian section rates in Robson's categories 01 and 03 have been included in the indicator, since the presence of a caesarian section childbirth may be a signal of inappropriateness. The third dimension considered relates to the territorial *basin* of the hospital, where the set of municipalities with at least 5% of its childbirths in the selected hospital defines the *basin* of the considered maternity hospital. For this dimension, the composite indicator includes both the total number of childbirths of women residing in the hospital's basin municipalities and the share of basin's childbirths in the hospital over the total number of basin's childbirths. As for the *travel time* dimension, the average travel time of total childbirths in the hospital has been included. Finally, for the *transfer* dimension, the share of childbirths with transfers within one day of both the mother and the newborn have been considered.

**Table 1:** Dimension, variables direction and weights of the composite indicator for maternity unit classification.

Dimension	Variables	Direction	Weight
Volumes	Total number of newborns ( <b>V1</b> )	(+)	0.10
	% of caesarean section births ( <b>V2</b> )	(-)	0.10
Complexity	% of caesarean section births in Robson = 01 or 03 ( <b>C1</b> )	(-)	0.10
Basin	Total newborns of women living in hospital's basin municipalities ( <b>B1</b> )	(+)	0.10
	% of basin's birth in the hospital ( <b>B2</b> )	(+)	0.15
Travel time	Average travel time from municipality of residence and hospital's municipality ( <b>T1</b> )	(-)	0.15
Transfer	% of transfers of the mother within one day from the birth ( <b>TR1</b> )	(-)	0.15
	% of transfers of the newborn within one day from the birth ( <b>TR2</b> )	(-)	0.15

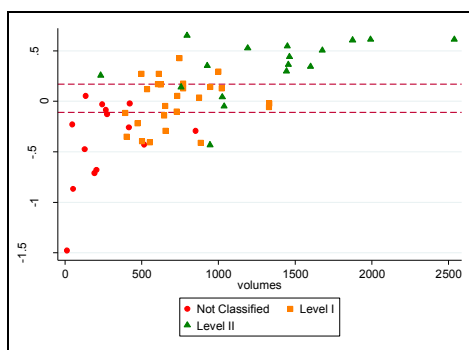
Once derived these information, a standardization procedure has been used for variable transformation, and the elementary indices have been aggregated through a weighted average, with the weights system indicated in Table 1. Weights have been derived after a sensitivity analysis with different weights systems and through technical meeting with experts. The produced indicator allowed for a classification of maternity hospitals into three categories. Finally, in order to evaluate the association between maternity unit's category and birth outcome, multiple logistic regression analyses were performed in which readmission of newborns has been

considered as an outcome measure, as a function of the proposed classification, and of other factors, such as the presence of neonatal intensive care unit and by controlling for other childbirth-related risk factors. For the definition of childbirth-at-risk the following criteria have been applied: mother's age below 20 or above 40 years old; very low birth weight (<1500 gr); gestational age lower than 37 weeks; multiple childbirth; small for gestational age (SGA).

### 3 Results

In 2014 a total of 44,436 newborns have been delivered in the 56 maternity hospitals of Sicily region. Nonetheless, after the application of the exclusion criteria, and record-linkage between Birth Certificate Records and Hospital Discharge Records, the valid cases resulted equal to 34,861. The average volume of activity resulted equal to 794, with a high degree of variability, with a minimum of 13 newborns in the case of Lipari Island hospital, and a maximum of 2538. Three hospitals have not been included in the analysis having less than 10 newborns in the year considered. A rather high degree of variability appeared also for all the other dimensions considered in the composite indicator.

In Figure 1 the relationship among delivery volumes, the score for the composite indicator proposed and the current regional classification are analyzed. Moreover, dashed lines indicates the cut points chosen for the categorization of the proposed indicator into three categories. It can be seen that, despite there is a direct relationship between delivery volumes and the proposed indicator, there are differences in terms of maternity units classification between the proposed classification and the current regional classification.



**Figure 1:** Delivery volumes, regional classification of maternity units, and composite indicator score, in Sicily, year 2014.

In Table 2 the results of logistic regression model are reported, in which the outcome variable assumes value 1 if the newborn has been readmitted within 30 days after the childbirth, and 0 elsewhere. In order to control for some risk-factors of the newborn and for the degree of severity, the presence of neonatal intensive care unit and other childbirth-related risk factors have been considered.

**Table 2:** Results of logistic regression model for infant readmission within 30 days after the childbirth by composite indicator categories, birth-at-risk and presence of neonatal intensive unit in the hospital

Variable	Category	OR	Inf CI 95%	Sup CI95%
Composite Indicator	High (Reference)			
	Medium	0.903	0.764	1.068
	Low	1.320**	1.065	1.636
Risk	Newborn-at-risk= No (Reference)			
	Newborn-at-risk = Yes	1.483***	1.267	1.736
UTIN	Neonatal intensive unit = No (Reference)			
	Neonatal intensive unit = Yes	4.431***	3.677	5.339
Constant		0.011***	0.009	0.013

\*, \*\* and \*\*\* indicate significance level at 0.1, 0.05 and 0.01 respectively.

From the analysis of the results in Table 2, a strong direct association between both the presence of risk factor (OR=1.48) and the presence of the Neonatal Intensive Unit (OR=4.43) with infant readmission within 30 days after the birth event appears. Nonetheless, after adjusting for these factors, low-category maternity units show a higher risk of readmission compared to births happened in high- and medium-level hospitals (OR=1.320). By considering that more complicated situations should be assisted by higher-level hospitals, this result calls for monitoring actions aimed at exploring the reasons of these readmissions in low-level maternity units. Moreover, the low-level category of the proposed classification comprises not only low-volumes maternity units, but also level I and one level II maternity units, according with the current regional classification criterion.

## 4 Discussion and conclusion

The classification of maternity units represents an important issue, not only from the scientific perspective, related to the analysis of the determinants of health outcomes, but especially from the health policy perspective [2]. It has been showed that several elements such as geographic location, transfers, post-partum length of stay, and childbirth risk factors, should be considered in evaluating the quality of service delivery [1,5,8,9]. The proposed approach tries to overcome some of the limits of the current classification system of maternity units, which is based mainly on delivery volumes. The proposed indicator takes into account for a number of

these dimensions and it may constitute the basis for classifying maternity units into different levels based on a set of dimensions which take into account for both some volume-related aspects of service delivery, and for some quality and territorial aspects related to the childbirth event.

The results showed that the proposed classification may predict comorbidity conditions, as measured by 30-days infant readmissions after the birth event. Further research is required to better highlight the determinants of these readmission, jointly with the analysis of the relationship between the proposed classification, and other potential proxies of quality of care (e.g. mother readmissions, mother and infant mortality).

## References

1. Bartels DB, Wypij D, Wenzlaff P, Dammann O, Poets CF. Hospital volume and neonatal mortality among very low birth weight infants. *Pediatrics* 2006; 117: 2206-14.
2. Falster MO, Roberts CL, Ford J, Morris J, Kinnear A, Nicholl M. Development of a maternity hospital classification for use in perinatal research. *NSW Public Health Bull.* 2012; 23(1-2): 12-16.
3. Ferrante M, Scodotto S, De Luca G, Fantaci G, Pollina-Addario S. Distance from the nearest hospital and mortality for acute myocardial infarction (AMI) in Sicily Region (Southern Italy). *Epidemiologia e Prevenzione*, 2014; 38(6): 373-378.
4. Halm EA, Lee C, Chassin MR. Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. *Ann Intern Med*, 2002; 137(6): 511-520.
5. Hemminki E, Heino A, Gissler M. Should births be centralised in higher level hospitals? Experiences from regionalised health care in Finland. *BJOG: An International Journal of Obstetrics & Gynaecology*, 2011; 118(10): 1186-1195.
6. Hollenbeak CS, Rogers AM, Barrus B, Wadiwala I, Cooney RN. Surgical volume impacts bariatric surgery mortality: a case for centers of excellence. *Surgery*, 2008; 144(5): 736-743.
7. Phibbs CS, Bronstein JM, Buxton E, Phibbs RH. The effects of patient volume and level of care at the hospital of birth on neonatal mortality. *JAMA*, 1996; 276(13): 1054-1059.
8. Pilkington H, Blondel B, Papiernik E, Cuttini M, Charreire H, Maier RF, Petrou S, Combier E, Kunzel W, Bréart G, Zeitlin J. Distribution of maternity units and spatial access to specialised care for women delivering before 32 weeks of gestation in Europe. *Health & place*, 2010; 16(3): 531-538.
9. Rautava L. The Effect of the Birth Hospital and the Time of Birth on the Outcome of Finnish Very Preterm Infants. *Annales Universitatis Turkuensis*, ser D 902. Turku: University of Turku, 2010.
10. Ravelli ACJ, Jager KJ, de Groot MH, Erwich JJHM, Rijninks-van Driel GC, Tromp M, Eskes M, Abu-Hanna A, Mol BWJ. Travel time from home to hospital and adverse perinatal outcomes in women at term in the Netherlands. *BJOG: An International Journal of Obstetrics & Gynaecology*, 2011; 118(4): 457-465.
11. Tracy SK, Sullivan E, Dahlen H, Black D, Wang YA, Tracy MB. Does size matter? A population-based study of birth in lower volume maternity hospitals for low risk women. *BJOG* 2006; 113: 86-96.

# Change of Variables theorem to fit Bimodal Distributions

## *Il teorema del Cambio di Variabile per modellare distribuzioni Bimodali*

Ferretti Camilla, Ganugi Piero, and Zammori Francesco

**Abstract** Bimodality is observed in empirical distributions of variables related to materials (glass resistance), companies (productivity) and natural phenomena (geyser eruption). Our proposal for modeling bimodality exploits the change of variables theorem requiring the choice of a *generating density function* which represents the main features of the phenomena under analysis, and the choice of the transforming function  $\varphi(x)$  that describes the observed departure from the expected behaviour. The novelty of this work consists in putting attention to the choice of  $\varphi(x)$  in two different cases: when bimodality arises from a slight departure from unimodality and when it is a proper structural feature of the variable under study. As an example we use the R "geyser" dataset.

**Abstract** *La bimodalità è osservata in distribuzioni empiriche legate alle proprietà dei materiali (resistenza del vetro), alla produttività delle imprese e a fenomeni naturali (eruzione di geysers). La nostra proposta per modellare distribuzioni bimodali si basa sul teorema del cambio di variabile, il quale richiede la selezione di una distribuzione generatrice che formalizza le caratteristiche strutturali del fenomeno in esame, e di una funzione trasformatrice  $\varphi(x)$  che descrive la distorsione osservata nei dati rispetto al comportamento atteso. La novità di questo lavoro consiste nello scegliere con particolare attenzione la funzione  $\varphi(x)$  in due casi differenti: quando la bimodalità è causata da una lieve distorsione di un fenomeno altrimenti unimodale, e quando essa è invece una caratteristica strutturale della variabile in esame. Come esempio utilizzeremo il dataset "geyser" di R.*

**Key words:** bimodal density function, change of variables theorem.

---

Ferretti C.,  
Dept. of Economic and Social Sciences, Univ. Cattolica del Sacro Cuore, Piacenza, e-mail: camilla.ferretti@unicatt.it

Ganugi P., Zammori F.,  
Dept. of Engineering and Architecture, Univ. degli Studi, Parma, e-mail: piero.ganugi@unipr.it, francesco.zammori@unipr.it

## 1 The problem

Bimodal distributions are observed in datasets arising from different research fields, for instance: 1) the glass resistance (to a given amount of pressure); 2) the firm productivity measured in number of units produced per unit of time (e.g. the Tuscan CAAF as in [1]); 3) the default probability for rated firms ([9]); 4) the waiting time among consequent eruptions of the Old Faithful geyser taken from the "geyser" dataset contained in the R package MASS ([10]).

Bimodality has been treated chiefly using the mixture approach. Mixtures represent a suitable model when two or more distinct groups with specific distributions are considered as a single set. Conversely, when original groups composing the mixture are not recognizable or the specific density function of each group is not known, or also when the observed phenomena is structurally bimodal, mixtures could not provide a good data fitting. Consequently, it is worth using an alternative approach to obtain a bimodal distribution for fitting the original data.

### 1.1 Change of variables theorem and choice of generating and transforming functions

The basic tool for modeling bimodality regards the use of the well-known change of variable theorem ([3, 5], among others):

**Theorem 1.** *Given a continuous r.v.  $Y$  with known density function  $f(y)$ , and given a transformed r.v.  $X = h(Y)$  with  $h(\cdot)$  monotone, the density function of  $X$  has the following formula:*

$$g(x) = \left| \frac{d}{dx} h^{-1}(x) \right| \cdot f(h^{-1}(x)). \quad (1)$$

Let  $X$  be the variable under study, having unknown density function. Such theorem permits to obtain a formula for the density function  $g(\cdot)$  through the following steps:

1. Choice of a suitable r.v.  $Y$  whose density function  $f(y)$  is known and represents the main structural features of the phenomena under analysis. The function  $f$  is called the *generating density function*.
2. Choice of a *transformation function*  $Y = \varphi(X)$  describing the relationship between the observed data and the expected theoretical behavior given by  $f(y)$ .

The theorem will be then applied on  $X = h(Y) = \varphi^{-1}(Y)$ .

This procedure has been used for proposing the fitting of unimodal data ([7]), assuming that the generating function coincides with the Standard Normal  $Z$  with density function  $\phi(z)$ , and three possible data transformations are proposed as in the following table:



**Table 1** Three choices for  $\varphi(x)$  proposed in [7] and the resulting density function  $g(x)$ .

$\varphi(x)$	Density function of $X$
$z = \ln(x), x \in (0, +\infty)$	$g(x) = \frac{1}{x\sqrt{2\pi}} \cdot e^{-\frac{1}{2}[\ln(x)]^2}$
$z = \ln\left(\frac{x}{1-x}\right), x \in (0, 1)$	$g(x) = \frac{1}{x(1-x)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left[\ln\left(\frac{x}{1-x}\right)\right]^2}$
$z = \ln(x + \sqrt{x^2 + 1}), x \in R$	$g(x) = \frac{1}{\sqrt{2\pi(x^2+1)}} \cdot e^{-\frac{1}{2}\left[\ln(x + \sqrt{x^2+1})\right]^2}$

Summarizing, in [7] it is assumed that the behavior of observed data is due to a departure from Normality which is formally described by the transformation  $\varphi(x)$ . Besides, the three transformation functions in Tab. 1 allows to model almost all the unimodal density function encountered in empirical applications.

## 2 Analyzing bimodal distributions

The novelty of this work consists in applying the mentioned procedure to the case of data displaying a bimodal behavior. According to this, we stress the following facts:

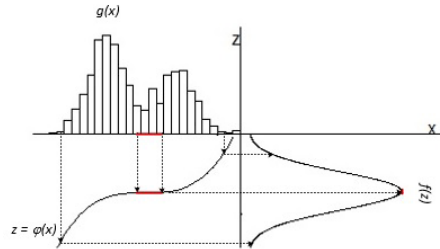
1. The choice of the generating function is related to the inner structure in the observed variable. If we choose an unimodal generating function (e.g. the Normal function) we are actually assuming that bimodality is slight and due only to some kind of moderate perturbation affecting the unimodal underlying model.
2. On the other side, if the variable is known to be structurally bimodal, the suitable choice is a bimodal generating function. In this case, the main problem is the poor menu of existing bimodal density functions.
3. Given the generating function, it is necessary to find a suitable transformation function  $\varphi(\cdot)$ . The choice of  $\varphi$  is fundamental to bring back observed data with the expected theoretical model. In literature, we observe mainly two approaches for the choice of  $\varphi(x)$ :
  - a. The regression analysis applied on the Q-Q plot of the observed percentiles w.r.t. the theoretical percentiles from the generating function as in [11].
  - b. The production of a differential equation whose solutions form a family of suitable transformation functions as in [4, 8].

## 3 Bimodality deriving from a Normal generating function

As a first step, we assume that the empirical variable is structurally unimodal, and we choose the Standard Normal distribution as generating function. Fig. 1 displays

a simulated example illustrating the form we expect for the transformation function. Indeed, the bimodality can be interpreted as a polarization of statistical units, originally located close to the median of the distribution, and shifting in opposite directions. In this light, a suitable transformation function should have a reversed-S shape as illustrated in Fig. 1 (bottom left panel).

**Fig. 1** Graphical representation of the transformation of bimodal data (upper panel) to obtain an unimodal (Normal) distribution (lower right panel), through reversed-S-shaped transformation function.



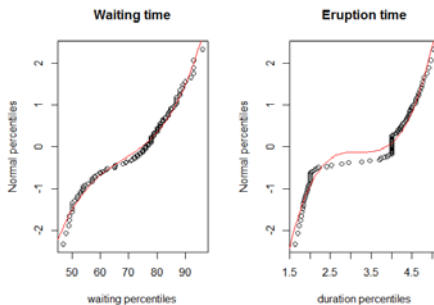
## 4 Empirical illustration

As an empirical example, we use the dataset "geyser" contained in the R package MASS. Data regard the observed values (in minutes) of the Old Faithful geyser eruptions duration, and the waiting time between two consecutive eruptions. Both the variables show bimodality, as shown by the histograms based on 25 classes in Figg. 3 and 4. The analysis of the Q-Q plot obtained comparing the percentiles observed using the given variables in comparison with the Normal percentiles confirms the reversed-S shaped transformation as explained in the previous section, for both "duration" and "waiting" (see Fig. 2).

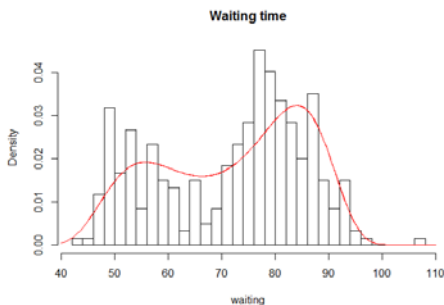
As a first attempt, we choose  $\varphi(x) = Ax^3 + Bx^2 + Cx + D$  for both the variables, given the flexibility of third-order polynomials, and we estimate parameters using OLS. We stress the necessity to improve the fitting of the Q-Q plot. In particular we note that prominent bimodality as in the "duration" variable is more challenging due to the lack of observations near the 50th percentile. With this aim, we will follow two approaches:

1. We substitute polynomial regression with third or fourth-order B-spline regression ([2]) which should improve the fitting and make possible to impose monotonicity constraints on  $\varphi(\cdot)$ , as required by the Theorem 1, and avoid collinearity drawbacks that instead affect the polynomial regression.
2. From a theoretical point of view, we aim to supply a differential equation for constructing a family of reversed-S-shaped transformation functions. The advantage of this choice consists in the possibility to make easier the economic or physical

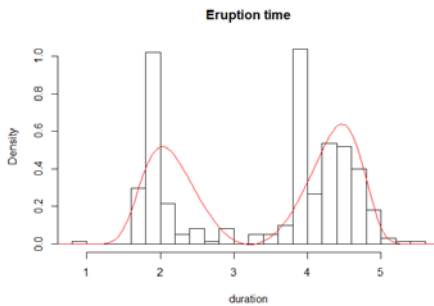
**Fig. 2** Q-Q plots obtained comparing the empirical percentiles observed using "waiting" and "duration" in the R dataset "geyser" with the Standard Normal percentiles.



**Fig. 3** Histogram of "waiting" compared with the fitted bimodal curve ( $D = -27.67$ ,  $C = 1.13$ ,  $B = -0.016$ ,  $A = 0.00008$ ).



**Fig. 4** Histogram of "duration" compared with the fitted bimodal curve ( $D = -15.27$ ,  $C = 14.13$ ,  $B = -4.39$ ,  $A = 0.45$ ).



interpretation of the parameters with respect to the spline regression approach, and with this, the explanation of the hidden causes of bimodality.

## 5 Conclusions and further research

In this work we focus on the problem of modeling bimodal data. We face the problem using the change of variable theorem, which requires a suitable transformation function such that the transformed data have a known shape. As an example, we apply this procedure to geyser data, choosing a Normal generating function and a polynomial transformation. The fit we obtain is not yet satisfactory, we aim then to improve the procedure using the B-spline regression. Alternatively we will try to produce a family of transformation functions, all of them mirroring a seemingly statistical regularity (the reversed-S-shaped Q-Q plot associated to bimodality). The advantage of the second approach is the possibility to make easier economic or physical interpretation to parameters.

As a further research step, we aim to choose a bimodal generating function. Since research on theoretical bimodal distributions is scarce until now (see for example some recent work as [6]), this choice will require the production, *ex-novo*, of a suitable bimodal density function to be used as structural model.

## References

1. Soliani, F.: Cost and productivity analysis and production planning in large size CAAF. Master Thesis Dissertation, Dept. of Industrial Engineering, Parma (2016).
2. Bollaerts, K. and Eilers, P. and Aerts, M.: Quantile regression with monotonicity restrictions using P-splines and the  $L_1$  norm. *Statistical Modeling*, 6: 189 - 207 (2006).
3. Billingsley, P.: *Probability and Measure*. Wiley series in Probability and Mathematical Statistics, Wiley: University of Michigan (1979).
4. D'Addario, R.: Ricerche sulla curva dei redditi. *Giornale degli Economisti ed Annali di Economia*, 1/2: 91 - 115 (1949).
5. Stirzaker, D.: *Elementary Probability*, 2nd ed, Cambridge University Press (2003).
6. Hassan, M.Y. and Hijazi, R.H.: A bimodal exponential power distribution. *Pakistan Journal of Statistics*, 26(2): 379 - 396 (2010).
7. Johnson, N.L.: Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*, 1/2: 149 - 176 (1949).
8. Kleiber, C. and Kotz, S.: *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley Probability and Statistics, Wiley-Interscience (2003)
9. Schuermann, T. and Hanson, S. G.: Estimating probabilities of default. Staff Report no. 190, Federal Reserve Bank of New York (2004).
10. Venables, W. N. and Ripley, B. D.: *Modern Applied Statistics with S*. Fourth Edition. Springer, New York (2002).
11. Vianelli, S.: *Manuale di metodologia statistica: metodologia descrittiva per la ricerca empirica*, Bologna: Calderini (1966).

# Space-time clustering for identifying population patterns from smartphone data

## *Clustering spazio-temporale per dati smartphone sulla distribuzione della popolazione*

Francesco Finazzi and Lucia Paci

**Abstract** In this work we aim at studying spatio-temporal patterns of the population movement across a large city. We exploit the information on people position collected by the smartphone application of the Earthquake Network project and we adopt a dynamic model-based clustering approach to identify the patterns. The approach is applied to smartphone data collected in Santiago (Chile) over the period February-April 2016. Some preliminary results are presented and discussed.

**Abstract** *L'obiettivo di questo lavoro è studiare i pattern spazio-temporali di movimento della popolazione su una grande città. Sfruttiamo l'informazione sulla posizione delle persone raccolta dall'applicazione smartphone del progetto Earthquake Network ed applichiamo un approccio di clustering dinamico per identificare i gruppi. L'approccio è applicato ai dati smartphone raccolti per la città di Santiago (Cile) lungo il periodo febbraio-aprile 2016. Alcuni risultati preliminari sono presentati e discussi.*

**Key words:** Finite mixture models, Markov chain Monte Carlo, spatio-temporal modeling, state-space, crowd-sourcing data

## 1 Introduction

Detecting population dynamics over short periods (e.g. daily movements) may provide the public with useful information to improve traffic infrastructure associated with spatio-temporal commuting patterns, upgrade accessibility or attractiveness of

---

Francesco Finazzi  
Department of Management, Information and Production Engineering, University of Bergamo,  
Dalmine, Italy, e-mail: [francesco.finazzi@unibg.it](mailto:francesco.finazzi@unibg.it)

Lucia Paci  
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy e-mail:  
[lucia.paci@unicatt.it](mailto:lucia.paci@unicatt.it)

areas interested by less people than others, enhance public transportation according to infrastructure/open space utilization. Indeed, population patterns are characterized by drastic changes during the day according to several activities such as education, working, recreation, visiting and shopping activities, among others.

Customary, population studies are based on census data that do not allow to capture population movements in short periods. Rather, mobile-based data collected over a given region at high temporal scale offers new opportunities to study population distribution and movement patterns over such region. For instance, Secchi et al (2015) proposed a non-parametric method for the analysis of spatially dependent functional mobile network data to identify subregions of the metropolitan area of Milan sharing a similar pattern along time, and possibly related to activities taking place in specific locations and/or times within the city.

Alternatively, we can identify potential partitions of the space and study their evolution over time to extract useful and concise information from smartphone-based data that is helpful to investigate population dynamics. Recently, Paci and Finazzi (2017) proposed a model-based approach to identify clusters in data collected at fixed spatial locations and time steps. Within finite mixture modeling, spatio-temporally varying mixing weights are introduced to allocate observations at nearby locations and consecutive time points with similar cluster's membership probabilities. As a result, a clustering varying over time and space is accomplished. Conditionally on the cluster's membership, a state-space model is deployed to describe the temporal evolution of the sites belonging to each group.

In this work we employ the dynamic space-time clustering approach to explore population dynamics and motion patterns over the city of Santiago (Chile) using data coming from the Earthquake Network project ([www.earthquakenetwork.it](http://www.earthquakenetwork.it)). The project implements a crowdsourced earthquake early warning system based on smartphones networks (Finazzi and Fassò, 2016) and it requires to collect the precise location in space of smartphones at regular time steps. Here, it is assumed that the smartphone location is also the position in space of its owner.

## 2 Bayesian space-time mixture modeling

Let  $y_t(\mathbf{s})$  be a response variable observed at time  $t$  ( $t = 1, \dots, T$ ) and location  $\mathbf{s} \in \mathbb{R}^2$ . We assume that observation  $y_t(\mathbf{s})$  comes from a finite mixture model, that is

$$f(y_t(\mathbf{s}) | \boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{k=1}^K \pi_{t,k}(\mathbf{s}) f(y_t(\mathbf{s}) | \boldsymbol{\Theta}_k) \quad (1)$$

where  $K$  is the number of components. The distribution under the  $k$ -th component ( $k = 1, \dots, K$ ) is denoted by  $f(\cdot | \boldsymbol{\Theta}_k)$  where  $f$  is a density function of specified form and  $\boldsymbol{\Theta}_k$  denotes the set of parameters of each component distribution. The mixing probability  $\pi_{t,k}(\mathbf{s})$  is the probability that the location  $\mathbf{s}$  belongs to component  $k$  at time  $t$  and it satisfies  $\pi_{t,k}(\mathbf{s}) > 0$  with  $\sum_{k=1}^K \pi_{t,k}(\mathbf{s}) = 1$  for each  $\mathbf{s}$  and  $t$ .

As usual in Bayesian analysis, a hierarchical formulation of the mixture model is exploited to facilitate the computation. For each observation, we introduce a latent allocation variable,  $w_t(\mathbf{s})$ , that identifies the component membership of  $y_t(\mathbf{s})$ , that is  $Pr(w_t(\mathbf{s}) = k) = \pi_{t,k}(\mathbf{s})$ . In other words, we assume that the allocation variables  $w_t(\mathbf{s})$  are conditionally independently distributed given  $\pi_{t,k}(\mathbf{s})$  and they come from a multinomial distribution. Given the latent  $w_t(\mathbf{s})$ , the observations  $y_t(\mathbf{s})$  are independent with  $f(y_t(\mathbf{s}) | w_t(\mathbf{s}) = k, \Theta) = f(y_t(\mathbf{s}) | \Theta_k)$ . As customary in model-based clustering, we interpret each mixture component as a cluster, such that observations are partitioned into mutually exclusive  $K$  groups.

The mixing probabilities,  $\pi_{t,k}(\mathbf{s})$ , are allowed to vary from observation to observation, i.e., across space and over time. Space-time dependence in the observations is introduced through the prior distribution of the weights such that observations corresponding to nearby locations and consecutive time points are more likely to have similar allocation probabilities than observations that are far apart in space and time. For each location  $\mathbf{s}$  and time  $t$ , the weights take the form

$$\pi_{t,k}(\mathbf{s}) = \frac{\exp(\mathbf{x}'_t(\mathbf{s})\beta_k + \phi_{t,k}(\mathbf{s}))}{\sum_{l=1}^K \exp(\mathbf{x}'_t(\mathbf{s})\beta_l + \phi_{t,l}(\mathbf{s}))} \quad (2)$$

where  $\mathbf{x}_{t,k}(\mathbf{s})$  is a  $p \times 1$  vector of covariates,  $\phi_{t,k}(\mathbf{s})$  are spatio-temporal random effects and  $\beta_1 = 0$  and  $\phi_{t,1}(\mathbf{s}) = 0$  ( $t = 1, \dots, T$ ) to ensure identifiability. The logistic-type transformation in (2) guarantees that the two conditions mentioned in Section 2 are satisfied (Fernández and Green, 2002). When available, covariates may help in predicting group membership's probabilities while random effects provide adjustment in space and time to the explanation provided by covariates. Therefore, the response distribution is allowed to vary in flexible ways across time, space and covariate profiles.

To allow for dynamics over time and dependence over space we assume, for  $k = 2, \dots, K$ ,

$$\phi_{t,k}(\mathbf{s}) = \rho_k \phi_{t-1,k}(\mathbf{s}) + \zeta_{t,k}(\mathbf{s}) \quad (3)$$

where  $\zeta_{t,k}(\mathbf{s})$  are independent-in-time spatially correlated errors coming from a zero-mean Gaussian Process (GP) equipped with an exponential spatial covariance function. Although the  $K - 1$  spatio-temporal random effects  $\phi_{t,k}(\mathbf{s})$  are assumed to be independent, the corresponding weights are not independent given their definition in (2). The space-time structure of random effects  $\phi_{t,k}(\mathbf{s})$  allows to borrow strength information from nearby sites and consecutive time steps. As a result, similar outcomes at near space and time points are assigned with similar cluster membership's probabilities.

Model (1) requires the specification of the sampling density  $f(y_t(\mathbf{s}) | \Theta_k)$ . The approach pursued in this work is based on dynamic linear modeling, often referred to as state-space models. In particular, we assume a dynamic linear model to describe the temporal evolution of all the sites within component  $k$ .

Let  $\mathbf{y}_t = (y_t(\mathbf{s}_1), \dots, y_t(\mathbf{s}_n))'$  be the  $n \times 1$  observation vector at time  $t$ , where  $n$  is the number of locations. Conditionally on the allocation variables, the space-state model is provided by

$$\mathbf{y}_t = \mathbf{H}_t \mathbf{z}_t + \boldsymbol{\varepsilon}_t \quad (4)$$

$$\mathbf{z}_t = \mathbf{G} \mathbf{z}_{t-1} + \boldsymbol{\eta}_t \quad (5)$$

where  $\mathbf{z}_t = (z_{t,1}, \dots, z_{t,K})'$  is the  $K \times 1$  state vector,  $\mathbf{H}_t$  is a  $n \times K$  matrix defined below, and  $\mathbf{G}$  is a  $K \times K$  stable transition matrix. Finally,  $\boldsymbol{\varepsilon}_t \sim N(0, \sigma^2 I_n)$  is the  $n \times 1$  measurement error vector and  $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \Sigma_\eta)$  is the  $K \times 1$  innovation vector.

We now turn to matrix  $\mathbf{H}_t$ . Suppose that site  $\mathbf{s}$  belongs to component  $k$  at time  $t$ . Then, the  $i$ -th row of matrix  $\mathbf{H}_t$  contains a single element equal to one at position  $k$ , while all the other elements are filled with zeros (Inoue et al, 2007; Finazzi et al, 2015). Note that, the one-zero structure of matrix  $\mathbf{H}_t$  is allowed to vary over time according to mixing probabilities  $\pi_{i,k}(\mathbf{s})$ . Also, we benefit from the borrowing strength of information of all sites belonging to component  $k$  at time  $t$ , since they all contribute in estimating the common latent state  $z_{t,k}$ . Given the specification in (5), the desired temporal pattern of cluster  $k$  is represented by latent state  $z_{t,k}$ .

Full inference is provided under a Bayesian framework. The hierarchy of the model is completed by independent noninformative prior distributions for all the hyperparameters and Monte Carlo Markov Chain (MCMC) algorithms are employed to approximate the joint posterior distribution; see Paci and Finazzi (2017) for all fitting details and posterior computation. Model fitting is carried out using the MATLAB code DYSC available online at the web page <https://github.com/graspa-group/DYSC>.

### 3 Analysis of smartphone data

Smartphones taking part in the Earthquake Network project send a heartbeat signal to a central server every around 30 minutes. Signals include the geographic location of the smartphones that is used to estimate the state of the network at any given time.

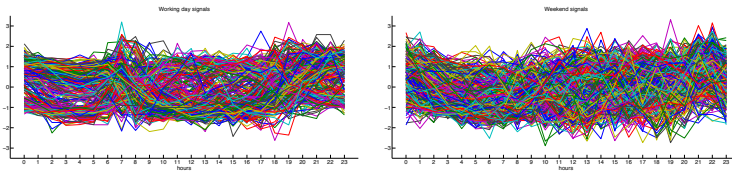
In this work, we exploit the information carried by the heartbeat signals to study the population movement across the city of Santiago. We consider 24'900 smartphones and we assume they are representative of the entire Santiago population. We partition the city of Santiago into a uniform lattice of  $N = 354$  sites and for each site we consider the number of signals on a hourly basis. For each hour of the day, we aggregate signals observed over the period February-April 2016, assuming that the daily motion patterns of the population are stable over the 3 months. Moreover, we distinguish between working days and weekend in order to investigate possible differences. The aggregation leads to two  $N \times T$  matrices for the working days and the weekend, respectively, with  $T = 24$ . Since we aim at studying the motion patterns independently from the number of signals, we standardize each time series with respect to its own mean and variance. This implies that sites are directly comparable. Hence, at each time step, the time series is interpreted in the following way: a negative value means that the number of signals coming from the site is below the site average, while a positive value means that the number of signals is above average.



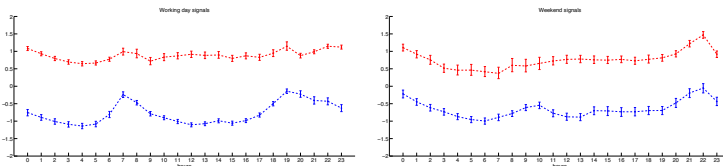
Figure 1 shows the standardized number of signals received from each site during working days (left panel) and weekend (right panel) over the study period.

At each time step, thus, we apply model described in Section 2 to cluster sites which behave in a similar way with respect to their average and then we explore how the clusters evolve over the 24 hours of the day. We employ the diagnostic tool provided by Paci and Finazzi (2017) to select the number of clusters. The analysis suggests that only two clusters are needed. This is a consequence of the fact that time series are standardized and the number of signals from each sites can be either below or above average. Figure 2 shows the Posterior 95% credible interval of the temporal patterns  $z_{t,k}$  for working day signals (left panel) and weekend signals (right panel), where each temporal pattern is related to a cluster. During working days, the separation between the temporal patterns is lower at 7 a.m. and 7 p.m., namely when people commute from home to work and vice-versa. During these hours, signals are more evenly distributed across city than in any other hour of the day. During the weekend, the same effect can be found at 10 a.m. and at midnight.

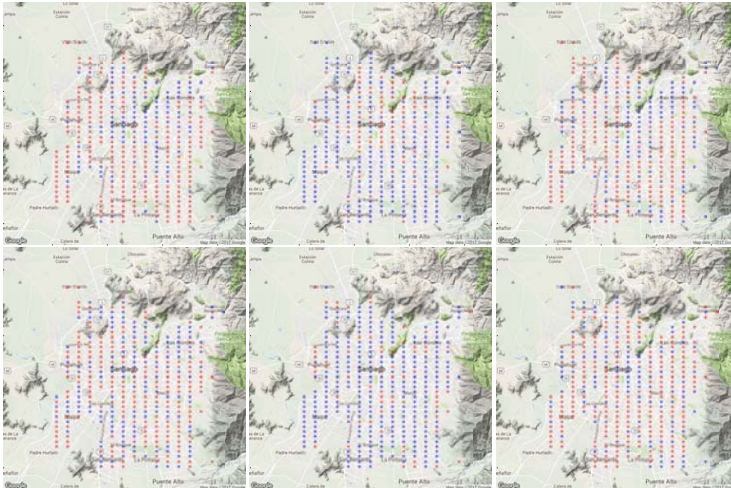
To provide the clustering, we assign each observations to their most likely group according to the maximum a posteriori probability (MAP) rule. In Figure 3 clustering result can be appreciated for 12 a.m., 8 a.m. and 8 p.m. and for both working days and the weekend. For any given hour of the day, blue and red points are sites with a number of signals below and above the average, respectively. During working days, the number of signals from the city center is below average at night and above average during the day. This pattern is disrupted during the weekend when people tend to move later in the morning and to return home later in the night.



**Fig. 1** Number of signals collected from each cell during working days (left panel) and weekend (right panel) over the period February-April, 2016.



**Fig. 2** Posterior 95% credible interval of the temporal patterns  $z_{t,k}$  for working day signals (left panel) and weekend signals (right panel).



**Fig. 3** Clustering result for working day (top row) and the weekend (bottom row) at 12 a.m. (left column), 8 a.m. (middle column) and 8 p.m. (right column). Blue and red dots refer to the blue and red temporal patterns in Figure 2, i.e., below and above the average, respectively.

## References

- Fernández C, Green PJ (2002) Modelling spatially correlated data via mixtures: A Bayesian approach. *J R Stat Soc Ser B* 64:805–826
- Finazzi F, Fassò A (2016) A statistical approach to crowdsourced smartphone-based earthquake early warning systems. *Stoch Environ Res Risk Assess* Doi:10.1007/s00477-016-1240-8
- Finazzi F, Haggarty R, Miller C, Scott M, Fassò A (2015) A comparison of clustering approaches for the study of the temporal coherence of multiple time series. *Stoch Environ Res Risk Assess* 29:463–475
- Inoue LYT, Neira M, Nelson C, Gleave M, Etzioni R (2007) Cluster-based network model for time-course gene expression data. *Biostatistics* 8:507–525
- Paci L, Finazzi F (2017) Dynamic model-based clustering for spatio-temporal data. *Stat Comput* DOI 10.1007/s11222-017-9735-9
- Secchi P, Vantini S, Vitelli V (2015) Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. *Stat Methods Appl* 24:279–300

# IT Solutions for Analyzing Large-Scale Statistical Datasets: Scanner Data for CPI

## *Soluzioni IT per l'analisi di dataset statistici di grandi dimensioni: scanner data per l'indice dei prezzi al consumo*

Annunziata Fiore, Antonella Simone and Antonino Virgillito

**Abstract** In this paper we present the issues and challenges related to dealing with datasets of big size such as those involved in the Scanner Data project at Istat. We describe the IT solutions introduced as part of a larger scope approach to the modernisation of tools and techniques used for data storage and processing in Istat, envisioning the future challenges posed by Big Data and Data Science in NSIs. We show how the IT architecture can help the methodological choices for the construction of consumer prices microindices by comparing different approaches to compute indices through an extensive analysis carried out over the entire data set.

**Abstract** *In questo paper vengono discusse le sfide legate all'utilizzo di dataset di grandi dimensioni come quelli nel progetto Scanner Data, attualmente in corso presso l'Istat. Vengono illustrate le soluzioni tecnologiche introdotte come parte di un approccio di portata più ampia alla modernizzazione dei tool e delle tecniche di memorizzazione e elaborazione dati in Istat. Mostriamo come la nostra architettura IT può sostenere le scelte metodologiche per la costruzione dei microindici dei prezzi al consumo. In particolare, presentiamo i risultati di una analisi, effettuata sull'intero dataset, mirata al confronto di diversi approcci al calcolo dell'indice.*

**Key words:** Scanner Data, Big Data, CPI

---

<sup>1</sup> Annunziata Fiore, Istat; [annunziata.fiore@istat.it](mailto:annunziata.fiore@istat.it)  
Antonella Simone, Istat; [ansimone@istat.it](mailto:ansimone@istat.it)  
Antonino Virgillito, Istat; [virgilli@istat.it](mailto:virgilli@istat.it)

## 1 Introduction

The Istat Scanner Data project started in 2015 [1] and it is currently going through a pre-production phase that will continue along the whole 2017. The objective of the project is to carry out a massive revision of the production process of Consumer Prices Indices in order to replace the on-field data collection for grocery products in supermarkets with the prices obtained from the scanner data source [3]. Once in production (scheduled for 2018), scanner data will be the first example at Istat of such a large data set being used as a source a production process.

A number of challenges are involved in handling the constant flow of data and storing it safely, thus a solid data processing pipeline had to be put in place by the IT sector, that allows the different sectors of the institute involved in the process (data collection and production) to control its correct evolution and the quality of the data.

However, in a modern data-driven organization IT tools are not only meant to back production processes but should be considered one of the main drivers behind the organization core business. In the era of Big Data, the capability of analysing large amounts of data necessarily requires IT solutions to be effective, governed and secure and also available not only to IT specialists but also to researchers.

Under these premises, the scanner data project was not only relevant for its main objective, the redesign of one of the most important production processes of Istat, but it also was the first important testbed of a general approach to the modernization of the tools used for statistical production.

In this paper we present the hybrid data architecture that has been realized to support the scanner data project, integrating a traditional RDBMS with a Big Data Processing Platform, that has been recently setup at Istat and has been used for the first time specifically for this project. An in-depth discussion is provided about the benefits and the trade-offs resulting from the use of Big Data technology for statistical production. We exploited the advanced capabilities of the platform to carry out extensive analysis on the whole dataset that could have not been possible through traditional tools. In particular, we simulated the implementation of two different methods of aggregating elementary indices (“static”, based on a yearly updated fixed basket, and “dynamic”, based on chaining prices over consecutive months) and provided the production sector with in-depth insights about the performance and the practical feasibility of the two methods.

## 2 Processing Scanner Data at Istat

In this section we give a general overview of the scanner data project and discuss the challenges that were related to the treatment of a large scale dataset, presenting the technical solutions that were adopted to store and process data.

## **2.1**    *The Scanner Data Dataset*

The periodical acquisition of scanner data from Istat is regulated by an agreement made by Istat with the Italian association of modern distribution, representing the main chains of modern retail trade. According to the agreement, Istat access to scanner data is mediated by a broker, which is the Nielsen company. Nielsen sends data to Istat on a monthly basis by uploading the data files on a dedicated web portal.

Currently, Istat has received data for 4 complete years related to 37 provinces, for a total of 1.4 billion records. Each record represents the weekly sum of turnover and quantity for a GTIN (Global Trade Item Number, formerly EAN code) sold during the week in a single store. The current provisioning of data includes 1470 stores while the final sample will be composed of 2100 stores. Also provided are classification tables for mapping GTINs to ECOICOP classification and determine the Elementary Aggregate they belong to. Data consists exclusively in grocery products, with 1579 EAs represented and 232,000 GTINs. Finally, the lists of stores and GTINs, integrated with additional information (descriptions, geographic location, etc.), are available.

A novel hybrid data architecture has been setup for storing scanner data. The architecture is composed by a traditional Relational Database Management System (RDBMS) and a Big Data Processing Platform (BDP). RDBMS stores only current year data and handles the cleaning and pre-processing of the acquired data, while BDP offloads the database, storing all the historical dataset. The motivation for the use of such architecture is explained in detail in the next section.

## **2.2**    *The Problem with Size*

The dimension of the scanner data dataset is not common for Istat, being one of the largest in terms of absolute size hosted in the institute. We experienced several issues as a consequence of this. Before setting up the BDP, we first loaded the datasets into the RDBMS. Besides the table with the raw data, a number of artefacts have been produced afterwards, including indexes, views and temporary tables used to store intermediate results of processing and analysis. The result is that the size of the whole tablespace exceeded 500Gb, which is the dimension over which some database administration tasks (like backups) start to become problematic.

We also experienced problems when analysing and processing data. The time required for analytic queries (that is, those involving aggregating a large number of records in order to compute distributions and totals) became unpredictable when considering queries spanning the whole dataset. In general, data access was slow, making it complex to setup and execute smooth analytics processes and pre-processing operations like the computation of indices or the cleaning of data. Moreover, researchers, for obvious reasons, could not operate with their familiar, desktop-based

tools and they were forced to sample portions of the dataset and/or work on partial views (e.g. a single province/market at a time).

It was clear from the early phases of the project that, while the RDBMS is still necessary for transactional operations on data, and performs well on datasets that are bounded in size, it is not suited to amass indefinite, fast-growing quantities of records, that should be analysed as a whole. All these issues motivated the need for an additional, and different, technology, to complement RDBMS in our data architecture.

### 2.3 *Big Data Technology*

We refer to a “big data tool” as one technological artefact specifically designed to cope with the features that differentiate so-called Big Data from traditional data, such as the size of datasets, the speed at which they are updated and the possible inclusion of non-structured content. Big Data tools are largely used in today’s data-heavy industry, where data can be produced in the order of terabytes per day. Although this order of magnitude is far from our requirements as statistical institutes, the growing attention towards the acquisition of new data sources involves size-related issues for which Big Data tools might represent a solution, as we discuss in the following of this section.

Big Data tools are based on the notion of distributed computing. The idea is to have clusters of interconnected machines working as a whole with the purpose of storing and processing data: data is spread on different nodes in the cluster and accessed in parallel by each machine. The de-facto standard for distributed computing is the open-source platform called Hadoop. Hadoop handles distribution by transparently managing inter-node communication. Users are unaware of parallelism and data files can be accessed like in a standard file system.

An Hadoop cluster can reach virtually unlimited scalability by growing in terms of space and processing power simply by adding new nodes to the cluster. The possibility of using non-specialized hardware make this solution the easier and more economical way to support large-scale computation.

An Hadoop-based system typically includes different components that constitute an eco-system for storing and analysing large-scale data. A distributed file system component (namely, HDFS) allows to store data, both structured and unstructured, organized in directories. Then, a directory can be wrapped in a table-like metadata structure, allowing analysts to query data using the familiar SQL language.

Different tools can be used to analyse the data stored on HDFS. Each tool can access the same copy of the data, allowing to pick the tool more suited for a given operation without having to make specific data extractions first. Two of the tools included in our Hadoop-based BDP were mostly used in the context of this project:

- Spark: a framework for developing programs that are executed in parallel over the Hadoop cluster.

- Massively Parallel Processing database (MPP) database: enables real-time querying capabilities on data stored on HDFS by exploiting parallelism and query optimizations.

## 2.4 Advantages of Big Data Tools

We tested our Big Data Platform by using Spark and the MPP database to implement the simulations described in Section 3. Their performance has been compared against the standard RDBMS used at Istat. We stress the fact that this is by no means intended to be a product benchmark (for this reason we decided to omit product names) that would have required putting all products in the best possible conditions in order to guarantee a fair and meaningful comparison. Our intention was to assess performance at the normal conditions in which we use our production systems, with no specific optimizations. The scanner data project was the perfect occasion to test the behaviour of both platforms, because the dataset is large enough to represent a significant testbed. Moreover we can have exact copies of data tables so that we could run the same analytical queries without modifications on both systems over tables containing the same data.

**Table 1:** Execution time of example operations, BDP vs. RDBMS

<i>Analytic query (cumulative turnover per item over a store/month/year)</i>		<i>Processing of indices at EA/store level over one year</i>	
MPP	RDBMS	Spark	RDBMS
62 sec.	699 sec.	55 min.	7 to 9 hours

Examples of execution times are showed in Table 1. We can see that through the BDP one can achieve significant improvements in performance, with a dramatic impact on the whole flow of the analytic activity: getting fast responses from a reactive data platform pushes researchers to issue more questions and to target more ambitious goals. Moreover, it makes it possible to operate on the whole dataset with no need to sample or to create extracts. On the other hand, it is important to point out that Hadoop is not meant to replace traditional data management solutions (like RDBMSs or statistical software) because its focus on large datasets introduces several trade-offs when applied to small/medium datasets.

## 3 Analyzing Scanner Data with Big Data Tools

In this section we present some of the results of an extensive analysis that we carried out in order to support the decisions of the production sector about the method that will be implemented in production for computing the indices based on scanner data.

In particular, following the Eurostat recommendation for processing scanner data [2] we considered the two approaches suggested for selecting the set GTINs that will contribute to the base indices at EA level:

- Static: follows the traditional method of considering a fixed basket of representative GTINs, to be followed over the entire year. Items that disappear along the year must be replaced.
- Dynamic: selects a representative set of items, considering the matching GTINs for each consecutive two months that have turnover over a certain threshold.

The comparison is targeted at assessing the feasibility of the two methods in the perspective of their use in production. In particular, since replacements in the static approach might be problematic if occurring for a large number of items, we are interested in evaluating the actual number of non-matching items per month, to understand if and how it could be possible to implement replacements. At the same time, we want to understand the behavior of the dynamic approach in terms of the global amount and monthly trend of matching items.

We simulated the complete process to compute EA indices in the two cases, strictly following the indications in the guidelines from Eurostat. In both cases data is firstly filtered to trim out extreme prices, then monthly prices are computed considering the arithmetic average of weekly prices, weighted by quantity, only picking the weeks that do not overlap two months. Then, additional cleaning is performed by removing all the products that belong to EAs that contribute for less than the 0.5% of the turnover for their segment or do not link to the ECOICOP classification.

### 3.1 *Static Approach*

Implementing the static approach consists in selecting the GTINs that contribute to the index. The rule we applied for this, following the guidelines, was to consider the set of GTINs that, within each store, in overall contribute to the 80% of turnover for each EA. We also consider an extraction with a 50% threshold. Analysis is carried out on 2016 data, using the total turnover in 2015 to identify the cut-off thresholds in each store/EA pair. The initial number of items in the selected set is 5,127,075 for 80% threshold and 2,014,320 for 50%, selected from a total number of items relative to December 2015 of 11,987,554.

Figure 1(a) shows the total number of non-matching items per month over all EAs, in the cases of 80% (dark grey) and 50% (light grey) thresholds. This is the number of items that should be replaced for computing the index. This number is clearly too high to be dealt with by a human operator, so some automatic method should be used.

**Figure 1:** Static approach – evaluation of unmatched items



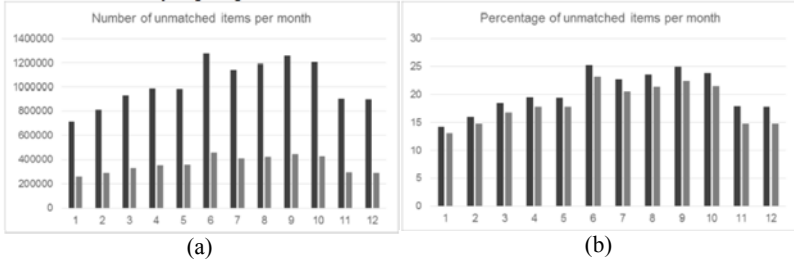


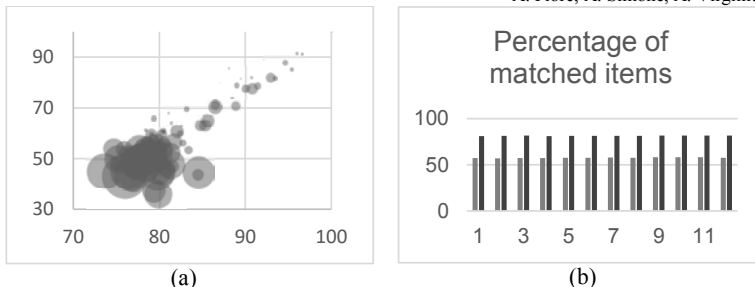
Figure 1(b) depicts the percentage of non-matching items in the two cases, highlighting an evident fluctuation in the item presence over the year. The skewness in the global distribution might be due to the influence of seasonal products and, in some case, to the lack of entire data store (data not sent by store to Nielsen), that should be estimated.

### 3.2 *Dynamic Approach*

The dynamic approach consists in the computation of micro-indices by comparing prices of products over two subsequent months. GTINs are selected on a monthly bases by applying a low-sale filter, defined in the guidelines (page 34), that should ensure that selected products contribute to a proportion of turnover in the 50%-80% range. Moreover, indices over 400% are removed. After filtering, the global number of items treated over the entire year is 122,273,323 (out of an initial amount of 238,295,930 items).

Figure 2(a) shows the average coverage of turnover per EA against the average percentage of selected products in the EA, broken by ECOICOP segment. Size of the bubbles is proportional to the total turnover generated by the segment. The graph shows that the empirical filtering rule suggested in the guidelines actually removes the expected portion of turnover, with around half of the products in the EA contributing to at least for the 70% of the turnover in most of the cases. Smaller EAs are naturally more concentrated around a small number of products, so less products are filtered out in proportion but the remaining ones contribute to almost the entire turnover. Figure 2(b) shows the percentage of matched GTINs over the year (light grey) and the corresponding percentage of turnover (dark grey). Please note that non-matching GTINs were not estimated or imputed in this phase of analysis for the sake of simplicity. An estimation of temporarily missing GTINs would ensure an even higher amount of matched GTINs and an improvement in index stability.

**Figure 2:** dynamic approach – evaluation of matched items



## 4 Discussion and Conclusions

The analysis carried out over the entire scanner data dataset allowed to derive some significant insights. Firstly, the overall number of replacements required in the static approach make this method feasible only if automatic replacements are considered. However, the realization of such an algorithm seems a rather complex task at the moment. Secondly, the dynamic approach showed a surprising stability in the portion of items matching at each couple of months, once the less relevant EAs and the less sold items were filtered out. Further analysis were carried out but were not included for lack of space, while others are planned and will be part of future work.

In general we can say the use of Big Data tools proved its benefits in terms of enhanced analytical capabilities when facing with a challenging dataset such as the scanner data one. Using a SQL dialect for accessing data allowed to make a smooth transitions to the new platform for trained database users and IT-bounded statistical analysts. The integration of our BDP with common statistical tools is nevertheless a necessary step in order to involve a broader user base of statisticians.

## References

1. Brunetti A.: Preliminary results of scanner data analysis and their use to estimate Italian Inflation. Eurostat Scanner Data Workshop, 1-2 October 2014
2. Eurostat: Practical Guide for Processing Supermarket Scanner Data (Draft), March 2017
3. Polidoro F., Virgillito A.: Italian experience and perspective of using big data to estimate inflation UNECE Meeting of the Group of Experts on Consumer Price Indices, May 2016

# Model-based Clustering with Sparse Covariance Matrices

## *Model-based Clustering con Matrici di Covarianza Sparse*

Michael Fop, Thomas Brendan Murphy and Luca Scrucca

**Abstract** We introduce *mixtures of Gaussian covariance graph models* for model-based clustering with sparse covariance matrices. The framework allows a parsimonious model-based clustering of the data, where clusters are characterized by sparse covariance matrices and the associated dependence structures are represented by graphs. The graphical models pose a set of pairwise independence restrictions on the covariance matrices, resulting in sparsity and a flexible model for the joint distribution of the variables. The model is estimated employing a penalised likelihood approach, whose maximisation is carried out using a genetic algorithm embedded in a structural-EM. The method is naturally extended to allow for Bayesian regularization in the case of high-dimensional data.

**Abstract** *In questo lavoro introduciamo una mistura di modelli grafici gaussiani per il clustering parametrico con matrici di covarianza sparse. La modellizzazione proposta permette una cluster analysis dei dati in cui i gruppi sono caratterizzati da matrici di covarianza sparse e le strutture di dipendenza tra le variabili vengono rappresentate da grafi. Il contesto dei modelli grafici permette di definire vincoli d'indipendenza tra le variabili, ottenendo modelli parsimoniosi e una rappresentazione flessibile delle distribuzioni congiunte delle variabili. Un approccio di massima verosimiglianza penalizzata è considerato per la stima del modello. La massimizzazione è effettuata tramite un algoritmo genetico incorporato in un algoritmo EM strutturale. Il modello proposto è facilmente estendibile all'analisi di dati di elevata dimensione tramite metodi di regolarizzazione bayesiana.*

**Key words:** Graphical models, genetic algorithms, model-based clustering, penalised likelihood, sparse covariance matrix.

---

Michael Fop · T. Brendan Murphy  
University College Dublin  
e-mail: michael.fop@ucdconnect.ie  
brendan.murphy@ucd.ie

Luca Scrucca  
University of Perugia  
e-mail: luca.scrucce@unipg.it

## 1 Introduction

Model-based clustering assumes that the data arise from a finite mixture of Gaussian distributions where each mixture component is associated to a cluster. In the model, the component means and covariance matrices define the characteristics of the clusters. The model complexity is led by the covariance terms and the number of parameters to be estimated grows quadratically with the number of variables in the data. To attain parsimony, a variety of methods has been proposed in the literature; for example [2, 5, 1]. However, all rely on matrix decompositions and none of them places sparsity directly on the entries of the covariance matrices.

Moreover, the model does not explicitly consider that some variables may be independent of each other in a given cluster. In fact, usually independence is obtained by considering mixture components with diagonal covariance matrices, with all the variables independent. Although parsimonious, this may be not a realistic assumption, because in many applications only some of the variables may be independent and the association structures may vary across the clusters.

## 2 Model framework

Gaussian graphical models determine a framework for estimating multivariate Normal distributions with sparse covariance matrices. In this section we incorporate this framework into model-based clustering to obtain a clustering of the data with sparse covariance matrices and groups with different dependence patterns.

### 2.1 Gaussian covariance graph model

A graph  $G$  is a mathematical object denoted as the pair  $G = (V, E)$ , where  $V$  is the set of vertices (or nodes) and  $E \subseteq V \times V$  is the set of edges.

Let us consider a graph whose vertex set represents a set of random variables  $\{X_1, \dots, X_j, \dots, X_V\}$  distributed according to a multivariate Gaussian distribution. A *covariance graph model* encodes marginal dependencies among the variables expressed in the covariance matrix  $\Sigma$  [6]. In fact, a missing edge between two nodes in the graph corresponds to marginal independence between the related variables. For a pair of variables  $(X_h, X_j)$  the following properties hold:

$$(h, j) \notin E \iff X_h \perp\!\!\!\perp X_j \iff \sigma_{hj} = 0,$$

with  $\sigma_{h_j}$  the covariance term in  $\Sigma$ . Therefore a given graph  $G$  poses a set of linear constraints on the off-diagonal entries of  $\Sigma$ , allowing the estimation of a sparse covariance matrix with the requirement of being positive definite and belonging to  $\mathcal{C}^+(G)$ , the cone of positive definite matrices induced by  $G$ .

## 2.2 Mixture of Gaussian covariance graph models

Let  $\mathbf{X}$  be the  $N \times V$  data matrix, where each observation  $\mathbf{x}_i$  is a realization of a  $V$ -dimensional vector of random variables  $\{X_1, \dots, X_j, \dots, X_V\}$ . In a mixture of Gaussian covariance graph models we assume that the density of each data point is defined as follows:

$$f(\mathbf{x}_i | \Psi, \mathbb{G}) = \sum_{k=1}^K \tau_k \phi(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, G_k) \quad \text{with } \boldsymbol{\Sigma}_k \in \mathcal{C}^+(G_k), \quad (1)$$

where  $\mathbb{G} = \{G_1, \dots, G_k, \dots, G_K\}$  is the set of graphs of the mixture components and  $\mathcal{C}^+(G_k)$  is the cone of positive definite matrices induced by graph  $G_k = (V, E_k)$ . In the model, the mixture components are characterized by different edge sets  $E_k$ , thus we allow the variables to have distinct association patterns across the clusters.

For the model in (1) we consider the penalised log-likelihood:

$$\ell(\mathbf{X}; \Psi, \mathbb{G}) = \sum_{i=1}^N \log \left\{ \sum_{k=1}^K \tau_k \phi(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, G_k) \right\} - \sum_{k=1}^K p(|E_k|), \quad (2)$$

where the penalty term  $\sum_{k=1}^K p(|E_k|)$  is a function of the number of edges  $|E_k|$  in each graph, i.e. the number of covariance parameters for each mixture component. Different penalisation terms correspond to different modeling strategies for the association among the variables and prior information can be included.

## 2.3 Model estimation

For a fixed number of components  $K$ , model estimation corresponds to estimation of mixture parameters  $\Psi$  and selection of graph structures  $\mathbb{G}$ . To accomplish the task we introduce a *structural EM algorithm* (S-EM), which allows to estimate parameters and infer graph configuration in presence of missing data [3].

The S-EM algorithm maximises a penalised version of the log-likelihood, where the penalisation term is some function of the edge set of the graph. The M step alternates the maximisation of the expected complete data log-likelihood with respect to parameters and graph configuration. In particular, the maximisation of this quantity with respect to the component graph structures is a combinatorial problem. Indeed, at each step of the algorithm the penalisation term permits to define a scoring rule for different edge sets and to search for the best graph. To tackle the problem we use a genetic algorithm [7] where the fitness function is the penalised expected complete data log-likelihood and the edge sets are expressed as binary strings. The algorithm makes use of standard genetic operators, such as mutation and crossover, and allows for an efficient search through the graph space.

## 2.4 Bayesian regularization

In the case of high-dimensional data or when the sample size is relatively small compared to the number of variables, singularities may arise in the estimation of the covariance matrix. Following [4] we propose a Bayesian regularization approach where the maximum (penalised) likelihood estimator is replaced by a maximum (penalised) a posteriori estimator. Standard prior distributions are assumed for the parameters and hyperparameters are selected appropriately for clustering.

## 3 Discussion

We introduced a framework for model-based clustering with sparse covariance matrices where clusters can be characterized by different association patterns among the variables.

The method is applied to simulated data and benchmark clustering datasets, where it is shown to give good clustering performance and insights about the relationships between the variables.

## References

1. Biernacki, C. and Lourme, A.: Stable and visualizable Gaussian parsimonious clustering models. *Statistics and Computing*. **24**(6), 953-969 (2014)
2. Celeux, G. and Govaert, G.: Gaussian parsimonious clustering models. *Pattern Recognition*. **28**(5), 781-793 (1995)
3. Friedman, N.: Learning belief networks in the presence of missing values and hidden variables. In: Fisher, D. (ed.) *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 125-133. Morgan Kaufmann (1997)
4. Fraley, C. and Raftery, A.E.: Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*. **24** 155-181 (2007)
5. McNicholas, P. and Murphy T.B.: Parsimonious Gaussian mixture models. *Statistics and Computing*. **18**(3), 285-296 (2008)
6. Richardson, T. and Spirtes, P.: Ancestral graph markov models. *The Annals of Statistics*. **30**(4), 962-1030 (2002)
7. Sivanandam, S.N. and Deepa, S.N.: *Introduction to Genetic Algorithms*. Springer-Verlag, Berlin (2007)

# Quantile Regression for Functional Data

## *Regressione Quantile per Dati Funzionali*

Maria Franco-Villoria and Marian Scott

**Abstract** Quantile regression allows estimation of the relationship between response and explanatory variables at any percentile of the distribution of the response (conditioned on the explanatory variables). We extend quantile regression to the functional case, rewriting the quantile regression model as a generalized additive model where both the functional covariates and the functional coefficients are parametrized in terms of B-splines. Parameter estimation is done using a penalized iterative reweighted least squares (PIRLS) algorithm. We evaluate the performance of the model by means of a simulation study.

**Abstract** La regressione quantile permette di stimare la relazione fra una variabile risposta e delle covariate considerando un qualsiasi percentile della distribuzione (condizionata alle covariate) della risposta. In questo lavoro si estende la regressione quantile al caso di dati funzionali, riscrivendo il modello di regressione come un modello additivo generalizzato dove sia le covariate funzionali che i coefficienti funzionali vengono parametrizzati attraverso B-splines. La stima dei parametri viene effettuata attraverso un algoritmo iterativo di minimi quadrati pesati. La performance del modello valutata in uno studio di simulazione.

**Key words:** B-splines, functional coefficient, generalized additive model, PIRLS

## 1 Introduction

Linear regression has the goal of estimation of the expected value of the response variable and its dependence on any set of explanatory variables. However, there

---

Maria Franco-Villoria  
University of Torino, Italy e-mail: maria.francovilloria@unito.it

Marian Scott  
University of Glasgow, UK e-mail: Marian.Scott@glasgow.ac.uk

might be situations in which the mean of the distribution is not informative, e.g. if one is interested in the high values of a given variable. Quantile regression [7] allows estimation of the relationship between response and explanatory variables at any percentile of the distribution of the response (conditioned on the explanatory variables). As a result, rates of change in the response variable can be estimated for the whole distribution and not only in the mean. Quantile regression is widely used and has been applied in different fields such as finance, medicine or the environment. On the other hand, growing dimensionality of data available has stimulated the development of models for functional data [10], where the observed data are considered as a discrete realization of an underlying smooth function, i.e. a curve. In this work, we extend quantile regression to the functional case. However, the definition of a quantile in a functional data setting is not straightforward given the lack of a distribution function. An interesting proposal to define functional quantiles is that of Lopez-Pintado and Romo [8], who propose to order the curves based on their depth, where the deepest curve would correspond to the median. Quantile regression for functional data is a relatively new area of research that has only been explored in recent years, hence literature available is very limited. Cardot, Cambres e Sarda [1, 2] and Kato [6] have extended functional linear regression models to the case of quantile regression considering functional covariates and scalar response. A non-parametric version was proposed by Dabo-Niang e Laksaci [5], while Crambes, Gannoun and Henchiri [3, 4] use support vector machine methods for fitting quantile regression models where the covariates are functional and the response is scalar.

Regression models for functional data need to be addressed differently depending on whether the response variable is scalar or functional. In section 2 we discuss how the quantile regression coefficients can be estimated when the response variable is scalar, while in Section 3 we present preliminary results from a simulation study. Extension to the functional response case is briefly discussed in Section 4.

## 2 The Model

For  $\tau \in (0, 1)$  fixed, a quantile regression model:

$$Q_Y(\tau|x(t)) = \alpha + \int_T \beta(t)x(t)dt = \alpha + \langle \beta, x \rangle$$

where  $Y$  is a scalar response variable,  $x(t)$  is a functional covariate,  $Q_Y(\tau|x(t))$  is the  $100\tau^{th}$  quantile of the distribution of  $Y|x(t)$  and  $\langle \cdot, \cdot \rangle$  is the inner product. The parameters  $\alpha$ ,  $\beta(t)$  can be estimated by minimizing the objective function:

$$R(\alpha, \beta(t)) = \sum_{i=1}^n \rho_\tau(y_i - (\alpha - \int_T \beta(t)x_i(t)dt))$$

where  $\rho_\tau(u) = u(\tau - I(u < 0))$  is the check function and  $I$  is an indicator function.



The quantile regression model can be rewritten as a generalized additive model where both the functional covariate and the functional coefficients are parametrized in terms of B-spline basis functions. The objective function to be minimized is a sum of asymmetrically weighted absolute residuals; in the quantile regression literature, linear programming methods are used to estimate the unknown regression parameters. Instead, we approximate the absolute residuals with the squared residuals and adjust the weights accordingly. This way the regression coefficients can be estimated using a penalized iterative reweighted least squares (PIRLS) algorithm.

### 3 Preliminary results

We evaluate the performance of the estimating algorithm by means of a simulation study, where we consider different sample sizes, two levels of noise and various forms of complexity for the functional coefficient. We evaluate the performance at four different quantiles  $q_{0.2}$ ,  $q_{0.5}$ ,  $q_{0.7}$  and  $q_{0.9}$ . The simulated data are built as

$$y_i^{sim} = \alpha + \int_T \beta(t)x_i(t)dt + \varepsilon_i.$$

The functional covariate  $x(t) = \sum_{j=1}^{10} \xi_j B_j(t)$ , where  $B_j(t)$  are B-spline basis functions evaluated at  $t \in T = [0, 1] \subset \mathbb{R}$ ,  $j = 1, \dots, 10$  and the spline coefficients  $\xi_j \sim N(0, 1)$ . The random errors  $\varepsilon_i$  are simulated from a normal distribution  $N(q_\tau, \sigma^2)$  with  $q_\tau$  the  $100\tau^{th}$  quantile of the  $N(0, \sigma^2)$ ; values of  $\sigma$  were chosen to ensure a signal to noise ratio of 2 and 4.

To evaluate how well  $\beta(t)$  is estimated, we consider two indicators, the distance ( $L_2$  norm) between the simulated and estimated coefficient and the proportion of negative and positive residuals. Results from a preliminary simulation study suggest that the method performs well; when the sample size is small ( $n = 50$ ) distance values range from 0.01 to 0.45 when the functional coefficient is linear and from 0.1 to 1.2 when the functional coefficient is non-linear. Results improve with increasing sample size and the closer we get to the median, as expected. The percentages of positive and negative residuals were very close to the expected  $100(1 - \tau)\%$  and  $100\tau\%$  respectively. Convergence was reached after 4 to 29 iterations.

### 4 Discussion and Future Work

In this work, we propose a quantile regression model when the covariates are functional and the response is scalar. Preliminary results from a first simulation study suggest good performance for a range of different quantiles. The model can be easily extended to incorporate more covariates keeping the computational cost low thanks to the use of sparse matrix computation.

We are currently working on the case of a quantile regression model where the response is functional too. In this case, the residuals themselves are functional data and working out the weights is not as straightforward as in the scalar response case. A possibility would be to consider the distance from the zero curve as a proxy for the size of each residual, while the sign of the residual could be worked out using some sort of curve ordering technique such as band depth or a more recent proposal based on epigraphs and hypographs [9].

In particular, quantile regression for functional data could prove useful in solving the problem of uncertainty evaluation of a predicted curve, where the 2.5% and 97.5% quantiles could be used to build a functional confidence band.

## References

1. Cardot, H., Crambes, C., Sarda, P.: Conditional quantiles with functional covariates: an application to ozone pollution forecasting. In: Antoch, J. (ed.) *Compstat 2004 Proceedings*, pp. 769–776. Physica-Verlag (2004)
2. Cardot, H., Crambes, C., Sarda, P.: Quantile regression when the covariates are functions. *Journal of Nonparametric Statistics* **17**(7), 841–856 (2005)
3. Crambes, C., Gannoun, A., Henchiri Y.: Support vector machine quantile regression approach for functional data: Simulation and application studies. *Journal of Multivariate Analysis* **121**, 50–68 (2013)
4. Crambes, C., Gannoun, A., Henchiri Y.: Modelling functional additive quantile regression using support vector machines approach. *Journal of Nonparametric Statistics* **26**(4), 639–668 (2014)
5. Dabo-Niang, S., Laksaci, A.: Nonparametric Quantile Regression Estimation for Functional Dependent Data. *Communications in Statistics - Theory and Methods* **41**(7), 1254–1268 (2012)
6. Kato, K.: Estimation in functional linear quantile regression. *The Annals of Statistics* **40**(6), 3108–3136 (2012)
7. Koenker, R.: *Quantile Regression*. Cambridge University Press (2005)
8. Lopez-Pintado, S., Romo, J.: On the concept of depth for functional data. *J. Amer. Statist. Assoc.* **104**(486), 718–734 (2009)
9. Martin-Barragan, B., Lillo, R.E., Romo, J.: Functional boxplots based on epigraphs and hypographs. *Journal of Applied Statistics* **43**(6), 1088–1103 (2016)
10. Ramsay, J.O., Silverman, B.W.: *Functional Data Analysis*. pringer, Dordrecht (2005)

# Three-way compositional data: a multi-stage trilinear decomposition algorithm

## *Dati composizionali a tre vie: un algoritmo per la decomposizione trilineare a più stadi*

Gallo M., Simonacci V., and Di Palma M.A.

**Abstract** The CANDECOMP/PARAFAC model is an extension of bilinear PCA and has been designed to model three-way data by preserving their multidimensional configuration. The Alternating Least Squares (ALS) procedure is the preferred estimating algorithm for this model because it guarantees stable results. It can, however, be slow at converging and sensitive to collinearity and over-factoring. Dealing with these issues is even more pressing when data are compositional and thus collinear by definition. In this talk the solution proposed is based on a multi-stage approach. Here parameters are optimized with procedures that work better for collinearity and over-factoring, namely ATLD and SWATLD, and then results are refined with ALS.

**Abstract** *Il modello CANDECOMP/PARAFAC è una generalizzazione per matrici a tre indici della ACP. Per stimare i parametri di tale modello la procedura di stima più usata è l'Alternating Least Squares (ALS). Tale algoritmo è il più usato in quanto garantisce risultati stabili, tuttavia, presenta anche degli inconvenienti, quali essere lento e sensibile alla multicollinearità e alla sovra-fattorizzazione. Affrontare questi problemi diventa poi particolarmente impegnativo quando i dati sono multi-collineari per costruzione, come nel caso dei dati composizionali. Come soluzione di tali problemi, nel presente lavoro si propone un approccio multi-stadio in cui i parametri sono prima ottimizzati con procedure che funzionano meglio quando vi è collinearità e sovra-fattorizzazione, cioè ATLD e SWATLD, e successivamente i risultati finali sono individuati con l'ALS.*

---

Michele Gallo  
DISUS University of Naples "L'Orientale", Largo S. Giovanni Maggiore 30, 80134 Naples, Italy.  
e-mail: mgallo@unior.it

Violetta Simonacci  
DISUS University of Naples "L'Orientale", Largo S. Giovanni Maggiore 30, 80134 Naples, Italy.  
e-mail: vsimonacci@unior.it

Maria Anna Di Palma  
DISUS University of Naples "L'Orientale", Largo S. Giovanni Maggiore 30, 80134 Naples, Italy.  
e-mail: madipalma@unior.it

**Key words:** PARAFAC-ALS, ATLD, SWATLD, Compositional Data, Log-ratios

## 1 Introduction

Observations over a set of variables can be recorded in different occasions, such as time or location. These data present a tridimensional structure and the only way to obtain a low rank approximation without confusing the variability of two dimensions together is using multi-linear techniques such as the CANDECOMP/PARAFAC (CP) model [2, 10]. This model estimates three separate sets of parameters, one for each mode of the analysis, thus is highly complex and the search for innovative ways to improve its efficiency without compromising accuracy of results is of great relevance.

The most widely used algorithm for the CP model is currently PARAFAC-ALS (ALS) thanks to the merit of granting stable results, a least square solution and an always monotonically decreasing fit. It does, however, present some problematic aspects such as slow convergence and sensitiveness to over-factoring, multicollinearity and factor collinearity. These issues are even more significant when dealing with data that present particular challenges such as Compositional Data (CoDa) [1, 11]. CoDa can be defined as positive vectors with a purely multicollinear structure as their elements describe the parts of a whole and thus only carry relative information.

Given these considerations, in [9] an alternative way to overcome these difficulties in a compositional framework is presented. Specifically it is suggested that in order to mitigate ALS inefficiencies this procedure can be integrated by adding an initialization/recovery stage where parameters are optimized through the Self-Weighted TriLinear Decomposition (SWATLD). In this manner a novel two-stage procedure is implemented (INT-1).

SWATLD proposed by [3] was chosen amongst other alternatives because it can be seen as complementary to ALS given that its strengths are fast convergence and robustness to over-factoring and collinearity while its fallacies are finding a solution in a non-least-square sense and unstable results [5, 12, 14, 16].

INT-1 appears to work quite well in the simulations presented in the cited article, however several ways to improve its performance and reliability were suggested in future developments but not yet verified. In this perspective the purpose of this contribution is to explore the possibility of improving the performance of INT-1 by trying to answer two unresolved queries.

The first question is the consequence of a methodological comparison with [15] where it is argued that the Alternating TriLinear Decomposition (ATLD) proposed in [13] works better than SWATLD for initializing random numbers, multicollinearity and speediness. We thus wondered if ATLD could be considered as an initialization step. To resolve this, a second multi-stage procedure (INT-2) was devised, this time with three stages, to see if adding an ATLD step to start off could improve performance.

The second problem concerns the identification of an optimal transition point from one stage to the next, i.e. are there optimal convergence criteria capable of making INT-1 and INT-2 perform at their best? This question is addressed in a simulation study on stage transition parameters.

Once these two aspects are dealt with, a new comparative study can be carried out to verify three points of interest: 1) how INT-1 and INT-2 perform with respect to ALS for compositional data; 2) which between INT-1 and INT-2 is a better alternative; and 3) how do data characteristics such as noise level and factor collinearity influence results.

## 2 Compositions in a CP model

Let us consider a three-way array  $\underline{\mathbf{V}} (I \times J \times K)$  with generic positive element  $v_{ijk}$  where  $i = 1 \dots I$ ,  $j = 1 \dots J$ , and  $k = 1 \dots K$ . If its row vectors  $\mathbf{v}_{ik} = [v_{i1k}, \dots, v_{iJk}]$  present a biased covariance structure due to an implicit or explicit sum constraint  $v_{i1k} + \dots + v_{iJk} = \kappa$ , where  $\kappa$  is a positive constant, the array has a compositional structure and should be processed with compositional methodology.

This bounded covariance imposes a purely multicollinear structure to the data since the elements of a compositional vector are not linearly independent. As a consequence the covariance matrix for each of the  $K$  frontal slabs  $\mathbf{V}_k(I \times J)$  of the array  $\underline{\mathbf{V}}$  will be singular.

From a geometric stand point these row vectors are forced in a subspace of  $\mathfrak{R}_+^J$  known as simplex and defined as:

$$S^J = \{(v_{i1k}, \dots, v_{iJk}) : v_{i1k} \geq 0, \dots, v_{iJk} \geq 0; v_{i1k} + \dots + v_{iJk} = \kappa\} \quad (1)$$

To operate within this subspace a non-Euclidean set of rules, known as Aitchison geometry, is used to identify a linear vector space [11]. Compositional vectors can, however, be converted into Euclidean space coordinates by using log-ratio transformations: pairwise, centered, additive [1] or isometric [4].

For the purpose of this contribution we will only be referring to centered log-ratio (*clr*) coordinates which can be expressed as:

$$\mathbf{z}_{ik} = clr(\mathbf{v}_{ik}) = \left[ \ln \frac{v_{i1k}}{g(\mathbf{v}_{ik})}, \dots, \ln \frac{v_{iJk}}{g(\mathbf{v}_{ik})} \right] \quad \text{with } g(\mathbf{v}_{ik}) = \sqrt[ J ]{ \prod_{j=1}^J v_{ijk} } \quad (2)$$

By applying this transformation the tridimensional array of compositions  $\underline{\mathbf{V}}$  can easily be changed into an array of *clr*-coordinates  $\underline{\mathbf{Z}}$  so that standard algorithms can be applied as long as results are interpreted in compositional terms [6, 8]. It is important to note that *clr*-coordinates by providing an  $S^J$  to  $\mathfrak{R}^J$  projection, do not remove the collinearity problem.

An array of *clr*-coordinates  $\underline{\mathbf{Z}}$  can be decomposed with the CP model in three sets of parameters, one for each mode of the analysis. Let  $F$  be the number of considered

factors, using a slab-wise notation we can write:

$$\mathbf{Z}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^t + \mathbf{E}_k \quad k = 1, \dots, K \quad (3)$$

where  $\mathbf{A}$  ( $I \times F$ ) and  $\mathbf{B}$  ( $J \times F$ ) are the loading matrices for the first and second mode, respectively;  $\mathbf{D}_k$  is a diagonal matrix containing the  $k$ th row of  $\mathbf{C}$  ( $K \times F$ ), loading matrix of third mode;  $\mathbf{Z}_k$  ( $I \times J$ ) is the  $k$ th frontal slab of  $\underline{\mathbf{Z}}$ ; and  $\mathbf{E}_k$  ( $I \times J$ ) is the corresponding frontal slab of the error array  $\underline{\mathbf{E}}$ .

### 3 Estimating procedures

Different algorithms can be used to fit the data to the model. The most common one is ALS. This is an iterative procedure where sets of parameters are estimated in three successive least-square steps. On the other hand, ATLD and SWATLD are also three-step iterative procedures but do not follow a least-square approach and are characterized by the use of three distinct objective function, one for each mode, which focus on prioritizing the trilinear structure of the data.

The described algorithms all present some qualities and weaknesses directly derived from the properties of their loss functions. ATLD is the fastest at converging and it is robust to over-factoring, collinearity and initial values. It does not, however, find a least-square solution, it may not monotonically decrease, it is sensitive to noise and often does not converge properly.

On the opposite end there is ALS, the slowest at converging, stable in its results, capable of finding a solution in the least square sense but sensitive to collinearity and over-factoring. SWATLD occupies a middle ground: it is more stable than ATLD but not quite as reliable as ALS, it is pretty fast at converging but slower than ATLD while still robust to over-factoring and collinearity. In addition it may still not have a monotonically decreasing fit and not converge to a least square solution.

Given these considerations two multi-stage procedures were devised to try and maximize the advantages and counter-balance the inefficiencies of these algorithms.

INT-1 is structured in the following manner: in a first stage (recovery stage) parameters are estimated by SWATLD with the purpose of identifying the correct underlying components in case of over-factoring, to deal better with multicollinearity and to speed up the procedure; successively in a second stage the solution is adjusted through ALS steps (refinement stage) to obtain a least square solution and avoid SWATLD instabilities.

INT-2 presents a similar outline but also includes an additional initialization ATLD stage, which could help when dealing with multicollinearity and bad initial values. A schematic overview of the procedures is displayed in Fig.1. In both cases step transition can be user defined in terms of relative fit and number of iterations. However these transition parameters can hugely hinder or improve performance of both INT-1 and INT-2, thus ideal values will be identified through a threshold simula-

tion study. It is also important to note that for both algorithms at least one iteration has to be performed at each stage. Once optimal parameters are found, they will be included as defining elements of the procedures.

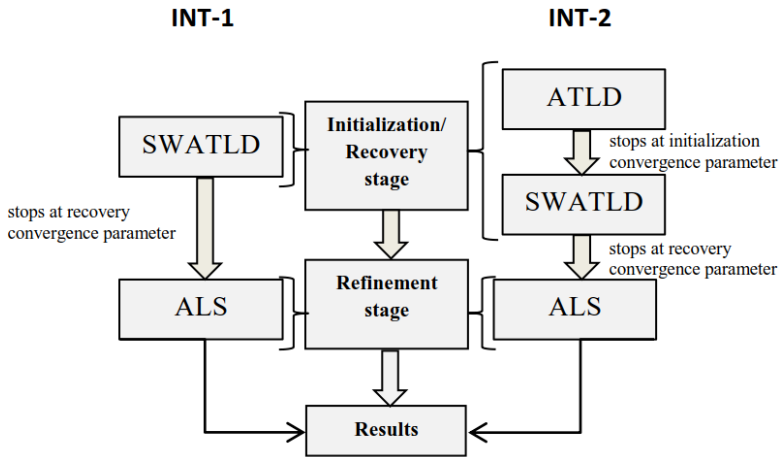


Fig. 1 Multi-stage procedures outline

## 4 Conclusion

With the purpose of further developing the findings presented in [9], where a two stage SWATLD-ALS algorithm was introduced, this contribution proposes two important advancements: 1) devising a three-step INT-2 procedure to see if initializing with ATLD grants additional benefits; and 2) setting up a study to identify ideal stage transition parameters for both INT-1 and INT-2.

To test the goodness of the proposed modifications, a comparative simulation study between INT-1, INT-2 and ALS will then be carried out in a compositional setting.

Given that only partial results are available at this stage, we can make the following considerations. In terms of ideal transition parameters, there is a trade-off between accuracy and efficiency: stricter relative fit convergence criteria ( $10^{-3}$  or  $10^{-4}$ ) render the algorithms more efficient but less stable. On the other hand looser criteria are less fast but more reliable ( $10^{-1}$  or  $10^{-2}$ ) and for this reason generally preferable.

In terms of comparative results we expect to see INT-1 and INT-2 (set up with ideal parameters) performing similarly to ALS in terms of reliability for correct factor estimation and better in case of over-factoring while being far more efficient. INT-1

will probably be slightly more reliable but a little slower than INT-2. Complete and in-depth results will be discussed during presentation.

## References

1. Aitchison, J.: *The Statistical Analysis of Compositional Data*. Chapman & Hall (1986)
2. Carroll, J. D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. *Psychometrika* 35(3), 283–319 (1970)
3. Chen, Z.P., Wu, H.L., Jiang, J.H., Li, Y., Yu, R.Q.: A novel trilinear decomposition algorithm for second-order linear calibration. *Chemometrics and Intelligent Laboratory Systems* 52(1):75–86 (2000)
4. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras G., Barcelo-Vidal C.: Isometric logratio transformations for compositional data analysis. *Mathematical Geology* 35(3):279–300 (2003)
5. Faber, N.K.M., Bro, R., Hopke, P.K.: Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometrics and Intelligent Laboratory Systems* 65(1):119–137 (2003)
6. Gallo, M.: Log-ratio and parallel factor analysis: an approach to analyze three-way compositional data. In: *Advanced dynamic modeling of economic and social systems*, Springer, pp 209–221 (2013)
7. Gallo, M., Buccianti, A.: Weighted principal component analysis for compositional data: application example for the water chemistry of the arno river (Tuscany, central Italy). *Environmetrics* 24(4):269–277 (2013)
8. Gallo, M., Simonacci V.: A procedure for the three-mode analysis of compositions. *Electronic Journal of Applied Statistical Analysis* 6(2):202–210 (2013)
9. Gallo, M., Di Palma, M.A., Simonacci V.: Integrated SWATLD-ALS algorithm for Compositional Data (2016). Submitted
10. Harshman, R.A.: *Foundations of the PARAFAC procedure: Models and conditions for an 'exploratory' multi-modal factor analysis* (1970)
11. Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado R.: *Modeling and analysis of compositional data*. John Wiley & Sons (2015).
12. Tomasi, G., Bro, R.: A comparison of algorithms for fitting the PARAFAC model. *Computational Statistics & Data Analysis* 50(7):1700–1734 (2006)
13. Wu, H.L., Shibukawa, M., Oguma, K.: An alternating trilinear decomposition algorithm with application to calibration of for simultaneous determination of overlapped chlorinated aromatic hydrocarbons. *Journal of Chemometrics* 12(1), (1998)
14. Yu, Y.J., Wu, H.L., Nie, J.F., Zhang, S.R., Li, S.F., Li, Y.N., Zhu, S.H., Yu, R.Q.: A comparison of several trilinear second-order calibration algorithms. *Chemometrics and Intelligent Laboratory Systems* 106(1):93–107 (2011)
15. Yu, Y. J., Wu, H. L., Kang, C., Wang, Y., Zhao, J., Li, Y. N., Liu, Y.J., Yu, R. Q.: Algorithm combination strategy to obtain the second order advantage: simultaneous determination of target analytes in plasma using three dimensional fluorescence spectroscopy. *Journal of Chemometrics*, 26(5), 197–208 (2012)
16. Zhang, S.R., Wu, H.L., Yu, R.Q.: A study on the differential strategy of some iterative trilinear decomposition algorithms: PARAFAC-ALS, ATLD, SWATLD, and APTLD. *Journal of Chemometrics* 29(3): (2015)



# Nonparametric shared frailty model for classification of survival data

## *Modelli con termine random condiviso per la classificazione di dati di sopravvivenza*

Francesca Gasperoni, Francesca Ieva, Anna Maria Paganoni, Chris Jackson and Linda Sharples

**Abstract** In this work, we propose an innovative model for fitting grouped survival data and for detecting a second level of clusters among groups. In order to achieve this goal, we start from a classical semiparametric Cox model and we add a nonparametric discrete random term as a multiplicative factor. This research question arose from a project about healthcare management of Regione Lombardia. We analyze a rich administrative database, where several information about patients is collected (i.e. dates of hospitalizations, death, comorbidities, procedures etc.). In this framework, patients are the statistical units and hospitals are the known groups. Through the application of this new model, we are able to detect hidden populations among hospitals and we provide a clustering tool for survival data.

**Abstract** *In questo lavoro proponiamo un modello innovativo per dati di sopravvivenza raggruppati, al fine di identificare una possibile classificazione dei gruppi. Per raggiungere l'obiettivo, siamo partiti da un modello semiparametrico di Cox e abbiamo aggiunto un termine moltiplicativo aleatorio. Questa domanda di ricerca è sorta a partire da un progetto di Regione Lombardia focalizzato sul manage-*

---

Francesca Gasperoni

MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: francesca.gasperoni@polimi.it

Francesca Ieva

MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: francesca.leva@polimi.it

Anna Maria Paganoni

MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: anna.paganoni@polimi.it

Chris Jackson

MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge (UK), CB2 0SR, e-mail: chris.jackson@mrc-bsu.cam.ac.uk

Linda Sharples

Department of Medical Statistics, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, e-mail: Linda.Sharples@lshtm.ac.uk

*ment delle strutture ospedaliere. Abbiamo analizzato un database amministrativo, in cui sono state raccolte diverse informazioni sui pazienti (date dei ricoveri, date di morte, comorbidità, operazioni etc.). In questo caso, i pazienti sono le unità statistiche e gli ospedali sono i gruppi noti. Grazie all'applicazione di questo modello innovativo, siamo in grado di identificare popolazioni latenti fra gli ospedali e forniamo un metodo di clustering per dati di sopravvivenza.*

**Key words:** Hierarchical clustering, time-to-event data, administrative databases.

## 1 Introduction

Classical survival models implicitly assume that statistical units are homogenous, meaning that all individuals have the same risk of experiencing the event of interest. However, this is a strong assumption. Indeed, it is almost impossible to know all relevant risk factors, because of time or, often, because of financial issue. This negligence of covariates leads to unobserved heterogeneity, which means different risks for different subjects. A possible way to model this heterogeneity consists in including an additional random term, commonly known as frailty term. A frailty model is a random effects model for time-to-event data, where the random effect (frailty) has a multiplicative effect on the baseline hazard, [13]. Usually, univariate frailty models are chosen for capturing the heterogeneity due to unobserved covariates, as it was introduced by Vaupel et al., [12].

Another, completely different, use of the random term faces the issue of independence among different observations when the statistical units are clustered in several groups. Indeed, these frailty models provide an interesting way to capture the dependence between observations within a group and/or the heterogeneity among groups. These models, also known as shared frailty models, are hazard models with a multiplicative random term, common to all members of the same group. This random term is the realization of a specified distribution implying the need to fix a priori the distribution which fits best the data.

The most common distributions for a frailty term are Gamma and log-Normal, and this is a consequence of the simplicity of the computations and the availability of the software. Indeed, for the Gamma frailty term, Therneau and Grambsch, [10], describe a penalized partial likelihood approach which accelerates the classical computation made through the EM algorithm. This method is implemented in the function `coxph`, which is part of `survival` package, [9, 10]. The log-Normal distributed frailty was introduced for the first time by McGilchrist, [8], and it is implemented efficiently in a specific package named `coxme`, [11].

However, these are just two options for the distribution choice, since there is quite a wide variety of other distributions, introduced by Hougaard, [6, 7] and by Aalen [1]. A deep overview of these frailty models is given by Duchateau and Janssen in their book [4], where theoretical computations and examples are reported

also for the less used Positive Stable, Power Variance and Compound Poisson distributions.

Then, a key issue is how to choose among all these distributions. There is not any guideline available right now for answering this question. So, it is common to see analysis in which the performances of different frailty distributions are compared.

The aim of this work is to introduce a nonparametric frailty term, which allows us to avoid the choice of a specific shape for our frailty and to cluster observations, at the same time. Indeed, the introduction of a nonparametric random term makes us able to detect a possible latent structure in the dataset. The proposal of this in-built clustering technique is the most innovative aspect of this paper since there is no way of clustering time-to-event data, to the best of our knowledge.

Moreover, only few works dealt with a discrete frailty term. For example, Guo and Rodriguez, [5], proposed a nonparametric frailty term with two masses for estimating the death hazard rate for children in Guatemala. Caroni et al., [2], proposed three different discrete distributions for the frailty term, Poisson, Negative Binomial and Geometric, while Dos Santos et al., [3], proposed a comparison between a model with parametric baseline and nonparametric frailty and a model with nonparametric baseline and parametric frailty. In all these papers the baseline is parametric, usually a Weibull. No one of them proposed a model combining the two critical points: a nonparametric baseline and a nonparametric frailty term.

In this work, we want to introduce for the first time a new model for the hazard rate which considers a semi-parametric Cox model together with a group specific nonparametric frailty term.

## 2 Models and Methods

The aim of this paper consists in investigating a possible dependence among known groups, through detecting a second clustering level. *Groups* are the first clustering level, while *hidden populations* are the second clustering level. In order to achieve this goal, we propose a semiparametric Cox model with a nonparametric discrete frailty term.

$T_{ij}$  is the random variable which models time-to-event for the  $i$ -th statistical unit in the  $j$ -th group, where  $i \in \{1, \dots, n_j\}$  and  $n_j$  is the number of statistical units collected in group  $j$ .  $C_{ij}$  represents the censoring time and  $\delta_{ij} = 1_{\{T_{ij} \leq C_{ij}\}}$  is the status index.  $X_{ij}$  is a  $p$ -dimensional vector of covariates,  $\beta$  is the parameters vector and  $\lambda_0(t_{ij})$  represents the nonparametric baseline hazard. Then, we suppose the existence of  $K$  hidden populations, each one characterized by a frailty level  $w_k$ , and we introduce a new random variable  $Z_{jk}$  which is equal to 1 if the  $j$ -th cluster belongs to the  $k$ -th population,  $Z_{jk} \stackrel{i.i.d}{\sim} Be(\pi_k)$ . To sum up, we have the following extra parameters in the model:  $K$ , the number of hidden populations, with  $K \leq J$ ;  $\pi = [\pi_1, \pi_2, \dots, \pi_K]$ , the probability vector such that  $\sum_{k=1}^K \pi_k = 1$ , and  $\mathbf{w} = [w_1, w_2, \dots, w_K]$ , the frailty levels vector. In this case the hazard rate becomes:

$$\lambda(t_{ij}) = \lambda_0(t_{ij})w_k \exp(X_{ij}^T \beta) \quad i \in \{1, \dots, n_j\} \quad j \in \{1, \dots, J\} \quad k \in \{1, \dots, K\}$$

Starting from the hazard rate, we are able to explicitly write the full likelihood of our model as:

$$L_{full} = \prod_{k=1}^K \prod_{j=1}^J \prod_{i=1}^{n_j} \left\{ [\lambda_0(t_{ij})w_k \exp(X_{ij}^T \beta)]^{\delta_{ij}} \cdot \exp[-\Lambda_0(t_{ij})w_k \exp(X_{ij}^T \beta)] \right\}^{z_{jk}} \cdot \pi_k^{z_{jk}}.$$

Finally, we compute the parameters estimates of the proposed model through a proper Expectation-Maximization algorithm. In particular, the  $\beta$ , the cumulative baseline hazard  $\Lambda(t_{ij})$ ,  $\pi$  and  $w_k/w_1$ , with  $k = 2 : K$ , are estimated inside the EM. We estimate the ratio of frailty levels instead of the single frailty level because of an identifiability issue. On the contrary, the value of  $K$  is estimated outside the EM algorithm, through a backward selection method.

### 3 Application to clinical administrative database

The dataset is extracted from the clinical administrative database of Regione Lombardia, and concerns patients that have been hospitalized with a diagnosis of chronic heart failure between 2005 and 2012. We have a total of 164,384 events (admission in, discharge from hospital and death) for a total of 43,998 patients, identified by anonymous codes. The total number of hospitals is 307.

For this specific application, we focus our attention only on the second admission to hospital, omitting those patients that died before the second admission. So, the final cohort is composed by 24,075 patients and the recorded hospitals are 291. We apply our algorithm on the selected cohort, considering hospitals as known groups, and we include three covariates in the model: gender, age and comorbidity index. Finally, we detect four different populations, with the following vector of proportions:  $\pi = [0.291, 0.273, 0.274, 0.162]$ . The estimates of the regression parameters are  $\beta_{GENDER} = 0.26$ ,  $\beta_{AGE} = 0.04$  and  $\beta_{COM} = 0.37$ . Speaking of the frailty ratios, we obtain that  $w_2/w_1$  is 1.60,  $w_3/w_1$  is 1.99 and  $w_4/w_1$  is 2.35. Then, we can conclude that being hospitalized in a hospital that belongs to the second population rather than being hospitalized in a hospital which belongs to the first population leads to a higher instantaneous risk of being readmitted. Similar observations can be done for population 3 and 4.

In order to visualize the latent structures, we plot the Kaplan-Meier estimates of the split cohort, see Fig.1. It is immediate to see how the survival curves linked to different populations are distant one from each other (neither the confidence intervals overlap). It is also important to notice the order of curves, indeed we can see that the risk increases from the first population to the fourth one, as it is expected from the estimated frailty ratios.

## 4 Conclusions

To the best of our knowledge, the idea of detecting a second level of clustering structure through a nonparametric discrete frailty term has never been investigated in survival research field. Moreover, there is no available software that allows to implement discrete frailty term. Finally, the application to a clinical administrative database is very powerful, since it should have a great impact on healthcare management policies.

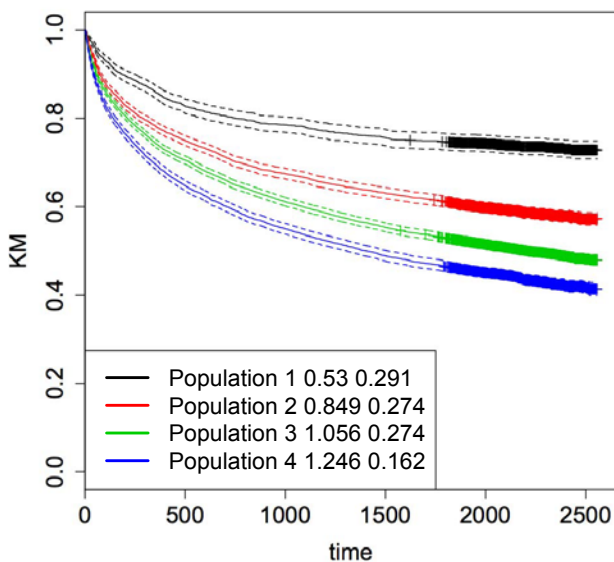


Fig. 1 Kaplan-Meier estimates computed for the four populations identified by our algorithm.

## References

1. Aalen, O.O.: Modelling Eterogeneity in Survival Analysis by the Compound Poisson Distribution. *Ann. Appl. Probab.* **2**, 951–972 (1992)
2. Caroni, C., Crowder, M., Kimber, A.: Proportional hazards models with discrete frailty. *Life-time Data Anal.* **16**, 374–384 (2010)

3. dos Santos, D. M., Davies, R. B., Francis, B.: Nonparametric hazard versus non parametric frailty distribution in modelling recurrence of breast cancer. *J. Stat. Plan. Inference.* **47**, 111–127.
4. Duchateau, L., Janssen, P.: *The frailty model.* Springer Science & Business Media (2007)
5. Guo, G., Rodriguez, G.: Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala. *JASA*, **87**, 969–976 (1992)
6. Hougaard, P.: A class of multivariate failure time distributions. *Biometrika*. **73**, 671–678 (1986a)
7. Hougaard, P.: Survival models for heterogeneous populations derived from stable distributions. *Biometrika*. **73**, 387–396 (1986b)
8. McGilchrist, C. A.: REML estimation for survival models with frailty. *Biometrics*. **49** 221–225 (1993)
9. Therneau T.: A Package for Survival Analysis in S. version 2.38, <http://CRAN.R-project.org/package=survival>. (2015)
10. Therneau, T. M., Grambsch, P. M.: *Modeling survival data: extending the Cox model.* Springer Science & Business Media. (2000)
11. Terry M. Therneau: *coxme: Mixed Effects Cox Models.* R package version 2.2-5. <http://CRAN.R-project.org/package=coxme> (2015)
12. Vaupel, J. W., Manton, K. G., Stallard, E.: The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*. **16**, 439–454 (1979)
13. Wienke, A.: *Frailty models in survival analysis.* CRC Press (2010)

# Clustering landmark-based shapes using Information Geometry tools

## *La Geometria dell'Informazione per la cluster analysis delle forme basate sui landmark*

Stefano A. Gattone and Angela De Sanctis

**Sommario** In this talk we shall describe a method for clustering shapes configurations in two dimensions. Variation in the shape space is obtained by introducing deformations carrying individual landmarks from one to another. The framework, provided by the Information Geometry, is the following. A shape is represented by a probability distribution. Then, a Riemannian metric is defined on the shape space and the length of the geodesics with respect to this metric is used to measure differences in shape.

**Sommario** *In questa talk descriveremo un metodo per condurre analisi di cluster delle forme in due dimensioni utilizzando la Geometria dell'Informazione. La variabilità nello spazio della forma viene trattata attraverso deformazioni da un landmark all'altro. L'ambito di lavoro fornito dalla Geometria dell'Informazione è il seguente. Una forma è rappresentata da una distribuzione di probabilità. Una metrica Riemanniana è definita sullo spazio della forma e la lunghezza delle geodetiche calcolate rispetto a questa metrica viene utilizzata per misurare le differenze tra le forme.*

**Key words:** Information Geometry, Geodesics, Fisher-Rao distance, Wasserstein distance

## 1 Introduction

Shapes clustering is of interest in various fields such as geometric morphometrics, computer vision and medical imaging. In the clustering of shapes is important to select an appropriate measurement of distance among observations. The aim of this talk is to model and clustering shapes configurations in two dimensions using Information Geometry tools. Information Geometry combines geometry and statistics

---

Stefano A. Gattone, Angela De Sanctis

University G. d'Annunzio of Chieti-Pescara, Italy, e-mail: gattone@unich.it, a.desanctis@unich.it

(Amari and Nagaoka, 2000; Murray and Rice, 1984) and it can be used in Shape Analysis (Dryden and Mardia, 1998) to describe mathematically patterns from complex systems and their changes in time.

We consider objects whose shapes are based on landmarks (Bookstein, 1991; Cootes et al, 1995; Kendall, 1984). The objects can be obtained by medical imaging procedures, curves in space obtained by manually or automatically assigned feature points or by a discrete sampling of the object contours.

Since the shape space is invariant under similarity transformations, that is translations, rotations and scaling, an Euclidean distance function on such a space is not really meaningful. In order to apply standard clustering algorithms to planar shapes, the Euclidean metric has to be replaced by the metric of the shape space. Examples are provided in Amaral et al. (2010) and Stoyan and Stoyan (1990) where the Procrustes distance was integrated in standard clustering algorithms such as the  $k$ -means. Similarly, Lele and Richtsmeier (2001) applied standard hierarchical or  $k$ -means clustering using dissimilarity measures based on the inter-landmark distances. In a model-based clustering framework Huang et al. (2016) and Kume and Welling (2010) developed a mixture model of offset-normal shape distributions.

Statistical manifolds are the objects of study in Information Geometry. They are families of probability density functions with their local coordinates defined by the model parameters. Rao proved that the Fisher information matrix induces a Riemannian metric on a statistical manifold. Geodesics with respect to this metric can be used to measure differences in shape. Applications of geodesics to shape clustering techniques are provided, in a landmark-free context, by Srivastava et al. (2005) and Mio et al. (2007).

With the aim of clustering shapes, we first describe each landmark using a bivariate Gaussian model, where the means are the landmark coordinates while the variances reflects the variability across a family of patterns. Next, we define distances between objects associated with different Riemannian metrics. These distances are induced by the geodesics of the metrics (geodesic distances). In general, computing the geodesic distance requires numerical solutions. We, rather, focus on cases where analytical expressions are available: the Fisher-Rao (Costa et al., 2015) and the Wasserstein metrics (Takatsu, 2011) for Gaussian distributions.

## 2 The method

Suppose we are given a planar shape configuration,  $C$ , consisting of a fixed number of labeled landmarks  $K$

$$C = \{\mu_1, \mu_2, \dots, \mu_K\}$$

with generic element  $\mu_k = \{\mu_{k1}, \mu_{k2}\}$  for  $k = 1, \dots, K$ . Following De Sanctis and Gattone (2016) and Peter and Rangarajan (2009), the  $k$ -th landmark may be represented by a bivariate Gaussian density as follows:



$$f(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-1} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \quad (1)$$

with  $\mathbf{x}$  being a generic 2-dimensional vector and  $\boldsymbol{\Sigma}_k$  given by

$$\boldsymbol{\Sigma}_k = \boldsymbol{\sigma}_k^2 \mathbf{I}_2 = \text{diag}(\boldsymbol{\sigma}_{k1}^2, \boldsymbol{\sigma}_{k2}^2) \quad (2)$$

where  $\{\boldsymbol{\sigma}_{k1}^2, \boldsymbol{\sigma}_{k2}^2\}$  is the vector of the variances of the  $k$ -th landmark coordinates, for  $k = 1, \dots, K$ . The variances capture uncertainties that arise in landmark placement and/or the natural variability across a population of shapes. Equation (1) represents the  $k$ -th landmark coordinates on a 4-dimensional manifold, say  $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)$ . The space of landmarks can be parameterized through the  $\boldsymbol{\theta}$ 's identifying them so that two shapes  $S$  and  $S'$  can be defined as follows:  $S = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$  and  $S' = (\boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_K)$ . For every  $k$ , let  $\gamma_k(t)$  with  $t \in [0, 1]$  be a path of the manifold such that  $\gamma_k(0) = \boldsymbol{\theta}_k$  and  $\gamma_k(1) = \boldsymbol{\theta}'_k$ . From differential geometry we know that a given Riemannian metric  $g$  induces an inner product  $\langle \cdot, \cdot \rangle_g$  on the tangent space of the manifold such that the length of  $\gamma_k(t)$  is defined as follows

$$l(\gamma_k) = \int \|\dot{\gamma}_k(t)\|_g^2 dt. \quad (3)$$

The distance between the  $k$ -th landmarks of the two shapes is given by the minimum length of the trajectory  $\gamma_k(t)$

$$d_g(\boldsymbol{\theta}_k, \boldsymbol{\theta}'_k)_g = \inf_{\gamma_k} \{ \sqrt{l(\gamma_k)} : \gamma_k(0) = \boldsymbol{\theta}_k, \gamma_k(1) = \boldsymbol{\theta}'_k \}. \quad (4)$$

Finally we use the matrix of pairwise distances between landmarks as distance of the two shapes  $S$  and  $S'$ .

In the space of Gaussians distributions, we will consider two different Riemannian metrics which in turn induce two types of geodesic distances. One metric is the Fisher-Rao metric  $g_f$ , defined by the Fisher information matrix  $g$ , with generic  $(i, j)$  entry given by

$$g_{ij} = \int f(\mathbf{x} | \boldsymbol{\theta}) \frac{\partial}{\partial \theta^i} \log f(\mathbf{x} | \boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \log f(\mathbf{x} | \boldsymbol{\theta}) d\mathbf{x}. \quad (5)$$

The other metric considered in this talk is the Riemannian metric  $g_w$ , which induces the Wasserstein distance (Takatsu, 2011). For Gaussian measures it is given by

$$d_{g_w}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \|\boldsymbol{\mu} - \boldsymbol{\mu}'\| + \text{tr}(\boldsymbol{\Sigma}) + \text{tr}(\boldsymbol{\Sigma}') - 2\text{tr}(\sqrt{\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\Sigma}' \boldsymbol{\Sigma}^{\frac{1}{2}}}). \quad (6)$$

The geodesic distances induced by these two metrics can be used to define a shape distance. The discriminative power of these shape distances will be evaluated, in the extended version of this manuscript, in the context of shapes clustering on both simulated and real data sets.

## Riferimenti bibliografici

1. Amaral, G.J., Dore, L.H., Lessa, R.P., Stolic, B.: k-Means Algorithm in Statistical Shape Analysis. *Commun Stat-Simul C* **39**, 1016–1026 (2010)
2. Amari, S., Nagaoka, H.: *Methods of Information Geometry*. AMS & Oxford University Press, Providence (2000)
3. Bookstein, F.L.: *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press (1991)
4. Costa, S.I.R., Santos, S.A., Strapasson, J.E. : Fisher information distance: A geometrical reading. *Discrete Appl Math* **197**, 59–69 (2015)
5. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. *Comput Vis Image Und* **61**, 38–59 (1995)
6. De Sanctis, A., Gattone, S.A. : Methods of Information Geometry to model complex shapes. *Eur Phys J ST* **225**, 1271–1279 (2016)
7. Dryden, I.L., Mardia, K.V.: *Statistical Shape Analysis*. John Wiley & Sons, London (1998)
8. Huang, C., Styner, M., Zhu, H.: Clustering High-Dimensional Landmark-based Two-dimensional Shape Data. *J Am Stat Assoc* **19**, 702–723 (2016)
9. Kendall, D.G. : Shape manifolds, procrustean metrics and complex projective space. *Bull London Math Soc* **16**, 81–121 (1984)
10. Kume, A., Welling, M.: Maximum likelihood estimation for the offset-normal shape distributions using EM. *J Comput Graph Stat* **19**, 702–723 (2010)
11. Lele, S., Richtsmeier, J.: *An invariant approach to statistical analysis of shapes*. Chapman & Hall/CRC, New York (2001)
12. Mio, W., Srivastava, A., Joshi, S.H. : On Shape of Plane Elastic Curves. *Int J Comput Vision* **73**, 307–324 (2007)
13. Murray, M. K., Rice, J. W.: *Differential Geometry and Statistics*. Chapman & Hall (1984)
14. Peter, A., Rangarajan, A. : Information Geometry for landmark shape analysis: unifying shape representation and deformation. *IEEE T Pattern Anal* **31**, 337–350 (2009)
15. Srivastava, A., Joshi, S.H., Mio, W., Liu, X.: Statistical shape analysis: Clustering, learning, and testing. *IEEE T Pattern Anal* **27**, 590–602 (2005)
16. Srivastava, A., Joshi, S.H., Mio, W., Liu, X. : Statistical Shape analysis: Clustering, learning, and testing. *IEEE T Pattern Anal* **27**, 590–602 (2005)
17. Stoyan, D., Stoyan, H. : A further application of d.g. kendalls procrustes analysis. *Biometrical J* **32**, 293–301 (1990)
18. Takatsu, A. : Wasserstein geometry of Gaussian measures. *Osaka J Math* **48**, 1005–1026 (2011)

# Space and circular time log Gaussian Cox processes with application to crime event data

## *Processi di Cox log-Gaussiani spaziali e tempo-circolari per lo studio della criminalità*

Alan E. Gelfand and Shinichiro Shirota

**Abstract** We view the locations and times of a collection of crime events as a space-time point pattern modeled as either a nonhomogeneous Poisson process or a more general log Gaussian Cox process. We need to specify a space-time intensity. Viewing time as circular, necessitates a valid separable and nonseparable covariance functions over a bounded spatial region crossed with circular time. Additionally, crimes are classified by crime type and each crime event is marked by day of the year which we convert to day of the week. We present marked point pattern models to accommodate such data. Our specifications take the form of hierarchical models which we fit within a Bayesian framework. We consider model comparison between the nonhomogeneous Poisson process and the log Gaussian Cox process as well as separable vs. nonseparable covariance specifications. Our motivating dataset is a collection of crime events for the city of San Francisco during the year 2012.

**Abstract** *In questo lavoro si studiano gli episodi di criminalità attraverso processi di punto di Poisson non omogenei e processi di Cox log-Gaussiani marcati. Questi modelli richiedono di specificare l'intensità spazio-temporale. Inoltre, l'interpretazione del tempo come variabile circolare richiede di specificare funzioni di covarianza separabili e non separabili valide sul dominio spaziale e temporale-circolare. Si presentano i modelli per processi di punto adatti a descrivere questi dati. Si propone una formulazione gerarchica del modello secondo l'impostazione Bayesiana. I dati analizzati sono relativi agli eventi di criminalità avvenuti a San Francisco nell'anno 2012.*

**Key words:** derived covariates; hierarchical model; marked point pattern; Markov chain Monte Carlo; separable and nonseparable covariance functions; wrapped circular variables

---

Alan E. Gelfand

Department of Statistical Science, Duke University, Durham, USA e-mail: alan@stat.duke.edu

Shinichiro Shirota

Department of Statistical Science, Duke University, Durham, USA e-mail: ss571@stat.duke.edu

## 1 Introduction

The times of crime events can be viewed as circular data. That is, working at the scale of a day, we can imagine event times as wrapped around a circle of circumference 24 hours (which, without loss of generality, can be rescaled to  $[0, 2\pi)$ ). Furthermore, over a specified number of days, we can view the set of event times, consisting of a random number of crimes, as a point pattern on the circle. Suppose, additionally, that we attach to each crime event its spatial location over a bounded domain. Then, for a bounded spatial region, we have a space-time point pattern over this domain, again with time being circular.

The contribution here is to develop suitable models for such data, motivated by a set of crime events for the city of San Francisco in 2012. The challenges we address involve (i) clustering in time - event times are not uniformly distributed over the 24 hour circle; (ii) spatial structure - evidently, some parts of the city have higher incidence of crime events than others; (iii) crime type - characterization of point pattern varies with type of crime so different models are needed for different crime types; (iv) incorporating covariate information - we anticipate that introducing suitable *constructed* spatial and temporal covariates will help to explain the observed point patterns; (v) the need for spatio-temporal random effects - the constructed spatial and temporal covariates will not adequately explain the space-time point patterns; (vi) the availability of marks - in addition to a location and a time within the day, each event has an associated day of the year which we convert to a day of the week. We propose a range of point pattern models to address these issues; fortunately, our motivating dataset is rich enough to investigate them.

We focus on the problem of building a log Gaussian Cox process (LGCP) which includes, as a special case, a nonhomogeneous Poisson process (NHPP), over space and circular time. We need to build a suitable intensity surface which is driven by a realization of a log Gaussian process incorporating a valid covariance function over space and time. Typically, time is modeled linearly, leading to a large literature on point patterns over bounded time intervals (see, e.g., [1] and [2]). Adding space, [3] offer development of a space-time LGCP. [4] consider a space-time process convolution model for modeling of space time crime events.

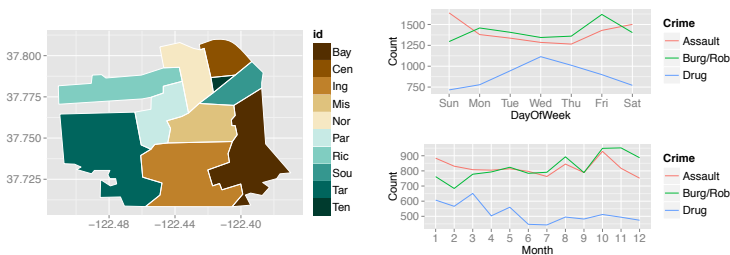
In fact, in this context, it is important to articulate the difference between viewing time in a *linear* manner vs. a *circular* manner. With linear time there is a past and a future. We can condition on the past and predict the future, we can incorporate seasonality and trend in time. With circular time, as with angular data in general, we only obtain a value once we supply an orientation, e.g., the customary midnight with time, although, below, we argue to start the day at 02:00. So, we have no temporal ordering of our crime events except within a defined 24 hour window. We are only interested in modeling the intensity over space and circular time. For event times during a day, wrapping time seems natural. Again, these times only arise given an orientation. However, crimes at 23:55 and 00:05 are as temporally close as crimes at 23:45 and 23:55.

Our data consists of a set of crime events in San Francisco (SF) during the year 2012. Each event has a time of day and a location. In fact, we also have a classifica-

tion into crime type and we also have assignment of each crime to a district, arising by suitable partitioning of the city. Lastly, we know the day of the year for the event, enabling consideration of day of the week effects.

## 2 The dataset

Our dataset consists of crime events in the city of San Francisco in 2012. We have three crime type categories: (1) assault, (2) burglary/robbery, and (3) drug. Each crime event has a time (date, day of week, time of day) and location (latitude and longitude) information. Spatial coordinates (latitude and longitude) were transformed into eastings and northings. Each crime event is also classified into a district. In particular, there are 10 districts in San Francisco: (1) Bayview, (2) Central, (3) Ingleside, (4) Mission, (5) Northern, (6) Park, (7) Richmond, (8) Southern, (9) Taraval, (10) Tenderloin (see Figure 2, left panel). Figure 2 (right panel) shows the counts of crime events for day of week<sup>1</sup>. Counts for crime types show different patterns. Assault events happen more on weekends, but burglary/robbery events happen most on Friday.



**Fig. 1** The map of San Francisco (left) and crime counts on each day of week (right)

Figure 2 shows the data by type and by day of the week ( $3 \times 7$  plots) in the form of ‘rose’ diagrams. This figure reveals differences among crime types and also differences across day of the week. For example, drug-related crime events are observed more from 5 to 7 pm, while burglary/robbery crime events are observed later in the day. Overall, the circular time dependence of crime events is seen, i.e., large counts from evening to late night and small counts from early morning through the middle of the day. In the point pattern model construction below, we model each crime type separately and, within crime type, incorporate day of week as a mark.

<sup>1</sup> Below, we take day of the week as 02:00 to 02:00. This definition interprets crime events on, e.g., Saturday night as including the early hours of Sunday morning.

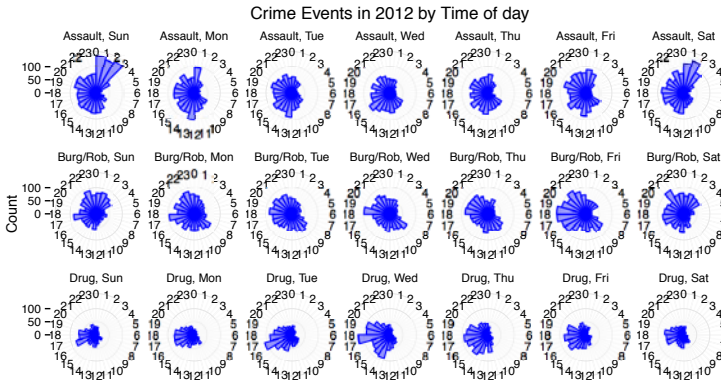


Fig. 2 Histograms of crime events by type and by day of the week

### 3 Modeling and Theory

Observations on a circle lead us to the world of directional data, as illustrated in Figure 2. Once an orientation has been chosen, the circular observations are specified using the angle from the orientation to the corresponding point on the unit circle. Here, we are only concerned with point patterns on a circle. For the nonhomogeneous Poisson process and log Gaussian Cox process models we only need to specify intensities over  $D \times S^1$  where  $S^1$  is the unit circle.

#### 3.1 The nonhomogeneous Poisson process (NHPP) and Log Gaussian Cox process (LGCP)

Again, since the crime events are random both in number and in space-time location, we think of them as a random point pattern over space and time. We consider the two most common models for such a setting: the NHPP and the LGCP. The LGCP dates at least to [5]. As a spatial process, it is defined so that the log of the intensity is a Gaussian process (GP), i.e.,

$$\log \lambda(\mathbf{s}) = X(\mathbf{s})^T \beta + Z(\mathbf{s}), \quad Z(\mathbf{s}) \sim \mathcal{G}\mathcal{P}(\mathbf{0}, C). \tag{1}$$

Here,  $Z(\mathbf{s})$  is a zero mean stationary, isotropic GP over  $D$  with covariance function  $C$ , which provides spatial random effects for the intensity surface, pushing up and pulling down the surface, as appropriate. If we remove  $Z(\mathbf{s})$  from the log intensity, we obtain the associated NHPP.

Again, we consider a three dimensional Gaussian process with a two dimensional location, and one dimensional circular time. In general, we seek  $Z(\mathbf{s}, t) \sim \mathcal{GP}(0, C)$ ,  $(\mathbf{s}, t) \in \mathbb{R}^2 \times S^1$ . We need to specify valid correlation functions over  $\mathbb{R}^2 \times S^1$ .

[6] proposes families of circular correlation functions (CCF's) based on truncation of familiar spatial correlation functions. He shows that the completely monotone functions are strictly positive definite on spheres of any dimension, e.g., powered exponential, Matérn, generalized Cauchy, and Dagum families. Another example in [6] which we adopt is the generalized Cauchy family,

$$C_{GC}(u) = \left(1 + (\phi u)^\alpha\right)^{-\tau/\alpha}, \quad \text{for } u \in [0, \pi] \quad \alpha \in (0, 1], \quad \tau > 0. \quad (2)$$

where  $\tau$  is a shape parameter which doesn't affect the positive definiteness as long as  $\tau > 0$ . This function is positive definite for any dimension if  $\alpha \in (0, 1]$ . It may be surprising that restriction of familiar spatial correlation functions to the spherical domain maintains positive definiteness on the sphere.

In the context of the LGCP model, we need to specify the covariance function for the latent Gaussian process  $Z(\mathbf{s}, t)$ . Separable space time covariance functions are often adopted due to convenient specification and computational simplification [7]. The separable specification arises as a product of a valid space and a valid time covariance function, i.e.,

$$C_{s,t}(\mathbf{h}, u) = C_s(\mathbf{h})C_t(u) \quad (3)$$

So, we can define a valid space-time covariance function merely by choosing as  $C_s$  any valid covariance function on  $R^2$  and multiplying it by any of the foregoing valid CCF's. The resulting covariance matrix for a set of  $(\mathbf{s}, t)$ 's with  $N$   $\mathbf{s}$ 's by  $M$   $t$ 's will have a Kronecker product form  $C_s \otimes C_t$  where  $C_s$  and  $C_t$  are  $N \times N$  and  $M \times M$  covariance matrices. Simplified inverse, determinant, and Cholesky decomposition result, making the separable specification computationally efficient and tractable in high dimensional cases.

It is evident that the separable covariance specification is restrictive for real data applications because it precludes space-time interaction of the sort we mentioned in the Introduction. Various versions of nonseparable covariance functions have been proposed for the case where space is again  $R^2$  and time is linear. We need a nonseparable version with circular time. In the full paper we develop such specifications and fit data using these specifications. Details are omitted here but, disappointingly, we found essentially no improvement in model performance for the nonseparable choice.

We employ constructed space and time covariates. For the spatial covariates, we identify a set of landmarks. These landmarks are referred to as *crime attractors* and are selected from centers of commercial activity, i.e., places with high population density, high human exposure. Examples might include malls, market streets, and amusement centers. For a given landmark, we employ a directional Gaussian kernel function as the distance measure from crime location to landmark. That is, inverse distance measures risk; the smaller the distance the larger the risk. To form a temporal covariate we need a function whose support is the unit circle. Since crime events occur more frequently in the evening and night hours, less in the morning and afternoon hours, the most elementary constructed covariate which reflects this would have two levels. Here, we let

$$\kappa(t) = \mu(1 + \delta 1(t \in [4\pi/3, 2\pi))). \quad (4)$$

On the 24 hour scale, this choice of  $\kappa$  would be interpreted as adopting level  $\mu$  for times between 02:00 and 18:00 and level  $\mu(1 + \delta)$  for times between 18:00 and 02:00 in the morning.  $\mu$  and  $\delta$  become model parameters; alternative windows could be explored.

Full model specification, full prior details and full elaboration of the model fitting are provided in the full paper. Also, details of the out-of-sample model adequacy and comparison leading to preference for the LGCP over the NHPP are provided.

## 4 Results

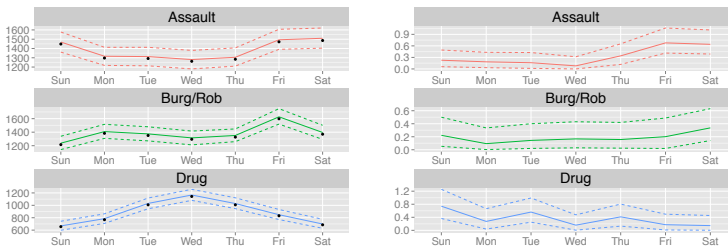
Table 1 shows the estimation results for the space by circular time LGCP for the three crime categories in 2012. With day of week specific  $\mu_w$  and  $\delta_w$ ,  $\mu_0$  and  $\delta_0$  are set to the means of them over the days of week, yielding  $\mu_w - \mu_0$  and  $\delta_w - \delta_0$  as deviations. See Figure 4 below for inference on the  $\delta_w$  across day of the week. The spatial covariates  $\mathbf{m}\beta$  are *positively* significant. In particular,  $\beta_1$  and  $\beta_2$  for drug crimes show larger values than those for the other crime types. This result suggests that drug events are more concentrated around landmarks  $L_1$  and  $L_2$ .

Figure 4 shows the posterior mean and 95% CI of  $\sum_{j=1}^J \lambda(s_j^*, t_j^*, w) \Delta_{s,t,w}$  against counts on each day of week. For a given  $w$ ,  $\sum_{j=1}^J \lambda(s_j^*, t_j^*, w) \Delta_{s,t,w}$  is approximately the expected number of crime events on day  $w$  a year. The left panel demonstrates that the posterior mean of  $\sum_{j=1}^J \lambda(s_j^*, t_j^*, w) \Delta_{s,t,w}$  traces the observed counts on days of week accurately. The right panel displays the posterior mean and 95% CI of  $\delta_w$ . Although the variance of  $\delta_w$  is large, this figure shows that  $\delta_w$  varies with day of week; for assault, weekend  $\delta$ 's are larger. Since all of the  $\delta_w$ 's are positive, regardless of day of week or type of crime, we find elevated risk in the evening hours.



**Table 1** Estimation results for space by circular time LGCP for the full region with separable covariance: assault (left), burglary/robbery (middle) and drug (right)

	Assault		Burglary/Robbery			Drug		
	Mean	95%CI	IF Mean	95%CI	IF	Mean	95%CI	
$\mu_0$	36.94	[20.62, 60.12]	70	36.18	[21.45, 57.79]	70	32.43	[19.04, 58.32]
$\delta_0$	0.342	[0.201, 0.613]	68	0.188	[0.069, 0.424]	69	0.344	[0.137, 0.672]
$\beta_1$	1.654	[1.146, 2.023]	69	2.646	[2.038, 3.542]	73	3.874	[2.146, 5.045]
$\beta_2$	1.202	[0.614, 1.778]	70	0.470	[0.048, 0.823]	65	3.745	[2.117, 4.399]
$\sigma^2$	5.598	[5.064, 6.471]	73	5.756	[5.331, 6.219]	72	8.424	[7.868, 8.984]
$\phi_s$	0.011	[0.010, 0.013]	70	0.005	[0.005, 0.007]	70	0.027	[0.025, 0.030]
$\phi_t$	0.137	[0.123, 0.151]	58	0.178	[0.163, 0.196]	59	0.161	[0.140, 0.182]
$\sigma^2 \phi_s$	0.066	[0.059, 0.072]	61	0.033	[0.028, 0.037]	65	0.231	[0.207, 0.259]
$\sigma^2 \phi_t$	0.769	[0.681, 0.864]	61	1.027	[0.887, 1.164]	65	1.362	[1.182, 1.540]



**Fig. 3** Posterior mean and 95% CI of counts (left: dotted points are observed counts) and  $\delta_w$  (right) on each day of week for the full region: dashed lines are 95% CI

## 5 Summary

We have looked at times and locations of crime events for the city of San Francisco. We have argued that these data should be treated as point patterns in space and time where time should be treated as circular. We introduced derived spatial covariates (using distance from landmarks) and temporal covariates (using day of the week). We have looked at NHPP and LGCP models for such data. For the latter, we have proposed valid space and circular time Gaussian processes, both separable and nonseparable, for use in the LGCP. We have shown that the LGCP outperforms the NHPP for the SF crime data. However, strong support for nonseparability is not seen.

## References

1. Daley, D.J. and Vere-Jones D.: An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods, 2nd ed. Springer-Verlag, New York (2003)

2. Daley, D.J. and Vere-Jones D.: *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*, 2nd ed. Springer-Verlag, New York (2008)
3. Brix, A. and Diggle P.J.: Spatiotemporal prediction for log-gaussian cox processes. *J. Roy. Stat. Soc. B. Met.* **63**, 823–841 (2001)
4. Rodriguez, A. and Diggle P.J.: Bayesian estimation and prediction for inhomogeneous spatiotemporal Log-Gaussian Cox Processes using low-rank models, with application to criminal surveillance. *J. Am. Stat. Assoc.* **107**, 93–101 (2012)
5. Møller, J and Syversveen A.R. and Waagepetersen R.P.: Log Gaussian Cox Processes. *Scand J. Stat.* **25**, 451–482 (1998)
6. Gneiting, T.: Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19**, 1327–1349 (2013)
7. Banerjee, S. and Carlin B.P. and Gelfand A.E.: Chapman & Hall/RC (2014)

# Blind source separation

## *Cieco separazione alla fonte*

Abdelghani Ghazdali

**Abstract** The paper introduces a new notion of Blind Source Separation (BSS) in instantaneous mixtures of both independent or dependent source component signals. This approach is based on the minimization of a criterion between copula densities. This latter takes advantage of the copula to model the structure of the dependence between signal components. Simulation results are presented showing the convergence and the efficiency of the proposed algorithms.

**Abstract** *La carta introduce una nuova nozione di cieco separazione alla fonte in miscele istantanee di segnali indipendenti o dipendenti. Questo approccio si basa sulla minimizzazione di un criterio tra densità di copula. Quest'ultimo approfitta della copula per modellare la struttura della dipendenza tra i componenti del segnale. Sono presentati i risultati di simulazione mostrando la convergenza e l'efficienza degli algoritmi proposti.*

**Key words:** Blind source separation; instantaneous mixtures; Copulas; Mutual information; divergence between copulas.

## 1 Introduction

The blind source separation problem is a fundamental issue in different fields, such as signal and image processing, medical data analysis, communications, medical imaging... etc. The BSS aims to recover unknown source signals, out of a set of observations which are unknown, and being linear mixture of the sources. It was introduced and formulated by Bernard Ans, Jeanny Herault and Christian Jutten [1] since the 80's, describing a biological problem. In order to separate the data set, different assumptions on the sources have to be made. The most common assumptions are statistical independence of the sources, and the condition is that at most

---

Univ Hassan 1, Laboratoire LIPOSI, ENSA Khouribga, e-mail: [a.ghazdali@gmail.com](mailto:a.ghazdali@gmail.com)

one of the components is gaussian, which leads to the field of Independent Component Analysis (ICA), see for instance [2]. Then many methods of BSS have been proposed, using second or higher order statistics [3, 4], maximizing likelihood [5], minimizing the mutual information [6, 7], minimizing the criteria of  $\phi$ -divergences [8], ... etc. A good overview of the problem can be found in [9]. Recently, it has been shown in [10, 11, 12, 13] that, based on copula without the assumption of the independence of the sources, we can still determine the sources (up to scale and permutation indeterminacies) of both independent and dependent sources components.

The rest of this paper is organized as follows: Section 2 indicates some definitions and introduces the problem of BSS. In section 3, we describe our approach. Section 4 illustrates some numerical results. Finally, we conclude the paper and give some further research directions.

## 2 Problem formulation

BSS can be modeled as follows. Denoting  $A$  the mixing operator, the relationship between the observations and sources is

$$x(t) := A[s(t)] + b(t), \quad t \in \mathbb{R}, \quad (1)$$

where  $x$  is a set of observations,  $s$  is a set of unknown sources, and  $b$  is an additive noise. In this paper, we consider the linear BSS model with instantaneous mixtures, the operator  $A$  corresponds then to a scalar matrix, and the additive noise is either considered as an additional set of sources, or it is reduced by applying some form of preprocessing [8]. We assume that the number of sources is equal to the number of observations. Then we introduce the model as the following

$$x(t) := A s(t), \quad \forall t \in \mathbb{R}, \quad (2)$$

where  $x \in \mathbb{R}^p$  represents the observed vector,  $s \in \mathbb{R}^p$  is the unknown vector of sources to be estimated, and  $A$  is the unknown mixing matrix. The goal of BSS, is therefore to estimate the unknown sources  $s(t)$  from the set of observed mixtures  $x(t)$ . The estimation is performed with no prior information about either the sources or the mixing process  $A \in \mathbb{R}^{p \times p}$  (i.e. we are not in the bayesian paradigm). Specific restrictions are made on the mixing model and the source signals in order to limit the generality. The separating system is defined by

$$y(t) := B x(t), \quad \forall t \in \mathbb{R}. \quad (3)$$

The vector  $y(t) \in \mathbb{R}^p$  is the output signal vector (estimated source vector) and  $B \in \mathbb{R}^{p \times p}$  is called the separating operator. In other words, the problem is to obtain an estimator  $\hat{B}$  closing to the ideal solution  $A^{-1}$  using only the observation  $x(t)$ , which leads to accurate estimation of the source  $s(t)$

$$\widehat{y}(t) := \widehat{B} x(t) \simeq \widehat{s}(t). \quad (4)$$

### 3 The approach

The discrete version of the original problem (2) writes

$$x(n) := As(n), \quad n = 1, \dots, N. \quad (5)$$

The source signals  $s(n)$ ,  $n = 1, \dots, N$ , will be considered as  $N$  copies of the random source vector  $S$ , and then  $x(n), y(n) := Bx(n)$ ,  $n = 1, \dots, N$  are, respectively,  $N$  copies of the random source vector  $X$  and  $Y := BX$ .

The aim is to reconstruct an estimated source signal  $y(t)$  from the denoised observed signal  $x(t)$ . It has been shown in [10, 11, 12, 13] that if we dispose of some prior information about the density copula of the random source vector  $s(t)$ , we can detect both the mixing matrix and the sources uniquely for both independent and dependent sources. Let  $Y := (Y_1, \dots, Y_p)^\top \in \mathbb{R}^p$ ,  $p \geq 1$ , a random vector, with cumulative distribution function (c.d.f.)

$$F_Y(\cdot) : y \in \mathbb{R}^p \mapsto F_Y(y) := F_Y(y_1, \dots, y_p) := \mathbb{P}(Y_1 \leq y_1, \dots, Y_p \leq y_p), \quad (6)$$

and continuous marginal functions

$$F_{Y_i}(\cdot) : y_i \in \mathbb{R} \mapsto F_{Y_i}(y_i) := \mathbb{P}(Y_i \leq y_i), \quad \forall i = 1, \dots, p. \quad (7)$$

The mutual information of  $Y$  is defined by

$$MI(Y) := \int_{\mathbb{R}^p} -\log \frac{\prod_{i=1}^p f_{Y_i}(y_i)}{f_Y(y)} f_Y(y) dy_1, \dots, dy_p. \quad (8)$$

It is called also the modified Kullbak-Leibler divergence ( $KL_m$ ), between the product of the marginal densities and the joint density of the vector. Note also that  $MI(Y) := KL_m \left( \prod_{i=1}^p f_{Y_i}, f_Y \right)$  is nonnegative and achieves its minimum value zero iff  $f_Y(\cdot) = \prod_{i=1}^p f_{Y_i}(\cdot)$  i.e., iff the components of the vector  $Y$  are statistically independent. To clarify more precisely the BSS step, we will study separately, the case where the source components are independent, and the case where the source components are dependent.

#### 3.1 A separation procedure for independent sources.

Recall that the relationship between the probability density function and copula density is given by

$$f_Y(y) = \prod_{i=1}^p f_{Y_i}(y_i) c_Y(F_{Y_1}(y_1), \dots, F_{Y_p}(y_p)). \quad (9)$$

Assume that the source components are independent. Using the relation (9), between and applying the change variable formula for multiple integrals, we can show that  $MI(Y)$  can be written via copula densities as

$$MI(Y) := \int_{[0,1]^p} -\log\left(\frac{1}{c_Y(u)}\right) c_Y(u) du =: KL_m(c_{\Pi}, c_Y), \quad (10)$$

where  $c_Y(u)$  is the density copula of  $Y$ , and  $c_{\Pi}(u) := 1_{[0,1]^p}(u)$  is the product copula density. Moreover,  $KL_m(c_{\Pi}, c_Y)$  is nonnegative and achieves its minimum value zero iff  $c_Y(u) = c_{\Pi}(u)$ ,  $\forall u \in [0, 1]^p$ , namely, iff the components of the vector  $Y$  are independent.

Our approach consists in minimizing with respect to  $B$ , the following separation criterion:

$$KL_m(c_{\Pi}, c_Y) := \mathbb{E} \left[ \log \left( \frac{c_Y(F_{Y_1}(Y_1), \dots, F_{Y_p}(Y_p))}{c_{\Pi}(F_{Y_1}(Y_1), \dots, F_{Y_p}(Y_p))} \right) \right], \quad (11)$$

where  $\mathbb{E}(\cdot)$  denotes the mathematical expectation. The function  $B \mapsto KL_m(c_{\Pi}, c_Y)$  is nonnegative and attains its minimum value zero at  $B = DPA^{-1}$ , where  $D$  and  $P$  are, respectively a diagonal and permutation matrix. In other words, the separation is achieved in  $B = \arg \min_B KL_m(c_{\Pi}, c_Y)$ .

### 3.2 A separation procedure for dependent sources.

In the case where the source components are dependent, we assume that we dispose of some prior information about the density copula of the random source vector  $s$ . Note that this is possible for many practical problems, it can be done, from realizations of  $s$ , by a model selection procedure in semiparametric copula density models  $\{c_{\theta}(\cdot); \theta \in \Theta \subset \mathbb{R}^d\}$ , typically indexed by a multivariate parameter  $\theta$ , see [23]. The parameter  $\theta$  can be estimated using maximum semiparametric likelihood, see [24]. We denote by  $\hat{\theta}$ , the obtained value of  $\theta$  and  $c_{\hat{\theta}}(\cdot)$  the copula density modeling the dependency structure of the source components. Obviously, since the source components are assumed to be dependent,  $c_{\hat{\theta}}(\cdot)$  is different from the density copula of independence  $c_{\Pi}(\cdot)$ . Hence, we naturally replace in (10),  $c_{\Pi}$  by  $c_{\hat{\theta}}$ , then we define the separating criterion

$$\begin{aligned}
KL_m(c_{\hat{\theta}}, c_Y) &:= \int_{[0,1]^p} -\log\left(\frac{c_{\hat{\theta}}(u)}{c_Y(u)}\right) c_Y(u) du \\
&:= \mathbb{E} \left[ \log\left(\frac{c_Y(F_{Y_1}(Y_1), \dots, F_{Y_p}(Y_p))}{c_{\hat{\theta}}(F_{Y_1}(Y_1), \dots, F_{Y_p}(Y_p))}\right) \right].
\end{aligned} \tag{12}$$

Moreover, we can show that the function  $B \mapsto KL_m(c_{\hat{\theta}}, c_Y)$ , is nonnegative and attains its minimum value zero at  $B = DPA^{-1}$ . The separation for dependent source components, is reached in  $B = \arg \min_B KL_m(c_{\hat{\theta}}, c_Y)$ .

## 4 Simulation results

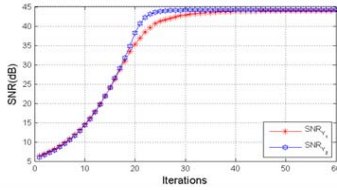
In this section, we present representative simulation results for the proposed method. We will limit ourselves to the case of 2 mixtures 2 sources. We start by illustrating the performance of BSS-copula with a simple experiment on independent sources. Then we move to use BSS-copula to separate dependent sources. The results will be compared with the classical mutual information (MI) criterion, see, [6], for the same data. The 2 sources are mixed with the matrix  $A := [1 \ 0.8; 0.8 \ 1]$ . The gradient descent parameter is taken  $\mu = 0.1$ . And the number of samples is  $N = 2000$ , and all simulations are repeated 20 times. The accuracy of source estimation is evaluated through the signal-noise-ratio (SNR), defined by

$$SNR_i := 10 \log_{10} \left( \frac{\sum_{k=1}^N \hat{y}_i(k)^2}{\sum_{k=1}^N (\hat{y}_i(k)^2 |_{s_i(k)=0})} \right), i = 1, 2. \tag{13}$$

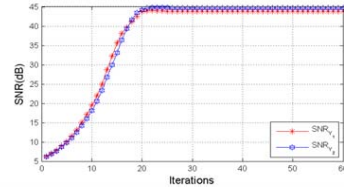
### 4.1 Independent source components:

In this experiment, we consider two mixed signals of two kinds of sample sources: uniform i.i.d with independent components FIGURE 1; i.i.d sources with independent components drawn from the 4-ASK (Amplitude Shift Keying) alphabet FIGURE 2. we observe from FIGURE 1 and FIGURE 2, that the proposed method gives good results for the standard case of independent component sources.

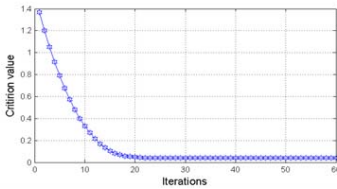
FIGURE 3 shows the criterion value vs iterations. We can see that our criterion converges to 0 when the separation is achieved.



**Fig. 1** Average output SNRs versus iteration number : Uniform independent sources.



**Fig. 2** Average output SNRs versus iteration number : ASK independent sources.



**Fig. 3** The criterion value vs iterations : uniform independent sources.

## 4.2 Dependent source components:

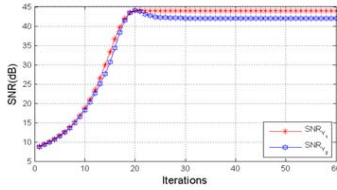
In this subsection we show the capability of the proposed method (Algorithm ?? for dependent sources) to successfully separate two dependent mixed signals, we dealt with instantaneous mixtures of tree kinds of sample sources:

- 1 i.i.d.(with uniform marginals) vector sources with dependent components generated from Ali-Mikhail-Haq (AMH) copula with  $\hat{\theta} = 0.8$ .
- 2 i.i.d.(binary phase-shift keying(BPSK)-marginals) vector sources with dependent components generated from Fairlie-Gumbel-Morgenstern (FGM) copula with  $\hat{\theta} = 0.85$ .
- 3 i.i.d.(with uniform marginals) vector sources with dependent components generated from Clayton copula with  $\hat{\theta} = 2.5$ .

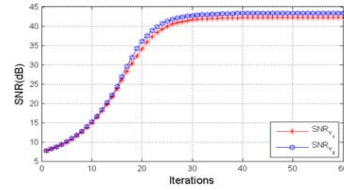
In figure ( 4)-( 6), we have shown the SNRs for each kind of sample sources. It can be seen from the simulations that the proposed method is able to separate, with good performance, the mixtures of dependent source components.

Moreover, figure ( 7) shows the criterion value versus iterations for AMH copula. We can see that our criterion converges to 0 when the separation is achieved.

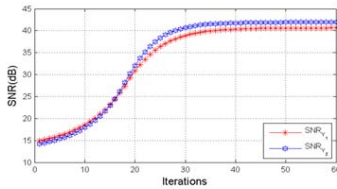




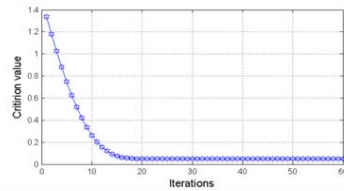
**Fig. 4** Average output SNRs versus iteration number : Uniform dependent sources from AMH-copula.



**Fig. 5** Average output SNRs versus iteration number : Bpsk dependent sources from FGM-copula.



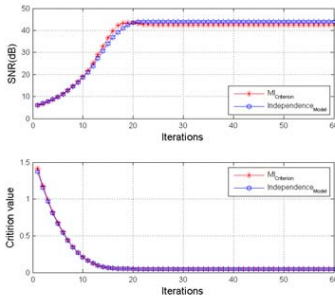
**Fig. 6** Average output SNRs versus iteration number : Uniform dependent sources from Clayton-copula.



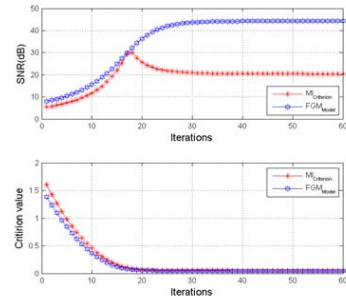
**Fig. 7** The criterion value vs iterations : Uniform dependent sources from AMH-copula.

### 4.3 Comparison

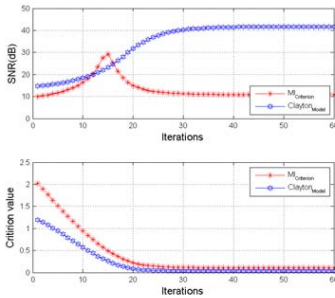
In this section, both independent and dependent signal sources are tested to confirm the performance of our proposed method, and compared with the MI method proposed by [6] for instantaneous linear mixture, under the same conditions. At the top of figure (8)-(10), we have shown the means of the SNRs of two sources for each kind of sample sources. It can be seen from the simulations of figure 8 (the standard case of independent component sources), that the method proposed achieves the separation with same similar accuracy as [8]. Likewise in the case of dependent component sources, one can see from the simulations of figures (9-10) that our method exhibits better performance than the MI one. At the bottom of figures (9)-(10), we show the criterion value vs iterations. As we can see, the both criteria of the two methods converges to zero when the separation is achieved. But the proposed method gives two well separate sources, unlike the MI one provides two independent sources very far from the sources. And that, is clearly seen at the top of figures (9)-(10), representing, the means of the SNRs of the two sources for each kind of sample sources.



**Fig. 8** Average output SNRs versus iteration number: uniform independent sources.



**Fig. 9** Average output SNRs versus iteration number: BPSK dependent sources from FGM-copula.



**Fig. 10** Average output SNRs versus iteration number: uniform dependent sources from Clayton-copula.

## 5 Conclusions

We have presented a new BSS algorithm. The approach is able to separate instantaneous linear mixtures of both independent and dependent source components. In Section 4, the accuracy and the consistency of the obtained algorithms are illustrated by simulation, for  $2 \times 2$  mixture-source. It should be mentioned that our proposed algorithms based on copula densities, rather than the classical ones based on probability densities, are more time consuming, since we estimate both copulas density of the vector and the marginal distribution function of each component. The present approach can be extended to deal with convolutive mixtures, that will be addressed in future communications.

## References

1. J. Hérault, C. Jutten, B. Ans, Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé, *GRETSI*, **2** (1985), 1017–1022.
2. P. Comon, Independent component analysis, A new concept?, *Signal Processing*, **36** (1994), no. 3, 287–314.
3. D.T. Pham, Blind separation of instantaneous mixture of sources based on order statistics, *IEEE Trans. on Signal Processing*, **48** (2000), no. 2, 363–375.
4. M. Castella, S. Rhioui, E. Moreau, J. Pesquet, Quadratic Higher Order Criteria for Iterative Blind Separation of a MIMO Convolutional Mixture of Sources, *IEEE Trans. on Signal Processing*, **55** (2007), no. 1, 218–232.
5. J.-F. Cardoso, Blind signal separation: statistical principles, *Proceedings of the IEEE*, **86** (1998), no. 10, 2009–2025.
6. D.T. Pham, Mutual information approach to blind separation of stationary sources, *IEEE Trans. on Information Theory*, **48** (2002), no. 7, 1935–1946.
7. M. El Rhabhi, G. Gelle, H. Fenniri, G. Delauna, A penalized mutual information criterion for blind separation of convolutional mixtures, *Signal Processing*, **84** (2004), no. 10, 1979–1984.
8. M. El Rhabhi, H. Fenniri, A. Keziou, E. Moreau, A robust algorithm for convolutional blind source separation in presence of noise, *Signal Processing*, **93** (2013), no. 4, 818–827.
9. P. Comon, C. Jutten, *Handbook of blind source separation : independent component analysis and applications*, Elsevier, 2010.
10. A. Keziou, H. Fenniri, A. Ghazdali, E. Moreau, New blind source separation method of independent/dependent sources, *Signal Processing*, **104** (2004), 319–324.
11. A. Ghazdali, M. El Rhabhi, H. Fenniri, A. Hakim, A. Keziou, Blind noisy mixture separation for independent/dependent sources through a regularized criterion on copulas, *Signal Processing*, **131**, February (2017), 502–513
12. A. Ghazdali, A. Hakim, A. Laghrib, N. Mamouni, S. Raghay, A new method for the extraction of fetal ECG from the dependent abdominal signals using blind source separation and adaptive noise cancellation techniques, *Theoretical Biology and Medical Modelling* **12**, December 2015
13. A. Ghazdali, A. Hakim, Blind source separation using measure on copulas, *Mathematics and Computer Science Series*, **42**, (2015), 104–116
14. I. Csiszár, Eine informations theoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten, *Magyar. Tud. Akad. Mat. Kutató Int. Közl.*, **8**, (1963), 85–108
15. I. Csiszár, Information-type measures of difference of probability distributions and indirect observations, *Studia Scientiarum Mathematicarum Hungarica*, **28**, (1966), 131–142
16. R. Beran, Minimum Hellinger distance estimates for parametric models, *Annals of Statistics* **5** (1977) 445–463.
17. R. Jimenez, Y. Shao, On robustness and efficiency of minimum divergence estimators, *Test* **10** (2001) 241–248.
18. M. Sklar, Fonctions de répartition à  $n$  dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris*, **8** (1959), 229–231.
19. R.B. Nelsen, *An introduction to copulas*, Springer, 2006.
20. H. Joe, *Multivariate models and dependence concepts*, Chapman & Hall, 1997.
21. B.W. Silverman, *Density estimation for statistics and data analysis*, Chapman & Hall, 1986.
22. M. Omelka, I. Gijbels, N. Veraverbeke, Improved kernel estimation of copulas: weak convergence and goodness-of-fit testing, *Ann. Statist.*, **37** (2009), no. 5B, 3023–3058.
23. X. Chen, Y. Fan, Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification, *Journal of Econometrics*, **135** (2006), no. 1–2, 125–154.
24. H. Tsukahara, Semiparametric estimation in copula models, *Canad. J. Statist.*, **33** (2005), no. 3, 357–375.
25. A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso et E. E. Moulines. A Blind Separation Technique Using Second-Order Statistics. *IEEE Trans. Signal processing* **45** (1997), 434–444.



# An innovative approach for Opinion Mining : the Plutchick analysis

## *Un approccio innovativo per l'Opinion Mining: la Plutchick analysis*

Massimiliano Giacalone, Antonio Ruoto, Davide Liga, Maria Pilato, Vito Santarcangelo

**Abstract** In this work we introduce an innovative approach for “Sentiment Analysis” or Opinion Mining, that is classically based on the concept that some words have positive or negative meanings. Infact, introducing the Plutchick score, it is possible to achieve an Emotional Analysis, that is a deeper analysis over the polarity. The original contribution of the paper is to present a program on Italian Emotional analysis of social networks hashtag mainly as part of “InfoSphere”. For this scope we introduce AIN\_EMOTION, an evolution of AIN Thesaurus, that is the first italian thesaurus for Emotional Analysis. This analysis gives a ratio of emotional hashtag on shared by social network users, can produce a behavioral trend and could be applied to any other language simply by changing the “emotional thesaurus”.

**Abstract** *In questo lavoro si introduce un approccio innovativo per la “Sentiment Analysis” o Opinion Mining, che si basa sul concetto classico che alcuni parole assumono significati positivi o negativi. In particolare, introducendo il punteggio Plutchick, e’ possibile realizzare un’analisi emotiva, cioe’ un’analisi piu’ approfondita sulla polarita’. Il contributo originale del lavoro e’ quello di presentare un programma su Emotional in italiano utilizzando un analisi delle reti sociali hashtag principalmente come parte di InfoSphere. Per questo scopo abbiamo introdotto AIN\_EMOTION, un’evoluzione del AIN Thesaurus, che e’ il primo thesaurus italiano per l’analisi emozionale. Questa analisi da’ un rapporto di hashtag emotiva condiviso da utenti del social network, in grado di produrre una tendenza comportamentale e potrebbe essere esteso ed applicato a qualsiasi altra lingua semplicemente cambiando il “thesaurus emozionale”.*

**Key words:** Sentiment analysis, Instagram, Social Network, Sentiment Thesaurus, Emotion Projection

---

Massimiliano Giacalone, University of Naples, e-mail: massimiliano.giacalone@unina.it  
Antonio Ruoto, iInformatica S.r.l.s., e-mail: marketing@iinformatica.it  
Davide Liga, iInformatica S.r.l.s., e-mail: web@iinformatica.it  
Maria Pilato, iInformatica S.r.l.s., e-mail: mariapilato@icloud.com  
Vito Santarcangelo, University of Catania, e-mail: santarcangelo@dm.unict.it

## 1 Introduction

Social networks push people to emphasize their own identity: people want to be part of certain kind of groups and, naturally, the wish of being part of some groups can be seen as the wish of not being part of many other groups. The aspect above mentioned means that social networks generate a world in which groups grow up becoming more and more separated if not openly opposed, a world in which identities are day by day clustered [8].

This sort of clustered Network Society presents some important concerns. Firstly, the more a society presents clustered identities the worse the risk of social conflicts becomes. Secondly, it seems that public opinion derived from social networks generates debates that are more emotional than reasonable. As a result, this kind of debates could be a major step towards a further consolidation of the Post-Truth politics. If the Web is the place where everybody can express his/her own opinion, and where groups develop becoming increasingly isolated or separated, social conflicts could arise[9].

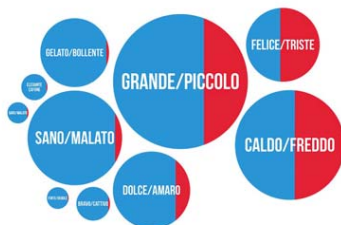
The reason of such a concern could be the lack of communication between different points of view: groups are closed and self-referential, so there is less space for a true and sensible debate. This is a huge issue because the Web is luckily to be that place where public opinion will be generated in the future. Moreover, the more groups become isolated and self-referential the more debates become emotional. In this regard, it should also be noticed that overload information can exercise a significant influence on individual choices and groups. It seems that this kind of information reduces the awareness of ideas, pushing also people to rely on groups and collective identities or even swapping from one group to another. If public opinion becomes emotional and volatile, political powers will try to adjust according to this lack of awareness. For example, that could be one of the reasons why western societies are facing the so-called populisms [10].

As can be seen, there is a good reason to believe that it is vital to develop an in-depth understanding of how human emotions and Network Society are related. As said, public opinion is becoming more emotional and volatile, in addition to this, societies are more clustered and potentially conflictual, and all these aspects seem to be politically relevant. In this scenario, social networks are luckily to become increasingly important in the generation of the common sentiment: they are the place where individual choices can be influenced and public opinion can be interpreted [11].

## 2 The social network in the infosphere semantic space

The following research aims to outline a survey methodology that can map the emotional level of the great Italian conversation on the network as part of the broader context of the infosphere. Specifically, it will be given an application of this methodology grafted on the social network “Instagram” with the aim of understanding with

regard to the Italian language users: 1) In what direction it propagates the generated emotional strength; 2) What are the main conveyed emotional projections; 3) What are the 10 most significant emotional biases conveyed. The term in the information



**Fig. 1** The 10 main polarizations

infosphere philosophy means the totality of the information space. The infosphere is “the semantic space constituted by the totality of the documents, the agents and their operations”, where “documents” means any data, information and knowledge, codified and implemented in any semiotic format, “agents” are any system able to interact with an independent document (such as a person, organization or web software robots) and the term “operations” includes any kind of action, interaction and transformation that can be done by an agent and which can be presented in a document. It is an environment in which the organisms are formed as interconnected cells [1].

It is immediately evident, within the perspective of contemporary Network Society, as the current dynamics of the Great Conversation network is configured as an essentially structural part of the infosphere. The spread of the Internet, mobile communications and digital media, along with a wide range of social software tools are driving the development of interactive and horizontal communication networks that connect, at any time, local and global. The communication system of industrial society revolved around the mass media, characterized by the mass distribution of a one-way message one-to-many, one to many. The communication foundation of the network society is the global system of horizontal communication networks, which include the multimodal exchange of many-to-many interactive messages, or many to many, synchronous and asynchronous. If we consider the public sphere as the space in which form public opinion, the analysis of the dynamics at play within the great conversation on the network, affected by the events and the agenda setting issues, can be useful to read the major changes taking place, producing predictive projections on generation of common sense and construction mechanisms of collective and individual imaginary [2].

However, it must take note that public opinion structured along the great network conversation is giving an opinion I can essentially emotional. The sharing of infor-

mation overload and Emotional effects are transforming the public debate in a more emotional than rational debate [3, 4].

### 3 Methodology

In the Network Society then the only effective message is an emotional message, which starts from something emotionally powerful. The amplification of the emotional sphere becomes the basis on which the circulation of information in the sphere of mass self communication. A movement that runs along the paths laid out by the emotional contagion mechanism. It is the confirmation of the claims of almost three centuries by philosopher David Hume is the reason to be the slave of the emotions, and not viceversa [5].

If then you become aware that in this context at the base of the social movements decisions there are processes that involve the transformation of the social emotions, emotional intelligence operation allows you to frame and define the extent of the emotional dynamics in place able to compete in processes of creation of common sense and influence the construction of collective and individual imaginary [7].

The mapping of the emotions conveyed in the large network conversation is carried out in this work through the methodology OSINT, proceeding to an activity of gathering information by consulting publicly available sources. Specifically, they are classified on the basis of an emotional score major adjectives Italian language as used by users of Instagram and hashtag products from 10/10/2010 (the year of the launch platform in Italy) to 10/01/2016 with a recurrence greater than or equal to 100,000. Emotional score attributed to each hashtag is based sull'AIN thesaurus, among the most comprehensive thesaurus for analysis sentiments of the Italian language, which attaches to the main adjectives of the Italian language a positive or negative emotional polarity following a scoring system based on the following ladder: + 2 (very positive), + 1.5, + 1, 0.5;

0 (neutral);

- 0.5, - 1, - 1.5, - 2 (very negative).

In addition to clarifying the emotional polarity of each hashtag analyzed, these have been classified through the emotional scale proposed by the American psychologist Robert Plutchik [6].

AIN\_EMOTION is the first thesaurus of "emotional" type for the Italian language. The AIN Thesaurus which is the thesaurus for opinion mining (positive, negative, neutral) the Italian language has been enriched the speech "EMOTION" categorizing each adjective with the emotional scale Plutchik. The following is an example, where for each adjective was associated one or more clusters of Plutchik. For each word (abandoned, lowered, down, propertied, combined, plentiful, tanned, abominable, endearing) there is a relative cluster associated (pain, sadness, acceptance, sadness, serenity, interest, admiration, disgust).

Thanks to information of the emotional scale it is possible to make an opinion upper



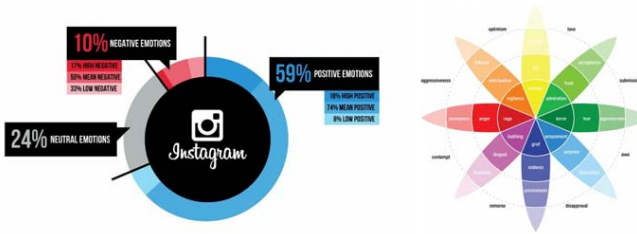


Fig. 2 Sentiment Analysis on Instagram and Plutchick Wheel of Emotions

level mining, not limited to one polarity but adding semantic information of particular importance, in order to carry out an investigation more focused on the emotional state of the author. In fact, this approach introduces an “emotional” analysis, thanks to the use of the granularity of the Plutchick score. The traditional sentiment analysis considers only 3 classes (negative, neutral, positive) with a possible weight for a better analysis. In the Plutchick approach we have 32 possible classes, 8 with 3 different degrees (e.g. admiration, trust, acceptance) and 8 intermediate (love, submission, awe, disapproval, remorse, contempt, aggressiveness, optimism) as shown in Fig. 2.

#### 4 Discussion of results and conclusions

From the surveys conducted it is summarized possible to state that, with regard to users of Italian language, the social network Instagram: 1) conveys average positive emotional biases: among analyzed hashtag, those used by users most often have an emotional rather than a positive polarity; very few, in parallel, are in effect the hashtag with an average negative emotional polarity; 2) It constitutes an environment in which the most represented social emotional classes are those relating to dell’ “expectation sphere” and the “pain”. Almost nonexistent turn out to be the emotions that belong to the emotional classes of “anger” and “disgust”.

Similar results should be read in light of an observation: the form of the content and architecture of social environments are more important than the content: from communication processes switching to narratological environments. And the stories conveyed by Instagram users, because of the nature of the medium, are on average functional to one goal: the digital packaging of the self. In this sense the self-production of content carried converges towards selfmarketing with the aim, more or less conscious and said, to show up on the market relations in a positive light as much as possible. A trend that seems to respond to a need for performance in a society where it seems that you have to be seen to exist. No wonder then that the emotion conveyed through the hashtag analyzed are crushed on emotional biases of almost exclusively positive type. It is a very precise seduction strategy of nego-

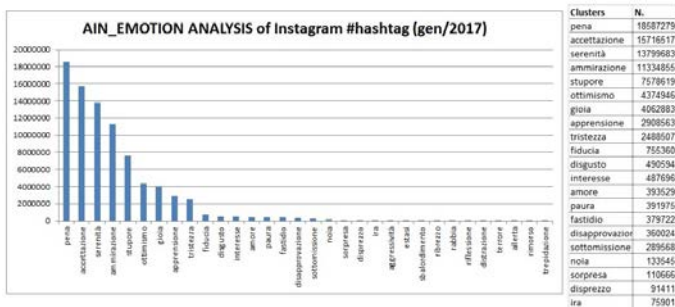


Fig. 3 Emotions extracted by M. Di Lecce tool

tiating type: convey positive emotions to receive positive in return. In this sense then on Instagram it seems to be banned any type of really critical about the real comparison. Evidence that is even more overwhelming if you take into account a relational dynamic imprinted on a omofiliaco logic. In this sense the emotional dynamics conveyed through Instagram eventually attributed to likeability, the principle of “pleasure”, the only true God to be served in the development of a strategy of attention conveyed through its social narratives, much of the power creation of common sense.

## References

1. Floridi L., (2012), La rivoluzione dell’informazione, Codice Edizioni, Torino.
2. Castells M., (2009), Comunicazione e potere, Università Bocconi Editore, Milano.
3. Lovink G.,(2012), Ossessioni collettive. Critica dei social media, Università Bocconi Editore, Milano.
4. Carr N.,(2011), Internet ci rende stupidi?, Raffaello Cortina Editore, Milano.
5. Westen D., (2008), La mente politica, il Saggiatore, Milano.
6. Plutchik R.,(2002), Emotions and Life: Perspectives from Psychology, Biology, and Evolution, Washington, DC: American Psychological Association.
7. Santarcangelo, V., Oddo, G., Pilato, M., Valenti, F., Fornaro, C., (2015). Social Opinion Mining: an approach for Italian language, SNAMS2015 at FiCloud2015.
8. Santarcangelo, V., Pilato, M. et al. (2015). An opinion mining application on OSINT for the reputation analysis of public administrations, Choice and preference analysis for quality improvement and seminar on experimentation, Bari.
9. Iorio, E., Ruoto, A., (2015). Nessun Tempo. Nessun Luogo. La comunicazione pubblica italiana all’epoca delle Reti.
10. Pilato, M., Santarcangelo, G., Santarcangelo, V., Oddo, G., (2015). AIN Thesaurus, RCE MULTIMEDIA.
11. Giacalone M. (2009) Manuale di Statistica Giudiziaria “Bel Ami edizioni” Roma.
12. Ruoto A., Santarcangelo V., Liga D., Oddo G., Giacalone M., Iorio E. (2017). The sentiment of the infosphere: a sentiment analysis approach for the big conversation on the net. In: (a cura di): Lauro C. Amatore E. Grassia M.G. Aragona B. Marino M. (Eds.), STUDIES IN CLASSIFICATION, DATA ANALYSIS, AND KNOWLEDGE ORGANIZATION, Springer.

# A G.E.D. method for market risk evaluation using a modified Gaussian Copula

## *Un metodo G.E.D. per la valutazione del rischio di mercato usando una Copula Gaussiana modificata*

Massimiliano Giacalone and Demetrio Panarello

**Abstract** In this paper, we show some results regarding the evaluation of Value-at-Risk (VaR) of some portfolios using a Gaussian Copula, modified by introducing the Generalized Correlation Coefficient, and assuming a Generalized Error Distribution (G.E.D.) for the single returns in the portfolios. In the literature, various authors considered the Copula function approach to evaluate market risk. In our proposal we consider a  $Lp_{\min}$  algorithm to estimate  $p$ , the shape parameter of the distribution. Finally, we compare the classical RiskMetrics method with our G.E.D. method based on a modified Gaussian Copula.

**Abstract** *In questo lavoro vengono mostrati alcuni risultati riguardanti la valutazione del Valore a Rischio (VaR) di alcuni portafogli utilizzando una Copula Gaussiana, modificata introducendo il Coefficiente di Correlazione Generalizzato, ed assumendo che i singoli rendimenti dei portafogli siano distribuiti secondo una Generalized Error Distribution (G.E.D.). Nella letteratura, vari autori hanno affrontato il tema della valutazione del rischio di mercato considerando l'approccio della funzione Copula. Nella nostra proposta consideriamo un algoritmo  $Lp_{\min}$  per stimare  $p$ , il parametro di forma della distribuzione. Infine, confrontiamo il classico metodo RiskMetrics con il nostro metodo G.E.D. basato su una Copula Gaussiana modificata.*

**Key words:** Value-at-Risk, Gaussian Copula, RiskMetrics Method, Generalized Error Distribution, Generalized Correlation Coefficient.

---

Massimiliano Giacalone  
Department of Economics and Statistics, University of Naples 'Federico II', e-mail: massimiliano.giacalone@unina.it

Demetrio Panarello  
Department of Economic and Legal Studies, Parthenope University of Naples, e-mail: demetrio.panarello@uniparthenope.it

## 1 Introduction

One of the most important issues in finance is to correctly measure the riskiness of a portfolio, which is fundamental to preserve its value over time. Since asset returns are usually fat-tailed, the use of Gaussian processes leads to an underestimation of the risk (Rachev et al., 2005).

Value-at-Risk (VaR) is used to quantify the risk of loss of an asset or a portfolio. The most straightforward method to calculate the (1-c)% Value-at-Risk is the Risk-Metrics one (Longestay et al., 1996), where it is hypothesized that the returns  $R_i$ , with  $i=1,2,\dots,N$ , of the  $N$  assets of a portfolio are jointly distributed according to a Gaussian multivariate (Kasch & Caporin, 2013). However, this hypothesis is simplifying, since it only considers the first two moments, neglecting the fact that the variations in asset returns usually have a leptokurtic and asymmetric behavior (Caporin, 2003).

For a better calculation of the risk, one of the proposals (e.g. Malevergne & Sornette, 2003) is to model the returns' interdependence of the assets in a portfolio by means of Copula functions. Here, the problem is to identify the marginal distributions that best model the returns of the single assets, and to define the Copula which is more suitable to represent the returns' interdependence structure.

## 2 The Generalized Error Distribution and the Gaussian Copula

The Generalized Error Distribution (G.E.D.) family was introduced by Subbotin (1923) and has been employed by various authors with different names and parameterizations. In our paper, we will use the Vianelli (1963) parameterization, which is:

$$f(x; \mu, \sigma_p, p) = \frac{1}{2\sigma_p p^{p-1} \Gamma(1/p)} \exp\left(-\frac{1}{p} \left| \frac{x-\mu}{\sigma_p} \right|^p\right) \text{ for } -\infty < x < \infty \quad (1)$$

where  $\mu$  is the location parameter,  $\sigma_p = [E|x - \mu|^p]^{1/p}$  is the scale parameter and  $p \geq 1$  is the shape parameter.

A Copula is a simple function that associates univariate marginal distributions to their joint ones (Jaworski, 2010). There are various Copula functions in the literature (McNeil et al., 2015) and others can be introduced, in order to capture the different dependence structures among stochastic variables.

In the bivariate case, the Gaussian Copula is:

$$C(u, v | \rho) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{(r^2-2\rho rs+s^2)}{2(1-\rho^2)}\right\} dr ds$$

where  $\Phi^{-1}$  is the inverse of Gaussian distribution function and  $\rho$  is the Pearson's correlation coefficient.

The Gaussian Copula considered here is indicated with  $C(\rho_p)$ , since the  $\rho$  parame-

ter is replaced by the Generalized Correlation Coefficient  $\rho_p$ , introduced by Taguchi (1974) as the correlation parameter of a bivariate Generalized Error Distribution, and defined as (Agrò & Martorana, 2002):

$$\rho_p = \frac{\text{codisp}^{(p)}(X,Y)}{\sigma_p(X)\sigma_p(Y)}, \text{ with } -1 \leq \rho_p \leq 1$$

where

$$|\text{codisp}^{(p)}(X,Y)|^p = |E[(Y - \mu_Y)|X - \mu_X|^{p-1} \text{sign}(X - \mu_X)]| \cdot |E[(X - \mu_X)|Y - \mu_Y|^{p-1} \text{sign}(Y - \mu_Y)]|,$$

$$\sigma_p(X) = [E|X - \mu_X|^p]^{1/p},$$

$$\sigma_p(Y) = [E|Y - \mu_Y|^p]^{1/p},$$

$\mu_X$  and  $\mu_Y$  power means of order p.

### 3 The algorithm

The joint density required for calculating portfolio's Value-at-Risk is obtained as (Agrò, 2008):

$$1 - c = \int \int_{s+t \leq -\text{VaR}} f(s,t) ds dt$$

The parameters  $\mu$ ,  $p$  and  $\sigma$  are estimated using the  $Lp_{min}$  method (Giacalone, 1996; Giacalone & Richiusa, 2006).

The  $\rho_p$  parameter is estimated using the Exponentially Weighted Moving Average recursive formula:

$$\rho_p = \frac{\text{codisp}_{t+1}^{(p)}(x,y)}{D_{t+1}^{(p)}(x)D_{t+1}^{(p)}(y)} \quad (2)$$

where

$$\text{codisp}_{t+1}^{(p)}(x,y) = (|(\mu_{y/x}^{(p)})_{t+1}| |x| |(\mu_{x/y}^{(p)})_{t+1}|)^{1/p} \times \text{sign}[(\mu_{y/x}^{(p)})_{t+1} + (\mu_{x/y}^{(p)})_{t+1}]$$

$$(\mu_{y/x}^{(p)})_{t+1} = (1 - \lambda) \sum_{i=1}^n \lambda^{n-i} y_i |x_i|^{p-1} \text{sign}(x_i)$$

$$(\mu_{x/y}^{(p)})_{t+1} = (1 - \lambda) \sum_{i=1}^n \lambda^{n-i} x_i |y_i|^{p-1} \text{sign}(y_i)$$

$$[D_{t+1}^{(p)}(x)]^p = (1 - \lambda) \sum_{i=1}^n \lambda^{i-1} |x|_{t-1}^p$$

$$[D_{t+1}^{(p)}(y)]^p = (1 - \lambda) \sum_{i=1}^n \lambda^{i-1} |y|_{t-1}^p$$

The fundamental steps in the algorithm are:

- estimation of the parameters  $\mu_i, \rho_i, \sigma_i$  for the two series of returns;
- estimation of the  $\rho_p$  parameter, with  $p = \sum_{i=1}^2 \rho_i / 2$ ;
- generation of  $(x, y)$  pairs, which is the realization of the double stochastic variable  $(X, Y)$  having G.E.D. marginals and relation of dependence expressed by a  $C(\rho_p)$ ;
- calculation of the distribution function of the returns in the portfolio;
- identification of Value-at-Risk of the return distribution.

## 4 Application and results

In order to evaluate and compare the performances of the two considered VaR methods, two portfolios were constructed, as is described below:

1. a Bond-ETF Portfolio, made up of a BTP-1FB37 4% Italian Bond, a BTP-1MZ21 3.75% Italian Bond and a LYXOR Exchange-Traded Fund. The data used are the daily prices for the years 2012-2016, for a total of 1267 data (data source: Teleborsa.it);
2. an Exchange indices Portfolio, made up of three indices on stock exchanges: the Euro-US Dollar (EUR-USD), the Pound Sterling-US Dollar (GBP-USD) and the Swiss Franc-Yen (CHF-JPY). The data used in this case refer to the daily quotations 2012-2016 for a total of 1305 data (data source: Investing.com).

Each time series of daily prices  $p_t$  was transformed into a series of logarithmic returns according to the relationship:

$$R_t = \log(p_{t+h}) - \log(p_t)$$

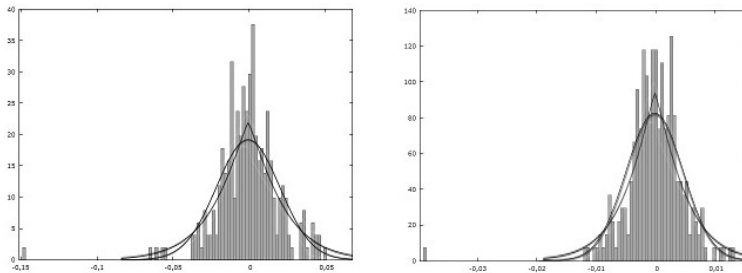
where  $h$ , the temporal interest interval, is set as one day.

The estimates of the  $p$  shape parameter were  $\hat{p} < 2$ ; hence the distributions of the returns are leptokurtic and more fat-tailed than the Gaussian one. Figure 1 shows the returns of the two portfolios, with the adaptation of a Gaussian distribution and a G.E.D. one.

The reliability of the methods for calculating Value-at-Risk is evaluated by means of a backtest, the (1-c)% VaR prediction data being compared with the values of profits and losses effectively recorded in the market.

The RiskMetrics method (VaR-R.M.) and the G.E.D. method, here called VaR-G.E.D., were applied to the two portfolios.

The backtest applied to the Bond-ETF and Exchange indices portfolios to predict 2.5% VaR and 0.5% VaR highlighted the better predictive capacity of the method which considers the leptokurtosis of marginal distributions. The G.E.D. VaR gives



**Fig. 1** Bond-ETF (left) and Exchange indices (right) Portfolios,  $\hat{p} = 1.4$

predictions which are closer to the real losses, also in relation to the extreme losses that are present in the time series of returns. Moreover, the extreme event, i.e. the exceptional loss or profit, is not somatized in a short time but influences subsequent predictions, which hence are of a cautionary type (overestimation of the risk). The number of VaR violations can be seen as a binomial stochastic variable in which the probability of success  $p$  is the percentage of VaR violations predicted (for example 5%) and the number of trials  $m$  is the number of days used for the backtest.

	Bond-ETF Port.		Exchange indices Port.	
	2.5%	0.5%	2.5%	0.5%
VaR-R.M.	6	4	5	3
Var-G.E.D.	5	2	3	2
Confidence intervals	1-9	0-3	1-9	0-3

**Table 1** VaR violations and 95% confidence intervals

Tab. 1 gives the number of VaR violations recorded for the two methods and the relative 95% confidence intervals in a number of observations  $m = 200$ . It shows that, for both portfolios, the VaR violations with the VaR-G.E.D. method are lower than the ones with the VaR-R.M. method and are within the confidence range.

## 5 Conclusions

In our application, we made a comparison between the Gaussian and the G.E.D. Copula. The reason is that the variation of the  $p$  shape parameter allows the G.E.D. to represent all the symmetric distributions that are described in the literature. That is, after estimating  $p$ , we are able to use the Copula which best fits our data: all the

Copula functions can be obtained as particular cases of the G.E.D. Copula. Among the different methods proposed in the literature for calculating Value-at-Risk, we took into account the well-known RiskMetrics method. We proposed a G.E.D. method and evaluated its performance compared to the RiskMetrics one. The two methods were evaluated by backtest, in order to examine the ability of predicting the potential loss of a portfolio.

The results obtained confirm the higher performance of the G.E.D. method, while the assumption of normality of the returns' distribution determines confidence intervals with the lowest predictive power. The assumption of normality, subject to verification, was rejected as the returns of all stocks examined have kurtosis characteristics which are neglected by the RiskMetrics method. It does seem that the VaR-G.E.D. method can constitute a valid generalization of the VaR-R.M., which it is close to in the case of Gaussian marginal distributions, while it moves away from it if the distributions are more fat-tailed.

All the necessary calculations have been implemented and processed on the statistical environment R.

## References

1. Agro', G.: On VaR using modified Gaussian Copula. *Annali della Facolta di Economia dell'Universita di Palermo* (2008)
2. Agro', G., Martorana, G.: VaR e Copula Normale di ordine p. *Atti del XXVI Convegno AMASES, Verona, 19-22 (2002)*
3. Caporin, M.: The trade off between complexity and efficiency of VaR measures: a comparison of RiskMetric and GARCH-type models (Vol. 3). GRETA, working papers (2003)
4. Giacalone, M.: Parameter evaluation of an exponential power function by simulation study. In: *Shorts Communications and Posters. Compstat 96, Barcelona (1996)*
5. Giacalone, M., Richiusa, R.: Lp-norm estimation: some simulation studies in presence of multicollinearity. *Student, 5, 235-246 (2006)*
6. Jaworski, P., Durante, F., Hardle, W. K., Rychlik, T.: *Copula theory and its applications*. Springer (2010)
7. Kasch, M., Caporin, M.: Volatility Threshold Dynamic Conditional Correlations: An International Analysis. *J.Financial Econom., 11, 706-742 (2013)*
8. Longestaey, J., Zangari, P.: *RiskMetrics Technical Document*. J.P. Morgan, Fourth Edition, New York (1996)
9. Malevergne, Y., Sornette, D.: Testing the Gaussian Copula hypothesis for financial assets dependences. *Quantitative Finance, 3, 231-250 (2003)*
10. McNeil, A. J., Frey, R., Embrechts, P.: *Quantitative risk management: Concepts, techniques and tools*. Princeton University Press (2015)
11. Rachev, S. T., Menn, C., Fabozzi, F. J.: *Fat-tailed and skewed asset return distributions: implications for risk management, portfolio selection, and option pricing*. John Wiley & Sons (2005)
12. Subbotin, M.: On the Law of Frequency of Error. *Mathematischeskii Sbornik (1923)*
13. Taguchi, T.: On Fechner's thesis and statistics with Norm-p. *Annals of the Institute of Statistical Mathematics, 26(1), 175-193 (1974)*
14. Vianelli, S.: La misura della variabilita condizionata in uno schema generale delle curve normali di frequenza. *Statistica, 33, 447-474 (1963)*



# Labour market dynamics and recent economic changes: the case of Italy

## *Dinamiche nel mercato del lavoro e recenti cambiamenti economici; il caso italiano*

Chiara Gigliarano and Francesco Maria Chelli

**Abstract** Aim of the paper is to analyse the differences in survival times of job contracts among subgroups of workers, based on different socio-demographic characteristics such as age, gender, educational level, geographical area. We in particular also the well-known Gini index to the measurement of concentration in survival times within groups of workers, and as a way to compare the distribution of survival times across such groups. We consider a test for differences in the heterogeneity of survival distributions, which may suggest the presence of a differential covariates effect on the job contract survival. The analysis is based on the Italian Compulsory Communications system data, which record all the activations, transformations, fixed-term extensions and anticipated terminations of employment relationships between any worker and employer in Italy.

**Abstract** *Il lavoro analizza le differenze nella durata di contratti di lavoro tra gruppi di lavoratori, basati su caratteristiche socio-demografiche quali età, sesso, livello di istruzione ed area geografica. In particolare, mediante l'indice di Gini si misura la concentrazione nei tempi di sopravvivenza per gruppi di lavoratori, al fine di confrontarne le distribuzioni. L'analisi si basa sui dati delle Comunicazioni Obbligatorie, che registrano tutte le attivazioni, trasformazioni, estensioni e cessazioni di contratti di lavoro dipendente in Italia.*

**Key words:** Labour market, Survival analysis, Gini index, Compulsory Communications system data

---

<sup>1</sup> Chiara Gigliarano, Dipartimento di Economia, Università degli Studi dell'Insubria, Varese, e-mail: chiara.gigliarano@uninsubria.it

Francesco Maria Chelli, Dipartimento di Scienze Economiche e Sociali, Università Politecnica delle Marche, Ancona, e-mail: f.chelli@univpm.it

## 1 Introduction

The Gini index is one of the most important statistical indices employed in social sciences for measuring concentration in the distribution of a positive random variable; it is mainly used in economics as a measure of income or wealth inequality among individuals or households (see, e.g., Gini 1912, 1914). Recently, the Gini coefficient has been used to describe concentration in levels of mortality, or in length of life, among different socio-economic groups, and to evaluate inequality in health and in life expectancy (see, e.g., Hanada 1983; Bonetti et al. 2009).

Aim of this paper is to analyse the differences in survival times of job contracts among subgroups of workers, from the point of view of concentration. We examine the differences both in the length of the first job contract and in the waiting time between the end of the first contract and the beginning of a new one. We apply the well-known Gini index to measure concentration in survival times within groups of workers, and as a way to compare the distribution of survival times across such groups. We consider a test for differences in the heterogeneity of survival distributions, which may suggest the presence of a differential covariates effect on the job contract survival. The analysis is based on the Italian Compulsory Communications system data, which record all the activations, transformations, fixed-term extensions and anticipated terminations of employment relationships between any worker and employer in Italy since January 2009 until June 2012. The target population is made up by the young workers, between 18 to 35 years old.

The rest of the paper is structured as follows: in Section 2 we briefly review the Gini test for survival data; in Section 3 we analyse the Italian labour market from the point of view of concentration; in Section 4 we conclude.

## 2 The Gini index for survival data: a brief review

The Gini index measures concentration in the distribution of a positive random variable. Bonetti et al. (2009) propose to apply the Gini index in survival analysis in order to measure concentration in survival times within groups of subjects. In particular, they apply a restricted version of the Gini index to right-censored survival data in order to detect differences in concentration (heterogeneity) between the survival time distributions of two groups.

A number of nonparametric statistical tests exist in the literature to test the difference in survival distribution functions between groups. Common tests are in the class of weighted linear rank tests, including the log-rank test (LR test), the Wilcoxon test (W test), the Gray and Tsiatis test (GT test); see, e.g., Harrington and Fleming 1982; Gray and Tsiatis 1989. Testing for differences between survival distributions via a concentration measure may prove more powerful than these methods, for example when one is far from the proportional hazard structure.

The Gini coefficient of concentration for a positive random variable  $X$  with cumulative distribution function  $F$  and survival function  $S$  is defined as

$$G = 1 - \frac{\int_0^{\infty} [1 - F(x)]^2 dx}{\int_0^{\infty} \Pr(X > x) dx} = 1 - \frac{\int_0^{\infty} [S(x)]^2 dx}{\int_0^{\infty} S(x) dx} ;$$

see Hanada, 1983. In survival analysis subjects have usually a finite follow-up time, so we consider the restricted version of the Gini index:

$$G_t = 1 - \frac{\int_0^t [S(x)]^2 dx}{\int_0^t S(x) dx} ,$$

where  $t$  represents the longest follow-up time in the data.

Minimum value of  $G_t$  is reached when all subjects have the same survival time, while maximum value is obtained when one individual has the maximum survival time and the rest of the population experiences the event immediately.

Bonetti et al. (2009) and Gigliarano and Bonetti (2013) propose a test based on the restricted Gini index  $G_t$  for comparing two survival functions related to two different groups. Their Gini test is aimed to test for differences in two survival distributions from the point of view of concentration. The Gini test statistic is

$$T = \frac{(\hat{G}_{1,t} - \hat{G}_{2,t})^2}{V\hat{a}r(\hat{G}_{1,t}) + V\hat{a}r(\hat{G}_{2,t})}$$

where  $\hat{G}_{j,t}$  is the estimator of the restricted Gini index for censored data referred to the group  $j$  and  $V\hat{a}r(\hat{G}_{j,t})$  is the estimator of the approximate variance of  $\hat{G}_{j,t}$ , for group  $j, j = 1, 2$ .

Bonetti et al. (2009) prove that under the null hypothesis of equality of the two survival distributions, the statistic  $T$  has an approximate chi-squared distribution with 1 degree of freedom, while, under any alternative to the null hypothesis,  $T$  is distributed as an approximate noncentral chi-squared distribution.

### 3 Data description

The empirical illustration is based on a sample of the Compulsory Communications ("Comunicazioni Obbligatorie") data provided by Italian Ministry of Labour and Social Policies.<sup>1</sup>

The Compulsory Communications (henceforth, CC) data include all activations, transformations, fixed-term extensions, early-anticipated terminations of a working relationship, either public or private.

The sample refers to all Italian workers born on 15 January, 15 April, 15 July and 15 October of any year. Our database therefore includes about 1 out of 91 of all workers who have been involved in the CC system over the period between January 2009 and June 2012.

The population of interest are the 18-35 aged workers who activated a contract in 2009. Individuals who entered the CC database for the first time after December 31, 2009 are excluded from the analysis.

The CC data have as unit of observation the contract ("contratto di lavoro"), defined as a working relationship between an employer and an employee and characterized by a starting date. However, in the context of mobility analysis, the key concept is the worker rather than the contract; therefore, the worker's history needs to be reconstructed starting from the original CC data, so that the observation unit becomes the individual.

For more details on the data preparation and cleaning process we refer to Lilla and Staffolani (2011), while further information on the methodology for joining different contracts corresponding to same individual can be found in Picchio and Staffolani (2013).

CC data provides information on the daily occupational status of an individual. Here for simplicity a monthly unit of time is considered, and for each month he prevalent contract is selected (according to type and length of contract).

The variable of interest is the occupational status. Four are the types of occupational status considered, that are ordered as follows: (i) not in employment, (ii) temporary contract, including fixed-term contract ("contratto a tempo determinato"), parasubordinate contract ("contratto di collaborazione coordinata e continuativa"), internship contract ("contratto di stage"), interim contract ("lavoro interinale"), (iii) apprenticeship contract ("contratto di apprendistato"), (iv) permanent contract, that is the open-ended contract ("contratto a tempo indeterminato").

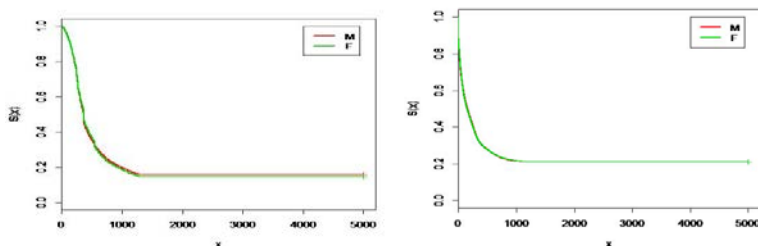
We apply the Gini test discussed above to the measurement of concentration in survival times within groups of workers, and as a way to compare the distribution of survival times across such groups.

Analysis of the differences in survival times of job contracts has been performed among subgroups of workers, based on gender, educational level and geographical area.

---

<sup>1</sup> The Compulsory Communication Data are used with the permission of the Ministry of Labour and Social Policies thanks to the agreement between the Department of Economics and Social Sciences of Marche Polytechnic University and General Department for the Innovation Technology of the Ministry of Labour and Social Policies. The authors are grateful to Stefano Staffolani and Matteo Picchio for the data preparation.

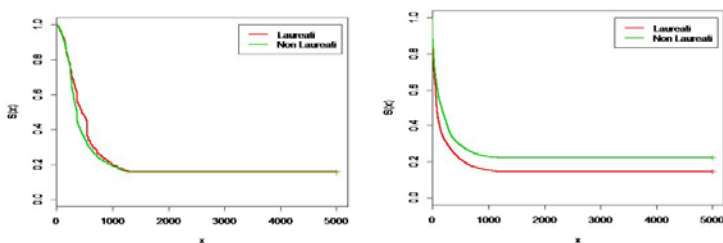
In particular, we have analysed differences both (i) in the length of the first job contract and (ii) in the waiting time between the end of the first contract and the beginning of the second one. The results are summarised in Table 1 and illustrated in Figures 1 to 4.



**Figure 1:** Male versus female. Left-hand side: Length of the first job. Right-hand side: Waiting time for a new first job.

A first analysis is aimed at determining whether there are gender differences in the Italian labour market. Figure 1 and Table 1 reveals that there exists no significant difference between young males and young females in the waiting time between the end of the first contract and the beginning of a new one, while significant differences emerge in the length of the first job contract, which is longer for males and females.

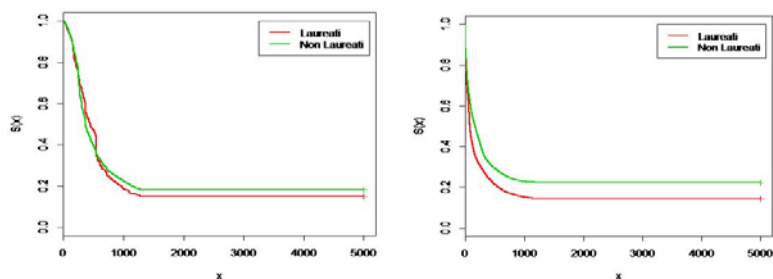
We also test for the presence of significant impact of the educational level on the Italian labour market: Table 1 and Figure 2 shows that tertiary education helps in finding quickly a new job, while it seems not so relevant for activating permanent contracts. With a particular focus on the tertiary economic sector, if a worker has tertiary education he will find quicker a job at the end of the first contract, but the length of his first contract will be shorter, in comparison to workers in the same economic sector but without tertiary education (see Table 1 and Figure 3).



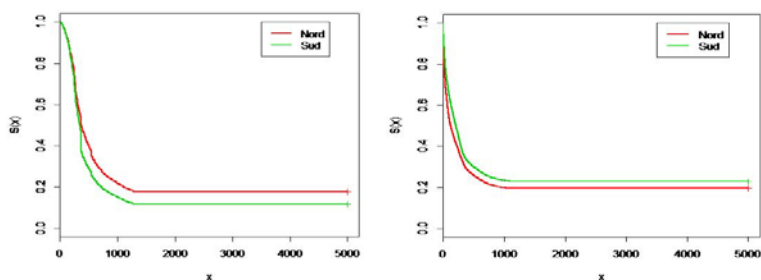
**Figure 2:** Tertiary education versus non-tertiary education. Left-hand side: Length of the first job. Right-hand side: Waiting time for a new job.

**Table 1:** P-values of Gini, Gray-Tsiatis (GT), Log Rank (LR) and Wilcoxon (W) tests for different groups comparisons.

		Gini	GT	LR	W
GENDER (Male versus female)	Length of the first job	0.0152	0.0051	0.4041	0.4997
	Waiting time for new job	0.8366	0.7629	0.9687	0.9865
EDUCATION (Tertiary versus non tertiary)	Length of the first job	0.0000	0.4646	0.0000	0.0000
	Waiting time for new job	0.0000	0.0000	0.0000	0.0000
EDUCATION IN TERTIARY SECTOR (Tertiary versus non tertiary)	Length of the first job	0.0000	0.0000	0.5114	0.0463
	Waiting time for new job	0.0000	0.0000	0.0000	0.0000
GEOGRAPHICAL AREA (North versus South)	Length of the first job	0.8834	0.0000	0.0000	0.0000
	Waiting time for new job	0.0000	0.0000	0.0000	0.0000

**Figure 3:** Tertiary education versus non-tertiary education within the tertiary economic sector. Left-hand side: Length of the first job. Right-hand side: Waiting time for a new job.

Finally, we compare the Italian macro areas (North, Center and South): no statistically significant differences emerge between North and Center of Italy (data are not shown), while differences emerge between North (or Center) and South of Italy. Table 1 and Figure 4 reveals that the labour market in the North of Italy is characterized by higher percentage of permanent contracts and by shorter waiting time for the activation of the second contract, if compared to the South of Italy.



**Figure 4:** North versus South of Italy. Left-hand side: Length of the first job. Right-hand side: Waiting time for a new job.

## 4 Concluding remarks

In this paper we have examined the Italian labour market dynamics from a novel point of view, based on the concentration analysis.

The empirical analysis revealed that there exists no significant difference between male and female in the waiting time between the end of the first contract and the beginning of a new one. Gender differences emerge, instead, in the length of the first job contract, which appears to be significantly longer for males than for females.

Significant differences emerge also among geographical areas: the North of Italy has the highest percentage of permanent contracts and also the shortest waiting time for the second contract.

Finally, different levels of education have different impact on the Italian labour market: tertiary education helps in finding quickly a new job, while it seems not so relevant for activating permanent contracts.

## References

1. Bonetti M., Gigliarano C., Muliere P., 2009. The Gini concentration test for survival data. *Lifetime Data Analysis*, Vol. 15, pp. 493-518.
2. GIGLIARANO C., BONETTI M. (2013), Gini test for survival data in presence of small and unbalanced groups, *Epidemiology, Biostatistics and Public Health*, Volume 10, Number 2, DOI:10-2427/8762.
3. GINI C. (1912) Variabilità e mutabilità. Contributo allo studio delle distribuzioni e relazioni statistiche. *Studi Economico-Giuridici dell'Università di Cagliari III*
4. GINI C. (1914) Sulla misura della concentrazione e della Variabilità dei caratteri. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti LXXIII (part 2):1203–1248.*
5. Gray R.J., Tsiatis A.A., 1989. A linear rank test for use when the main interest is in differences in cure rates, *Biometrics*, Vol. 45, pp.899-904.
6. Hanada K., 1983. A formula of Gini's concentration ratio and its applications to life tables, *Journal of the Japan Statistical Society*, Vol. 19, pp.293-325.
7. Harrington D.P., Fleming T.R., 1982. A class of rank test procedures for censored survival data, *Biometrika* Vol. 69, No. 3, pp.553-566.
8. PICCHIO M. and STAFFOLANI S. 2013. Does Apprenticeship Improve Job Opportunities? A Regression Discontinuity Approach, IZA DP No. 7719.



# On the use of DISTATIS to handle multiplex networks

## *L'utilizzo di DISTATIS per l'analisi delle reti multiplex*

Giuseppe Giordano, Giancarlo Ragozini and Maria Prosperina Vitale

**Abstract** Multiplex networks arise when there exists more than one source of relationships for a common set of nodes. For such data, the usual approaches consist of dealing with each relationship separately or of merging the information in a unique network. In the present contribution, we propose using factorial methods to visually explore the complex structure in multiplex networks. Specifically, the derived adjacency matrices from one-mode multiplex networks are analyzed using the DISTATIS technique, an extension of multidimensional scaling to three-way data. This technique allows the representation of the different types of relationships in separate spaces for each layer and in a compromise space. How the analytic procedure works is illustrated using a real-world example.

**Abstract** *Una rete è detta multiplex quando lo stesso insieme di nodi è connesso attraverso diversi tipi di relazioni. Questo tipo di reti è usualmente analizzato considerando i legami singoli o derivando una rete unica che nasce come combinazione dei diversi tipi di relazione. L'obiettivo del presente lavoro è di estendere l'utilizzo dei metodi fattoriali per esplorare la struttura complessa delle reti multiplex. In particolare, le matrici di adiacenza derivate da reti multiplex sono analizzate con la tecnica DISTATIS, ovvero lo scaling multidimensionale per dati a tre vie. Questo metodo permette di rappresentare i diversi tipi di relazioni in spazi separati e in uno spazio unico, detto compromesso. Le potenzialità dell'approccio proposto sono discusse attraverso un caso studio.*

**Key words:** AUCS data, DISTATIS, Multidimensional Scaling, Multiplex Network

---

G. Giordano

Department of Economics and Statistics, University of Salerno e-mail: [ggiordan@unisa.it](mailto:ggiordan@unisa.it)

G. Ragozini

Department of Political Science, University of Naples Federico II e-mail: [giragoz@unina.it](mailto:giragoz@unina.it)

M.P. Vitale

Department of Economics and Statistics, University of Salerno e-mail: [mvitale@unisa.it](mailto:mvitale@unisa.it)

## 1 Introduction

Multilayer network data arise when there exists more than one source of relationships for a common set or different sets of nodes. For instance, in social networks, one can consider several types of relationships of different actors: friendship, neighbors, kinship, membership, etc. A multiplex network is a special case of a multilayer network [4] that consists of a fixed set of nodes that interacts through different types of relationships. For this kind of data, the usual approaches consist of dealing with multiple relationships separately or of flattening the information embedded in all layers. The latter reduces the complexity of multiplex data and may lead to a loss of relevant information. To cope with this issue, it could be useful to propose analytic tools that can be used to adapt multivariate methods to network data [2]. In this regard, factorial methods have been proposed in the social network analysis (SNA) framework to explore different network structures [see, e.g., 3, 6, 10, 12], including attributes of nodes and events [7], or to analyze network-derived measures [9]. In the case of multiplex networks, canonical correlation analysis was adopted to identify dimensions along which networks are related to each other [2], and an analytical procedure was recently introduced for dimension reduction using cluster analysis [14].

To this end, the present contribution aims at extending the use of factorial methods to visually explore the hidden structure of multiplex networks preserving the inherent complexity. More specifically, we focus on one-mode networks, analyzing the corresponding set of adjacency matrices using the DISTATIS technique [1]. This represents an extension of the multidimensional scaling applied to a set of distance matrices derived on the same set of objects. It allows us to represent the different kinds of relationships (inter-structures) both in separate spaces and in a common space, called *compromise*. The proposed method enhances the visual exploration of: *i*) the network structure in terms of nodes' similarity in each single layer, *ii*) the common structure of all layers, *iii*) the nodes' variation across layers, and *iv*) the similarity among the structure of layers.

The paper is organized as follows. In Section 2, the concepts and notations for multiplex networks and the analytic procedure to handle multiplex network data using the DISTATIS method are briefly presented. Section 3 discusses a real-world example along with the main results obtained using the proposed analytic procedure. Section 4 concludes with suggestions for future lines of research.

## 2 Analyzing multiplex network data with DISTATIS

Multilayer networks explicitly incorporate multiple kinds of interactions among nodes and constitute a natural environment to describe complex systems in which different sets of nodes, or the same set of nodes as in the multiplex networks, could be connected according to different kinds of relational motifs, with each layer representing a motif. In multilayer networks, it is possible to observe two sets of links:

*i*) the intra-layer connections, that is, the edges that remain inside each layer, and *ii*) the inter-layer connections, that is, the edges that cross the layers.

More formally, a multilayer network  $\mathcal{M}$  is a pair  $(\mathcal{G}, \mathcal{E})$ , with  $\mathcal{G} = \{G_k\}_{k=1, \dots, K}$ , the collection of  $K$  networks, and  $\mathcal{E} = \{E_{kh}\}_{k, h=1, \dots, K}$ , the collections of intra-layer edges,  $E_{kk} \equiv E_k$ , and of inter-layer edges  $E_{kk'}$ . In each layer,  $G_k = (V_k, E_k)$ , with  $V_k = (v_{1k}, \dots, v_{nk})$  being the set of  $n$  nodes of each network, and  $E_k \subseteq V_k \times V_k$  being the set of edges.

Let  $\mathcal{M}$  be a multiplex network where the set of nodes is fixed, that is,  $V_1 = V_2 = \dots = V_K = V$ , and the inter-layer edges are constant, that is,  $E_{kh} = \{(v, v); v \in V\}$ ,  $\forall k \neq h$  [8], then we consider from the network  $G_k \in \mathcal{G}$  the corresponding adjacency matrix  $\mathbf{A}_k = (a_{ijk})$ , with  $a_{ijk} = 1$  if  $(v_i, v_j) \in E_k$ , and  $a_{ijk} = 0$  otherwise,  $\forall i \neq j$ . The set of the  $K$  adjacency matrices gives rise to a three-way relationship matrix  $\mathbb{A} = (\mathbf{A}_1, \dots, \mathbf{A}_K)$  that can be analyzed using one of the statistical methods designed for three-way data.

In the present contribution, we adopt the DISTATIS technique [1], that is, a generalization of multidimensional scaling in the STATIS approach [5] designed to explore a set of distance matrices. The different relationships can be considered as different facets of a common underlying relational structure (corresponding to the *compromise*). Indeed, the technique allows analyzing both the relational structure embedded in each single layer and the global relational structure derived as a linear combination of the layers with data-driven weights. Therefore, it provides a rich set of analytical and graphical results that also favor the comparison of the global structure and the single-layer structures.

In order to extend DISTATIS to the study of multiplex network data, a three-way distance matrix  $\mathbb{D} = (\mathbf{D}_1, \dots, \mathbf{D}_K)$  is derived from the adjacency matrix  $\mathbb{A}$ , with  $d_{ijk} = geo_k(v_i, v_j)$  being the geodesic distance between the nodes  $v_i$  and  $v_j$  in the layer  $k$  if  $geo_k(v_i, v_j) < \infty$ , that is, if the two nodes are reachable to each other – in the case of isolate nodes, we set  $d_{ijk} = 2 \max_{v_i, v_j} geo_k(v_i, v_j)$ . Other measures of distance or data transformations could be considered. For example, an alternative way to proceed is to calculate the complement of the adjacency matrix  $\mathbb{1} - \mathbb{A} - \mathbb{2}$ , with  $a_{ij} = 1$  for pairs of non-adjacent nodes, and  $a_{ij} = 0$  for adjacent nodes, and where  $\mathbb{1}$  is the all-ones three-way matrix. The effect of different distance measures on the proposed approach should be further investigated.

Here, we consider the matrix  $\mathbb{D}$  of geodesic distances and the procedure described in Abdi *et al.* [1], suitable adapted to handle of multiplex networks. The derived compromise matrix represents a weighted average of the distance matrices using a double system of weights: the  $\gamma_{1k}^{-1}$  coefficients express the relative importance of each layer in terms of inertia; whereas the  $\alpha_k$  coefficients measure such importance with reference to the similarity among the layers. The analytical results obtained by this technique will allow representing both the nodes related to each layer in the common reference space and each layer as a single point in the space defined by the first eigenvectors of the between-distance similarity matrix, highlighting the similarity of the layers in terms of the whole network structure.

### 3 A real-world example

Many multilayer networks are collected and used as examples to demonstrate the usefulness of new proposed methods [see, e.g., 4, 8, and references therein]. In order to illustrate how the DISTATIS technique works in practice for the treatment of multiplex networks, we consider a data set containing different kinds of online and offline relationships between 61 employees (out of 142) of the Computer Science Department at Aarhus University [AUCS data, 13], [4]. Five connections among employees were considered: co-authoring a publication [co-author]; being friends on Facebook [FB]; being involved in repeated leisure activities [leisure]; regularly eating lunch together [lunch]; and working together [work]. All relationships are undirected and unweighted. In addition, two attribute variables were measured for each employee, *research group* and *academic position* [i.e., professor, postdoc researcher, PhD student, and administrative staff].

Even if we observe each singleton dimension of collaboration among employees, with the proposed approach we can derive a unifying dimension of the underpinning concept as a whole. At the same time, every dimension (layer) tells us about local phenomena that can be analyzed and described in terms of an actor's position in the network. Therefore, the following two research questions must be addressed: 1) *Are there groups based on the position in the networks and on relational similarities?* 2) *How similar are the network structures achieved by the different types of relationships?*

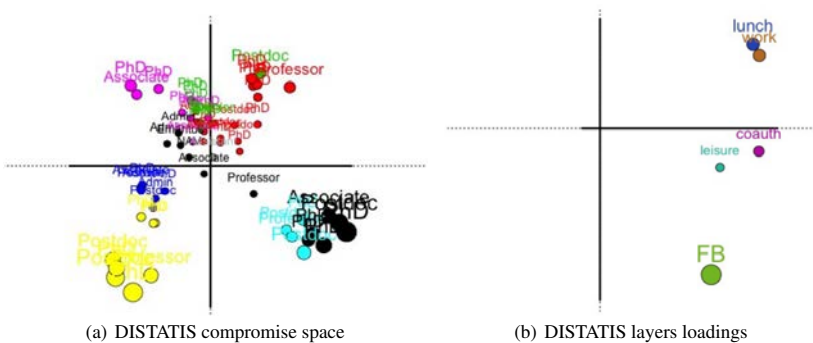
Starting with the five adjacency matrices of AUCS data, some DISTATIS results are summarized in Tables 1. The RV's coefficients matrix among layers, with coefficients that usually measure the congruence between two matrices [11], shows here the similarities between each pair of layers. The two layers *lunch* and *work* present the higher value. The factors' scores (F1, F2), the eigenvectors (V1, V2), and the  $\alpha$  weights (that are closely related to each other) indicate the importance of *lunch*, *co-authorship*, and *work* relationships in defining the compromise structure. In addition, the first two dimensions account for about 66% of the inter-layers dissimilarity.

	co-auth	FB	leisure	lunch	work	F1	F2	V1	V2	$\alpha$
co-auth	1.00	0.38	0.35	0.41	0.51	0.78	-0.11	0.50	-0.12	0.23
FB		1.00	0.23	0.22	0.25	0.55	-0.72	0.35	-0.77	0.16
leisure			1.00	0.32	0.27	0.59	-0.19	0.38	-0.21	0.17
lunch				1.00	0.58	0.75	0.41	0.48	0.44	0.22
work					1.00	0.78	0.36	0.50	0.39	0.23

**Table 1** DISTATIS results: RV's coefficients matrix among layers, factors' scores (F1, F2), eigenvectors (V1, V2) for the five layers in the first two dimensions, and  $\alpha$  weights

Based on the representation of the actors in the DISTATIS compromise factorial plan (Figure 1a), in which each employee is labelled by their academic position and colored according to the research group they belong to, some groups emerge

clearly. They are mostly consistent with the research group membership, even if some actors bridge different groups. The groups are instead mixed up with respect to the academic position. The factorial map in Figure 1b shows the role played by each layer in determining the final compromise space. Whereas every layer has an important (and positive) role in weighting the final configuration (let us look at the first component as a size-effect component), it is the second axis that reveals the real shape of our configuration. On the top, we can see the two prominent layers (lunch and work) that lie close and are separated by the co-authorship and leisure layers, located in the middle and opposite to the Facebook layer on the bottom.



**Fig. 1** DISTATIS representation of actors and layers in the compromise space of AUCS data

## 4 Concluding remarks

In this work, we proposed an analytic method to treat multiplex network data based on factorial techniques. The results of the illustrative example indicate the high explicative power of the DISTATIS technique in capturing similarities among relationships. The possibility of measuring the inter-dissimilarity between layers allows the definition of a suitable subspace where comparisons at both the layer and node levels can be made.

In conclusion, we provide some suggestions for future lines of research. The analytic results of the adopted approach are useful for the substantive interpretation of multiplex relationships. These findings could also be used to compute new measures for multiplex network data. Moreover, as network data allows for several ways of computing distances, a comparison of how different distance measures affect the results and the visualization of compromise space should be addressed. The analyzed real-world example considers dichotomous, undirected, one-mode networks, and

the attribute data have been exploited only to enrich the interpretation. Extensions of the methods to deal with directed one-mode networks and two-mode networks could be of interest, as could the inclusion of attribute data in defining the analytic procedure.

**Acknowledgements** The authors would like to thank Matteo Magnani (Uppsala University, Sweden) for data availability.

## References

- [1] Abdi, H., Valentin, D., Chollet, S., Chrea, C.: Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Qual. Prefer.* **18**, 627–640 (2007)
- [2] Carroll, C.: Canonical correlation analysis: Assessing links between multiplex networks. *Soc. Netw.* **28**, 310–330 (2006)
- [3] D’Esposito, M.R., De Stefano, D., Ragozini, G.: On the use of multiple correspondence analysis to visually explore affiliation networks. *Soc. Netw.* **38**, 28–40 (2014)
- [4] Dickison, M. E., Magnani, M., Rossi, L.: *Multilayer social networks*. Cambridge University Press, Cambridge (2016)
- [5] Escoufier, Y.: Objectifs et procédures de l’analyse conjointe de plusieurs tableaux de donnés. *Statistique et analyse des donnés* **10**, 1–10 (1985)
- [6] Faust, K.: Using correspondence analysis for joint displays of affiliation networks. In: Carrington, P.J., Scott, J., Wasserman, S. (eds.) *Models and methods in social network analysis*, pp. 117–147. Cambridge University Press, Cambridge (2005)
- [7] Giordano, G., Vitale, M. P.: On the use of external information in social network analysis. *Adv. Data Anal. Class.* **5**, 95–112 (2011)
- [8] Kivelà, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: *Multilayer networks*. *J. Complex Netw.* **2**, 203–271 (2014)
- [9] Liberati, C., Zappa, P.: Dynamic patterns analysis meets Social Network Analysis in the modeling of financial market behavior. In: *Proceedings 59th ISI World Statistics Congress Vol. 25*, pp. 2447–2452 (2013)
- [10] Ragozini, G., De Stefano, D., D’Esposito, M.R.: Multiple factor analysis for time-varying two-mode networks. *Netw. Sci.* **3**, 18–36 (2015)
- [11] Robert, P., Escoufier, Y.: A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient. *Appl. Stat.* **25** (3): 257–265 (1976)
- [12] Roberts, J.M.: Correspondence analysis of two-mode network data. *Soc. Netw.* **22**, 65–72 (2000)
- [13] Rossi, L., Magnani, M.: Towards effective visual analytics on multiplex and multilayer networks. *Chaos Soliton Fract.* **72**, 68–76 (2015)
- [14] Vörös, A., Snijders, T. A.: Cluster analysis of multiplex networks: Defining composite network measures. *Soc. Netw.* **49**, 93–112 (2017)

# Profiles of students on account of complex problem solving (CPS) strategies exploited via log-data

## *Profili di studenti basati sulle strategie di problem solving complesso esplorate attraverso log-data*

Michela Gnaldi, Silvia Bacci, Samuel Greiff and Thiemo Kunze

**Abstract** This paper aims at identifying profiles of students that are homogenous with regard to their ability to solve Complex Problem Solving (CPS) tasks, as assessed by the MicroDYN approach, a computer test made of 9 independent tasks, and administered to a sample of 6th and 9th grade Finnish students. For this aim, we estimate a discrete two-tier Item Response Theory (IRT) model. Results indicate that: (1) the conceptualisation of CPS as a three-dimensional variable is reasonable and (2) there are seven latent classes of students characterised by a specific profile with regard to the adopted CPS strategies, with students clustered in the higher latent classes having generally a higher CPS ability than the others, across the three CPS dimensions.

**Abstract** *L'obiettivo di questo articolo consiste nell'identificare profili omogenei di studenti rispetto alla loro capacità di risolvere problemi complessi, valutata con il test MicroDYN, un test al computer composto da 9 compiti indipendenti e somministrato a un campione di studenti finlandesi di sesto e nono grado. A questo fine, stimiamo un modello di Item Response Theory (IRT) multidimensionale a classi latenti [2]. I risultati indicano che: (1) è ragionevole concettualizzare il CPS come variabile tri-dimensionale e (2) sono osservabili sette classi latenti di studenti caratterizzate da uno specifico profilo in termini di strategia di risoluzione di compiti complessi adottata, con gli studenti raggruppati nelle classi latenti più alte che mostrano generalmente maggiori capacità degli altri, rispetto a ciascuna delle tre dimensioni di CPS.*

---

Michela Gnaldi

Department of Political Sciences, University of Perugia, e-mail: [michela.gnaldi@unipg.it](mailto:michela.gnaldi@unipg.it)

Silvia Bacci

Department of Economics, University of Perugia, e-mail: [silvia.bacci@unipg.it](mailto:silvia.bacci@unipg.it)

Samuel Greiff

Cognitive Science and Assessment, University of Luxembourg, e-mail: [samuel.greiff@uni.lu](mailto:samuel.greiff@uni.lu)

Thiemo Kunze

Cognitive Science and Assessment, University of Luxembourg, e-mail: [thiemo.kunze@uni.lu](mailto:thiemo.kunze@uni.lu)

**Key words:** Log-data, Complex Problem Solving, Discrete two-tier Item Response Theory (IRT) model, Profiles of students.

## 1 Introduction

Complex Problem Solving (CPS) can be conceptualized in terms of a multidimensional latent variable. Buchner [5] defines CPS as “the successful interaction with task environments that are dynamic (i.e., change as a function of user’s intervention and/or as a function of time) and in which some, if not all, of the environment’s regularities can only be revealed by successful exploration and integration of the information gained in that process”. Key aspects contributing to characterize CPS - and also to differentiate it from reasoning - are then: (i) dynamicity, as dynamic interactions are necessary to reveal previous unknown information and to achieve goals using subsequent steps that depend upon each other, (ii) not all information necessary to solve the problem is given at the outset, (iii) the testee has to apply adequate strategies in order to actively generate information, and (iv) procedural abilities have to be used in order to control a given system.

When interacting with a computer test to solve CPS tasks, students produce log-files, that is, finely grained data containing rich information on every single behavioral action they undertake. These pieces of information provide researchers, teachers, and policy makers with important insides about students’ proficiency and about how to support them in optimizing their cognitive potential [13]. The question of how log-files can be analysed to understand students’ levels of proficiency became central in 2012 when the computer-based assessment of complex problem solving was included in the Programme for International Student Assessment (PISA). Despite that, the exploitation of this rich resource through log-file analyses is still in its infancy [10].

In this contribution, we analyse log-data drawn from the MicroDYN, a computer test to assess CPS ability, administered to a sample of 6th and 9th grade Finnish students. The log-data at issue have been subsequently transformed into three types of dichotomous items, each reflecting the three underlying dimensions of CPS, over nine different tasks. With these data at hand, we aim at identifying profiles of students that are homogenous as regard to their ability in CPS, which can be conceptualized in terms of a multidimensional latent variable (for a similar applicative work see for instance [8]). For this aim, we have to account in a suitable way for: (i) the different dimensions that characterize CPS, that is, exploration behaviour, knowledge acquisition, knowledge application, (ii) the discrimination power and the difficulty of items that measure CPS, so as to evaluate the capacity of each item to distinguish between individuals with different levels of CPS ability. To investigate the above mentioned objectives we estimate a discrete two-tier Item Response Theory (IRT) model [2] allowing us to (i) characterise the latent classes of testees through the support points estimates over the accounted CPS dimensions and (ii) cluster testees



in the latent classes according to the maximum posterior class membership probabilities.

This paper is organised as follows. In Section 2 we describe the data used, in Section 3 the model and in Section 4 we provide results and a brief discussion of them.

## 2 The data

The MicroDYN test is a computer test to assess CPS ability. The test is organized in 9 independent tasks (named Lemonade, Drawing, Cat, Moped, Game, Gardening, Handball, Spaceship, and Aid, in line with the cover stories used for the different tasks), whose characteristics are varied in order to produce items across a broad range of difficulty [9]; each task lasts about five minutes, for a total testing time of less than one hour.

When working on MicroDYN, participants face three different aspects directly related to the three facets of problem solving ability [11]:

- exploration behaviour, denoting the use of adequate strategies;
- CPS knowledge acquisition, denoting the knowledge generated;
- CPS knowledge application, denoting the ability to control the system.

For any of the tasks of the MicroDYN, during knowledge acquisition the testee has to discover the relations between input variables, that can be manipulated, and output variables. Then, the testee has to draw his/her mental model on the screen. The match of this drawing with the real underlying model is the MicroDYN score for knowledge acquisition. After the model is drawn, the testee can click on a finish button, which leads to the knowledge application phase in which the whole complex system is reseted and the correct model provided to the students to avoid empirical dependency between knowledge acquisition and knowledge application. During knowledge application, the testee tries to reach given (and red indicated) target values on the output variables by changing the input variables until the maximum number of rounds are reached.

The MicroDYN adopts a Multiple-Item-Approach, in which multiple control rounds can be used in each task. In this approach:

- although participants work on a series of independent tasks with different goals, items in the test assessing knowledge acquisition gained during system exploration are related to the same underlying dimension and depend on one another. This is also the case for knowledge application items, such as using feedbacks in order to adjust behavior. Thus, variables within each of the dimensions (exploration behaviour, knowledge acquisition and knowledge application) are dependent on each other;
- Variables of the three kinds (exploration behaviour, knowledge acquisition, and knowledge application) within each task are likely to be dependent on one another;

- Additional knowledge that is eventually gained during knowledge application in a task does not help participants in resolving subsequent items in different tasks.

The MicroDYN was administered to a sample of 6th and 9th grade Finnish students. In each grade level, around 2,000 students were tested.

Each of the nine tasks in the MicroDYN test is characterized by three types of binary items, for a total of 27 items, as follows:

Items of type 1: a score of 1 was given when the testee applied the Vary One Thing At a Time strategy (VOTAT; [15]) on all of the MicroDYN input variables during the exploration phase in a certain task, and 0 if this was not done; these items are affected by the exploration behaviour dimension of CPS;

Items of type 2: a score of 1 was given if the model drawn by the testee and related to a certain task was completely correct, and 0 otherwise; these items are affected by the knowledge acquisition facet;

Items of type 3: a score of 1 was given if target areas of all variables in a certain task were reached, and 0 if this was not done; these items are affected by the knowledge application dimension.

### 3 The model

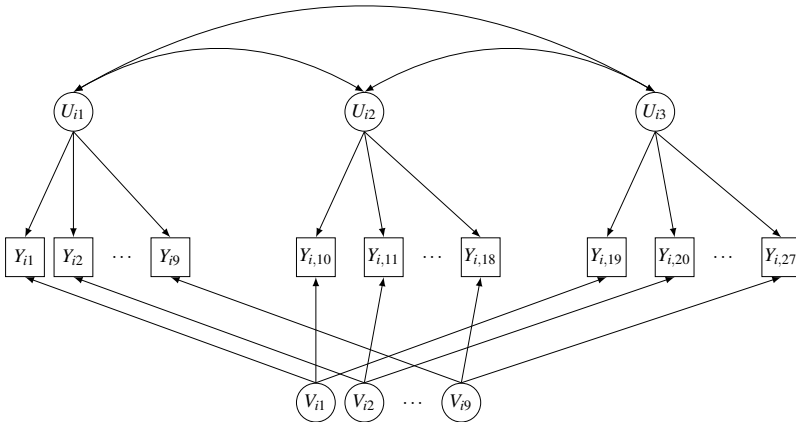
We aim at investigating the above mentioned objectives through the estimation of a discrete two-tier Item Response Theory (IRT) model [2]. The model at issue is characterized by two independent vectors of latent variables that are measured on each student  $i$  ( $i = 1, \dots, n$ ) through the responses on  $J = 27$  items related to 9 different tasks:

- latent variables  $U_i = \{U_{id_1}\}$  with  $d_1 = 1, \dots, D_1$ : in our case,  $D_1 = 3$  with  $U_{i1}$  denoting exploration behaviour,  $U_{i2}$  denoting CPS knowledge acquisition, and  $U_{i3}$  denoting CPS knowledge application;
- latent variables  $V_i = \{V_{id_2}\}$  with  $d_2 = 1, \dots, D_2$ : in our case,  $D_2 = 9$  with each  $V_{id_2}$  denoting a latent variable accounting for correlation among responses of individual  $i$  on items related to the *same task*  $d_2$ .

Latent variables in  $U_i$  are assumed to have a discrete distribution, with a finite number  $k_1$  of support points  $u_1, \dots, u_{k_1}$  and corresponding mass probabilities (or weights)  $\lambda_1, \dots, \lambda_{k_1}$ . Each support point  $u_{h_1} = (u_{h_1 1}, \dots, u_{h_1 D_1})'$  ( $h_1 = 1, \dots, k_1$ ) has a particularly nice interpretation, as it denotes an unobservable cluster (or latent class) of individuals having an homogenous profile in terms of attitude toward CPS. Besides, each weight  $\lambda_{h_1}$  denotes the latent class membership probability, that is,  $\lambda_{h_1} = p(U_i = u_{h_1})$ . Similarly, latent variables in  $V_i$  have a discrete distribution with  $k_2$  support points  $v_1, \dots, v_{k_2}$  and related weights  $\pi_1, \dots, \pi_{k_2}$ , with  $\pi_{h_2} = p(V_i = v_{h_2})$  ( $h_2 = 1, \dots, k_2$ ).

Items of type 1, 2, and 3 are related to the latent variables  $U_i$  and  $V_i$  through an IRT parameterization. In more detail, responses to items of type 1 depend on latent variable  $U_{i1}$ ; responses to items of type 2 depend on latent variable  $U_{i2}$ , and

responses to items of type 3 depend on latent variable  $U_{i3}$ . Moreover, each latent variable  $V_{id_2}$  is measured by the three types of items related to the same task  $d_2$  ( $d_2 = 1, \dots, 9$ ). These relations are formalized by introducing the disjoint sets  $\mathcal{U}_1, \dots, \mathcal{U}_{D_1}$  and  $\mathcal{V}_1, \dots, \mathcal{V}_{D_2}$ , where  $\mathcal{U}_{d_1}$  contains the indices of items affected by latent variable  $U_{id_1}$  ( $d_1 = 1, \dots, D_1$ ) and, similarly,  $\mathcal{V}_{d_2}$  contains the indices of items affected by latent variable  $V_{id_2}$  ( $d_2 = 1, \dots, D_2$ ). The detailed relations among latent variables and items are displayed in Figure 1, where  $Y_{ij}$  denotes the response of person  $i$  to item  $j$  ( $i = 1, \dots, n, j = 1, \dots, J$ ).



**Fig. 1** Path diagram of the discrete two-tier model with  $D_1 = 3$  latent variables  $U_i$ ,  $D_2 = 9$  latent variables  $V_i$ , and  $J = 27$  items (three types of item for each task).

Being all items dichotomously scored, we adopt a Two-Parameter Logistic (2-PL) parameterization [4], as follows ( $i = 1, \dots, n, j = 1, \dots, J$ )

$$\text{logit } p(Y_{ij} = 1 | U_i = u_{h_1}, V_i = v_{h_2}) = \gamma_{1j} \sum_{d_1=1}^{D_1} 1\{j \in \mathcal{U}_{d_1}\} u_{h_1 d_1} + \gamma_{2j} \sum_{d_2=1}^{D_2} 1\{j \in \mathcal{V}_{d_2}\} v_{h_2 d_2} - \beta_j, \tag{1}$$

where, as usual in the IRT parameterization,  $\gamma_{1j}$  and  $\gamma_{2j}$  are the discrimination parameters related to variables in  $U_i$  and  $V_i$ , respectively, and  $\beta_j$  is the difficulty parameter of item  $j$ ;  $1\{\cdot\}$  is the indicator variable. Parameters of model at issue may be estimated by means of the maximum likelihood approach, efficiently implemented through the Expectation-Maximization (EM) algorithm [6] in the R package `MLCIRTwithin` [3]. It has to be noted that the model identification requires the specification of suitable constraints on the item parameters and/or the support points; for details, see [2].

We also outline that formulation of equation (1) is completely general and it allows us for any number of components in  $U_i$  (other than in  $V_i$ ). More in detail, in the following we compare in terms of goodness-of-fit the proposed model characterized by  $D_1 = 3$  latent variables  $U_{i1}, U_{i2}, U_{i3}$  with the following nested alternatives: (i) model with  $D_1 = 2$  elements in  $U_i$ , being  $U_{i1}$  and  $U_{i2}$  collapsed in a same dimension; (ii) model with just  $D_1 = 1$  element in  $U_i$ , being all items affected by only one latent variable ( $U_{i1}, U_{i2}, U_{i3}$  collapsed).

It is also worth to be noted that the number of support points of  $U_i$  and  $V_i$ , that is,  $k_1$  and  $k_2$ , do not represent model parameters, but they have to be a priori fixed. For this aim, following the main stream of the literature (see mainly [12]) we base our choice on the Bayesian Information Criterion (BIC; [14]) and on the Akaike's Information Criterion (AIC; [1]). In both cases, as smaller the corresponding indices are, as better it is. In practice, we fix  $k_2 = 2$  as we are not particularly interested in clustering individuals according to latent variable  $V_i$  and we estimate model in (1) for increasing values of  $k_1$  until the index does not start to increase. Then, we select the previous value of  $k_1$  as the optimal number of latent classes, which guarantees the best compromise between goodness-of-fit and model parsimony.

To summarize, as main results of the model estimation we obtain:

- difficulty and discriminating parameters for each item, having the usual interpretation as in the traditional IRT models,
- support points and weights for the latent classes,
- posterior class membership probabilities for each individual, which are used to cluster individuals in the latent classes according to a suitable criterion (usually, the maximum a posteriori one).

## 4 Application and discussion of results

As stated in the previous section, within the model at issue, the latent variables in  $U_i$  are assumed to have a discrete distribution, with a finite number of latent classes,  $k_1$ . Thus, the first step of our application consists in selecting the number of latent classes  $k_1$ , that is, the number of classes of units (i.e., students) of our data. We make this selection by assuming 3 distinct dimensions,  $D_1 = 3$ , that is,  $U_{i1}, U_{i2}, U_{i3}$ . For this aim, we estimate the BIC index for an increasing number of latent classes, and we select as optimal number the one for which we observe the first lowest value for BIC. In our case, as the minimum value for BIC is observed in correspondence of the model for which  $k_1 = 7$  (i.e., 59236.58), we select this number of latent classes. Differently, for the other dimensions in  $V_i$ , which account for the possible correlation of the 3 items in each task, we assume for simplicity a fixed number of latent classes, that is  $k_2 = 2$ .

Afterwards, we check for the initial assumption of 3 distinct dimensions underlying CPS by comparing through BIC and AIC the proposed model characterized by  $D_1 = 3$  latent variables ( $U_{i1}, U_{i2}, U_{i3}$ ) with a model with  $D_1 = 2$  elements in  $U_i$ , being  $U_{i1}$  and  $U_{i2}$  collapsed in a same dimension (i.e., exploration is directly transferred

into knowledge acquisition) and with a model with  $D_1 = 1$  element in  $U_i$ , being  $U_{i1}$ ,  $U_{i2}$ ,  $U_{i3}$  collapsed in the same unique latent variable (i.e., CPS is a unidimensional latent variable). We select as best model the one for which we observe the minimum value for BIC and AIC, that is, a three-dimensional model, as shown in the table below. This choice implies that it is reasonable to assume that CPS is a multidimensional variable made of three distinct facets, that is, exploration behaviour, knowledge acquisition, and knowledge application.

**Table 1** Estimated BIC and AIC for the three models with  $D_1 = 3$  (CPS is a three-dimensional latent variable),  $D_1 = 2$  (CPS is a two-dimensional latent variable),  $D_1 = 1$  (CPS is a uni-dimensional latent variable). In boldface the smallest values.

	$D_1 = 3$	$D_1 = 2$	$D_1 = 1$
BIC	<b>59236.58</b>	59347.67	59478.31
AIC	<b>58708.66</b>	58850.09	59011.07

The key step of our application consists in estimating the support points  $(u_1, \dots, u_{k_1})$  and corresponding weights  $(\lambda_1, \dots, \lambda_{k_1})$ , denoting the class membership probabilities, given the  $D_1 = 3$  dimensions underlying CPS, as shown in Table 2.

**Table 2** Support points for the seven latent classes ( $h_1 = 1, \dots, 7$ ) and the three dimensions ( $U_{i1}$ ,  $U_{i2}$ ,  $U_{i3}$ ) of CPS.

	$h_1 = 1$	$h_1 = 2$	$h_1 = 3$	$h_1 = 4$	$h_1 = 5$	$h_1 = 6$	$h_1 = 7$
$U_{i1}$	-1,5000	-1,0000	-0,5000	0,0000	0,5000	1,0000	1,5000
$U_{i2}$	-0,2228	-0,6981	0,7389	1,1304	1,5816	2,1438	4,9642
$U_{i3}$	1,7671	-0,7089	1,0431	3,1116	1,8435	3,6384	4,7766
$\lambda_{k_1}$	0,3000	0,1849	0,0810	0,0837	0,1714	0,1481	0,0309

Support points provide a rich information, as the higher the support points, the higher the CPS ability of the testees clustered in each class to resolve the complex tasks involved in the three dimensions underlying CPS. It can be noticed that the support points tend to increase when moving from the lowest latent classes (i.e.,  $h_1 = 1$ ) to the highest (i.e.,  $h_1 = 7$ ), meaning the testees clustered in the highest latent classes have a higher ability than the others, over all the three dimensions. However, there are also cases of support points not respecting this general observed rule. For example, let us consider the first two latent classes ( $h_1 = 1$  and  $h_1 = 2$ ). In the first latent class ( $h_1 = 1$ ) there are students who are the poorest as regard  $U_{i1}$ , but (i) not the first poorest as regard  $U_{i2}$  (i.e., -0,2228 is the second lowest value) and (ii) close to average as regard  $U_{i3}$ . In the second latent class ( $h_1 = 2$ ) there are students who are the second poorest as regard  $U_{i1}$ , and the poorest as regard  $U_{i2}$  and  $U_{i3}$ . Overall, these analyses allow us to profile testees on account of the adopted strategies of CPS.

**Acknowledgements** We would like to particularly thank Mari-Pauliina Vainikainen and our Finnish colleagues at University of Helsinki for their great support in collecting the data.

## References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N., Csaki, F. (eds.) *Second International symposium of information theory*, pp. 267-281. Akademiai Kiado, Budapest (1973)
2. Bacci, S., Bartolucci, F.: Two-Tier Latent Class IRT Models in R. *The R Journal*, **8**, 139–166 (2016)
3. Bartolucci, F., Bacci, S.: MLCIRTwithin: Latent Class Item Response Theory (LC-IRT) Models under Within-Item Multidimensionality. R package version 2.1 (2016) Available via <https://cran.r-project.org/web/packages/MLCIRTwithin>. Cited 2 March 2017.
4. Birnbaum, A.: Some latent trait models and their use in inferring an examinee's ability. In: Lord, F. M., Novick, M. R. (eds.) *Statistical Theories of Mental Test Scores*, pp. 395-479. Addison-Wesley, Reading, MA (1968)
5. Buchner, A.: Basic topics and approaches to the study of complex problem solving. In P. A. Frensch, & J. Funke (Eds.), *Complex problem solving: The European perspective*. Hillsdale, NJ: Erlbaum, (1995).
6. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38 (1977)
7. Formann, A. K.: Mixture analysis of multivariate categorical data with covariates and missing entries. *Computational Statistics and Data Analysis*, **51**, 5236–5246 (2007)
8. Gnaldi, M., Bacci, S., Bartolucci, F.: A multilevel finite mixture item response model to cluster examinees and schools. *Advances in Data Analysis and Classification*, **10**, 53-70 (2015)
9. Greiff, S., Funke, J.: Systematische Erforschung komplexer Problem-lösefähigkeit anhand minimal komplexer Systeme [Some systematic research on complex problem solving ability by means of minimal complex systems]. *Zeitschrift für Pädagogik*, **56**, 216–227 (2010)
10. Greiff, S., Wüstenberg, S., Avvisati, F.: Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, **91**, 92-105, (2015).
11. Kroner, S., Plass, J. L., Leutner, D.: Intelligence assessment with computer simulations. *Intelligence*, **33**, 347–368 (2005)
12. McLachlan, G.J., Peel, D.: *Finite Mixture Models*. Wiley, New York (2000)
13. OECD.: PISA 2012 results: Creative problem solving. Paris: OECD Publishing. Osman, M. (2010). In, *Controlling uncertainty: a review of human behavior in complex dynamic environments*. *Psychological Bulletin*, **136**, 65-86.
14. Schwarz, G.: Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464 (1978)
15. Vollmeyer, R., Rheinberg, F.: Motivation and metacognition when learning a complex system. *European Journal of Psychology of Education*, **14**, 541–554 (1999)

# Characterising Italian municipalities according to the annual report of the prevention-of-corruption supervisor: a Latent Class approach.

## *Caratterizzazione dei comuni italiani sulla base delle relazioni annuali dei Responsabili Prevenzione Corruzione: un approccio a classi latenti.*

Michela Gnaldi and Simone Del Sarto

**Abstract** This work aims at characterising Italian municipalities according to what has been accomplished in terms of corruption prevention. The recent “anti-corruption law” of 2012 establishes a new plan for corruption prevention. It introduces a new figure, the prevention-of-corruption supervisor who reports if and how preventive measures are implemented within the public institution he/she represents, by filling in a standardised form, which has to be published in the institution website. We rely on these data – downloaded from each single municipality website – to apply a Latent Class model allowing us to identify groups of municipalities with a similar behaviour. Further, we qualify such classes on account of several covariates. First results show that *i.* there is a general tendency among municipalities to fulfil the prevention-of-corruption law and *ii.* virtuous municipalities are large municipalities experiencing at least one corruption event.

**Abstract** *L'obiettivo del presente lavoro è caratterizzare i comuni italiani in base a quanto realizzato in termini di prevenzione della corruzione. La recente legge “anti-corruzione” del 2012 stabilisce un nuovo programma per la prevenzione della corruzione e introduce la nuova figura del responsabile della prevenzione della corruzione. Mediante una scheda standardizzata, il supervisore riferisce le modalità con cui sono state implementate misure preventive della corruzione all'interno dell'amministrazione pubblica che rappresenta. I dati raccolti da tali schede – scaricate singolarmente dai siti istituzionali dei comuni campione – sono analizzati mediante un approccio a Classi Latenti, che consente di identificare gruppi di comuni con comportamenti simili. Inoltre, tali classi sono qualificate in base ad alcune covariate. I primi risultati mostrano i. una tendenza generale tra i comuni a*

---

Michela Gnaldi  
Department of Political Science, University of Perugia – Via Pascoli 20 06123 Perugia (Italy), e-mail: [michela.gnaldi@unipg.it](mailto:michela.gnaldi@unipg.it)

Simone Del Sarto  
Italian National Institute for the Evaluation of the Education System (INVALSI) – Via Ippolito Nievo 35 00153 Rome (Italy), e-mail: [simone.delsarto@email.com](mailto:simone.delsarto@email.com)

*rispettare la legge per la prevenzione della corruzione e ii. i comuni più virtuosi sono grandi comuni con esperienza di eventi corruttivi.*

**Key words:** prevention of corruption, Latent Class model

## 1 Introduction

Worldwide corruption is a plague affecting the economic, social and institutional development of a Country and political institutions have attempted to implement measures to contrast and contain this phenomenon [1]. Corruption is latent by nature. People involved in corruptive activities try to hide or falsify information and, as a consequence, its measurement is very complex [4]. However, there are measures of corruption, the so-called “perception-based” and “non-perceptual” measures. The former are based on the subjective perception of corruption provided by experts and are generally expressed at a country-level: an example is the well-know Corruption Perceptions Index (CPI). The latter are objective indexes based on proxies, i.e., market or statistical indicators tied with corruption – such as the price of input purchased by a public administration, the rate of criminal convictions to public officials for crimes related to corruption, and so on. The pros and cons of these two types of corruption measurement tools are commonly known, but these details are out of the scope of this work.

In Italy, with the recent law n.190 of 2012, named “Provisions for the prevention and repression of corruption and lawlessness in the public administration”, each public institution has to adopt a three-year plan for corruption prevention (“Piano Triennale per la Prevenzione della Corruzione”, PTPC), which provides an assessment of the different exposure levels of offices to the risk of corruption and specifies the organisational changes designed to prevent such risk. To this general aim, each institution selects a supervisor, indeed called prevention-of-corruption supervisor (“Responsabile per la Prevenzione della Corruzione”, RPC). Among his/her tasks, the supervisor has to fill in an annual report about the efficacy of the prevention measures defined by the PTPC. Such report is filled in through a questionnaire, made available in spreadsheet format by the Italian National Anti-Corruption Authority (ANAC) and has to be uploaded in the “Transparent administration” section of the public institution website.

In this work, we aim at classifying and qualifying a sample of municipalities according to what stated in the RPC form. Since it summarises what each institution has accomplished to prevent and contrast corruption, according to the PTPC, our purpose is to cluster municipalities in homogeneous groups as regards to the adopted anti-corruption measures. To this aim, we rely on the Latent Class (LC) model [2, 3], which allows us to cluster units with a similar behaviour on account of a latent and unobserved characteristic (i.e., corruption).

The following of this paper is organised as follows. In Section 2 the RPC form data considered here are briefly described, while the LC model is introduced in Section 3. Results are shown in Section 4 and some concluding remarks are given in Section 5.



## 2 The RPC form data

The data we consider in this work are collected through the RPC forms filled in in 2015. Each of them has been downloaded individually from the section “transparent administration” of the municipality institutional website. Overall, our sample size is made of 232 municipalities, comprising all Italian province municipalities, all the other municipalities with at least 40,000 inhabitants and particular “advised” municipalities, as stated by the ANAC act n.71 of 2013.

The RPC form has several sections reflecting different aspects about the efficacy of prevention measures, defined in the PTPC adopted by each institution. In this work we consider the responses to the opening questions of each section, requiring the institution at issue to state whether it has accomplished the required activities. The questions are related to the following contents (in square brackets it is reported the original label within the RPC form):

1. monitoring the sustainability of all measures, general and specific, identified in the PTPC [2A];
2. specific measures, in addition to mandatory ones [3A];
3. computerising the flow to fuel data publication in the “transparent administration” website section [4A];
4. monitoring data publication processes [4C];
5. training of employees, specifically dedicated to prevention of corruption [5A];
6. staff turnover as a risk prevention measure [6B];
7. checking the truthfulness of statements made by parties concerned with unfitness for office causes [7A];
8. measures to verify the existence of incompatibility conditions [8A];
9. prearranged procedures for issuing permits for assignments performance [9A];
10. reporting the collection of misconduct by public administration employees (whistleblowing) [10A].

Three possible answers can be provided to these questions: “Yes” (labelled as 1), “No, but expected by the PTPC” (2) and “No, not expected by the PTPC” (3). As can be noted, the second response is the least virtuous answer and the other two correspond to actions in line with the PTPC.

## 3 The Latent Class model

The Latent Class (LC) model is one of the most well-known latent variable models. It is often used to classify units of a sample in homogeneous groups, according to a set of categorical variables (e.g., the responses to questionnaire items). Such variables represent the observable manifestation of a unique underlying latent variable.

Considering a sample of  $n$  units, let  $\mathbf{Y}_i = [Y_{i1}, \dots, Y_{iJ}]^T$  denote the random vector of the responses provided by unit  $i$  to the  $J$  items of a questionnaire, with  $i = 1, \dots, n$ .

Each  $Y_{ij}$  is a categorical variable with  $l$  categories, generally labelled starting from 0 to  $l - 1$ . In this paper, we consider  $l = 3$  since three different response modalities can be provided to the selected RPC form items (see Section 2). However, we retain the original response labels (1,2,3).

The LC model assumes the existence of one discrete latent variable  $C_i$  with the same distribution for each unit  $i$ . This latent variable is based on  $k$  support points. Each point has a specific prior probability, denoted by  $\pi_c, c = 1, \dots, k$  and corresponds to a latent class in the population. Furthermore, the conditional probability that unit  $i$ , belonging to class  $c$ , provides response  $y$  to item  $j$  is:

$$\phi_{j|c}(y) = P(Y_{ij} = y | C_i = c), \quad j = 1, \dots, J, \quad y = 0, \dots, l - 1, \quad c = 1, \dots, k.$$

Moreover, it is assumed local independence between the response variables  $Y_{ij}$ : this hypothesis states that the response variables are conditionally independent given the latent class. This implies that the probability of observing the response vector  $\mathbf{y}_i = [y_{i1}, \dots, y_{iJ}]^\top$ , given that unit  $i$  is in latent class  $c$ , can be formulated as the product of each conditional probability reported above, over the  $J$  items. Specifically, we have:

$$P(\mathbf{y}_i | c) = P(\mathbf{Y}_i = \mathbf{y}_i | C_i = c) = \prod_{j=1}^J \phi_{j|c}(y_{ij}).$$

Then, the manifest probability of  $\mathbf{y}_i$  can be obtained as follows:

$$P(\mathbf{y}_i) = P(\mathbf{Y}_i = \mathbf{y}_i) = \sum_{c=1}^k P(\mathbf{y}_i | c) \pi_c.$$

It is often of interest to rely on an allocation rule, allowing to assign each sample unit to a particular latent class, given its response pattern. Such procedure is based on the posterior probability that unit  $i$  belongs to class  $c$ , given the response vector  $\mathbf{y}_i$ . It can be obtained using the Bayes' theorem, as follows:

$$P(c | \mathbf{y}_i) = P(C_i = c | \mathbf{Y}_i = \mathbf{y}_i) = \frac{P(\mathbf{y}_i | c) \pi_c}{P(\mathbf{y}_i)}, \quad c = 1, \dots, k. \quad (1)$$

In particular, each unit is assigned to the latent class according to its largest posterior probability.

## 4 Results

Relying on the BIC index, we can identify  $k = 2$  latent classes, for which the following prior probability estimates can be obtained:  $\hat{\pi}_1 = 0.482$  and  $\hat{\pi}_2 = 0.518$ .

In Table 1, the estimated conditional response probabilities  $\hat{\phi}_{j|c}(y)$  are reported for each item  $j$  and for each response category  $y$  (1,2,3). The conditional probabilities of a positive response ( $y = 1$ ) are often high ( $> 60\%$ ) for both the latent

**Table 1** Estimated conditional response probabilities  $\hat{\phi}_{j|c}(y)$  for the selected ten items of the RPC form.

$j$	$c$	$y = 1$	$y = 2$	$y = 3$	$j$	$c$	$y = 1$	$y = 2$	$y = 3$
1	1	0.875	0.036	0.089	6	1	0.549	0.185	0.266
	2	0.660	0.059	0.281		2	0.387	0.204	0.409
2	1	0.815	0.000	0.185	7	1	0.724	0.042	0.234
	2	0.615	0.075	0.311		2	0.077	0.187	0.736
3	1	0.811	0.055	0.134	8	1	0.817	0.001	0.181
	2	0.582	0.069	0.348		2	0.031	0.187	0.782
4	1	0.909	0.000	0.091	9	1	0.906	0.045	0.049
	2	0.886	0.052	0.062		2	0.804	0.033	0.163
5	1	0.926	0.051	0.023	10	1	0.850	0.092	0.058
	2	0.860	0.119	0.020		2	0.645	0.133	0.222

classes, meaning that such answer has been much frequent in our sample for almost all ten items. However, units belonging to the first latent class ( $c = 1$ ) have larger conditional probabilities to answer “Yes” than units in the second ( $c = 2$ ), since we can observe  $\hat{\phi}_{j|1}(1) > \hat{\phi}_{j|2}(1)$ , for  $j = 1, \dots, 10$ , even if the difference between these probabilities is sometimes negligible (see items 4 and 5). On the contrary, units in the second latent class generally exhibit a lower probability of affirmative responses and higher for the second and third response categories ( $y = 2$  and  $y = 3$ ). Then, units in this group have a higher probability not to accomplish the activities listed in the form than units in the first class, mainly because such activities are not expected by the PTPC, since  $\hat{\phi}_{j|2}(3) > \hat{\phi}_{j|2}(2)$ . Overall, looking at Table 1, it can be stated that the two subpopulations especially differentiate as regards the outcomes of items 7 and 8, since units in the first latent class answer “Yes” to these two items with a high probability, while for those in the second latent class the most likely response is the third.

Finally, according to the posterior probabilities computed using equation (1), each unit is assigned to one of the two latent classes. Such classification is crossed with some variables characterising the sample units, in order to further qualify the latent classes. Among others, two variables show an interesting relation: the occurrence of a corruptive event and the population size. The former is directly obtained from the RPC form (question 2B), while the latter is obtained dividing the population in quartiles according to municipality resident population.

As reported in Table 2, even if only 33 municipalities over 213 state at least one corruptive event (18.3%), we can observe that most of them (69.7%) belongs to the first latent class, while, among units with no events, the allocation between the latent classes is close to 50%. As far as the municipality population size is concerned, a trend in the latent class partition is observable along the quartiles of the population. In particular, the first quartile of municipalities (hence the least populated) is mostly

**Table 2** Cross-tables between latent class membership and occurrence of corruptive events within the municipality (a) and size of municipality (b). Values in parenthesis represent row-percentages.

(a)				(b)			
Latent class				Latent class			
Events	1	2	Total	Size	1	2	Total
none	79 (43.9)	101 (56.1)	180	1	20 (34.5)	38 (65.5)	58
at least one	23 (69.7)	10 (30.3)	33	2	27 (46.6)	31 (53.4)	58
Total	102	111	213*	3	29 (50.0)	29 (50.0)	58
				4	33 (56.9)	25 (43.1)	58
				Total	109	123	232

\*: 19 units have missing response in the occurrence of corruptive events.

characterised by units belonging to the second latent class (65.5%), while the last quartile (the most populated) has more units belonging to the first (56.9%).

## 5 Conclusions

This work is the first attempt to extensively analyse the richness of information included in the annual relations filled in by the prevention-of-corruption supervisors. Our purpose is to study the behaviour of a sample of Italian municipalities in adopting measures to contrast corruption. A Latent Class analysis is performed in order to cluster municipalities into homogeneous groups, according to their behaviour as regards corruption prevention. Two groups of municipalities are highlighted of which the first collects the most virtuous municipalities.

Furthermore, two variables can be considered important in qualifying the two latent classes: the occurrence of corruptive events within the institution and the municipality population size. In particular, the most virtuous class is characterised by the most populated municipalities, which experienced at least one corruptive event.

## References

1. Lambsdorff, J.G.: *The institutional economics of corruption and reform: theory, evidence and policy*. Cambridge University Press (2007)
2. Lazarsfeld, P.F.: *The logical and mathematical foundation of latent structure analysis*. In: Stouffer, S.A., Suchman, E.A., Guttman, L. (eds.) *Measurement and prediction*. Princeton University Press, New York (1950)
3. Lazarsfeld, P.F., Henry, N.W.: *Latent structure analysis*. Houghton Mifflin, Boston (1968)
4. Sampford, C., Shacklock, A., Connors, C., Galtung, F.: *Measuring corruption*. Routledge, New York (2006)

# A proposal of a discretization method applicable to Rasch measures

## *Proposta di un metodo di discretizzazione applicabile alle misure ottenute tramite il modello di Rasch*

Silvia Golia

**Abstract** The aim of this paper is to propose a discretization method which can be applied to measures obtained using the Rasch model or, more in general, a model belonging to the class of the IRT models. The motivation of this proposal lies in the fact that there are methodologies that work with discretized variables, one such example is the Bayesian Networks. The idea is to use the informations from the Rasch model in order to forecast the answer of a subject to a representative item, and this answer represents the category assigned to the subject in the categorized version of his/her latent trait. In order to verify the goodness of this proposal, the new discretized variable is compared with a global single-item measure, under the hypothesis that this item is a possible observed discretization of the latent variable.

**Abstract** *Lo scopo di questo lavoro è quello di proporre un metodo di discretizzazione applicabile a misure ottenute utilizzando il modello di Rasch o, più in generale, modelli di tipo IRT. La ragione di questa necessità risiede nella constatazione che vi sono metodologie che richiedono variabili categorizzate; un esempio è dato dalle Reti Bayesiane. L'idea è di sfruttare le informazioni del modello di Rasch per prevedere la risposta di un soggetto ad un item rappresentativo, e tale risposta rappresenta la categoria assegnata al soggetto nella versione categorizzata del suo aspetto latente. Per verificare la bontà della proposta, la nuova variabile discretizzata viene confrontata con le risposte ad un item globale, ipotizzando che questo rappresenti una possibile discretizzazione osservata della variabile latente.*

**Key words:** Discretization, Rasch measure, Global single-item

---

Silvia Golia

University of Brescia, Department of Economics and Management, C.da S.Chiera, 50 - 25122 Brescia, Italy e-mail: [silvia.golia@unibs.it](mailto:silvia.golia@unibs.it)

## 1 Introduction

Many statistical learning algorithms require only categorical variables or produce better models if the variables involved in the analysis are not continuous. However, many real databases include continuous as well as categorical variables, so it is necessary to reduce these continuous variables to discretized ones. In socio-economic and psychological contexts it is common to take into account latent variables that can not be measured in the standard way, as, for example, weight or height. In order to measure a latent trait in general an ad hoc questionnaire is prepared and administered to a sample of the target population. One approach that allows to estimate a measure of the latent trait is the so called item response theory (IRT) approach. The IRT approach is based on the idea that the probability of response in any one of two or more mutually exclusive categories of an item is a function of the subjects location on the latent continuum representing the latent trait of interest and of some estimable parameters characteristic of the item. The problem addressed in this paper is to find a good method able to discretize continuous measures estimated applying a model that follows the IRT approach. The resulting discretized variable must mimic the evidence that higher scores correspond to higher levels of latent trait, so the ordering of the categories matters. The measurement model considered in this paper is the Rating Scale Model (RSM) ([1]) which belongs to the family of Rasch models. It converts raw scores into linear and reproducible measurement and its distinguishing characteristics are: separable person and item parameters, sufficient statistics for the parameters and conjoint additivity; prerequisites are unidimensionality and local independence. If the data fit the model, then the obtained measures are objective and expressed in logits. ([7]). Following the RSM, given an item  $i$  with  $m + 1$  response categories ( $c = 0, 1, \dots, m$ ), the probability of the subject  $s$  with level of latent trait  $\beta_s$  to respond in category  $c$  is given by:

$$P(X_{si} = c) = p_{sic} = \frac{\exp \left\{ c(\beta_s - \delta_i) - \sum_{j=0}^c \tau_j \right\}}{\sum_{k=0}^m \exp \left\{ k(\beta_s - \delta_i) - \sum_{j=0}^k \tau_j \right\}} \quad (1)$$

where  $\delta_i$  represents the difficulty of item  $i$  and  $\tau_j$  is called threshold ( $\tau_0 \equiv 0$  and  $\sum_{j=1}^m \tau_j = 0$ ).

The idea underlying this paper is to use the information obtained from the application of equation 1, that is the estimates of the  $\beta_s$  and  $\tau_k$  parameters, in order to forecast the answer of the subjects to a representative item. These answers represent the categories assigned to the subjects in the categorized version of their latent trait.

## 2 Discretization of a Rasch measure

In order to discretize a continuous variable so that the obtained categories are ordinal, common methods are the equal-width and the equal-frequency discretization.

After being sorted the  $n$  observations of the variable, the equal-width discretization algorithm consists in dividing the range of the variable  $x$ , represented by the interval  $[x_1, x_n]$ , into a  $k$  predefined number of equal width discrete intervals and assigning the level  $j$  to the subject  $i$  if and only if  $x_1 + \frac{(j-1)(x_n-x_1)}{k} < x_i \leq x_1 + \frac{j(x_n-x_1)}{k}$ . The equal-frequency discretization algorithm consists in dividing the range of the variable  $x$  according to a user-defined number of intervals,  $k$ , delimited by the  $1/k, 2/k, \dots, (k-1)/k$  empirical quantiles  $Q$  so that the subject  $i$  is assigned to the level  $j$  if and only if  $Q_{j/k} < x_i \leq Q_{(j+1)/k}$ .

The method proposed in this paper originates from the consideration that the above methods keep no trace of the way with which the measure has been obtained. The measure was estimated using joint maximum likelihood estimation under the constraint that the sum of the difficulty parameters  $\delta_i$  was set equal to zero. An item with difficulty equal to zero corresponds to an item with average difficulty, so this item seems to be a representative one, able to discriminate the subjects.

Given the estimated thresholds  $\{\tau_k\}_{k=1}^m$ , the estimated measure of the latent trait  $\{\hat{\beta}_s\}_{s=1}^n$  and  $\delta_i = 0$ , the equation 1 allows one to calculate the response probability record  $\{p_{sic}\}_{c=0}^m$  for each subject  $s$ . The discretized version of  $\hat{\beta}_s$  for the subject  $s$ ,  $b_s$ , is represented by the most probable response category, that is

$$b_s = \arg \max_c p_{sic}. \tag{2}$$

### 3 Evaluation of the goodness of a discretization method

In order to verify the goodness of the proposed discretization, the new, discretized variable has been compared with a Rasch single-item measure. Intuitively, the global single-item measure can be seen as an approximation of the discretized version of the latent trait measured by the set of items that are proposed for this purpose; for example, the global item for the job satisfaction is "How satisfied are you with your job as a whole?" and it is reasonable to admit that the respondent can implicitly make a synthesis of his/her job satisfaction when answers to this question. Moreover, there is literature that has explored the use of a global single-item as a measure of latent constructs (see for example [4]); depending on the nature of the construct operationalized, global single-item measure is often adequate for the purpose.

In order to evaluate how the discretized measure performs with respect to the global single-item measure, it is necessary to fix a metric able to quantify the resemblance between global single-item and discretized measures. Let  $O_s$  be the response to the global single-item given by the subject  $s$  and let  $D_s$  be the discretization of the estimated measure  $\hat{\beta}_s$  for the subject  $s$ . A first indicator of resemblance is the percentage of perfect match between the  $O_s$  and the  $D_s$ . Let  $S_s$  be the indicator of perfect match for the subject  $s$  so that  $S_s = 1$  if  $O_s = D_s$ ; the percentage of perfect match is given by  $(\sum_{s=1}^n S_s/n) * 100$ , where  $n$  indicates the sample size. This indicator does not take into account the ordering of the categories, so two other measures

of resemblance can be considered. The first one is the Mean Absolute Difference between the  $O_s$  and the  $D_s$ s, that is:  $\sum_{s=1}^n |O_s - D_s|/n$ , and it gives an idea of the mean distance between the two variables. The second one is the Similarity index proposed by Gower ([5]); it is a normalized indicator with values near 1 indicating high degree of similarity.

#### 4 First evidences from real data and discussion

This section reports the first evidences that the proposed method performs better than the two standard ones considered. Five psychological constructs were taken into account and analyzed. The data regarding worker satisfaction and distributive fairness come from the Survey on the Italian Social Cooperatives carried out in 2007 ([3]). The respondents were paid workers employed in Italian social cooperatives of type A and B. The burn-out data come from a survey held in 2009 concerning social workers working in Veneto (Northern Italy) ([2]). The data concerning the avoidant attachment come from a survey carried out between March and June, 2016, at three nursing homes located in Lombardia (Northern Italy). The respondents were auxiliary nurses. The data regarding life satisfaction come from the Opinions and Lifestyle Survey ([6]); the respondents were components of household aged 16 and over living in Great Britain. The data were collected between April and May, 2015.

Table 1 reports summary information about the estimates of the measures of the latent traits considered and the estimated threshold parameters  $\hat{\tau}_k$ .

**Table 1** Summary of the information derived from the application of RSM

Latent Trait	Number Subjects	$E(\hat{\beta})$ (std)	Skewness - Kurtosis of $\hat{\beta}$	Thresholds
Worker Satisfaction	3980	0.89 (1.42)	0.87 - 4.79	-1.71; -0.94; 0.11; 2.54
Distributive Fairness	3666	-0.68 (1.88)	0.13 - 3.90	-2.86; -2.09; -1.18; 2.03; 4.11
Burn-out	770	-1.55 (1.90)	0.52 - 5.14	-9.93; -0.73; 0.08; 4.57
Avoidant Attachment	107	-1.20 (1.37)	-0.94 - 3.52	-1.24; -0.52; -0.04; 1.80
Life Satisfaction	2042	1.47 (1.43)	0.54 - 4.02	-1.72; -1.30; -0.09; 3.11

Table 2 reports the values of the three measures of resemblance considered. The discretization method proposed in this paper outperforms the two standard ones with respect to perfect match as well as similarity between discretized and "observed" measures.

These first results suggest that the developed method has chances to give a good discretization of the underlying latent variable measured by a Rasch model. Further investigation is needed, exploring new datasets as well as other discretization methods, to confirm what shown in this paper.



**Table 2** Measures of resemblance between the global single-item and discretized measures

Latent Trait / Discretization Method	% Perfect Match	Mean Absolute Difference	Similarity Index
<b>Worker Satisfaction</b>			
Rasch Discretization	56.4	0.511	0.872
Equal-width Discretization	20.4	1.005	0.749
Equal-frequency Discretization	28.5	1.199	0.700
<b>Distributive Fairness</b>			
Rasch Discretization	60.5	0.452	0.910
Equal-width Discretization	44.1	0.622	0.876
Equal-frequency Discretization	37.1	0.783	0.843
<b>Burn-out</b>			
Rasch Discretization	60.6	0.440	0.890
Equal-width Discretization	53.1	0.495	0.876
Equal-frequency Discretization	44.9	0.697	0.826
<b>Avoidant Attachment</b>			
Rasch Discretization	56.1	0.579	0.807
Equal-width Discretization	16.8	1.383	0.654
Equal-frequency Discretization	29.0	1.009	0.748
<b>Life Satisfaction</b>			
Rasch Discretization	70.5	0.315	0.921
Equal-width Discretization	24.0	0.883	0.779
Equal-frequency Discretization	26.4	1.207	0.698

## References

1. Andrich, D.: A rating formulation for ordered response categories. *Psychometrika* **43**, 561–573 (1978)
2. Bressan F., Pedrazza M., Neve E. (eds): *Il percorso formativo dell'assistente sociale. Autovalutazione e benessere professionale*. Franco Angeli, Milano (1992)
3. Carpita, M., Golia, S.: Measuring the quality of work: the case of the Italian social cooperatives. *Qual. Quant.* **46**, 16591685 (2012)
4. Fuchs, C., Diamantopoulos, A.: Using single-item measures for construct measurement in management research. *Die Betriebswirtschaft* **69(2)**, 195–210 (2009)
5. Gower, J.C.: A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857–874 (1971)
6. Office for National Statistics, University of Manchester. Cathie Marsh Institute for Social Research (CMIST). UK Data Service. SN: 7913 . Opinions and Lifestyle Survey, Well-Being Module, April-May 2015: Unrestricted Access Teaching Dataset. 2nd Edition (2016) <http://doi.org/10.5255/UKDA-SN-7913-2>
7. Wright, B.D., Master, G.N.: *Rating scale analysis*, MESA Press, Chicago (1982)



# Tree-based Non-linear Graphical Models

## *Modelli Grafici non lineari basati su alberi*

Anna Gottard

**Abstract** Graphical models are statistical models that are associated to graphs whose nodes represent variables of interest. The absence of an edge between two nodes corresponds to a conditional independence between the variables. In this work, I propose a class of graphical models for non-linear systems, where the shape of dependence is modelled by a Bayesian additive regression tree model. The proposed models are able to detect nonparametrically both non-linearities and interactions and are suitable for high dimensional data.

**Abstract** *I modelli grafici sono modelli statistici associati a grafi i cui nodi rappresentano le variabili di interesse. L'assenza di connessione tra due nodi implica una indipendenza condizionata tra le rispettive variabili. In questo lavoro, propongo una classe di modelli grafici per sistemi non lineari in cui la forma della dipendenza è dettata da un modello bayesiano additivo di alberi di regressione. Questi modelli consentono di cogliere in modo non parametrico sia non linearità che interazioni tra le variabili e sono adeguati anche per dati a grandi dimensioni.*

**Key words:** Graphical models, graph learning, non-linear systems, Bayesian additive regression trees.

## 1 Introduction

Graphical models (see [9] and [5], among others) have been widely utilised in many domains to study the conditional independence structure of a set of random variables. Random variables could concern, for instance, expression values of genes, metabolites, personal opinions on specific topics and so on.

In the graph, random variables are represented as nodes and conditional independence as missing edges. Graphs are characterised by the type of their edges as

---

Anna Gottard

DiSIA, University of Florence, V.le Morgagni 59, Firenze e-mail: gottard@disia.unifi.it

undirected, directed and mixed. Moreover, they are characterised by the type of variables, as continuous, discrete and mixed.

Most of the literature on learning the structure of the graph from data focused on concentration graph models, undirected graph models with random variables following a multivariate normal distribution. This assumption implies that the relationship among variables is linear and learning the graph corresponds to assess which partial correlation coefficient is zero.

The challenge of learning graphs when the relationship among variables is not linear has been tackled recently by many authors. In particular, [11, 12] propose graphical models for Nonparanormal random variables, a semiparametric Gaussian copula. In the case of directed acyclic graphs, [15] the use of non-linear structural equation models. Further interesting work connects graphs with regression trees (see [13] and [6]). Tree-based models are very attractive as they can model both non-linearities and interactions, but special caution has to be paid on the algorithm to detect the variable importance. As a matter of fact, a greedy search can sometimes bring to misleading results. A cumbersome situation is given, for instance, when the *true* graph is the one presented in Figure 1, for particular values of the parameters.

In this work, I consider the problem of Bayesian estimation and learning an undirected graph for continuous or mixed variables in the presence of non-linearities and interactions. As [13] and [6], my proposal is based on a tree ensemble model, Bayesian Additive Regression Trees (BART) models [4, 2]. The model is nonparametric and specifies the expected value as the sum of trees. The Bayesian Backfitting Monte Carlo Markov Chain procedure [7] for estimation avoids the necessity of greedy search algorithms. This approach provides full posterior inference, including credible intervals for model parameters. The structure learning of the undirected graph adjacency matrix is achieved by node-wise regression and the local Markov property. The graph learned according to this procedure is sometimes called *dependency network* [8].

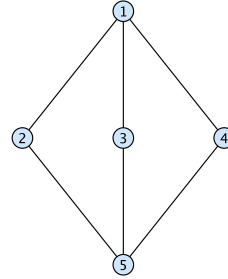
## 2 The proposal

Consider the collection of random variables  $X_v$ ,  $v \in V = \{1, \dots, p\}$  whose conditional independence structure is depicted in a graph  $\mathcal{G} = (V, E)$ . The set  $V$  collects the nodes of the graph and the set  $E$  collects the edges connecting the nodes.

The graph is ruled by a set of Markov properties. The local Markov property for undirected graphs assesses that each variable  $X_v$  in the graph is conditionally independent of all the other nodes given its neighbours. The set of *neighbours* of a node  $X_v$ , denoted with  $ne(v)$ , gathers all the variables whose nodes have an edge connecting them to  $v$ . The set containing  $v$  and its neighbours is called *closure* and denoted by  $cl(v)$ . Then, the local Markov property for undirected graphs can be written as

$$X_v \perp\!\!\!\perp X_{V \setminus cl(v)} \mid X_{ne(v)} \quad \forall v \in V.$$

**Fig. 1** Example of undirected graph with five nodes.



In the graph in Figure 1, for instance, according to the local Markov property  $X_2 \perp\!\!\!\perp (X_3, X_4) \mid (X_1, X_5)$ , as the neighbours of  $X_2$  are  $X_1, X_5$  and  $X_3, X_4$  are the only nodes not in the closure of  $X_2$ .

Learning the conditional independence structure of an undirected graphical model consists of learning the  $p$  by  $p$  adjacency matrix  $\mathcal{A}$ , with entry  $\mathcal{A}_{ij} = 1$  when  $X_i$  and  $X_j$  are neighbours and  $\mathcal{A}_{ij} = 0$  otherwise. A node-wise graph selection procedure consists of recovering the adjacency matrix line-by-line. Hence, learning the structure of the graph involves  $p$  separate steps, each one concerning the conditional distribution of a single variable given the others. The idea behind the neighbourhood selection goes back to [1]. See also [14], among others. When one can link the parameters of each conditional distribution with some of those of the other conditional distributions a better approach is based on the pseudo-likelihood function. See, for example, [10].

Now, suppose that the collection of random variables  $\mathbf{X}_V = \{X_1, \dots, X_p\}$  follows an unknown joint distribution  $F_V(\mathbf{X}; \Theta)$ , with finite first moment. Assume that dependence occurs through the conditional expectation, meaning that  $X_v \perp\!\!\!\perp X_j \mid \mathbf{X}_{-v,-j}$  iff

$$F_v(X_v \mid \mathbf{X}_{-v}) = F_v(X_v \mid \mathbf{X}_{-v,-j}) \iff E[X_v \mid \mathbf{X}_{-v}] = E[X_v \mid \mathbf{X}_{-v,-j}],$$

with  $\mathbf{X}_{-v,-j} = \mathbf{X}_V \setminus \{X_v, X_j\}$ . Consequently

$$X_v \perp\!\!\!\perp \mathbf{X}_{V \setminus cl(v)} \mid \mathbf{X}_{ne(v)} \iff E[X_v \mid \mathbf{X}_{V \setminus cl(v)}, \mathbf{X}_{ne(v)}] = E[X_v \mid \mathbf{X}_{ne(v)}].$$

For each node  $v \in V$ , the dependence structure of  $X_v$  on its neighbours is described by a Bayesian Additive Regression Trees (BART) [4], as follows

$$X_v = \sum_{j=1}^m \mathcal{T}_j^v(X_{V \setminus v}; T_j^v, M_j^v) + \varepsilon_v, \tag{1}$$

where  $\varepsilon_v$  has a certain distribution, for instance  $N(0, \sigma_v^2)$ . Here  $T_j^v$  is the structure of a  $j^{th}$  tree. Moreover,  $M_j^v$  is the subset of  $\Theta^v$  containing the parameters for the

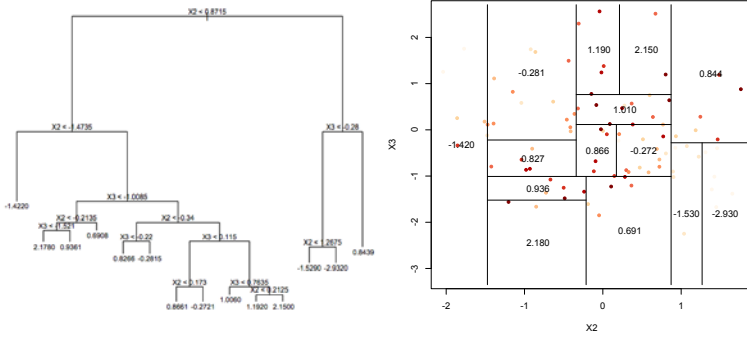


Fig. 2 Example of a regression tree with two variables and the corresponding partition.

tree  $\mathcal{T}_j^v$ , i.e. the mean vector over its terminal nodes, when regressing  $X_v$ . In the following, the suffix  $v$  will be omitted when not necessary.

Let  $\mathcal{X}_{V \setminus v} = \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_{v-1} \times \mathcal{X}_{v+1} \times \dots \times \mathcal{X}_p$  be the product state space of the variables in  $\mathbf{X}_{-v}$  and  $\mathcal{R}$  the space of all binary partitions  $R$  of  $\mathcal{X}_{V \setminus v}$ . A partition is a binary partition if it can be obtained by sequentially dividing  $\mathcal{X}$  into two parts by splitting only one component  $\mathcal{X}_j$ ,  $j \in V \setminus v$ . Each structure  $T_j$  detects exactly one partition, say  $R^j = \{R_1, \dots, R_{d_j}\} \in \mathcal{R}$ . Then  $M_j$  has length  $d_j$ . The tree  $\mathcal{T}_j$  in (3) can be written as

$$\mathcal{T}_j(\mathbf{X}_{-v}; T_j, M_j) = \sum_{m=1}^{d_j} M_{jm} I_{\{x \in R_m\}}. \quad (2)$$

Figure 2 provides an example of tree and its corresponding partition.

The Bayesian Backfitting Monte Carlo Markov Chain can be implemented to draw samples of  $T_j$  as suggested in [4]. This avoids the greedy search of the partition. Regularization prior distributions can be used for high dimensional settings. The importance of the variables can be computed via permutation inference as suggested by [2]. The inclusion of an additional linear component in (3) sets these models in the class of quasi-linear systems [16]

$$X_v = \beta \mathbf{X}_{-v} + \sum_{j=1}^m \mathcal{T}_j^v(\mathbf{X}_{-v}; T_j^v, M_j^v) + \varepsilon_v. \quad (3)$$

The Bayesian Backfitting Monte Carlo Markov Chain can be implemented also in this case. By the addition of a linear component, the depth of the trees can be reduced, diminishing the risk of overfitting.

### 3 Conclusions

For many high dimensional problems, the assumption of joint gaussianity to study the association structure of the variables of interest may be inadequate. An interesting aspect of the approach I am proposing is that it can handle nonlinearities, interactions and also the case of mixed-type data. BART models are a tree-based method able to produce a proper inference and credible intervals. This aspect makes them more attractive than ordinary tree-based models and random forests.

### References

1. Besag, J.: Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **24**(3), 179–195 (1975)
2. Bleich, J., Kapelner, A., George, E.I., Jensen, S.T.: Variable selection for BART: an application to gene regulation. *The Annals of Applied Statistics*, **8**(3), 1750–1781 (2014)
3. Chen, S., Witten, D. M.: Selection and estimation for mixed graphical models. *Biometrika*, **102**(1), 47–64 (2015)
4. Chipman H.A., George E.I., McCulloch R.E.: BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **4**(1), 266–298 (2010)
5. Cox, D.R., Wermuth, N.: *Multivariate dependencies: Models, analysis and interpretation*. CRC Press (1996)
6. Fellinghauer, B., Bhlmann, P., Ryffel, M., Von Rhein, M., Reinhardt, J.D.: Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis* **64**, 132–152 (2013)
7. Hastie, T., Tibshirani, R.: Bayesian backfitting (with comments and a rejoinder by the authors). *Statistical Science*, **15**(3), 196–223 (2000)
8. Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., Kadie, C.: Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, **1**, 49–75 (2000)
9. Lauritzen, S.L.: *Graphical models*. Clarendon Press (1996)
10. Leppä-aho, J., Pensar, J., Roos, T., Corander, J.: Learning Gaussian graphical models with fractional marginal pseudo-likelihood. *International Journal of Approximate Reasoning*, **83**, 21–42 (2017)
11. Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L.: High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics* **40**(4), 2293–2326 (2012)
12. Liu, H., Lafferty, J., Wasserman, L.: The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10**, 2295–2328 (2009)
13. Liu, Y., Zayas-Castro, J.L., Fabri, P., Huang, S.: Learning high-dimensional networks with nonlinear interactions by a novel tree-embedded graphical model. *Pattern Recognition Letters* **49**, 207–213 (2014)
14. Meinshausen N., & Bühlmann P.: High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**(3), 1436–1462 (2006)
15. Peters, J., Mooij, J. M., Janzing, D., Schölkopf, B.: Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, **15**(1), 2009–2053 (2014)
16. Wermuth, N., Cox, D.R.: On association models defined over independence graphs. *Bernoulli*, **4**(4), 477–495 (1998)





# Sentiment Analysis for micro-blogging using LSTM Recurrent Neural Networks

## *Sentiment Analysis per il micro-blogging utilizzando LSTM Recurrent Neural Networks*

Sara Hbali, Youssef Hbali, Mohamed Sadgal and Abdelaziz El Fazziki

**Abstract** In this paper we study a novel method that consider multiple information. Not only textual properties, but also visual entries. By combining these information such proposed method offers more information to feed the model for sentiment analysis. Most of the papers used only textual properties for sentimental analysis, whilst our contribution is to the add the visual property to achieve better movie review classification.

**Abstract** *In questo documento, studiamo una nuova metodologia che prende in considerazione diversi tipologie di informazioni. Non solamente proprietà testuali, ma anche input visivi. Combinando queste informazioni, tale metodo offre un maggior numero d'informazioni per alimentare il modello per l'analisi del sentimento. La maggior parte degli studi precedenti usano solo proprietà di tipo testuale per l'analisi sentimentale, mentre il nostro contributo consiste nell'aggiungere proprietà visive in modo tale da ottenere una migliore classificazione di recensione del film.*

**Key words:** Sentimental analysis, features extraction, LSTM, CNN

## 1 Introduction

The principal of micro-blogging is based on users opinion, feedback and experiences about any chosen topic, the difference between regular blogs and micro is the content size, for instance twitter is one of the most popular social networking and communication platforms used to exchange and expressing ideas. However one tweet is only 140 characters with the possibility to add image or video, thus all previous works concentrate only on the textual con-

---

Sara Hbali e-mail: hbali.sara@gmail.com · Youssef Hbali · Mohamed Sadgal · Abdelaziz El Fazziki

Cadi Ayyad University. B.P. 2390, Avenue Prince My Abdellah, Marrakech, Morocco

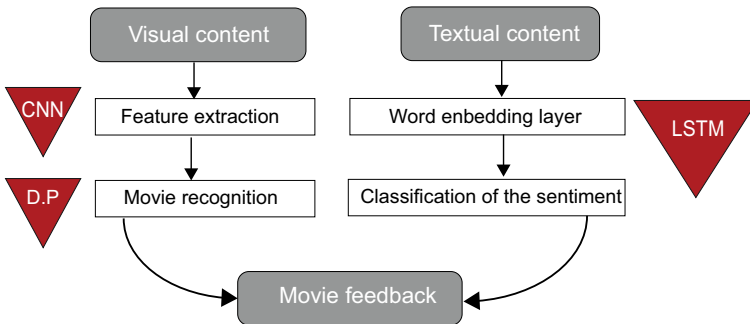
tent, which leaves the visual information inexploitable and which could offers more detailed information to use in sentiment analysis.

The sentiment analysis, also known as opinion mining, is the process to detect whether the characteristics of tweets tends towards neutral, positive or negative. This research area exists before twitter and micro-blogging, since it can be applied to different domains, in Refs [9] that aims to retrieve little know communities but yet are relevant or [8] where they focus more on public opinion and news article to apply on the stock markets. In Refs [7] focus on tweets to evaluate the impact on votes for political candidates.

In order to demonstrate the effectiveness of the method including and visual content, we will apply our approach to movies review classification.

## 2 General approach

The proposed model aims to process a large movie reviews for classification, therefor we consider that each movie review is a entry and one variable sequence of words and the sentiment of each movie review must be classified.



**Fig. 1** Complete model for movie classification.

### Movie review classification

In these part of the paper we used model proposed in [5, 3, 4, 1, 2] to implements standard LSTM model with variant modification The equations below describe how a layer of memory cells is updated at every timestep  $t$  in these equations

- $x_t$  is the input to memory cell layer at time  $t$ .
- $W_i, W_f, W_c, W_o, U_i, U_f, U_c, U_o$  and  $V_o$  are weight matrices.
- $b_i, b_f$  and  $b_o$  are bias vectors

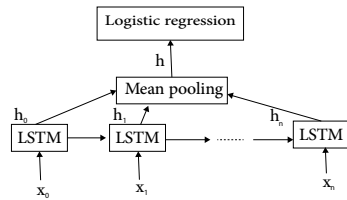
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (1)$$

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (2)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (3)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (4)$$

**Fig. 2** Illustration of the model used in this paper. It is composed of single LSTM layer followed by mean pooling over time and logistic regression.



Equations (1), (2), (3) and (4) are performed in parallel to make the computation more efficient. This is possible because none of these equations rely on a result produced by the other ones. It is achieved by concatenating the four matrices  $W_*$  into a single weight matrix  $W$  and performing the same concatenation on the weight matrices  $U_*$  to produce the matrix  $U$  and the bias vectors  $b_*$  to produce the vector  $b$ . Then, the pre-nonlinearity activations can be computed with :

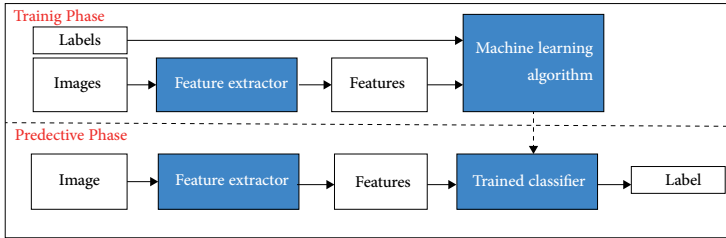
$$z = Wx_t + Uh_{t-1} + b \quad (5)$$

The result is then sliced to obtain the pre-nonlinearity activations for  $i$ ,  $f$ ,  $\tilde{C}_t$ , and  $o$  and the non-linearities are then applied independently for each.

## Image processing

The entry step : analyzing attached image to identify the movie category In this step we extract features using CNN layer as explain in .. When we train a deep neural network in Caffe [6] to classify images, we specify a multilayered neural network with different types of layers like convolution, softmax loss and so on.

The last layer is the output layer that gives us the output tag with the corresponding confidence value. In this model we used various layers for features vectors of the input image.



**Fig. 3** Illustration of the CNN model used in this paper.

To match between our training data and the input image, we use dot product of two features vectors :

$$\text{dot}(a, b)[i, j, k, m] = \sum (a[i, j, :] * b[k, :, m]) \quad (6)$$

The result of the Dot product (6) of  $a$  and  $b$  is equivalent to matrix multiplication of the two vectors of training phase and predictive phase

### 2.0.1 Experiment

In this section we first introduced the dataset used in our experience and the implementation details and finally discuss results.

### 2.0.2 Dataset

Usually deep learning architectures need large training dataset. For this purpose, we used IMDB dataset for movie reviews and images collected from IMDB website, containing more than 25 000 polar movie reviews for training and 25 000 for testing and more than 400 000 images (celebrities profiles and and scene from movies..) used in [11, 10] for apparent age from single images.

### 2.0.3 Implementation details

In order to demonstrate the effectiveness of the method, we apply already pre-trained models on our example for both textual and visual content.

The entry data for our model is the image, we can predict multiple information as explained in figure 4 which can give us hint about the textual content. such as the category of the movie or actors.



**Fig. 4** Extracted labels using visual content.

**Table 1** Achieved results of our the LSTM model

	Phase 1	Phase 2	Phase 3
<b>Loss</b>	0.5570	0.3530	0.2559
<b>Accuracy</b>	0.7149	0.8577	0.9019
<b>Training time</b>	107s	107s	107s

Training our model using IMDB dataset provides a very discriminative sentiment analysis as shown in table 1. We challenged our model with complete additional information, the image recognition model achieves a detection rate of 89% using the features learned from a convolutional neural network model.

## 2.0.4 Conclusion

We propose optimization of sentiment analysis problem for IMDB movie reviews such that each review is labeled with either positive or negative sentiment. then we combine review sentiment analysis with image analysis to give a complete overview of the requested content.

One of the weaknesses of the proposed model that is it requires time and memory to have effective results.

For future work, we can try various deep learning models for classification and explore other information such as location or video.

## References

1. Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio.

- Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*, 2012.
2. James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: A cpu and gpu math compiler in python. In *Proc. 9th Python in Science Conf*, pages 1–7, 2010.
  3. Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.
  4. Alex Graves. Supervised sequence labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*, pages 5–13. Springer, 2012.
  5. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
  6. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
  7. Sanne Kruike-meier. How political candidates use twitter and the impact on votes. *Computers in Human Behavior*, 34:131–139, 2014.
  8. Qing Li, TieJun Wang, Ping Li, Ling Liu, Qixu Gong, and Yuanzhu Chen. The effect of news and public mood on stock movements. *Information Sciences*, 278:826–840, 2014.
  9. Denis Parra, Christoph Trattner, Diego Gómez, Matías Hurtado, Xidao Wen, and Yu-Ru Lin. Twitter in academic events: a study of temporal usage, communication, sentimental and topical patterns in 16 computer science conferences. *Computer Communications*, 73:301–314, 2016.
  10. Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.
  11. Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*, July 2016.

# How to Exploit Big Data from Social Networks: a Subjective Well-being Indicator via Twitter

Stefano Maria Iacus, Giuseppe Porro, Silvia Salini and Elena Siletti

**Abstract** In our research we apply a new technique of opinion analysis over Twitter data to propose a new indicator of perceived and subjective well-being: The SWBI examines many dimension of individual and social life. In the purpose to investigate whether SWBI and its single components may adequately represent the reaction of a community to changes in everyday life conditions, we propose a comparative analysis, among the Italian provinces, of perceived well-being, measured with SWBI, with objective well-being, measured with the *Il Sole 24 Ore* QoL Index. The idea is to create a composite well-being indicator which mixes stable official statistics and fluctuating social media data.

**Abstract** *Nella nostra ricerca applichiamo una nuova tecnica di analisi dei dati provenienti da Twitter per proporre un nuovo indicatore di benessere percepito e soggettivo: L'SWBI considera molte dimensioni della vita individuale e sociale. Per indagare se l'SWBI e i suoi singoli componenti possano rappresentare in modo adeguato la reazione di una comunità ai cambiamenti delle condizioni di vita di tutti i giorni, proponiamo un'analisi comparativa, tra le province italiane, del benessere percepito, misurato con l'SWBI, e del benessere oggettivo, misurato con l'indice della qualità della vita de il Sole 24 Ore. L'idea è di creare un indicatore composito di benessere che integri le statistiche ufficiali e i dati provenienti dai social media.*

**Key words:** well-being, social indicators, big data, social networks, sentiment analysis

---

Stefano Maria Iacus, Silvia Salini, Elena Siletti  
Department of Economics, Management and Quantitative Methods, University of Milan

e-mail: stefano.iacus, silvia.salini, elena.siletti@unimi.it

Giuseppe Porro

Department of Law, Economics and Culture, University of Insubria

e-mail: giuseppe.porro@uninsubria.it

## 1 Introduction: Theoretical Frameworks

In the last decades scholars have become increasingly interested in new measures of quality of life. A milestone in 2009, when the so-called Stiglitz Commission proposed to build a system of objective and subjective indicators, with a strong influence in further studies: different indicators, with different structures, considering a great variety of dimensions and for many purposes are now considered. For subjective indicators, self-reports have been extensively used, forgetting that they are often misleading (9) and despite the efforts made it remains much uncertainty using them (6). The two main limitations: the influence that a single question can have, and the limited frequency of the surveys, that may fail in capturing the trend changes and in distinguishing between the short-run “emotional” and the structural component (“life evaluation” or “life satisfaction”).

Social networks offers a new rich source of information, which is available without any survey, they simply allow to listen to. They host an open, enormous amount of data that allow to study social dynamics from an unseen perspective. Analysing them allows to listen to what people say: with well-being this means to be able to measure feelings in real-time, mapping its fluctuation (5). In the last years researchers have used these data for a wide range of applications including monitoring influenza and other health outbreaks, predicting the stock market, and understanding sentiment about products or people. There exists a wide set of works aiming at tracking happiness through Twitter, for the Italian provinces, (5) propose the iHappy index, that is measured with an innovative statistical techniques on millions of tweets.

Social media data enable to collect longitudinal data and to measure phenomena more frequently. Skeptics have questioned whether enthusiasts’ claims are overly optimistic (4), and whether any form of non-probability sampling as this new analysis is too risky (1). Others noted that media data may introduce new kind of bias (2), which raises the question of whether they are sufficiently reliable. We need to understand, to solve the new challenges: we can not ignore this new and rich source of information. While big data are unlikely to replace high quality surveys, they could be useful when there are not. The two methods can serve complementary functions.

Sentiment analysis is the core aspect, despite many limitations (4), if correctly performed, it seems to be a useful framework to exploit when the constraints of standard survey methodology may be too strong (8). On one hand there are no questions to pose, all that the analyst has to do is to listen to and classify the opinions expressed accordingly; on the other hand, the available information is updated in real time and hence the frequency can be as high as desired, allowing for separating the volatile/emotional component from the permanent/structural one.

With the SWBI (Social Well-being index) we make a new proposal, relying on Twitter data and on one of the most recent techniques for sentiment



analysis. This approach disentangles the main methodological issues raised in the literature on well-being measurement, and produces a set of indicators that span the wide range of well-being perceptions.

## 2 The SWBI

The SWBI is a multidimensional indicator derived from a new human supervised technique (iSA-Integrated Sentiment Analysis (3)) designed to capture several aspects. In iSA algorithm the human part is essential because information can be retrieved from texts without relying on dictionaries of special semantic rules. Human just read a text and associate a topic ( $D = \text{“satisfied at work”}$ ) to it. Then, the computer learn the association between the whole set of words used in a text to express that opinion and extends the same rule.

Formally, let us denote by  $\mathcal{D} = \{D_0, D_2, \dots, D_M\}$  the set of possible categories (i.e. opinions). The target of interest is  $\{P(D), D \in \mathcal{D}\}$ , i.e. the distribution of opinions in a corpus of  $N$  texts.  $D_0$  refers to Off-topic or not relevant texts (i.e. *noises*). Let  $S_i$ ,  $i = 1, \dots, K$ , be a vector of  $L$  possible stems which identifies one of the texts in a corpus. More than one text in the corpus can be represented by the same  $S_i$  and is such that each element is equal to 1 if that stem is contained in a text, or 0 in absence. Formalized data set is  $\{(s_j, d_j), j = 1, \dots, N\}$  where  $s_j \in \mathcal{S}$  (the space of possible vectors  $S_j$ ) and  $d_j$  can either be “NA” or one of the hand coded categories  $D \in \mathcal{D}$ .

The “traditional” approach includes machine learning methods and statistical models; predict the outcome of  $\hat{d}_j = D$  for the texts with  $S = s_j$  belonging to the test set; when all data have been imputed, estimated categories  $\hat{d}_j$  are aggregated to obtain an estimate of  $\hat{P}(D)$ . We can write

$$P(D) = \underset{M \times 1}{P(D|S)} \underset{M \times K}{P(S)} \underset{K \times 1}{P(S)} \quad (1)$$

where  $P(D|S)$  is a  $M \times K$  matrix of conditional probabilities, and  $P(S)$  is a vector with the distribution of  $S_i$  over the corpus. Once  $P(D|S)$  is estimated from the training set with, say,  $\hat{P}(D|S)$ , then for each document in the test set with stem vector  $s_j$ , the opinion  $\hat{d}_j$  is estimated with the simple Bayes estimator as the maximizer of the conditional probability, i.e.  $\hat{d}_j = \arg \max_{D \in \mathcal{D}} \hat{P}(D|S = s_j)$ . This approach does not work if  $P(D_0)$  is very large compared to the rest of the  $D_i$ 's. iSA follow the idea by (7) of changing the point of view but goes one step further in terms of computational efficiency and variance reduction. Instead of (1), one can consider this new equation

$$P(S) = \underset{K \times 1}{P(S|D)} \underset{K \times M}{P(D)} \underset{M \times 1}{P(D)} \quad (2)$$

where now  $P(S|D)$  is a matrix whose elements  $P(S = S_k|D = D_i)$  represent the frequency of a particular stem  $S_k$  given the set of texts which actually express the opinion  $D = D_i$ . The solution of the problem is

$$\text{(inverse problem)} \quad \underset{M \times 1}{P(D)} = [\underset{M \times M}{P(S|D)^T P(S|D)}]^{-1} \underset{M \times K}{P(S|D)^T} \underset{K \times 1}{P(S)} \quad (3)$$

Equation (3) is such that the direct estimation of the distribution of opinion  $P(D)$  is obtained but individual classification is no longer possible. In fact, this is not a limitation as the accuracy of (3) with respect to (1) is vastly better. Moreover, researchers are comprehensibly more interested in the aggregate distribution of opinions than in the estimation of individual opinion (3).

To define SWBI, we inspired by NEF (New Economic Foundation) and their Happy Planet Index. It has eight dimensions concerning three different well-being areas. Each component is defined through the hypothetical question one might find: no questions, the sentiment is extracted from the text. Here the components: **Personal well-being**: *emotional well-being-(emo)*, *satisfying life-(sat)*, *vitality-(vit)*, *resilience and self-esteem-(res)*, *positive functioning-(fun)*; **Social well-being**: *trust and belonging-(tru)*, *relationships-(rel)*; **Well-being at work**: *quality of job-(wor)*.

Each tweet has been classified according to the scale -1, 0, 1, where -1 is for negative, 0 is neutral and 1 is positive feeling. To enhance the action of human supervision, additional rules have been introduced:

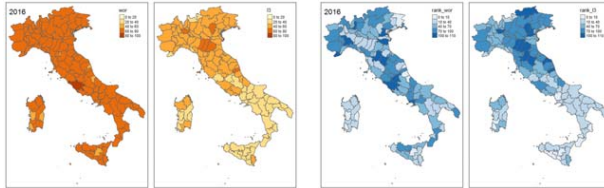
- Each tweet can be classified along one or more dimensions;
- Only self-expressed or individual expression of well-being or own views of the tweeter are considered;
- Re-tweet are considered, because the tweeters share the same view;
- Off-Topic texts are marked appropriately;
- If the encoders are not fully convinced about the semantic context they do not classify the text, just skip it and classify another one.

Our data source are tweets written in Italian language from Italy, accessed through Twitter's public API. Around 1 to 5% each day tweet contain geo-reference information which allows to build indicators at province level. From February 2012 we have stored and analysed more than 180 millions of tweets.

### 3 The SWBI and the *Il Sole 24 Ore* QoL Index in the Italian Provinces

Since 1990, the Italian business newspaper *Il Sole 24 Ore* publishes an index of the quality of life (QoL) for all the Italian provinces. Since 2016, the composite indicator has six components based on a simple arithmetic mean of 42 normalized indicators. To analyse its components according to

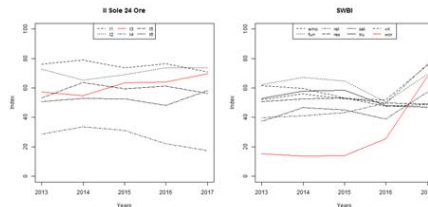
the SWBI, we rescaled from 0 to 100. Here the components: I1-Income, Savings, Consumption; I2-Environment, Services, Welfare; I3-Business, Work, Innovation; I4-Justice, Security, Crime; I5-Demographics, Family, Integration; I6-Culture, Leisure, Participation. As one can see, the *Il Sole 24 Ore* QoL



**Fig. 1** All the Figure refer to 2016, with red shades the original index, with blue shades the ranking of the Italian provinces

index cover only material quality of life and, for this reason, has become a benchmark indicator for objective well-being. Despite efforts to improve the quality, the index, in addition to having a low frequency with only an annual data, often shows delayed information. This is a serious flaw when decision-makers want to base their choices on such information. As we noticed, SWBI has the twofold advantage to be a high frequency instrument, which can be updated almost in real time. On the other hand, SWBI is an index of subjective well-being, and the differences between the two dimensions (objective and subjective) clearly emerge from the comparison of the two indicator.

As an example, we compare the SWBI component on well-being at work (*wor*) to the I3 (Business, Work and Innovation) component of *Il Sole 24 Ore* QoL index, where the quality of work and labour market is evaluated by objective quantities (total employment rate, exports in % of GDP, number of innovative start-ups per 1000 enterprises, number of registered enterprises per 100 inhabitants, loans on deposits ratio, patent applications per 1000 inhabitants, rate of youth unemployment 15-24 years). Clearly the informa-



**Fig. 2** SWBI and *Il Sole 24 Ore* Index Components in Milan, in red lines respectively, the I3 and *wor* component

tion conveyed by the two indicators is not the same. First of all (see Fig. 1,

left panels) shows a strong polarization: Northern and Central Italy have I3 values significantly higher compared to the Southern provinces; (*wor*), on the other side, is more stable across provinces and does not show appreciable concentration phenomena. The evidence is confirmed by the ranking of provinces according to (*wor*) and I3 values, respectively (see Fig. 1, right panels).

Moreover, even if we polish out the volatility of (*wor*) due to its high frequency and compare the annual average values of (*wor*) and I3, different trends must be pointed out. Let us examine, for example, the indicators for the city of Milan since 2013 (see Fig.2): while I3 shows a slightly increasing trend, (*wor*) exhibits a remarkable increase starting from 2015, and the same behaviour is shown by almost all the SWBI components since 2016. Maybe that the feeling of a recovery of the economic conditions and an improved confidence in personal and collective future have an impact on perceived well-being even beyond the possibility to observe these improvements in current, traditional and objective economic indicators.

## References

- [1] Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J., Tourangeau, R.: Summary report of the aapor task force on non-probability sampling. *Journal of Survey Statistics and Methodology* **1**(2), 90 (2013)
- [2] Biemer, P.P.: Total survey error: Design, implementation, and evaluation. *The Public Opinion Quarterly* **74**(5), 817–848 (2010)
- [3] Ceron, A., Curini, L., Iacus, S.M.: isa: a fast, scalable and accurate algorithm for sentiment analysis of social media content. submitted pp. 1–30 (2015)
- [4] Couper, M.P.: Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods* **7**(3), 145–156 (2013)
- [5] Curini, L., Iacus, S., Canova, L.: Measuring idiosyncratic happiness through the analysis of twitter: An application to the italian case. *Social Indicators Research* **121**(2), 525–542 (2015)
- [6] Feddersen, J., Metcalfe, R., Wooden, M.: Subjective wellbeing: why weather matters. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **179**(1), 203–228 (2016)
- [7] Hopkins, D., King, G.: A method of automated nonparametric content analysis for social science. *American Journal of Political Science* **54**(1), 229–247 (2010)
- [8] King, G.: Preface: Big data is not about the data In: R.M. Alvarez (ed.) *Computational Social Science: Discovery and Prediction*, chap. 1, pp. 1–10. Cambridge University Press, Cambridge (In Press)
- [9] Schwarz, N.: Self-reports: how the questions shape the answers. *American psychologist* **54**(2), 93–105 (1999)

# Network Analysis of Comorbidity Patterns in Heart Failure Patients using Administrative Data

## *Analisi delle Reti di co-patologie in pazienti affetti da Scompenso Cardiaco*

Francesca Ieva

**Abstract** In this work, we investigate the pattern of comorbidities in patients affected by Heart Failure (HF) through network analysis. Specifically, the pathologies are kept as the nodes of the network, while the links represent connections between two of them (a link is present if a patient is affected by both the pathologies in his/her last hospitalization). Thus, we study the comorbidity pattern of HF patients hospitalized in Lombardy between 2006 and 2012 using administrative data. We also applied techniques of community detection in order to detect groups of diseases which are more strongly connected. The application of network analysis to such data enables a new perspective in the study of heart failure disease.

**Abstract** *In questo lavoro, studiamo i pattern di comorbidità nei pazienti affetti da scompenso cardiaco attraverso un'analisi di rete (network analysis). Più precisamente, le patologie sono considerate come i nodi della rete, mentre i link rappresentano le connessioni tra due di loro (un link è presente se un paziente è affetto da entrambe le patologie durante la sua ultima ospedalizzazione). Studiamo quindi i pattern di comorbidità dei pazienti affetti da scompenso cardiaco che sono stati ricoverati in Lombardia tra il 2006 e il 2012 utilizzando dati amministrativi. Abbiamo inoltre applicato le tecniche di community detection al fine di rilevare gruppi di malattie che sono più fortemente connesse. L'applicazione della network analysis su tali dati rappresenta un approccio innovativo per lo studio dello scompenso cardiaco.*

**Key words:** Heart Failure, Comorbidities, Administrative Data, Network Analysis, Community Detection

---

Francesca Ieva

MOX - Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, via Bonardi 9, e-mail: francesca.leva@polimi.it

## Introduction

Congestive heart failure (CHF) is a disease that occurs when the heart muscle cannot pump properly the blood into the vessels. The term *congestive* is used since a common symptom of HF is congestion, i.e., too much fluid in tissues and veins is retained. The most common symptoms of HF mainly affect lungs and respiratory system in general as well as circulation. On the other hand, the HF can also cause liver enlargement and coagulopathy. From a clinical and social perspective, HF is one of the main public health issues and it still carries substantial morbidity and mortality, with 5-year mortality that rival those of many cancers. Moreover, the prevalence of HF in the world seems to show an increasing trend due mainly to the enlargement of life expectancy. In addition to this scenario, HF is often part of a comorbidity and this contributes to the worsening of the quality of life of HF patients.

On the other hand, the importance of complex networks in mathematical modeling is growing very fast in the recent years [1]. This is due to their flexibility and to the more and more consistent presence of relational data collected nowadays in context like social networks, interacting dynamical systems, social media, web pages, spreading processes, transportation systems, biological interactions and many others.

The aim of this work is to adopt a network approach in the study of the comorbidities recorded in HF patients. Specifically, we wish to investigate relationships among morbidities accompanying HF. Moreover, we target this goal for the first time in literature using administrative data in Italy. According to this approach, the perspective is shifted from the patients to their diseases. This allows us to study the pathologies taking advantage of techniques of network analysis. The administrative data supporting the study regard HF hospitalizations occurred in Lombardy (the most populated Italian region) between 2006 and 2012. The advantages of using administrative data are various: indeed, they are population based and their updates continues over time. There are also some drawbacks linked to the use of administrative data: in particular, they are not collected for epidemiological/statistical purposes. Moreover, the criteria of codifications for pathologies may vary over time and this complicates the merging of different administrative databases.

## 1 Data

### 1.1 Heart failure (HF) data

Originally, we deal with a big dataset made up of 503,247 hospitalizations regarding 142,587 different patients hospitalized in Lombardy between 2006 and 2012. In particular, all the hospitalizations concerned with congestive heart failure (that we will call from now on simply HF) are considered.

For each record (i.e., hospitalization) 76 variables are available. They can be divided into four categories or blocks:

1. patient personal information (e.g., his/her regional code, their total number of hospitalizations, sex and age)
2. details of the hospitalizations (e.g., all the surgical procedures and the comorbidities observed for that patient)
3. pharmacological treatments (e.g., the drugs taken by the patient after his/her stay in hospital)
4. ambulatory services the patients made use of before their stay in hospital.

Since we want to investigate the comorbidities patterns of patients, the most important block for our purposes is the second, composed by twenty dummy variables for each record. Each dummy indicates whether or not a disease affects the patient. Figure 1 shows the list of diseases we are going to consider (on the left the keyword is presented, on the right the corresponding disease or disorder).

Apart from the information about the comorbidities summarized by the dummy variables, we kept only a few other variables such as the identification number of each patient (ID), the number of current hospitalization per patient (adm number), the sex of the patient, their age, the date in which the patient leaves the study (dateOUT), the date of discharge (dateDISCHARGE), the number of days between the admission to the hospital and the leaving of the study (timeADMtoOUT) and the dummy variable which indicates whether or no a patient died (DEATH) without distinguishing if the death occurred in hospital or not. We decided to omit all the others in this work since they were not useful for our purpose that is investigating if, and also how, the patterns of comorbidities evolve during the years of the study through the usage of proper networks for representing the data.

Beyond the choices about the variables that can be useful for us, we also need a proper pre-processing and reshaping of our data in order to apply network analysis to them.

Keyword	Corresponding disease/disorder
metastatic	metastatic cancer
chf	congestive heart failure
dementia	dementia
renal	renal insufficiency
wtloss	unintentional weight loss
hemiplegia	hemiparesis
alcohol	alcoholism
tumor	any kind of non-metastatic cancer
arrhythmia	cardiac arrhythmia
pulmonarydz	chronic pulmonary disease
coagulopathy	coagulopathy
compdiabetes	complicated diabetes
anemia	deficiency anemia
electrolytes	fluid and electrolyte disorders
liver	liver disease
pvd	peripheral vascular disease
psychosis	psychosis
pulmicirc	pulmonary circulation disorders
hiv aids	HIV infection/AIDS
hypertension	hypertension

**Fig. 1** List of diseases with corresponding acronyms

## 1.2 Data pre-processing

In order to properly represent the data within a network, we extracted only the *last hospitalization* of each patient. This choice is due to the way the morbidity load is computed in the dataset: once a morbidity appears, it remains “active” also in the following hospitalizations of the relative patient. Moreover, since the last hospitalization will present the worst case for each of the patients, this is a good indicator for comorbidities diffusion. We also extracted *age* and *sex* as covariates for each patient and then we computed the average of the age and the percentage of men affected by each morbidity. These features will be considered as nodes attributes in our networks. In addition, we chose to consider the death as a morbidity, not distinguishing if it took place in hospital or not. This choice is aimed at discovering which morbidities are directly connected to the death, since they will be more perilous than the others.

Given the network, our aim is to investigate whether, and potentially how, the patterns of comorbidities evolve in time. We then built one network per year. We decided to make each patient contribute only to one network, that is the one related to the year of the patient’s last hospitalization. This approach guarantees the fact that if a patient is recorded in the network regarding one year, that patient cannot be found also in another network representing a different year.

## 2 Building the Network

Separating the years of hospitalizations and making patients contribute only to the network containing their last hospitalization, we get seven networks, one for each year between 2006 and 2012. In other words, we made a photography of the morbidity patterns of the patients year by year, not inducing dependencies between the different years. We also considered a patient as contributing to the death edge only if his/her death happens during the year his/her last hospitalization happens.

Building the networks with this approach, we still obtained very dense graphs, so in order to reduce the amount of connections considering only the edges whose weights are significant, we decided to consider only the positive correlations and also to impose a threshold on them empirically.

In Figure 2 it is possible to see an example of the network for year 2007. This network consists on twenty nodes that are nineteen different morbidities (HF is excluded) plus the death, while the links are weighted by  $\phi$ -correlations (see [4] for details). In particular, we considered only links that had a  $\phi$ -correlation [5] greater than 0.02.

Notice that we chose a small threshold in order to keep as many as possible connections among the ones with a positive associated weight. Nodes have different shapes: they are circular shaped if the disease is mainly observed in women rather than men, otherwise they are square shaped. On the other hand, the edges widths are proportional to the corresponding weights. Finally, each vertex has a color represent-



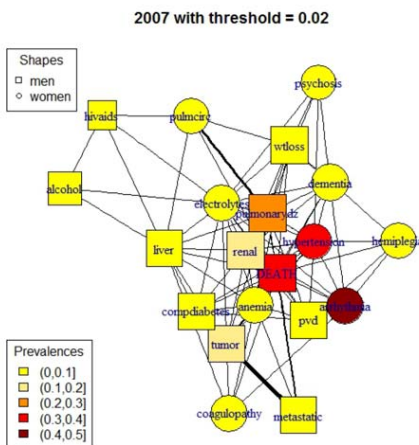
ing the prevalence of the disease, that is the proportion of the population presenting that morbidity.

### 3 Methods and Results

A preliminary descriptive analysis of the networks described in Section 2 was carried out, according to the literature on network analysis (see, among others, [2], [3] and [4]). We observed that the most of the diseases have a small prevalence apart from arrhythmia, hypertension, renal and pulmonarydz. We have also noticed that the death vertex is a very connected and important node. We remind that the prevalences of the death coincide with the percentages of patients whose last hospitalization occurred in that year and that also died in the same year.

The observed high values of the global transitivity indicators highlight that the pathologies tend to form complex agglomerations of morbidities, i.e., rarely HF appears alone. Indeed, HF is often present with a comorbidity load composed by three other diseases. Moreover, in our networks, we noticed that the least connected pathologies are also the ones that tend to take part in more complex clinical condition.

Locally, the pathologies can be easily ranked thanks to centrality measures: in particular, we ranked the diseases accordingly to their spread in HF patients (prevalence), their propensity to form connections and/or strong connections (degree and strength) and their closeness to all the others. In practice, as an innovative result we proposed a mixed index for the ranking of the pathologies. This approach led



**Fig. 2** Network of comorbidities related to year 2007. Vertex shapes are related to gender prevalence, colors talk about the prevalence of the disease for the current year.

us to focus on the important diseases in all the years of the study. Thanks to this, we discovered peculiar diseases of HF patients such as pulmonary disease, hypertension and renal insufficiency. Similarly, the  $\phi$ -correlations can be instead used in order to rank the pairwise comorbidities, which were represented by the links in the networks. We observed that among the heaviest edges we could find many edges connected to the node DEATH, but other pathologies were involved as well. We showed the trends of the most important pairwise comorbidities of HF patients.

Finally, community detection algorithms [6] allowed the agglomeration of morbidities to be pointed out, indeed we identified some groups of diseases that are more connected to each other with respect to the diseases of other groups, and suitable interpretations can be made.

## 4 Conclusions

In this work we showed a novel approach to the analysis of comorbidity patterns in patients affected by HF using networks. It represents a new method that can be adopted for different kind of analyses and pathologies.

The conclusions reached in this work are only a starting point for deeper analyses that can be done in the future. However, this work showed that networks are powerful and promising tools for investigating such kind of problems.

**Acknowledgements** This work has been developed within the FARB project *Public Management Research: Health and Education Systems Assessment*, funded by Politecnico di Milano. The author wishes to thank dott. Daniele Bitonti, for boosting this analysis with the studies carried out during his MD thesis.

## References

1. Barabasi, A.L.: Network Science. Cambridge University Press, First edition (2016).
2. Csardi, G., Nepusz, T.: The igraph software package for complex network research. *International Journal of Complex Systems*, 1695 (2006).
3. Kolaczyk, E.D.: Statistical Analysis of Network Data - Methods and Models. Springer, New York (2009).
4. Kolaczyk, E.D., Csardi, G.: Statistical Analysis of Network Data with R. Springer, New York (2014).
5. Newman, M.E.J.: Networks - An Introduction. Oxford University Press, New York (2010).
6. Opsahl, T., Panzaras, P.: Clustering in Weighted Networks. *Social Networks*, 31 (2), 155-163 (2009).

# Automatic variable and components weighting systems for Fuzzy cmeans of distributional data

## *Sistemi automatici di pesi di variabili e componenti per il Fuzzy cmeans di dati distribuzionali*

Antonio Irpino, Francisco de A.T. De Carvalho, Rosanna Verde

**Abstract** A distributional variable describes an object by a 1-D probability or frequency density function. While in standard clustering algorithms all the variables contribute to the clusters definition with the same importance, subspace clustering aims at finding a subspace, as a linear combination of the original variables, where clusters are well represented. This is done by weighting variables automatically and accordingly to their capacity of being discriminant for the clusters. Considering a decomposition of the squared  $L_2$  Wasserstein distance for distributional data, and using the notion of adaptive distance, we extend a fuzzy subspace clustering for automatically computing relevance weights associated with variables as well as with their components. This is done for the whole dataset or cluster-wisely. An application shows the advantages of using such algorithms.

**Abstract** *Una variabile distribuzionale permette di descrivere un oggetto attraverso una funzione di densità di probabilità o di frequenza. Mentre negli algoritmi standard di clustering tutte le variabili contribuiscono allo stesso modo alla definizione dei gruppi, le tecniche di subspace clustering cercano di individuare un sottospazio, come combinazione lineare delle variabili originarie, dove i gruppi siano ben rappresentati. Ciò è ottenuto attraverso l'individuazione automatica di un sistema di pesi per le variabili derivante dalla loro capacità discriminativa. Utilizzando un'aparticolare decomposizione della distanza  $L_2$  di Wasserstein per dati distribuzionali, e utilizzando la nozione di distanza adattativa, proponiamo delle estensioni di un algoritmo fuzzy di subspace clustering che permetta di calcolare automaticamente*

---

Antonio Irpino

Dip. di Matematica e Fisica, Università degli Studi della Campania "L. Vanvitelli", Viale Lincoln 5, 81100 Caserta e-mail: antonio.irpino@unicampania.it

Francisco de A.T. De Carvalho

Centro de Informática, Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes s/n - Cidade Universitaria, CEP 50740-560, Recife-PE, Brazil e-mail: fatc@cin.ufpe.br

Rosanna Verde

Dip. di Matematica e Fisica, Università degli Studi della Campania "L. Vanvitelli", Viale Lincoln 5, 81100 Caserta e-mail: rosanna.verde@unicampania.it

*dei pesi associati alle variabili o alle loro componenti. I pesi possono riferirsi o all'intero insieme di dati o al singolo gruppo. Un'applicazione mostra i vantaggi degli algoritmi proposti.*

**Key words:** Distributional data, Fuzzy cmeans, subspace clustering, automatic weights, Wasserstein distance.

## 1 Introduction

A distributional (or distribution-valued) data is observed when an object is described by a distributional variable, since its realizations are 1-D frequency (or probability) density functions. Such kinds of data can be observed in many practical situation. For example, official statistics institute, in order to preserve the privacy of respondents, cannot diffuse microdata collected from a territorial unit, but only a summary of such data. A similar case occurs with repeated data observed on individuals collected, for example, from a bank or an hospital. Also in this case, only a summarized version of such data can be available. Empirical parametric or non-parametric density estimates are useful tools for this aims. Clustering aims to organize a set of objects into groups such that those within a given cluster are more similar with respect the ones of a different clusters. Partitioning and hierarchical clustering are two possible approaches. According to how much an object belongs to a cluster, in hard clusterings an object is assigned to a cluster, while in fuzzy[?] ones an object may belong, according to a membership degree, to more than one cluster at the same time.

The most clustering algorithms proposed for histogram data are partitioning *hard* clustering methods[14, 18, 11, 15, 16, 17].

However, particular structure of the observed distributional data could give clusters not well separated and with a high internal variability due to the presence of some data that are forced to belong to only one cluster. In presence of this kind of problem, the fuzzy clustering algorithm is a suitable choice. This paper extends Refs. [11, 15, 16, 17] by proposing a fuzzy c-means clustering algorithm.

Another main issue in clustering analysis is to consider the different contribution of the several variables in the clustering process. Generally, clustering methods do not take into account the different relevance of the variables in the analysis. However, in most applications, some variables may be more discriminant of the clusters than others; in some other applications each cluster may have a different set of more relevant variables to group together the data. The approach of subspace clustering [1] aims at finding a subspace of the original descriptor space using a linear combination of the original variables. Generally, when data are described by a large number of variables, subspace clustering act as feature selection method, too. The subspace (if all the data are considered) or the subspaces (if a subspace is generated for each cluster) produced by such algorithms are optimal with respect a criterion that maximizes the homogeneity of clusters and/or maximize the separation among

them. A similar result was already reached in [6], where the use of adaptive distances was proposed for clustering standard data.

In the framework of Symbolic Data Analysis, [4, 5, 3] proposed several adaptive distances, based on Hausdorff, City-Block and Euclidean distances in dynamic clustering algorithm of set-valued data. Recently [13] a partitioning hard clustering algorithm using an adaptive distance based on the  $L_2$  Wasserstein metric has been proposed. The authors propose two novel adaptive distances based on clustering schemes able to compute automatically the relevance of each histogram variable during the partitioning of the data set. Starting from a decomposition of the  $L_2$  Wasserstein distance [12] and considering the variability measure introduced in [17], the distance between two distributional data can be shared in two components: one related to the variability of averages of the distributions and another related to the different variability of the compared distributions. In all the algorithms based on the approach of adaptive distances of [6], a k-means-like algorithm is proposed, where the minimization of an homogeneity criterion is subject to a constraint on the product of relevance weights. On the other hand, considering the subspace clustering approaches reviewed in [1], a constraint on the sum of relevance weights is considered.

In this paper, we consider to extend a subspace fuzzy c-means algorithm to distributional data using adaptive  $L_2$  Wasserstein distance. Taking into consideration the  $L_2$  Wasserstein distance decomposition in two additive components [17] we propose adaptive distances that take into account the two components of the variability of a set of distributions. We propose to associate two sets of weights with each variable and with each component, such that the sum of weights for the whole dataset or for each cluster is equal to one. The proposed fuzzy clustering algorithm, based on adaptive distances, alternates three steps that estimates the membership of the objects to the clusters, the weights for each variable and/or each component, and the cluster prototypes.

## References

1. Deng, Z., Choi, K., Jiang, Y., Wang, J., Wang, S.: A survey on soft subspace clustering. *Information Sciences*, **384**, 84–106 (2016)
2. H. H. Bock and E. Diday. *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, Berlin, 2000.
3. De Carvalho, F. A. T., De Souza, R. M. C. R.: Unsupervised pattern recognition models for mixed feature-type symbolic data. *Pattern Recognition Letters*, **31**, 430–443 (2010)
4. De Carvalho, F. A. T., Lechevallier, Y.: Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recognition*, **42**(7), 1223–1236 (2009)
5. De Carvalho, F. A. T., Lechevallier, Y.: Dynamic clustering of interval-valued data based on adaptive quadratic distances. *Trans. Sys. Man Cyber. Part A*, **39**(6), 1295–1306 (2009)
6. Diday, E., Govaert, G.: Classification automatique avec distances adaptatives. *RAIRO Informatique/Computer Science*, **11**(4), 329–349 (1977).
7. Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

8. Warren Gilchrist. *Statistical Modelling with Quantile Functions*. CRC Press, Abingdon, 2000.
9. Clark R. Givens and Rae Michael Shortt. A class of wasserstein metrics for probability distributions. *Michigan Math. J.*, 31(2):231–240, 1984.
10. J.Z. Huang, M.K. Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):657–668, 2005.
11. Irpino, A., Verde, R.: A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In V. Batanjeli, H.H. Bock, A. Ferligoj, and A. Ziberna, editors: *Data Science and Classification*, 185–192. Springer, Berlin, (2006)
12. Antonio Irpino A., Romano E.: Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. *Revue des Nouvelles Technologies de l'Information, RNTI-E-9*, 99–110 (2007)
13. Irpino, A., Verde, R., De Carvalho, F.A.T.: Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. *Expert Systems with Applications*, **41**(7), 3351 – 3366 (2014)
14. Y. Terada, Y., Yadohisa, H.: Non-hierarchical clustering for distribution-valued data. In Lechevallier, Y., Saporta, G., editors, *Proceedings of COMPSTAT 2010*, pages 1653–1660. Springer, Berlin, (2010)
15. Verde, R., Irpino, A.: Dynamic clustering of histograms using Wasserstein metric. In A. Rizzi and M. Vichi, editors: *Proceedings in Computational Statistics, COMPSTAT 2006*, 869–876, Physica Verlag, Heidelberg (2006)
16. Verde, R., Irpino, A.: Comparing histogram data using a Mahalanobis-Wasserstein distance. In P. Brito, editor: *Proceedings in Computational Statistics, COMPSTAT 2008*, 77–89, Springer Verlag, Heidelberg (2008)
17. Verde R., Irpino, A.: Dynamic clustering of histogram data: using the right metric. In P. Brito, P. Bertrand, G. Cucumel, and F. De Carvalho, editors.: *Selected contributions in data analysis and classification*, 123–134, Springer, Berlin, (2008)
18. Vrac, M., Billard, L., Diday, E., Chedin, A.: Copula analysis of mixture models. *Computational Statistics*, **27**, 427–457 (2012)

# A Bayesian oblique factor model with extension to tensor data

## *Un modello fattoriale obliquo bayesiano con estensione a dati tensoriali*

Michael Jauch, Paolo Giordani, and David Dunson

**Abstract** In this short paper, we discuss a novel way of constructing prior distributions for correlation matrices and an associated approach to inference. We construct a prior penalizing large correlations, which we incorporate into an oblique factor model and a Candecomp/Parafac model for three-way data. We argue that this choice of prior for the factor correlation matrix, combined with a shrinkage prior for elements of the factor loadings matrix, leads to interpretable solutions. At the meeting we will demonstrate this through applications to real data.

**Abstract** *In questo short paper discutiamo un nuovo modo di costruire distribuzioni a priori per matrici di correlazione ed i relativi aspetti inferenziali. La distribuzione a priori, costruita in maniera tale da penalizzare correlazioni elevate, viene inserita all'interno di un modello di analisi fattoriale obliqua e del modello Candecomp/Parafac per dati a tre vie. Riteniamo che questa scelta della a priori per la matrice di correlazione fattoriale, combinata con una a priori shrinkage per gli elementi della matrice dei loading fattoriali permette di ottenere soluzioni interpretabili. Al convegno dimostreremo il nostro assunto mediante applicazioni a dati reali*

**Key words:** oblique factor model, prior for correlation matrices, tensor decomposition, three-mode factor analysis

---

Michael Jauch

Department of Statistical Science, Duke University, e-mail: michael.jauch@duke.edu

Paolo Giordani

Department of Statistical Sciences, Sapienza University of Rome, e-mail: paolo.giordani@uniroma1.it

David B. Dunson

Department of Statistical Science, Duke University, e-mail: dunson@duke.edu

## 1 Introduction

Factor analysis aims to explain the covariance structure between observed variables as arising from a smaller number of unobserved latent factors. A Gaussian factor model with factor dimension  $S$  has the form

$$\mathbf{y}_i | \mathbf{B}, \mathbf{f}_i, \Sigma \sim N(\mathbf{B}\mathbf{f}_i, \Sigma), \quad \mathbf{f}_i \sim N(\mathbf{0}, \Omega) \quad (1)$$

where  $\mathbf{y}_i$  is the centered vector of observed variables corresponding to the  $i$ th observation,  $\mathbf{B}$  is the factor loadings matrix,  $\mathbf{f}_i$  is the  $S$ -dimensional vector of latent factors for observation  $i$ ,  $\Omega$  is the covariance matrix of the latent factors, and  $\Sigma$  is a diagonal positive definite matrix. Marginalizing out the latent factors yields

$$\mathbf{y}_i | \mathbf{B}, \Sigma \sim N(\mathbf{0}, \mathbf{B}\Omega\mathbf{B}^T + \Sigma). \quad (2)$$

As is well-known, the factor model is not identifiable without further restrictions on  $\mathbf{B}$ ,  $\Omega$ ,  $\Sigma$ . Identifiability assumptions are important in Bayesian computation as a means to ensure that estimation based on posterior samples is meaningful. See [14] for a discussion of identifiability of the oblique factor model. We impose the usual restriction that  $\Omega$  be a correlation matrix.

If  $\Omega$  is diagonal then the latent factors are uncorrelated (and thus independent) and we obtain the conventional orthogonal factor analysis model. If  $\Omega$  is not diagonal, the latent factors are correlated and we obtain the so-called oblique factor model. Many authors have argued that the restriction to uncorrelated factors is too strict. For example, discussing application of factor analysis in psychology, Thurstone [19] remarks

It seems just as unnecessary to require that mental traits shall be uncorrelated in the general population as to require that height and weight be uncorrelated in the general population.

A large body of methodology has been developed for the oblique factor model. For a detailed discussion, see Chapter 12 of Harman [11].

In traditional applications of factor analysis, interest lies in interpreting the latent factors as distinct and scientifically meaningful quantities. Interpretation proceeds by examination of the factor loadings matrix, which relates the latent factors to the observed variables. Interpretation is made easier if the factor loadings matrix possesses a simple structure. An example of a simple structure is near sparsity, in which the factor loadings matrix has a small number of large entries and a large number of small entries. Typically, allowing correlated factors allows for a simpler structure in the factor loadings matrix. However, correlated factors present their own difficulties to interpretation. If two factors are highly correlated, then it becomes impossible to interpret them as distinct quantities. When allowing for correlated factors, we should recognize the tradeoff between factor correlation and complexity of the loadings matrix. We tolerate some of the former if it buys us less of the latter, and vice versa. This idea is captured in a quote from [11]:

It is clear that a certain simplicity of interpretation is sacrificed upon relinquishing the standard of orthogonality. This disadvantage may be offset, however, if the linear descriptions



of the variables in terms of correlated factors can be made simpler than in the case of uncorrelated ones. Generally this is possible.

We can address this tradeoff in a Bayesian oblique factor analysis setting through the choice of prior distributions for  $\mathbf{B}$  and  $\mathbf{\Omega}$ . For those elements of the factor loadings matrix  $\mathbf{B}$  which are not restricted to be zero for the sake of identifiability, we can take advantage of local-global shrinkage priors which will result in a nearly sparse estimate for  $\mathbf{B}$  [15]. For the factor correlation matrix  $\mathbf{\Omega}$ , we need a prior on the set of correlation matrices which penalizes factor correlations. Defining such a prior distribution that lends itself to relatively simple and scalable inference is challenging. For a recent approach in the context of Bayesian factor analysis with correlated factors, see [9] which provides extensive references to earlier relevant works.

A main contribution of this short paper is what we believe to be a novel approach to constructing priors for correlation matrices. The construction is based on the observation that, for any  $N \times P$  matrix  $\mathbf{X}$  with unit norm columns, the product  $\mathbf{X}^T \mathbf{X}$  is a correlation matrix. Assigning each column of  $\mathbf{X}$  a probability distribution having support on the unit sphere then induces a probability distribution on correlation matrices. In the special case that each column of  $\mathbf{X}$  is independent and uniformly distributed on the unit sphere, we obtain closed-form densities for the correlations which match priors discussed previously by [1, 9], and others. The proposed prior does indeed penalize correlation, and the penalty increases as  $N$  increases. In this short paper, we only discuss the special case of independent columns uniformly distributed on the unit sphere, but future work may consider other choices, leading to more flexible distributions for correlation matrices. Inference for parameters lying on the unit sphere can be performed using geodesic Monte Carlo [6], a scalable Markov chain Monte Carlo (MCMC) method which can accommodate parameters lying on manifolds embedded in Euclidean space.

We define a Bayesian oblique factor model using the aforementioned prior for the factor correlation matrix and a global-local shrinkage prior for elements of the factor loadings matrix. We discuss extension of the factor model to tensor valued data with an emphasis on the three-way case. For the conference presentation, we will show applications to real data.

## 2 A Bayesian model for oblique factor analysis

Suppose we have  $I$  observations of  $J$  variables. We let  $\mathbf{y}_i$  be the vector of  $J$  centered variables corresponding to observation  $i$ . As before, we suppose that

$$\mathbf{y}_i = \mathbf{B} \mathbf{f}_i + \mathbf{e}_i, \quad \mathbf{f}_i \sim N(\mathbf{0}, \mathbf{\Omega}) \quad i = 1, \dots, I \quad (3)$$

where  $\mathbf{B}$  is the  $J \times S$  factor loadings matrix,  $\mathbf{f}_i$  is the  $S \times 1$  vector of latent factors for observation  $i$ ,  $\mathbf{\Omega}$  is a  $S \times S$  correlation matrix, and  $\mathbf{e}_i$  is a  $J \times 1$  vector of errors. The errors are independent and identically-distributed  $N(\mathbf{0}, \mathbf{\Sigma})$  where  $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_J^2)$  is a diagonal positive definite matrix. In matrix form, we have

that

$$\mathbf{Y} = \mathbf{F}\mathbf{B}^T + \mathbf{E} \quad (4)$$

where  $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_I)^T$  is the  $I \times J$  data matrix,  $\mathbf{F} = (\mathbf{f}_1, \dots, \mathbf{f}_I)^T$  is the  $I \times S$  matrix of latent factors, and  $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_I)^T$  is the  $I \times J$  matrix of errors. To complete the Bayesian model specification, we need to define priors for our parameters.

## 2.1 The prior for $\boldsymbol{\Omega}$

As described in the introduction, let the matrix  $\mathbf{X}$  have columns  $\mathbf{x}_1, \dots, \mathbf{x}_P \in \mathcal{S}_{N-1}$ , the  $N - 1$ -dimensional unit sphere. Suppose that the columns of  $\mathbf{X}$  are independent and uniformly-distributed on  $\mathcal{S}_{N-1}$ . We then set  $\boldsymbol{\Omega} = \mathbf{X}^T \mathbf{X}$ .

Due to the simple construction for  $\boldsymbol{\Omega}$ , it is possible to derive closed-form expressions describing its distribution [8]. For instance, let  $\omega$  be an arbitrary off-diagonal element of  $\boldsymbol{\Omega}$ . Then the density of  $\omega$  is

$$p(\omega) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{N}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)} (1 - \omega^2)^{\frac{N-3}{2}}, \quad \omega \in [-1, 1], \quad (5)$$

the even-order moments are

$$\mathbb{E}(\omega^{2m}) = \prod_{j=1}^m \frac{2j-1}{N+2j-2}, \quad m = 1, 2, 3, \dots \quad (6)$$

and the odd-order moments are zero. As Fig. 1 makes evident,

$$\frac{\omega+1}{2} \sim \text{Beta}((N-1)/2, (N-1)/2) \quad (7)$$

and the prior places a penalty on correlations which increases with  $N$ .

The above properties make it clear that we have presented an alternate way of constructing a prior distribution for correlation matrices having the same marginal distributions for the correlations as the prior for correlation matrices discussed in [9] and the relevant references given there. However, our prior construction naturally leads to a wide variety of flexible generalizations (by choosing different distributions on the unit sphere) and allows for a different MCMC approach to inference based on Geodesic Monte Carlo [6].

## 2.2 Completing the prior specification

We would like to choose a prior for  $\mathbf{B}$  favoring a simple, nearly sparse structure. A variety of global-local shrinkage priors [15, 4] have been proposed which satisfy this requirement and have desirable posterior concentration properties. These glocal-

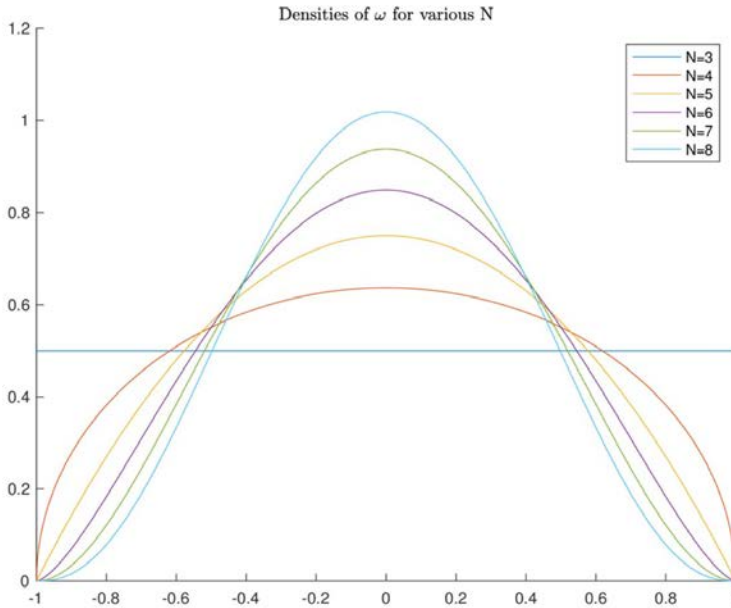
local shrinkage priors can typically be represented as scale mixtures of Gaussians, simplifying computation.

As mentioned in the introduction, identifiability assumptions are important in Bayesian computation because they ensure that estimation based on posterior samples is meaningful. We have already constrained  $\mathbf{\Omega}$  to be a correlation matrix. The article by Peeters [14] gives three additional conditions on  $\mathbf{B}$  which, under the usual regularity assumptions, guarantee identifiability of the oblique factor model. Decisions about how to satisfy those conditions should be made on a case by case basis.

We can assign conventional priors for variances to the diagonal elements of  $\mathbf{\Sigma}$ , e.g. inverse gamma or reference priors.

### 3 Extension to tensor data

When  $I$  observations of  $J$  variables are collected at  $K$  occasions we have a three-way array or tensor denoted by  $\underline{\mathbf{Y}}$  of order  $I \times J \times K$ . Occasions may refer to time or in general to different conditions. Three-way tensor data are characterized by three



**Fig. 1** Density of  $\omega$  for various values of  $N$ . The densities are shifted and scaled Beta  $((N-1)/2, (N-1)/2)$  densities.

modes, namely observation, variable and occasion modes. We let  $\mathbf{y}_i$  be the vector corresponding to observation  $i$ . In contrast to the standard two-way case,  $\mathbf{y}_i$  contains the scores of  $J$  centered variables at  $K$  occasions and thus has length  $JK$ .

In principle, the two-way factor model in (4) might still be applied for tensor data. In fact, it would be sufficient to juxtapose next to each other the observation-by-variable matrices collected at every occasion obtaining a matrix with rows given by  $\mathbf{y}_1, \dots, \mathbf{y}_I$ . Such a matrix, usually denoted by  $\mathbf{Y}_A$ , is the so-called observation mode matricization (or unfolding) of the tensor  $\underline{\mathbf{Y}}$ . However, in practice, the decomposition of  $\mathbf{Y}_A$  through the factor model in (4) is inappropriate because the interactions among the modes cannot be modelled.

A more sensible strategy is the three-mode factor analysis model known as Candecomp [7] or Parafac [12] which we will refer to as Candecomp/Parafac or, more briefly, CP. The CP model can be expressed as

$$\mathbf{Y}_A = \mathbf{F}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{E}_A \quad (8)$$

where  $\mathbf{B}$  and  $\mathbf{C}$  are the factor loadings matrices for the variables (of order  $J \times S$ ) and for the occasions (of order  $K \times S$ ), respectively. They capture the influences of the variables and occasions on the  $S$  latent factors. The symbol  $\odot$  denotes the Khatri-Rao product, the Kronecker product between pairs of columns ( $\mathbf{C} \odot \mathbf{B} = [\mathbf{c}_1 \otimes \mathbf{b}_1 | \dots | \mathbf{c}_S \otimes \mathbf{b}_S]$ , with  $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_S)$  and  $\mathbf{C} = (\mathbf{c}_1, \dots, \mathbf{c}_S)$ ).  $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_I)^T$  is the  $I \times JK$  matricization of the tensor of errors  $\underline{\mathbf{E}}$ . In contrast with the two-way case, under mild conditions (see, e.g., [13]) the solution of the CP model is identified up to trivial scaling and simultaneous permutation of the columns of  $\mathbf{F}$ ,  $\mathbf{B}$  and  $\mathbf{C}$ .

The CP model was originally proposed as an exploratory tool without probabilistic assumptions. The probabilistic version was developed in [5] and [2, 3]. Actually, such probabilistic counterparts were proposed for the so-called Tucker3 model [20], which we will refer to as the T3 model. The T3 model represents an alternative three-mode generalization of the two-way factor analysis model. It can be formulated as

$$\mathbf{Y}_A = \mathbf{F}\mathbf{G}_A(\mathbf{C} \otimes \mathbf{B})^T + \mathbf{E}_A. \quad (9)$$

In the T3 model, each mode has its own factors and different numbers of latent factors for each mode can be assumed. Hence,  $\mathbf{F}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  are matrices of order  $I \times P$ ,  $J \times Q$ , and  $K \times R$ , respectively, with  $P$ ,  $Q$  and  $R$  denoting the numbers of factors for each mode. The triple interactions among the factors of the three modes are captured by the  $P \times Q \times R$  core tensor  $\underline{\mathbf{G}}$ , the generic element of which,  $g_{pqr}$ , expresses the strength of the interaction among factor  $p$  for the observation mode, factor  $q$  for the variable mode, and factor  $r$  for the occasion mode. Note that in (9)  $\mathbf{G}_A$  denotes the observation mode matricization of  $\underline{\mathbf{G}}$ .

The CP and T3 models are closely related. If  $P = Q = R = S$  and  $\underline{\mathbf{G}}$  has a superidentity structure ( $g_{pqr} = 1$  when  $p = q = r$  and 0 otherwise), then it is easy to see that formulas (8) and (9) coincide. Therefore, the CP model can be seen as a constrained version of the T3 model where the same number of latent factors is assumed for all the modes and each factor of a certain mode interacts with exactly one factor of the other modes. This produces some relevant distinctions between the two

models. The T3 model is more general than the CP model, but the solution is not identified. Equally well-fitting solutions can be found by rotating the factor matrices and compensating for such rotations in the core. On the other hand, the CP model has a more parsimonious structure and the solution, as mentioned, is identified. For this reason, we focus our attention on the CP model.

If the latent factors of  $\mathbf{F}$  are correlated, the covariance structure of the data induced by the CP model takes the form (see also [17])

$$(\mathbf{C} \odot \mathbf{B})\mathbf{\Omega}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{\Sigma}. \quad (10)$$

Hence, under the same assumptions adopted in the two-way case, the model for the  $i$ th observation in the tensor case is

$$\mathbf{y}_i | \mathbf{B}, \mathbf{C}, \mathbf{\Omega}, \mathbf{\Sigma} \sim N(\mathbf{0}, (\mathbf{C} \odot \mathbf{B})\mathbf{\Omega}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{\Sigma}). \quad (11)$$

A Bayesian CP model can be developed as a natural generalization of the two-way model presented previously. The prior of Section 2.1 can again be used for the factor correlation matrix  $\mathbf{\Omega}$ . For the elements of  $\mathbf{B}$  and  $\mathbf{C}$ , we can again use global-local shrinkage priors which favor a nearly sparse structure. The uniqueness of the CP solution up to scaling and simultaneous permutation of the columns of  $\mathbf{B}$  and  $\mathbf{C}$  implies that we only need to worry about column switching in the posterior samples, since the prior distributions for  $\mathbf{B}$  and  $\mathbf{C}$  fix their respective scales. Two solutions are a relabeling scheme in the style of [18] or simply fixing particular elements of  $\mathbf{B}$  or  $\mathbf{C}$  to be zero so that the columns can no longer be permuted. The Bayesian CP model enjoys the same advantages as the two-way oblique factor model. In particular, we hope to demonstrate that allowing (but penalizing) latent factor correlation and applying a shrinkage prior on the factor loadings leads to a general yet interpretable CP model.

## 4 Inference

For inference, we use geodesic Monte Carlo [6]. Geodesic Monte Carlo extends Hamiltonian Monte Carlo [16] to certain special manifolds which can be embedded in Euclidean space, such as the simplex, the unit sphere, or the Stiefel manifold. Like Hamiltonian Monte Carlo, geodesic Monte Carlo can generate distant Metropolis-Hastings proposals with a high probability of acceptance, ideally leading to a rapidly-mixing Markov chain with low autocorrelation. As described in [6], sometimes parallel tempering [10] is required to move between isolated modes of the posterior distribution.

## References

1. Barnard, J., McCulloch, R., Meng, X.-L.: Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Stat. Sin.* **10**, 1281–1311 (2000)
2. Bentler, P.M., Lee, S.Y.: Statistical aspects of a three-mode factor analysis model, *Psychometrika* **43**, 343–352 (1978)
3. Bentler, P.M., Lee, S.Y.: A statistical development of three-mode factor analysis, *Brit. J. Math. Stat. Psy.* **32**, 87–104 (1979)
4. Bhattacharya, A., Pati, D., Pillai, N.S., Dunson, D.B.: Dirichlet–Laplace priors for optimal shrinkage, *J. Am. Stat. Assoc.* **110**, 1479–1490 (2015)
5. Bloxom, B.: A note on invariance in three-mode factor analysis, *Psychometrika* **33**, 347–350 (1968)
6. Byrne, S., Girolami, M.: Geodesic Monte Carlo on embedded manifolds, *Scand. J. Stat.* **40**, 825–845 (2013)
7. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an  $n$ -way generalization of “Eckart-Young” decomposition, *Psychometrika* **35**, 283–319 (1970)
8. Cho, E.: Inner product of random vectors on  $S^n$ , *J. Pure Appl. Math.: Adv. Appl.* **9**, 63–68 (2013)
9. Conti, G., Frühwirth-Schnatter, S., Heckman, J.J., Piatek, R.: Bayesian exploratory factor analysis, *J. Econom.* **183**, 31–57 (2014)
10. Geyer, C.J.: Markov chain Monte Carlo maximum likelihood. In: *Computing Science and Statistics, Proc. 23rd Symp. Interface*, pp. 156163. Interface Foundation of North America (1991)
11. Harman, H.H.: *Modern Factor Analysis*. University of Chicago Press, Chicago (1967)
12. Harshman, R.A.: Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multi-modal factor analysis”, *UCLA Work. Pap. Phon.* **16**, 1–84 (1970)
13. Kruskal, J.B.: Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, *Linear Algebra Appl.* **18**, 95–138 (1977)
14. Peeters, C.F.W.: Rotational uniqueness conditions under oblique factor correlation metric, *Psychometrika* **77**, 288–292 (2012)
15. Polson, N.G., Scott, J.G., Clarke, B., Severinski, C.: Shrink globally, act locally: sparse bayesian regularization and prediction. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (eds.) *Bayesian Statistics 9*, Oxford University Press (2012)
16. Neal, R.M.: MCMC using Hamiltonian dynamics. In: Brooks, S., Gelman, A., Jones, G., Meng, X.-L. (eds.) *Handbook of Markov Chain Monte Carlo*, pp. 113–162. CRC Press (2011)
17. Stegeman, A., Lam, T.T.T.: Three-mode factor analysis by means of Candecomp/Parafac, *Psychometrika* **79**, 426–443 (2014)
18. Stephens, M.: Dealing with label switching in mixture models, *J. R. Stat. Soc. Series B: Stat. Methodol.* **62**, 795–809 (2000)
19. Thurstone, L.L.: *Multiple-factor Analysis: A Development and Expansion of The Vectors of Mind*. University of Chicago Press, Chicago (1947)
20. Tucker, L.R.: Some mathematical notes on three-mode factor analysis, *Psychometrika* **31**, 279–311 (1966)

# Statistical analysis for partially observed multilayered networks

## *Analisi statistica di reti multi-strato parzialmente osservate*

Johan Koskinen, Chiara Broccatelli, Peng Wang, and Garry Robins

**Abstract** Multilayered networks have been proposed as a joint representation of associations between multiple types of entities or nodes, such as people and organization, where two types of nodes gives rise to three distinct types of ties. The typical roster data collection method may be impractical or infeasible when the node sets are hard to detect or define or because of the cognitive demands on respondents. Multilayered networks allow us to consider a multitude of different sources of data and to sample on different types of nodes and relations. We consider modelling multilayered networks using exponential random graph models and extend a recently developed Bayesian data-augmentation scheme to allow partially missing data.

**Abstract** *Le reti multi-strato sono state proposte come una rappresentazione congiunta di associazioni tra diversi tipi di entità o nodi, quando due tipologie diverse di nodi generano tre diverse combinazioni di legame. I metodi tradizionali di raccolta dati non sempre sono utilizzabili, sia perché l'insieme di nodi potrebbe essere difficile da definire, sia in seguito ad eventuali esigenze cognitive degli intervistati. Le reti multi-strato offrono un vantaggio in queste situazioni poiché permettono di utilizzare più fonti d'informazione congiuntamente e semplificano il campionamento di diversi nodi e relazioni. Questo lavoro intende modellare le reti multi-strato tramite i modelli esponenziali per reti casuali (ERGM) ed estende una re-*

---

Johan Koskinen

Social Statistics Discipline Area, University of Manchester, Manchester M13 9PL e-mail: johan.koskinen@manchester.ac.uk

Chiara Broccatelli

Sociology, University of Manchester, Manchester M13 9PL e-mail: chiara.broccatelli@postgrad.manchester.ac.uk

Peng Wang, Centre for Transformative Innovation, Faculty of Business and Law, Swinburne University of Technology, Australia e-mail: pengwang@swin.edu.au · Garry Robins, Melbourne School of Psychological Sciences, The University of Melbourne, Australia e-mail: garrylr@unimelb.edu.au

*cente tecnica di data-augmentation fondata sul paradigma Bayesiano che consente di gestire dati parzialmente mancanti.*

**Key words:** ERGM, Exchange algorithm, Missing data, Multilevel networks, Social Networks

## 1 Introduction

Kivelä et al (2014) coined the term ‘multilayered networks’ as a general framework for jointly designating multiple types of network data, such as one-mode, two-mode, and multiplex networks, where researchers had typically dealt with each instance separately. Here we are primarily considering the extension of the exponential random graph (ERGM) family of distributions proposed by Wang et al. (2013) to the subclass of multilayered networks typically referred to as ‘multilevel networks’ (Lazega et al., 2008) even though the key ideas in dealing with partially observed data generalises to other extensions of ERGM, such as multiplex networks (Pattison and Wasserman, 1999). In situations where you are likely to have imperfect information on network ties, availing yourself of the full set of tools that may be derived from a wider framework for networks may prove beneficial.

## 2 Data Structure

We assume two distinct set of nodes:  $A = \{1, \dots, n\}$  and  $B = \{1, \dots, m\}$  where we might observe ties among all combinations of nodes type. A tie thus belong to either of the sets  $\binom{A}{2}$ ,  $A \times B$ , or  $\binom{B}{2}$ . In the sequel we will use  $AA$ ,  $AB$ , and  $BB$  as a notational shorthand for these edge-sets, with the corresponding incidence matrices  $\mathbf{X}_{AA}$ ,  $\mathbf{X}_{AB}$ , and  $\mathbf{X}_{BB}$ , respectively. The element  $X_{E,v}$  of matrix  $\mathbf{X}_E$  is equal to 1 if the edge  $v \in E$  belongs to the graph and 0 otherwise. The multilevel network may be represented as a one-mode network with a blocked, symmetric adjacency matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{AA} & \mathbf{X}_{AB} \\ \mathbf{X}_{BA} & \mathbf{X}_{BB} \end{pmatrix}$$

When extending binary one-mode networks to multiple relations (say ‘friendship’ and ‘advice’) it is convention to represent this as a collection of graphs or adjacency matrices, one for each relation. For multilevel networks we by definition have different relations for different combinations of node-sets. Let the number of relations be denoted by  $R_E$ , for  $E = AA, AB, BB$ , with incidence matrices being defined as  $\mathbf{X}_E^{(r)} = (X_{E,v}^{(r)})$ , where  $X_{E,v}^{(r)} = 1$  if there is a tie on relation  $r = 0, \dots, R_E - 1$  for edge-set  $E = AA, AB, BB$ . When the number of relations for  $E = AA, AB, BB$  differ, we



are not able to unambiguously define the multilayered network as a collection of one-mode network with blocked, symmetric adjacency matrices.

For  $AA$ ,  $AB$ , and  $BB$  define the binary indicator matrices  $\mathbf{D}_{AA}$ ,  $\mathbf{D}_{AB}$ , and  $\mathbf{D}_{BB}$ , each of which having elements  $D_{E,v}$  of  $\mathbf{D}_E$  equal to 1 or 0 depending on whether the corresponding tie-variable  $v$  is observed or not, respectively. For each  $E = AA, AB, BB$  the indicators extend straightforwardly to account for more than one relation. Thus, for example, if  $\mathbf{X}_{AA}^{(0)}$  represent friendship ties and  $\mathbf{X}_{AA}^{(1)}$  represent advice ties, the corresponding matrices  $\mathbf{D}_{AA}^{(0)}$  and  $\mathbf{D}_{AA}^{(1)}$  would indicate what friendship and advice ties were observed and which ones were not observed.

We follow the convention (Little & Rubin, 1987) of partitioning data  $\mathbf{X}$  into observed  $\mathbf{X}^{\text{obs}} = \{X_v : D_v = 1\}$  and unobserved  $\mathbf{X}^{\text{miss}} = \{X_v : D_v = 0\}$  data, conditional on an outcome  $\mathbf{D}$ . For a given  $\mathbf{D}$  we take  $(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss}})$  to denote  $\mathbf{X}$  reconstructed.

### 3 Model Formulation

Frank and Strauss (1986) derived ERGMs for one-mode networks from the so called Markov dependence assumption that posited that for any two pairs  $\{i, j\}$  and  $\{k, \ell\}$  of vertices of a graph, the tie-variables  $X_{i,j} \perp X_{k,\ell} | X_{-(i,j),(k,\ell)}$  if  $\{i, j\} \cap \{k, \ell\} = \emptyset$ . They proved that the Markov dependence assumption implied a log-linear model for the collection of tie-variables that has as its sufficient statistics counts of different network ‘configurations’ (incidentally echoing the conclusions drawn by Moreno and Jennings, 1938). Snijders et al. (2006) elaborated on the Markov model by proposing parameters derived from the so called social circuit dependence assumption. The general form of ERGM is

$$p(\mathbf{X}|\theta) = \exp\{q(\mathbf{X}; \theta) - \psi(\theta)\}$$

where the normalising constant  $\psi(\theta) = \sum_{Y \in \mathcal{X}} \exp\{q(\mathbf{Y}; \theta)\}$  and  $q(\mathbf{X}; \theta)$  is a potential dependent on the structure of the network and a vector  $\theta$  of statistical parameters. This general form is agnostic to the specific dependencies we may hypothesise for a particular type of network object. For undirected one-mode network, the model of Frank and Strauss (1986) has the potential written as a weighted sum of sufficient graph statistics

$$\log q(\mathbf{X}; \theta) = \sum_r^{P_S} \theta_{S_r} \binom{X_{i+}}{r} + \theta_T \sum_{(i,j,k) \in \binom{A}{3}} X_{ij} X_{ik} X_{jk}$$

where the statistics correspond to two distinct categories of statistics, namely stars and triangles (in the expression  $X_{i+} = \sum_j X_{ij}$ ). ERG models have been proposed for two-mode networks (Skvoretz and Faust, 1999; Agneessens and Roose, 2008; Wang et al., 2009) and multiplex networks (Pattison and Wasserman, 1999). The modelling family has also been extended to the joint analysis of ties between different types of nodes (Wasserman and Iacobucci, 1991) and for fully defined multilevel

networks by Wang et al. (2013). Wang et al. (2013) factor the function  $q(\mathbf{X}; \theta)$

$$\begin{aligned} \log q(\mathbf{X}; \theta) &= \theta_{AA}^\top z(\mathbf{X}_{AA}) + \theta_{BB}^\top z(\mathbf{X}_{BB}) + \theta_{AB}^\top z(\mathbf{X}_{AB}) \\ &= \theta_{AA, BB}^\top z(\mathbf{X}_{AA}, \mathbf{X}_{BB}) + \theta_{AA, AB}^\top z(\mathbf{X}_{AA}, \mathbf{X}_{AB}) + \theta_{BB, AB}^\top z(\mathbf{X}_{BB}, \mathbf{X}_{AB}) \\ &= \theta_{AA, BB, AB}^\top z(\mathbf{X}_{AA}, \mathbf{X}_{BB}, \mathbf{X}_{AB}) \end{aligned}$$

to explicitly allow for different dependencies depending on what edge-sets are considered. For example,  $z(\mathbf{X}_{AA})$  only involve statistics calculated on  $AA$  while  $z(\mathbf{X}_{AA, BB})$  involve crossed statistics, calculated for ties in  $\binom{A}{2} \times \binom{B}{2}$ . With multiple relations, statistics can be further partitioned, so that the linear predictors take into account dependencies between different types of ties between different types of nodes. Considering for example the interactions between ties in  $AA$  and  $AB$ , we have

$$\theta_{AA, AB}^\top z(\mathbf{X}_{AA}, \mathbf{X}_{AB}) = \sum_{s=0}^{R_{AA}-1} \sum_{t=0}^{R_{AB}-1} \theta_{AA, AB, st}^\top z(\mathbf{X}_{AA}^{(s)}, \mathbf{X}_{AB}^{(t)})$$

The interpretation is that a tie of type  $s$  among pairs in  $AA$  may depend on affiliation of nodes in  $A$  with nodes in  $B$  of type  $t$ . Conditional on a realisation  $\mathbf{X}$ , we assume an observation process

$$f(\mathbf{D}|\mathbf{X}, \zeta)\pi(\zeta)$$

where the parameter  $\zeta$  is distinct (Little & Rubin, 1987) from  $\theta$ . The observation process may be thought of equivalently as a missing data generating mechanism or a sampling design, such as snowball sampling, for purposes of inference (Handcock and Gile, 2010). If we assume that tie-variables are observed conditionally independently conditional on  $\mathbf{X}$ ,  $f(\cdot)$  can be modelled as a regular log-linear model with a standard link function. Given that  $\mathbf{D}$  has the same range-space  $\mathcal{X}$  as  $\mathbf{X}$ , the observation indicators can also be modelled using an ERGM. Inference for an informative, MNAR process will however be contingent on informative priors.

## 4 Estimation

We build on a recently proposed Bayesian data-augmentation scheme for doing inference for one-mode ERGM under the assumption of MAR (Koskinen et al., 2013). A Markov chain Monte Carlo (MCMC) scheme is constructed by drawing from the joint posterior of  $(\theta, \mathbf{X}^{\text{miss}}, \zeta)$  using updating steps that update from  $(\theta^{(t-1)}, \mathbf{X}^{\text{miss},(t-1)}, \zeta^{(t-1)})$  to  $(\theta^{(t)}, \mathbf{X}^{\text{miss},(t)}, \zeta^{(t)})$ . Conditional on  $\mathbf{D}$ ,  $\theta$  is updated using the approximate exchange sampler (Caimo and Friel, 2011):

- (a) Draw  $\eta$  from  $h(\eta|\theta^{(t-1)})$
- (b) Draw  $\mathbf{Y}$  from  $p(\mathbf{Y}|\eta) = \exp\{q(\mathbf{Y}; \eta) - \psi(\eta)\}$
- (c) With probability  $\min\{1, H\}$ , set  $\eta^{(t)} := \theta^{(t-1)}$  and  $\theta^{(t)} := \eta$  where

$$\begin{aligned}
 H &= \frac{p(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss},(t-1)} | \boldsymbol{\eta}) \pi(\boldsymbol{\eta}) h(\boldsymbol{\theta}^{(t-1)} | \boldsymbol{\eta}) p(\mathbf{Y} | \boldsymbol{\theta}^{(t-1)})}{p(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss},(t-1)} | \boldsymbol{\theta}^{(t-1)}) \pi(\boldsymbol{\theta}^{(t-1)}) h(\boldsymbol{\eta} | \boldsymbol{\theta}^{(t-1)}) p(\mathbf{y} | \boldsymbol{\eta})} \\
 &= \exp\{q(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss},(t-1)}; \boldsymbol{\eta}) + q(\mathbf{Y}; \boldsymbol{\theta}^{(t-1)}) \\
 &\quad - q(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss},(t-1)}; \boldsymbol{\theta}^{(t-1)}) - q(\mathbf{Y}; \boldsymbol{\eta})\} \pi(\boldsymbol{\eta}) / \pi(\boldsymbol{\theta}^{(t-1)})
 \end{aligned}$$

otherwise  $\boldsymbol{\eta}^{(t)} := \boldsymbol{\eta}$  and  $\boldsymbol{\theta}^{(t)} := \boldsymbol{\theta}^{(t-1)}$

In (a),  $h(\cdot)$  is a symmetric proposal distribution, typically a multivariate Gaussian distribution. In the exchange sampler (Murray et al., 2006), updating steps (a) and (b) are performed by drawing directly from the conditional distributions in a Gibbs update. Generally for ERGM (b) will have to be performed through MCMC, meaning that the algorithm for drawing from the posterior is not a proper MCMC scheme. Koskinen (2008) uses an on-line monitoring algorithm for appraising burn-in with properties similar to the automatic convergence guaranteed by perfect sampling (Propp and Wilson, 1996).

Koskinen et al. (2013) propose to update  $\mathbf{X}^{\text{miss}}$  under the assumption of missing at random (MAR) for one-mode networks. Whereas MAR implies

$$f(\mathbf{D} | \mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss}}, \boldsymbol{\zeta}) = f(\mathbf{D} | \mathbf{X}^{\text{obs}}, \boldsymbol{\zeta}),$$

we relax the assumption of MAR and allow for  $\mathbf{D}$  to depend on all of  $\mathbf{X}$ . The modification of the updating-step for missing data is to draw  $\mathbf{X}^{\text{miss}}$  given the rest from the full conditional posterior

$$\pi(\mathbf{X}^{\text{miss}} | \mathbf{X}^{\text{obs}}, \boldsymbol{\theta}) = \frac{\exp\{q(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss}}, \boldsymbol{\theta}) - \psi(\boldsymbol{\theta})\} f(\mathbf{D} | \mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss}}, \boldsymbol{\zeta}) \pi(\boldsymbol{\zeta})}{\sum_{\mathbf{Y}^{\text{miss}}} \exp\{q(\mathbf{X}^{\text{obs}}, \mathbf{Y}^{\text{miss}}, \boldsymbol{\theta}) - \psi(\boldsymbol{\theta})\} f(\mathbf{D} | \mathbf{X}^{\text{obs}}, \mathbf{Y}^{\text{miss}}, \boldsymbol{\zeta}) \pi(\boldsymbol{\zeta})}$$

The conditional distribution of  $\boldsymbol{\zeta}^{(t)}$  simplifies to a distribution proportional to  $f(\mathbf{D} | \mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss}}, \boldsymbol{\zeta}) \pi(\boldsymbol{\zeta})$ . If the distribution  $f(\cdot)$  is not fully tractable, draws of  $\boldsymbol{\zeta}$  cannot be made directly. Assuming that it is straightforward to draw  $\mathbf{D}$  from  $f(\cdot)$ ,  $\boldsymbol{\zeta}$  can be updated using steps (a), (b) and (c), with  $f(\cdot)$  playing the role of  $p(\cdot)$

## 5 Empirical illustration

We provide a brief empirical case-study using the so-called ‘Noordin Top’ Terrorist Network (Everton, 2012) as our assumed true network. The node set  $A$  consists of  $n = 79$  individuals and  $B$  of  $m = 129$  recorded events. The friendship ties reported in Everton serve as the ties in  $AA$  and the participation in events and operations listed by Roberts and Everton (2011) are the ties of the affiliation set  $AB$ . To construct ties  $BB$  among events, we have elaborated on the time-stamped version of Broccatelli, Everett and Koskinen (2016) and coded up the explicitly mentioned connections between different events and operations in the International Crisis Group Report

(International Crisis Group, 2006). For the purposes of illustration, the event-by-event network is considered fixed and exogenous. Furthermore, we condition on the overall activity of the network, fixing the number of ties in both  $AA$  and  $BB$ . Consequently, all analyses have to be interpreted conditionally on the overall number of event participations and total number of friendship ties.

The configurations  $z(\cdot)$  are illustrated in Figure 1 and are described in more detail in Wang et al. (2014). For the completely observed network, summaries of the posteriors for the corresponding parameters are provided in Table 2. Typical for one-mode network we find strong support for triadic closure (the 95% CI for ATA is (0.341, 1)) but also strong support for people taking part in events that are functionally related to other events that they take part in (the 95% CI for ATA is (0.786, 1.859)).

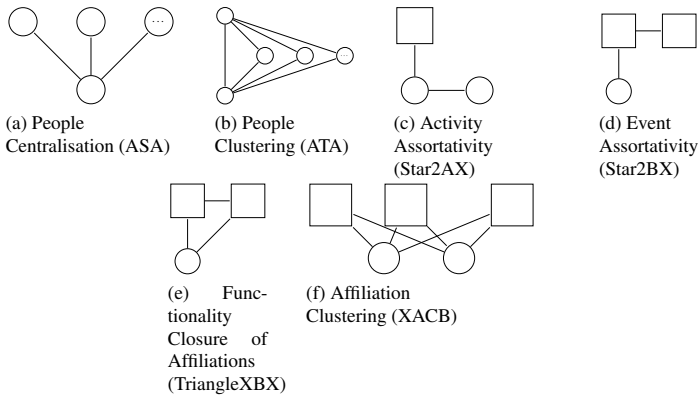


Fig. 1: Configurations of multilevel ERGM for Noordin Top (configurations a, b, and f are geometrically weighted)

To provide an example of multilevel snowball sampling, we snowball using Operation 3 as our seed (this is the 2004 Australian embassy bombing that took place on 9 September 2004 in Jakarta, Indonesia, killing 9-11 people and injuring more than 150 people). Anyone who participated in this operation is defined as being in wave 1, and anyone who is not in wave 1 but is tied to anyone in wave 1, belongs to wave 2. The result in Table 2 are qualitatively the same as for the model with completely observed data.

To provide a brief example of a MNAR observation process, for each tie-variable  $(i, j)$ , we define independently  $\Pr(D_{ij} = 1 | \mathbf{X}, \zeta) = \Pr(D_{ij} = 1 | h_{ij}(\mathbf{X}), \zeta)$ , where  $h_{ij}(\mathbf{X}) = \max\{d_i(\mathbf{X}), d_j(\mathbf{X})\}$ , where  $d_i(\mathbf{X})$  is the distance in  $\mathbf{X}$  between  $i \in A, B$  and Noordin Top (all ties in  $BB$  are assumed fixed and known). We model the probabilities  $\Pr(D_{ij} = 1 | h_{ij}(\mathbf{X}), \zeta)$  as in Table 1, with the interpretation that ties that are further from the leader Noordin Top are less visible than ties close to him. The results in Table 1 indicate that effects corresponding to clustering is attenu-

ated but degree-related effects are amplified (with the exception of XASA). These changes are a natural consequence of the observation process respecting distance but not necessarily clustering.

Table 1: Detection bias in MNAR observation mechanism for Noordin Top

$h_{ij}(x)$	1	2	3	4	> 4
$n_h(x)$	1122	6360	6090	1190	1750
$\Pr(D_{ij} x)$	0.99	0.75	0.5	0.25	0.15

Table 2: Posterior summaries for ERGM fitted to Noordin Top

Effect	no missing		snowball sample		MNAR	
	Mean	Std	Mean	Std	Mean	Std
ASA	0.162	0.215	0.160	0.229	0.662	0.264
ATA	0.673	0.169	0.637	0.177	0.29	0.201
Star2AX	0.106	0.020	0.106	0.020	0.129	0.021
Star2BX	-0.014	0.046	0.000	0.049	0.022	0.06
TriangleXBX	1.322	0.273	1.299	0.278	1.191	0.293
XASA	0.185	0.205	0.337	0.213	0.037	0.212
XACB	0.106	0.029	0.091	0.035	0.069	0.046

## 6 Conclusions and future directions

We have proposed a statistical approach for analysing the structure of multilayered networks that account for imperfections in data. We provide an illustrative example of analysis of a multilevel network for three types of observation processes. While the approach is consistent when the observation process is known, a MNAR process requires making a number of untestable assumptions and is most likely of use merely as a sensitivity analysis. Further work is needed in order to systematically investigate the sensitivity of MNAR to different plausible MNAR mechanisms.

**Acknowledgements** The work of Koskinen and Broccatelli is funded by the Leverhulme Trust Grant RPG-2013-140.

## References

1. Agneessens, F., Roose, H.: Local Structural Properties and Attribute Characteristics in 2-mode Networks:  $p$  Models to Map Choices of Theater Events. *The Journal of Mathematical Sociology* 32 (3), 204–237 (2008).
2. Broccatelli, C., Everett, M., and Koskinen, J.: Temporal dynamics in covert networks. *Methodological Innovations*, 9: 1-14 (2016).
3. Caimo, A., Friel, N.: Bayesian inference for exponential random graph models. *Social Networks* 33, 41–55 (2011).
4. Everton, S. F.: *Disrupting Dark Networks*, Cambridge University Press, NY. (2012)
5. Frank, O., Strauss, D.: Markov Graphs. *Journal of the American Statistical Association* 81 (395), 832?842 (1986)
6. Handcock, M.S., Gile, K.: Modeling networks from sampled data. *Annals of Applied Statistics* 4, 5–25 (2010)
7. International Crisis Group: *Terrorism in Indonesia: Noordin's Networks*, Kinshasa/Nairobi/Brussels (2006)
8. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A.: Multi-layer networks, *Journal of Complex Networks*, 2(3), 203–271 (2014)
9. Koskinen, J.: The Linked Importance Sampler Auxiliary variable (LISA) Metropolis-Hastings algorithm for distributions with intractable normalising constants. MelNet Social Networks Laboratory Technical Report 08-01, Department of Psychology, University of Melbourne, Australia (2008)
10. Koskinen, J. H., Robins, G. L., Wang, P., & Pattison, P. E.: Bayesian analysis for partially observed network data, missing ties, attributes and actors. *Social Networks*, 35(4), 514–527 (2013).
11. Lazega, E., Jourda, M.-T., Mounier, L., Stofer, R.: Catching up with big fish in the big pond? Multi-level network analysis through linked design. *Social Networks* 30 (2), 159–176 (2008).
12. Little, R.J.A., and Rubin, D.B.: *Statistical Analysis with Missing Data*. New York: Wiley (1987).
13. Moreno, J., Jennings, H.: Statistics of Social Configurations. *Sociometry* 1 (3/4), 342–374 (1938).
14. Murray, I., Ghahramani, Z., MacKay, D.: MCMC for doubly-intractable distributions. In: *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*. AUAI Press, Arlington, Virginia (2006)
15. Pattison, P. and Wasserman, S.: Logit models and logistic regressions for social networks: II. Multivariate relations. *Brit. J. Math. Stat. Psych.*, 52:169–193 (1999)
16. Propp, J., and Wilson, D.: Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics. *Random Structures and Algorithms*, 9, 232-252 (1996).
17. Roberts, Nancy, and Sean F. Everton: Strategies for Combating Dark Networks. *Journal of Social Structure* 12 (2) (2011)
18. Skvoretz, J., Faust, K.: Logit models for affiliation networks. *Sociological Methodology* 29 (1), 253–280 (1999)
19. Snijders, T. A. B., Pattison, P. E., Robins, G. L., Handcock, M. S.: New Specifications for Exponential Random Graph Models. *Sociological Methodology* 36 (1), 99?153 (2006).
20. Wang, P., Sharpe, K., Robins, G. L., Pattison, P. E.: Exponential random graph ( $p$ ) models for affiliation networks. *Social Networks* 31, 12–25 (2009).
21. Wang, P., Robins, G., Pattison, P., and Koskinen, J. (2014). MPNet, Program for the Simulation and Estimation of ( $p^*$ ) Exponential Random Graph Models for Multilevel Networks: USER MANUAL. Melbourne School of Psychological Sciences The University of Melbourne Australia (2014)
22. Wang, P., Robins, G. L., Pattison, P. E., Lazega, E.: Exponential random graph models for multilevel networks. *Social Networks* 35 (1), 96–115. (2013)
23. Wasserman, S., Jacobucci, D.: Statistical modelling of one mode and two mode networks: Simultaneous analysis of graphs and bipartite graphs. *British Journal of Mathematical and Statistical Psychology* 44 (1), 13–43 (1991).

# Copula-based segmentation of environmental time series with linear and circular components

## *Segmentazione di serie storiche ambientali con componenti lineari e circolari basata su copula*

Francesco Lagona

**Abstract** A novel segmentation method is proposed for the analysis of bivariate time series of intensities and angles that often occur in environmental applications. The model is based on a mixture of copula-based cylindrical distributions, whose parameters evolve according to a latent Markov chain. The model parsimoniously accommodates typical features of cylindrical time series such as circular-linear correlation, multimodality, skewness and temporal auto-correlation. A computationally efficient Expectation-Maximization algorithm is described to estimate the parameters and a parametric bootstrap routine is exploited to compute confidence intervals. These methods are illustrated on cylindrical time series of wave heights and directions in the Adriatic sea.

**Abstract** *Si propone una nuova procedura di segmentazione per serie storiche bivariate con componenti lineari e circolari, tipiche delle applicazioni di statistica ambientale. Il modello si basa su una mistura di distribuzioni cilindriche, definite attraverso una copula, i cui parametri variano secondo l'evoluzione di una catena markoviana latente. Il modello integra in modo parsimonioso caratteristiche tipiche delle serie cilindriche, quali la correlazione tra osservazioni lineari e circolari, l'asimmetria, la multimodalità e l'auto-correlazione temporale dei dati. Si propone un algoritmo di tipo EM computazionalmente efficiente per la stima dei parametri ed una procedura di tipo bootstrap per la stima degli intervalli di confidenza. L'applicazione a una serie storica di direzioni e altezze d'onda nell'Adriatico illustra la metodologia.*

**Key words:** clustering, copula, hidden Markov model, environmetrics, linear-circular data, segmentation, waves

---

Francesco Lagona  
University of Roma Tre, via G. Chiabrera 199, 00145 Rome,  
e-mail: francesco.lagona@uniroma3.it

## 1 Introduction

Time series of angles and intensities arise often in environmental research. Recent studies have focused, for example, on time series of wind directions and pollutant concentrations [2], wind directions and speeds [7], wave directions and heights [6]. Bivariate sequences of angles and intensities are often referred to as *cylindrical* time series, because the pair of an angle and an intensity can be represented as a point on a cylinder.

The analysis of cylindrical time series has been overlooked due to the special topology of the support on which the measurements are taken (the cylinder), and to the difficulties in modeling the cross-correlations between angular and linear measurements over time. Further complications arise from the skewness and the multimodality of the marginal distribution of the data. Indeed, intensities are typically negatively skewed and directional data are rarely symmetric; multimodality may arise as well as the data are often observed under heterogeneous conditions that vary over time.

This paper introduces a dynamic mixture of copula-based cylindrical distributions that parsimoniously accounts for the specific features of cylindrical time series. More precisely, we first introduce a cylindrical density as a joint distribution of a von Mises and a Weibull distribution by means of a circular copula and then approximate the data distribution with a mixture of these cylindrical densities, whose parameters depend on the states of a latent Markov chain. This approach flexibly extends previous proposals that are either based on mixtures of conditionally independent linear and circular densities [4] or based on mixtures of Abe-Ley cylindrical densities [6]. It provides a unified framework where distributions that are typically exploited in environmental applications are jointly integrated. It is additionally numerically tractable, by exploiting a suitable Expectation-Maximization (EM) algorithm for parameter estimation.

## 2 A copula-based cylindrical distribution

A cylindrical sample is a pair  $\mathbf{z} = (x, y)$ , where  $x \in [0, 2\pi)$  is a point in the circle and  $y$  is a point on the positive semi-line  $[0, +\infty)$ . Let  $I_0$  be the modified Bessel function of order 0. If the circular component  $x$  and the linear component  $y$  are respectively drawn from a von Mises distribution with density

$$f(x; \mu, \kappa) = \frac{\exp(\kappa \cos(x - \mu))}{2\pi I_0(\kappa)},$$

and from a Weibull distribution with density

$$f(y; \alpha, \beta) = \frac{\alpha}{\beta} \left(\frac{y}{\beta}\right)^{\alpha-1} \exp\left(-\frac{y}{\beta}\right)^\alpha,$$



and if  $g(u)$  is a von Mises density with parameters  $\mu_c$  and  $\kappa_c$ , say

$$g(u; \mu_c, \kappa_c) = \frac{\exp(\kappa_c \cos(u - \mu_c))}{2\pi I_0(\kappa_c)},$$

then the copula-based density

$$f(\mathbf{z}; \boldsymbol{\theta}) = 2\pi g(2\pi(F(x) + F(y))) f(x)f(y) \tag{1}$$

is defined on a cylindrical support up to the parameter vector  $\boldsymbol{\theta} = (\mu, \kappa, \alpha, \beta, \mu_c, \kappa_c)$  and has  $f(x; \mu, \kappa)$  and  $f(y; \alpha, \beta)$  as marginal densities [3].

### 3 Dynamic mixtures of cylindrical densities

Let  $\mathbf{z}_{0:T} = (\mathbf{z}_t, t = 0, \dots, T)$ ,  $\mathbf{z}_t = (x_t, y_t)$ ,  $x_t \in [0, 2\pi)$ ,  $y_t \in [0, +\infty)$  be a cylindrical time series. We assume that the distribution of the data is driven by the evolution of an unobserved Markov chain with  $K$  states, which represents (time-varying) latent classes and can be specified as a sequence  $\boldsymbol{\xi}_{0:T} = (\boldsymbol{\xi}_t, t = 0, \dots, T)$  of multinomial variables  $\boldsymbol{\xi}_t = (\xi_{t1} \dots \xi_{tK})$  with one trial and  $K$  classes, whose binary components represent class membership at time  $t$ . The joint distribution  $p(\boldsymbol{\xi}_{0:T}; \mathbf{p})$  of the chain is fully known up to a parameter  $\mathbf{p}$  that includes  $K$  initial probabilities  $p_k = P(\xi_{0k} = 1), k = 1, \dots, K, \sum_k p_k = 1$ , and  $K^2$  transition probabilities  $p_{hk} = P(\xi_{tk} = 1 | \xi_{t-1,h} = 1), h, k = 1, \dots, K, \sum_k p_{hk} = 1$ . Formally, we assume that

$$p(\boldsymbol{\xi}_{0:T}; \mathbf{p}) = \prod_{k=1}^K p_k^{\xi_{0k}} \prod_{t=1}^T \prod_{h=1}^K \prod_{k=1}^K p_{hk}^{\xi_{t-1,h} \xi_{tk}}. \tag{2}$$

The specification of the model is completed by assuming that the observations are conditionally independent, given a realization of the Markov chain. As a result, the conditional distribution of the observed process, given the latent process, takes the form of a product density, say

$$f(\mathbf{z}_{0:T} | \boldsymbol{\xi}_{0:T}; \boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_K) = \prod_{t=0}^T \prod_{k=1}^K f(\mathbf{z}_t; \boldsymbol{\theta}_k)^{\xi_{tk}}, \tag{3}$$

where  $f(\mathbf{z}; \boldsymbol{\theta}_k), k = 1, \dots, K$  are the  $K$  cylindrical densities defined by (1) and known up to a vector of parameters  $\boldsymbol{\theta}_k$ . The likelihood function of the model is therefore obtained by integrating the joint density of the observed data and the unobserved class memberships with respect to the segmentation  $\boldsymbol{\xi}_{0:T}$ , namely

$$L(\mathbf{p}, \boldsymbol{\theta}; \mathbf{z}_{0:T}) = \sum_{\boldsymbol{\xi}_{0:T}} p(\boldsymbol{\xi}_{0:T}; \mathbf{p}) f(\mathbf{z}_{0:T} | \boldsymbol{\xi}_{0:T}; \boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_K). \tag{4}$$

By computing the maximum likelihood estimate  $\hat{\theta}$ , the cylindrical time series can be then segmented according to the posterior probabilities of class membership  $\hat{p}_{tk} = P(\xi_{tk} = 1 | \mathbf{z}_{0:T}; \hat{\theta})$ , based on  $\hat{\theta}$ .

### 4 Estimation

An EM algorithm is proposed to maximize the likelihood function (4). It is based on the following complete-data log-likelihood function

$$\begin{aligned} \log L_{\text{comp}}(\theta, \xi_{0:T}, \mathbf{z}_{0:T}) &= \sum_{k=1}^K \xi_{0k} \log p_k + \sum_{t=1}^T \sum_{h=1}^K \sum_{k=1}^K \xi_{t-1,h} \xi_{t,k} \log p_{hk} \\ &+ \sum_{t=0}^T \sum_{k=1}^K \xi_{tk} \log f(\mathbf{z}_t; \theta_k). \end{aligned} \tag{5}$$

The algorithm is iterated by alternating an expectation (E) and a maximization (M) step. Given the estimates  $\hat{p}_s$  and  $\hat{\theta}_s$ , obtained at the end of the  $s$ -th iteration, the  $(s + 1)$ -th iteration is initialized by the E-step, which evaluates the expected value of (5) with respect to the conditional distribution of the missing values  $\xi_{tk}$  given the observed data.

The E step reduces to the computation of the univariate posterior probabilities of each latent state at time  $t$ ,  $\hat{p}_{tk} = P(\xi_{tk} = 1 | \mathbf{z}_{0:T}, \hat{p}_s, \hat{\theta}_s)$   $k = 1 \dots K, t = 0 \dots T$ , and the computation of the bivariate posterior probabilities of each pair of states in two adjacent times, say  $\hat{p}_{t-1,t,hk} = P(\xi_{t-1,h} = 1, \xi_{t,k} = 1 | \mathbf{z}_{0:T}, \hat{p}_s, \hat{\theta}_s)$   $h, k = 1 \dots K, t = 1 \dots T$ . The task of computing these posterior probabilities from an estimate  $(\hat{p}_s, \hat{\theta}_s)$  is generally referred to as the HMM-smoothing numerical issue and it is typically solved by specifying the posterior probabilities in terms of suitably normalized functions, which can be computed recursively, avoiding unpractical summations over the state space of latent Markov chain and numerical under- and over-flows [1].

The M-step of the algorithm updates the estimate  $(\hat{p}_s, \hat{\theta}_s)$  with a new estimate  $(\hat{p}_{s+1}, \hat{\theta}_{s+1})$ , by maximizing the expected value of (5), obtained from the previous E step. This expected value is the sum of functions that depend on independent sets of parameters and can therefore be maximized separately. Maximization with respect to the transition probabilities  $p_{hk}$ , under the constraints  $\sum_{k=1}^K p_{hk} = 1, h = 1 \dots K$ , provides the closed-form updating formula

$$\hat{p}_{hk(s+1)} = \frac{\sum_{t=1}^T \hat{p}_{t-1,t,hk}(\hat{p}_s, \hat{\theta}_s)}{\sum_{t=1}^T \hat{p}_{t-1,h}(\hat{p}_s, \hat{\theta}_s)}, \quad h, k = 1, \dots, K.$$

Maximization with respect to the parameters  $\theta_k$  of the  $K$  copula-based cylindrical components reduces to a traditional IFM (inference function for margins; [5]) routine, implemented on a weighted augmented datasets of  $n \times K$  observations, where

each observation is replicated  $K$  times and weighted by the  $K$  univariate posterior probabilities  $\hat{p}_{ik}$ , computed during the previous E step.

The procedure outlined above does not produce confidence intervals of the estimates, which however can be computed by taking a parametric bootstrap approach. In this paper, the model was re-fitted from  $R = 400$  bootstrap samples, simulated from the estimated model parameters, and the 2.5% and the 97.5% quantiles of the empirical distribution of each bootstrap estimate was computed.

**Table 1** Parameter estimates and bootstrap quantiles under three segmentation classes

Class 1 parameter estimate			
	2.5% quantile	97.5% quantile	
$\alpha$	2.61	2.16	2.75
$\beta$	0.81	0.72	0.86
$\kappa$	1.23	0.80	1.31
$\mu$	1.13	1.05	1.19
$\kappa_c$	0.98	0.89	1.08
$\mu_c$	0.52	0.49	0.62

Class 2 parameter estimate			
	2.5% quantile	97.5% quantile	
$\alpha$	2.41	2.26	2.58
$\beta$	2.99	2.81	3.17
$\kappa$	0.26	0.21	0.60
$\mu$	0.22	0.11	3.81
$\kappa_c$	2.18	1.98	3.19
$\mu_c$	0.72	0.49	0.92

Class 3 parameter estimate			
	2.5% quantile	97.5% quantile	
$\alpha$	2.18	2.05	3.26
$\beta$	2.10	1.92	2.47
$\kappa$	1.95	1.48	2.72
$\mu$	2.03	1.87	2.11
$\kappa_c$	4.01	2.98	5.01
$\mu_c$	0.12	0.04	1.01

origin	destination		
	Class 1	Class 2	Class 3
Class 1	0.975	0.018	0.007
Class 2	0.000	0.990	0.010
Class 3	0.010	0.016	0.974

## 5 Application: regimes of wave in the Adriatic sea

The proposed methods have been implemented to segment a time series of semi-hourly wave directions and heights, recorded in the period 2/15/2010 - 3/16/2010 by the buoy of Ancona, located in the Adriatic Sea at about 30 km from the coast. Segmentation of these data according to meaningful environmental regimes is often

required in studies of the drift of floating objects and oil spills, in the design of off-shore structures and in studies of sediment transport and coastal erosion.

A number of models have been estimated from these data, by varying the number  $K$  of components from 2 to 4, and the BIC statistic suggested to segment the data according to 3 regimes. Table 1 displays the estimates under these three latent classes, along with bootstrap percentiles, computed by simulating 400 samples.

The first component of the model (class 1) is associated with high waves coming from North. These waves are generated by northern Bora jets that blow along the major axis of the Adriatic basin. Under a Bora episode, most of the wind energy is transferred to the sea surface and, as a result, most of the data with the highest waves in the sample are clustered within this regime. The second component of the model (class 2) is associated with periods of calm sea. Under this regime, moderate waves are uniformly distributed around the circle of directions. The third component (class 3) is associated with Sirocco episodes. In this regime, waves travel southeasterly along the major axis of the Adriatic basin, driven by winds that blow from a similar directional angle.

The rows at the bottom of Table 1 include the estimated transition probabilities of the latent Markov chain. The transition probability matrix is essentially diagonal, reflecting the temporal persistence of the classes. In particular, the small off-diagonal transition probabilities between class 1 and 3 indicate that direct transitions between Sirocco and Bora episodes are very unlikely. The segmentation model hence confirms that the sea surface in the study area tend to alternate relevant marine events with periods of good sea conditions.

**Acknowledgements** This work is developed under the PRIN2015 supported-project "Environmental processes and human activities: capturing their interactions via statistical methods (EPHA-STAT)' funded by MIUR (Italian Ministry of Education, University and Scientific Research)"

## References

1. Bulla J, Lagona F, Maruotti A, Picone M (2012) A Multivariate Hidden Markov Model for the Identification of Sea Regimes from Incomplete Skewed and Circular Time Series. *Journal of Agricultural, Biological and Environmental Statistics*, 17: 544-567.
2. Garcia-Portuguès E, Crujeiras RM, González-Manteiga W (2013) Exploring wind direction and SO<sub>2</sub> concentration by circular-linear density estimation. *Stochastic Environmental Research and Risk Assessment* 27(5): 1055-1067.
3. Johnson RA, Wehrly TE (1978) Some angular-linear distributions and related regression models, *Journal of the American Statistical Association* 73: 602-606.
4. Holzmann H, Munk A, Suster M, Zucchini W (2006) Hidden Markov models for circular and linear-circular time series. *Environmental and Ecological Statistics* 13: 325-347.
5. Kim, G. Silvapulle, M.J., Silvapulle, P. (2007) Comparison of semiparametric and parametric methods for estimating copulas, *Computational Statistics & Data Analysis* 51: 2836-2850.
6. Lagona, F. Picone, M. and Maruotti, A. (2015) A hidden Markov model for the analysis of cylindrical time series. *Environmetrics* 26: 534-544.
7. Mastrantonio G, Maruotti A, Jona Lasinio G. (2015) Bayesian hidden Markov modelling using circular-linear general projected normal distribution. *Environmetrics* 26(2): 145-158.

# A Multiscale Approach to Manifold Estimation

## *Un Approccio Multiscala alla Stima di Varietà*

Alessandro Lanteri and Mauro Maggioni

**Abstract** In presence of high-dimensional data, sampled from an unknown distribution  $\mu$  on  $\mathbb{R}^D$ , it is common to assume that the support of  $\mu$  is approximately a  $d$ -dimensional set, for example the  $d$ -dimensional Riemannian manifold. We introduce a novel technique to estimate underlying structure of the data using an algorithm which approximates the manifold with a collection of hyperplanes. This is done in a multiscale fashion, using a subspace clustering algorithm iteratively. The proposed approach is data-adaptive and, by construction, provides a tree structure for the data. The performance of the proposed method is evaluated both with synthetic and real data showing promising results.

**Abstract** *In presenza di dati ad alta dimensionalità, estratti da una distribuzione  $\mu$  sconosciuta in  $\mathbb{R}^D$ , è usuale assumere che il supporto di  $\mu$  sia approssimativamente un set  $d$ -dimensionale, ad esempio una varietà Riemanniana  $d$ -dimensionale. In questo lavoro viene introdotta una nuova tecnica per stimare la struttura sottostante ai dati usando un algoritmo in grado di approssimare la varietà con una collezione di iperpiani. Il metodo viene eseguito in maniera multiscala utilizzando iterativamente un algoritmo di subspace clustering. L'approccio proposto è data-adattivo e, per costruzione, genera una struttura ad albero per i dati. Il metodo viene testato tramite l'utilizzo di dati sintetici e reali e produce risultati promettenti.*

**Key words:** Manifold Learning, Dictionary Learning, Multi-Resolution Analysis, Adaptive Approximation, Data Encoding

---

Alessandro Lanteri

Department of Applied Mathematics and Statistics, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA e-mail: alanter2@jhu.edu

Mauro Maggioni

Department of Applied Mathematics and Statistics, Department of Mathematics, The Institute for Data Intensive Engineering and Science, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA e-mail: mauro.maggioni@jhu.edu

## 1 Introduction

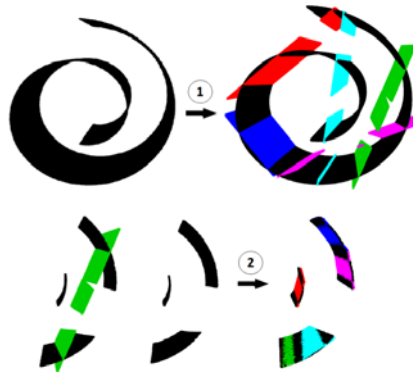
In recent years, due to the huge amount of data that new technologies provide, modern science has become dependent on reliable methods to deal with high-dimensional data. Several difficulties arise when dealing with this kind of data. One of the main problems is that when data dimension grows, the volume of the space grows exponentially, leading to the available data being scattered sparsely in space, with gaps between data points growing also exponentially in the dimension. This behavior may cause problems in statistical analysis, since the data needed to give statistical significance to the analysis usually grows exponentially with the dimension. This problem is commonly referred to as the *curse of dimensionality* and it is the main reason why high-dimensional data can not be analyzed efficiently with traditional methods. To tackle this issue new techniques for dimensionality reduction have been proposed in literature. Most of them are based on the assumption that there is a low-dimensional model which approximates properly the true high-dimensional distribution from which the data was generated. One of the most used methods which exploits a geometrical assumption on the data is principal component analysis (PCA) [10]. Notwithstanding its popularity in applications, the assumption of PCA that data lies close to a linear variety is often unsatisfied. In the past decade much work has focused on replacing the linear variety assumption with that of a nonlinear manifold  $\mathcal{M}$  in  $\mathbb{R}^D$  [5, 11, 2, 6, 4, 9]. In this work we focus on multiscale techniques for manifold learning and dictionary learning, in the same direction as Geometric Multi-Resolution Analysis [1, 9, 7].

## 2 The algorithm

Let  $X := \{x_i\}_{i=1}^n$  be a set of  $n$  samples from a probability measure  $\mu$  in  $\mathbb{R}^D$ . We assume that  $\mu$  is supported near a set  $\mathcal{M}$ , e.g. a manifold, of dimension  $d \ll D$ . Here for “near” we mean that the data may be corrupted by noise, or perhaps the data is not quite distributed on a manifold, but close to one (model error) [9]. Our goal is to learn a data-dependent dictionary that efficiently encodes data sampled from  $\mu$ . We proceed in a multiscale fashion, in order to have approximation at different scales, with increasing accuracy as the scales get finer. We describe our procedure in Algorithm 1. The intuition behind our method is that a Riemannian manifold can be locally approximated by a  $d$ -dimensional plane, thus the underlying structure of the data can be approximated by a suitable collection of planes. The main tool used in our method is the MAPA algorithm [3], a subspace clustering algorithm proposed to solve the plane arrangement problem using a Multiscale Singular Value Decomposition approach [8]. Given a set of samples lying on a collection of  $d$ -dimensional planes with different dimension, MAPA can reconstruct the model estimating the number of planes, their dimension and how they are arranged in the ambient space. However, when MAPA is applied to data sampled from a manifold  $\mathcal{M}$ , it fails to approximate properly the manifold because of curvature. Still, the

algorithm will produce a coarse approximation of  $\mathcal{M}$  using up to  $K$   $d$ -dimensional affine planes  $\{\pi_k\}_{k=1}^K$ . The data  $X$  may be clustered into  $K$  disjoint subsets  $\{X_{1,k}\}_{k=1}^K$  using the minimal point-plane distance, i.e. assigning  $x$  to  $\operatorname{argmin}_k d(x, P_k)$ , with  $P_k$  the orthogonal projection onto the affine subspace  $\pi_k$ . In rMAPA we now recurse on each of these subsets: MAPA is applied independently on each  $X_{1,k}$ , generating a further, finer scale family of subsets  $X_{2,k}$ , each approximated by another set of  $d$ -dimensional planes. Figure 1 provides a pictorial representation of this construction. This process can be iterated until a desired precision in the estimation of  $\mathcal{M}$  is achieved: in this way the process is made adaptive, meaning that if a region of the manifold is irregular it will be approximated by several planes, while a rather flat region will be approximated by a moderate number of planes. Examples of full manifold reconstruction are shown in Figure 2. This approach, in comparison to a uniform approach, leads to a reduction of the number of planes used to approximate the manifold while maintaining the same level of precision and, of course, decrease the running time of the algorithm. Another feature of the proposed method is that it produces a tree structure for the data. Each scale of approximation corresponds to a level of the tree, with each cluster represented by a tree node.

**Fig. 1** The multiscale nature of the Algorithm 1. In the first step, the Swissroll manifold is roughly approximated by five planes. In the second step, the manifold is divided in five subsets and each of them (here we show only one subset) is again approximated by another set of five planes in order to get a finer approximation. This process may be iterated as long as the number of points permits it, or until the desired precision is achieved.




---

**Algorithm 1** rMAPA

---

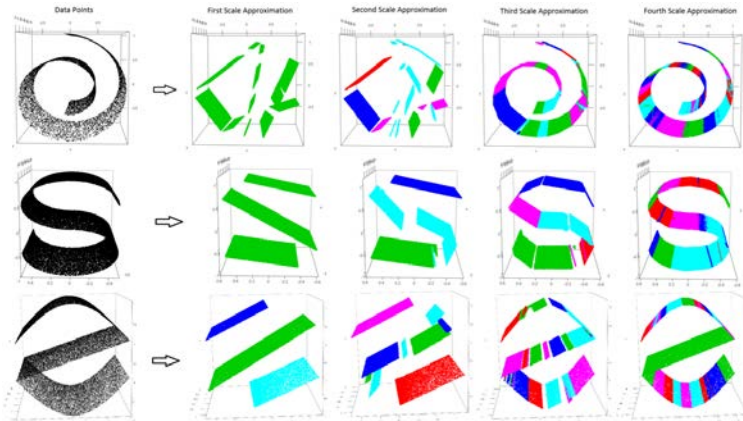
**Require:** data  $X$ , intrinsic dimension  $d$ , precision  $\kappa$ .

**Ensure:**  $\hat{P}_{s,k}$  : piecewise linear projectors  $k = 1, \dots, K_s$  for each scale  $s = 1, \dots, S$ .

- 1: Apply MAPA to  $X$  and obtain  $K_1$  piecewise linear projector  $\hat{P}_{1,j}$ , with  $j = 1, \dots, K_1$ .
  - 2: Form  $K_1$  clusters  $X_{1,j}$ , with  $j = 1, \dots, K_1$  from  $X$  using the minimal point-plane distance.
  - 3: Apply MAPA to the obtained clusters in order to obtain  $K_s$  clusters and their piecewise linear projections.
  - 4: Repeat Step 3 until, at scale  $S$ ,  $\|X_{S,k} - \hat{P}_{s,k}\|^2 \leq \kappa$  is obtained for each pair  $(S, k)$ .
-

### 3 Experiments

In this section we show some empirical results on the performance of the proposed algorithm on synthetic data. We draw  $10^6$  points uniformly on a  $d$ -dimensional manifold. These points are then embedded and randomly rotated in  $\mathbb{R}^D$  with  $D = 30$ . Each point is corrupted with additive Gaussian noise  $\eta_i \sim \frac{\sigma}{\sqrt{D}}N(0, I_D)$ . The method is tested in different settings, varying the manifold type, a Swissroll or an S-Manifold, and the noise level  $\sigma$ . In all settings we requested the precision parameter  $\kappa$  to be equal to  $\sigma$  in order to avoid overfitting. Results in Table 1 show that the methods always achieve a precision higher than the requested  $\kappa$ , this means that it acted as a denoiser in these settings. The number of planes used is rather small and the computational time is reasonable compared to other methods. For instance, we ran LLE [11], a standard algorithm for dimensionality reduction, on a Swissroll with only  $10^4$  point embedded in  $\mathbb{R}^5$  with  $\sigma = 10^{-4}$  and it took about 2 hours to reconstruct the manifold, while our method took an average of 11 seconds to complete the same task, on a data matrix six hundred times bigger than the one used on the LLE algorithm. From this experiment we also note that the number of planes needed to estimate the S-Manifold is lower than those needed for the Swissroll to achieve the same error. This follows from the fact that, as it is shown in Figure 2, the S-Manifold have more flat regions than the Swissroll and the algorithm exploits this feature and optimizes the number of planes needed for the approximation.



**Fig. 2** Reconstruction of three different manifold, a Swissroll, an S-Manifold and an SZ-Manifold, using Algorithm 1. Here the adaptive nature of the algorithm can be appreciated. This can be appreciated in particular at the last scale of the SZ-Manifold approximation, where the flat region is well approximated by a single plane, while the curvy regions are approximated by several planes. This feature is more visible at last scale approximation because at lower scales there is not yet a single plane which alone can approximate the flat region better than a union of planes.



**Table 1** Approximation error (MSE) in different settings. A Swissroll and an S-Manifold,  $n = 10^6$ ,  $D = 30$ , Gaussian noise  $\eta_i \sim \frac{\sigma}{\sqrt{D}}N(0, I_D)$  with  $\sigma \in \{0.001, 0.05, 0.1\}$  and precision  $\kappa = \sigma$ . Results are averaged over 100 iterations on a test set of sample size  $n$ . Values in parentheses are standard errors. For the sake of readability, MSE values are multiplied by  $10^3$ .

$\sigma \times 10^3$		Swissroll	S-Manifold
1	MSE $\times 10^3$	0.19 (0.13)	0.34 (0.31)
	Planes <sup>a</sup>	25.23 (9.56)	14.12 (4.83)
	Time <sup>b</sup>	70.22 (24.41)	41.25 (23.05)
5	MSE $\times 10^3$	1.70 (1.10)	3.07 (0.96)
	Planes <sup>a</sup>	19.14 (10.96)	6.85 (2.54)
	Time <sup>b</sup>	48.82 (29.17)	5.28 (9.77)
10	MSE $\times 10^3$	7.30 (2.27)	7.63 (0.96)
	Planes <sup>a</sup>	9.38 (8.02)	5.31 (1.27)
	Time <sup>b</sup>	10.09 (21.39)	2.23 (3.39)

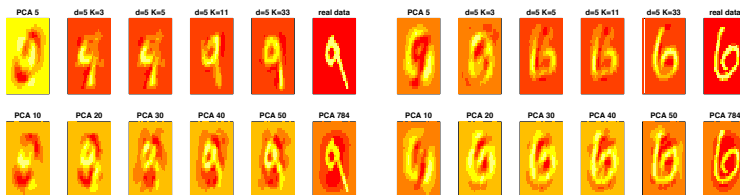
<sup>a</sup> Number of linear projectors used for approximation. <sup>b</sup> Running time in seconds.

## 4 Application to Real Data

We consider the MNIST data set from <http://yann.lecun.com/exdb/mnist/>, which contains images of handwritten digits, each of size  $28 \times 28$ , grayscale (i.e. in dimension 784). Sample size is 60000 for the train set and 10000 for the test set. This data set is very interesting since its intrinsic dimension varies for different digits and across scales, as it was observed in [8]. In Figure 3, we compare our method to a projection on the first several principal component of the train set. To approximate a digit in our multiscale approach, we project the point (i.e. the digit) on the 5-dimensional plane closer, in orthogonal projection. We chose  $d = 5$  because it has been shown in [8] and [7] that the dimension of each digit should not exceed this number. In Figure 3 it is appreciable how we can obtain good approximation, encoding data using several low-dimensional planes instead of one high-dimensional plane. Like we mentioned before, this feature is very important to tackle the curse of dimensionality when doing inference.

## 5 Conclusion

We proposed a data-adaptive multiscale algorithm for manifold learning which exploits a subspace clustering method. The algorithm also produces a tree structure for the data, which is a useful feature per se. Numerical experiments show that this methods is very fast to handle a big number of samples embedded in high dimension and succeeds in giving the requested approximation error encoding the data with a collection of linear projectors, even when the data is corrupted with noise. The number of projectors is limited, this thanks to the adaptive nature of the algo-



**Fig. 3** Approximation of two digits, a nine and a six, from the MNIST test set. First row is the reconstruction with the proposed multiscale approximation with planes dimension  $d = 5$  and different number of planes  $K$ . In second row are shown results from the projection of the digits on the first several principal components of the train set. Note that a PCA 5 is equivalent to the proposed algorithm with  $d = 5$  and  $K = 1$ .

rithm, which automatically finds flat regions which can be approximated with less planes. An application to the MNIST data showed how the proposed algorithm succeeds to approximate and encode the digits with a collection of low-dimensional planes with an accuracy comparable to a high-dimensional encoder which uses the first several principal components.

## References

1. W.K. Allard, G. Chen, and M. Maggioni. Multi-scale geometric methods for data sets ii: Geometric multi-resolution analysis. *Applied and Computational Harmonic Analysis*, 32(3):435–462, 2012.
2. M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.
3. G. Chen and M. Maggioni. Multiscale geometric and spectral analysis of plane arrangements. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2825–2832, 2011.
4. R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *P Natl Acad Sci Usa*, 102(21):7426–7431, 2005.
5. V. De Silva, J.C. Langford, and J.B. Tenenbaum. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
6. D.L. Donoho and C. Grimes. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. *Proc. Nat. Acad. Sciences*, pages 5591–5596, March 2003.
7. W. Liao and M. Maggioni. Adaptive geometric multiscale approximations for intrinsically low-dimensional data. *arXiv preprint*, (arXiv:1611.01179), 2017.
8. A.V. Little, M. Maggioni, and L. Rosasco. Multiscale geometric methods for data sets i: Multiscale svd, noise and curvature. *Appl. Comput. Harmon. Anal.*, (<http://dx.doi.org/10.1016/j.acha.2015.09.009>), 2016.
9. M. Maggioni, S. Minsker, and N. Strawn. Dictionary learning and non-asymptotic bounds for the Geometric Multi-Resolution Analysis. *J.M.L.R.*, 2014.
10. K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.
11. S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

# Using scanner and CPI data to estimate Italian sub-national PPPs

*L'uso degli scanner data e dei dati dell'indagine*

*tradizionale sui prezzi al consumo per la stima delle PPA a livello intra-nazionale in Italia*

Tiziana Laureti, Carlo Ferrante, Barbara Dramis

## Abstract

The recent availability of high-frequency electronic-point-of-sale “scanner data” together with the traditional Consumer Price Index (CPI) data has the potential to significantly change how to compile spatial price indexes. Indeed, one of the main issues when constructing sub-national Purchasing Power Parities (PPPs) is to obtain price data from multiple sources and outlets, which are representative of local consumption patterns and comparable on the basis of a set of price determining characteristics.

Within this framework, the aim of this paper is to suggest a new stochastic methodological approach to index numbers based on the Country-Product-Dummy (CPD) method for computing sub-national PPPs in Italy. This approach enables us to use both scanner data and traditional CPI data coming from large-scale retail trade.

By using millions of price records concerning food and grocery products collected in supermarkets and hypermarkets of the most important chains of modern distribution located in the 20 regional chief towns in Italy, we provide estimates of Basic Heading (BH) PPPs for the year 2015. By using the CPD based stochastic approach we are able to obtain reliability measures for sub-national PPPs as well as to address

---

<sup>1</sup> Tiziana Laureti, University of Tuscia, Department of Economics and Business;  
email:laureti@unitus.it

Carlo Ferrante, Living conditions and Consumer price Unit, Istat;  
email:ferrante@istat.it

Barbara Dramis, Living conditions and Consumer price Unit, Istat;  
email:dramis@istat.it

methodological issues concerning the choice of the aggregation methods above the BH level.

### **Riassunto**

*La recente disponibilità di "dati scanner" va ad arricchire le informazioni provenienti dall'indagine tradizionale per la costruzione degli Indici dei Prezzi al Consumo e può significativamente cambiare il modo di costruire gli indici spaziali dei prezzi.*

*Infatti, per la stima delle Parità del Potere di Acquisto (PPA) è necessario disporre di dati sui prezzi al consumo di prodotti e servizi che siano comparabili a livello spaziale e rappresentativi dei consumi locali.*

*L'obiettivo del presente lavoro è di suggerire un approccio metodologico di tipo stocastico basato sul metodo Country-Product-Dummy per il calcolo delle PPA a livello intra-nazionale in Italia. Tale metodo consente di utilizzare contemporaneamente sia i dati scanner che i dati tradizionali con riferimento alla Grande Distribuzione Organizzata (GDO).*

*Utilizzando milioni di osservazioni relative a prezzi e quantità di prodotti alimentari e di prodotti per l'igiene della persona e la pulizia della casa, rilevate nei supermercati e ipermercati delle più importanti catene della GDO situati nei 20 capoluoghi di regione, il lavoro presenta le stime delle PPA per sottoclassi di prodotto per l'anno 2015. Il metodo utilizzato consente di costruire intervalli di confidenza per tali indici, nonché di esaminare diversi metodi per aggregare le PPA delle sottoclassi di prodotto.*

**Key words:** scanner data, sub-national Purchasing Power Parities, Country Product Dummy models

## **1 Introduction**

Spatial price indexes that measure the differences in price levels across regions within a country are essential for comparing real income, standards of living and consumer expenditure patterns. In countries characterized by large territorial differences in consumer preferences as well as in quality of products and household characteristics, the calculation of sub-national Purchasing Power Parities (PPPs) acquires great importance.

The Italian National Institute of Statistics (Istat) is one of the few National Statistical Offices (NSOs) that has carried out official experimental sub-national PPP computations by using price data from Consumer Price Indexes (CPIs) and *ad hoc* surveys and focusing on comparing consumer prices across the 20 Italian regions.

Significant price differences were found in 2010 which encouraged Istat to confirm the project for producing sub-national PPPs on a regular basis (Biggeri et al., 2016). The recent availability of high-frequency “scanner data” in addition to other sources of data enables us to deal with the sub-national PPP issue from a renewed approach, thus increasing cost efficiency and reducing burden response. Indeed, the way in which CPI data are collected often complicates the estimation of spatial price differences as products collected for CPI may not be comparable or representative across different areas, especially when the areas within a country differ in terms of climate, tastes and preferences. Moreover, *ad hoc* surveys are generally very expensive for the NSOs and do not provide information on consumption expenditure as in the case of CPI data. Within the European Multipurpose Price Statistics project, Istat has been exploring the possibility of using scanner data for computing official CPIs since 2014 and recently for compiling PPPs.

However, scanner datasets provide both opportunities and challenges for price statisticians since they must deal with huge amounts of highly detailed data on consumer purchases with high variability of products sold among cities.

Therefore, nowadays it is essential to determine how best to use scanner data and how to combine them with other data from various sources in order to construct sub-national PPPs as there is growing interest worldwide in using these data for compiling official price statistics but as yet research has been mainly focused on using scanner data for compiling CPIs in order to measure inflation rates.

The aim of this paper is to contribute to the advancement of spatial price index literature by exploring this new source of price data together with CPI data in order to compute Italian sub-national PPPs at Basic Heading (BH) level. We propose using a stochastic approach to index numbers based on the Country-Product-Dummy (CPD) method since it enables us to use both scanner data and traditional CPI data obtained from the large-scale retail trade and obtain reliability measures for sub-national PPPs. Interesting results have been obtained from numerous experiments using both sources of data even if, due to lack of space, we will only focus on a few of them in order to illustrate the potential of the proposed methodology and highlight the possible informative results.

The paper is structured as follows. Features of Italian scanner and CPI data and the results of the analyses carried out are presented and discussed in Section 2. The methodology used is described in Section 3 while in Section 4 some of the results obtained using various CPD models are presented and discussed for a set of BHs.

## 2 Scanner and CPI data analyses

Scanner data obtained from electronic points of sale benefits from an impressive coverage of transactions along with the availability of information on sales, prices, quantities sold and quality characteristics of products sold (brand, size and type of outlet) provided at the level of the barcode or, more precisely, the GTIN (Global Trade Item Number, formerly EAN code).

Currently, scanner data predominantly replaces price collection in supermarkets, especially for food, beverages and personal and home care products.

However, there are also several potential drawbacks in using scanner data as they are characterized by a high attrition rate of goods and volatility of the prices and quantities due to sales. Indeed, this new source of data are able to capture frequent, and often large, shifts in quantities purchased in response to price changes. Moreover, using highly detailed data on consumer purchases implies that when computing price indexes it is essential to deal with the issue of aggregation of individual items from both a theoretical and a practical perspective.

Over the last decade an increasing number of studies have been carried out with the aim of analysing the impact of different aggregation methods on inflation estimates by using scanner data as various NSOs are interested in using these data in their official price statistics (see for example Ivancic et al, 2011).

However, to the authors' knowledge as yet few studies have used scanner data and carried out experiments on aggregation issues when comparing consumer prices across space (Heravi, et al, 2003; Laureti and Polidoro, 2016).

Although time dimension of aggregation should not be difficult to deal with when comparing price levels across areas within a country, much attention should be paid when aggregating transaction price and quantity data for constructing spatial price indexes as the choices made will reflect different implicit assumptions regarding consumer behaviour. By referring to a specified set of geographical areas within a country, i.e. regions or cities, transactions can be aggregated over different items, stores and time periods.

In Italy, scanner data is collected weekly (approximately 1 million records) and after a process of data cleaning and trimming outliers, they are used to compute unit value price per item code calculated as the total turnover for that item code divided by the total quantities sold over the week.

In order to understand how best to aggregate the detailed information contained in the Italian scanner data for constructing sub-national PPPs, several analyses were carried out. More specifically, by referring to each Italian regional chief town, we carried out ANOVA and t tests on a sample of items in order to verify if the price of the same item could reflect auxiliary services provided by the seller. Indeed, within each city, the same item is found in different supermarket chains and in different stores which belong to the same retail chain. Results show significant differences in prices of the same items thus suggesting product differentiation which is embodied in the range or quality of services offered by different retailers, both across chain and across stores within the same chain. Moreover, by calculating skewness measures and Gini coefficients, we found that the distribution of expenditures within a product category is usually highly skewed and a relatively small number of items account for the majority of expenditures (see Table 1). Therefore, with respect to item groupings, we decided to use the finest classification of item that is available within the BH, i.e. the product code, which is identical across the Italian territory. As regards the time dimension, since the International Comparison Program (ICP) relies on a national average price for each item below the BH, we decided to use annual regional average prices which are obtained by aggregating the weekly price of each

EAN code by considering outlet-type (hypermarket or supermarket of a specific chain) and modern distribution chains for the 20 regional chief towns and by using turnover and quantity as weights. In this way we can mitigate the effects of the large fluctuations in quantities purchased in response to price discounts which still emerge using monthly average prices. The dataset consists in 2,799,320 annual price quotes from the 20 regional chief towns concerning the six most important modern distribution chains.

All results obtained cannot be reported due to lack of space, therefore we have selected a few BHs and their Gini coefficients and the number of different items sold in each regional chief town are shown in Table 1. Significant differences in Gini coefficients and number of products can be observed both between the various BHs within the same regional chief town and across cities within the same BH.

**Table 1:** Descriptive statistics by regional chief towns and BHs

Regional chief towns	Mineral or spring water		Personal care products		Household Cleaning and maintenance products	
	N.Items	Gini	N.Items	Gini	N.Items	Gini
<b>North</b>						
Aosta	131	0.628	1180	0.539	709	0.468
Torino	254	0.757	2918	0.661	1459	0.604
Genova	83	0.716	1077	0.672	470	0.603
Milano	258	0.797	2930	0.76	1477	0.746
Trento	79	0.542	789	0.587	413	0.508
Venezia	197	0.724	2234	0.638	1216	0.573
Trieste	105	0.593	890	0.506	588	0.518
Bologna	204	0.771	2458	0.67	1189	0.652
<b>Centre</b>						
Firenze	147	0.827	1387	0.747	669	0.721
Ancona	189	0.727	1814	0.648	1077	0.578
Perugia	188	0.784	1412	0.731	768	0.682
Roma	270	0.752	2428	0.692	1223	0.623
<b>South and Islands</b>						
L'Aquila	149	0.698	984	0.594	564	0.467
Campobasso	117	0.666	703	0.587	456	0.455
Napoli	175	0.709	1589	0.678	877	0.622
Potenza	89	0.693	470	0.554	307	0.496
Bari	151	0.716	1611	0.673	787	0.595
Catanzaro	66	0.607	602	0.579	335	0.559
Palermo	122	0.678	1390	0.639	758	0.594
Cagliari	136	0.738	1795	0.611	887	0.565

The scanner data set excludes perishables and seasonal products such as vegetables, fruit and meat since these products are sold at price per quantity and are not pre-packaged with EAN codes. Therefore, we integrated the scanner data with prices for fruit and vegetables which are traditionally collected for CPI production in modern distribution. After data quality controls and preliminary analyses of the basket, the CPI dataset includes annual average prices for 151 vegetable products collected in the 20 regional chief towns. For both scanner and CPI data, the items for which price data were collected in a single regional chief town only were eliminated in order to ensure that the incomplete price tableau is connected and therefore the CPD method is feasible.

### 3 Methodology and Empirical strategy

PPP compilation is undertaken at two levels, viz., at BH level, which is defined as a group of similar well-defined goods or services, and at a more aggregated level. Price data are usually aggregated at BH level without weights to produce PPPs for various BHs which are then aggregated to obtain PPPs for higher level aggregates.

This paper focuses on the first stage of aggregation, thus obtaining estimates of sub-national PPPs at BH level, because BHs are the foundations of overall comparison and it is essential to obtain reliable PPP estimates (Biggeri *et al.*, 2016).

From a methodological point of view, we refer to the relatively new strand of the stochastic approach to spatial price indexes with its roots in hedonic and CPD regression-based methodology, which is also used by the ICP at the World Bank.

Several authors have demonstrated that, thanks to its econometric nature, the CPD method could be extended and generalized in order to provide a comprehensive framework for carrying out international and intra-national comparisons using price data from various sources and also allows for the computation of standard errors for the various spatial price index methods which can be derived from variants of the CPD model (Rao and Hajargasht, 2016).

Due to the type and characteristics of the scanner and CPI data as well as the results of our preliminary analyses it is advisable to use hedonic CPD models through which information on the type of outlet and retail chain may be considered when constructing sub-national PPPs. Moreover, in order to account for the economic importance of each item in its market, which is essential in index number literature as demonstrated also by our analyses, we estimated weighted hedonic CPD models using both expenditure share and quantity as weights.

Let us assume that we are attempting to make a spatial comparison of prices between  $R$  areas (i.e regional chief towns) at BH level and  $p_{knr}$  denotes the annual price of item  $n$  in outlet  $k$  of area  $r$  ( $n = 1, 2, \dots, N$ ;  $r = 1, 2, \dots, R$ ;  $k = 1, \dots, K_{nr}$ ). Assuming that  $Z_1, Z_2, \dots, Z_J$  represent the set of quality characteristics associated with each item, thus the hedonic CPD model estimates the following regression equation separately for each BH:

$$\ln p_{knr} = \sum_{r=1}^R a_r D_r + \sum_{n=1}^N b_n D_n^* + \sum_{j=1}^J c_j Z_j + v_{knr}, \quad (1)$$

where,  $D_r$  are dummies for the areas,  $D_n^*$  are dummies for the type of product,  $a_r$ ,  $b_n$  and  $c_j$  are, respectively, the difference of (fixed) effects associated to the areas, type of product and quality characteristics with respect to a specific item;  $v_{knr}$  are random disturbance terms which are independently and identically (normally) distributed with zero mean and variance  $\sigma^2$ .

With reference to the scanner data we used the hedonic weighted CPD, expressed as:

$$\sqrt{w_{ij}} \ln p_{knr} = \sum_{r=1}^R a_r \sqrt{w_{ij}} D_r + \sum_{n=1}^N b_n \sqrt{w_{ij}} D_n^* + \sum_{j=1}^J \lambda_j \sqrt{w_{ij}} Z_j + \sqrt{w_{ij}} v_{knr} \quad (2)$$



where  $w_{ij}$  are expenditure and quantity weights reflecting the relative importance of different items. The parameter  $a_r$  is interpreted as the average level of prices (over all items in the BH) in area  $r$  relative to other areas while  $b_n$  is to be interpreted as the average (over all areas) premium that item  $n$  is worth relative to an average item in this BH. If  $a_r$  is expressed relative to a reference area (in our case Rome=100), then the PPP for area  $r$  is given by  $PPP_r = e^{a_r}$  where  $a_r$  is the difference between the coefficient for area  $r$  and that corresponding to the reference area (Rome).

## 4 Results

We ran hedonic CPD and hedonic weighted CPD for all available BHs in scanner data using expenditure and quantity as weights. Sub-national PPP results are only reported for the “Personal care products” BH (Table 2). Moreover, Table 2 shows PPP results for the “Fresh and chilled vegetables” BH based on CPI data referring to the six most important modern distribution chains (first two columns of Table 2). Significant differences can be observed between the results obtained from the hedonic CPD and WCPD yet similar results can be observed for the two expenditure and quantity WCPD models. These results appear to be coherent with our expectations and the territorial characteristics of the Italian macro areas.

**Table 2** CPD Estimation results: PPP estimations for regional chief towns by BH, ROME=100

	Fresh and chilled vegetable				Personal care products			
	HEDONIC CPD		HEDONIC CPD		HEDONIC WCPD (weights=expenditures)		HEDONIC WCPD (weights=quantity)	
	PPPs	Sig.	PPPs	Sig.	PPPs	Sig.	PPPs	Sig.
<b>North</b>								
Aosta	100.44		101.43	***	102.66	***	102.02	***
Torino	101.50		96.37	***	97.68	***	97.73	***
Genova	108.20	**	103.96	***	100.31		100.37	
Milano	92.45	***	98.01	***	100.15		100.12	
Trento	94.14	***	103.26	***	102.39	***	102.38	***
Venezia	108.26	***	95.14	***	97.85	***	97.70	***
Trieste	99.42		102.16	***	102.92	***	102.72	***
Bologna	103.47		99.63	*	100.61	***	100.98	***
<b>Centre</b>								
Firenze	100.98		89.72	***	85.22	***	83.54	***
Ancona	108.22	***	100.04		100.16		100.49	*
Perugia	101.57		100.51	**	97.74	***	97.24	***
<b>South and Islands</b>								
L'Aquila	86.55	***	101.37	***	100.26		99.48	
Campobasso	84.63	***	100.78	**	99.21	*	98.49	**
Napoli	75.37	***	96.01	***	95.80	***	94.24	***
Potenza	88.33	***	97.08	***	94.83	***	94.39	***
Bari	97.47		94.62	***	95.15	***	94.59	***
Catanzaro	103.67	*	98.15	***	97.76	***	97.66	***
Palermo	103.70		98.57	***	97.79	***	96.83	***
Cagliari	102.87		98.57	***	100.15		100.23	
Obs.	3.327		66,604		66,604		66,604	
Root MSE	0.17573		0.1170		0.1017		0.1017	

Note: \* 10 %, \*\* 5 %, \*\*\* 1 %

**Table 3** WCPD Estimation results: PPP ranking of regional towns by BH (1=highest prices, 20= lowest prices)

Geographical area		Rice	Other cereals and flour	Bread	Other bakery products	pizza&quiche	Pasta products	breakfast cereals	Other cereal products	Average position
Aosta	North	2	2	12	3	9	2	3	14	6
Torino	North	8	5	17	17	10	11	6	19	12
Genova	North	1	1	4	11	15	5	20	1	7
Milano	North	4	3	15	8	11	4	15	18	10
Trento	North	13	4	9	5	2	9	7	4	7
Venezia	North	16	16	19	19	6	15	1	7	12
Trieste	North	12	9	3	1	3	1	14	2	6
Bologna	North	3	11	13	7	4	6	17	17	10
<b>North</b>		<b>7</b>	<b>6</b>	<b>12</b>	<b>9</b>	<b>8</b>	<b>7</b>	<b>10</b>	<b>10</b>	<b>9</b>
Firenze	Centre	20	20	20	20	20	20	12	20	19
Ancona	Centre	10	15	14	18	5	3	18	5	11
Perugia	Centre	18	18	10	16	17	13	9	10	14
Roma	Centre	6	7	7	9	12	10	8	8	8
<b>Centre</b>		<b>14</b>	<b>15</b>	<b>13</b>	<b>16</b>	<b>14</b>	<b>12</b>	<b>12</b>	<b>11</b>	<b>13</b>
L'Aquila	South & Islands	9	10	1	6	13	8	13	6	8
Campobasso	South & Islands	11	12	6	4	1	14	10	11	9
Napoli	South & Islands	15	14	8	15	7	12	5	12	11
Potenza	South & Islands	17	13	2	13	8	18	19	3	12
Bari	South & Islands	14	17	16	12	18	19	16	16	16
Catanzaro	South & Islands	19	19	5	2	19	17	11	15	13
Palermo	South & Islands	7	8	11	10	16	16	2	9	10
Cagliari	South & Islands	5	6	18	14	14	7	4	13	10
<b>South and Islands</b>		<b>12</b>	<b>12</b>	<b>8</b>	<b>10</b>	<b>12</b>	<b>14</b>	<b>10</b>	<b>11</b>	<b>11</b>

Table 3 reports the position in the ranking of the PPP values for the regional chief towns compiled using expenditure share WCPD for the BHs belonging to “Bread and cereals” class. Significant variability is found in the position of the regional chief towns among BHs within the same class of products which may reflect different consumer behaviors and characteristics of modern retail chains in the Italian regions. Our results provide valuable insight on how to aggregate PPPs above the BH level.

## References

- Biggeri, L., Laureti, T. and Polidoro, F. (2016) Computing Sub-national PPPs with CPI Data: An Empirical Analysis on Italian Data Using Country Product Dummy Models. *Social Indicators Research*, in press DOI 10.1007/s11205-015-1217-x
- Heravi, S., Heston, A., & Silver, M. (2003). Using scanner data to estimate country price parities: A hedonic regression approach. *Review of Income and Wealth*, 49(1), 1-21.
- Ivancic, L., Diewert, W. E., and Fox, K. J. (2011). Scanner data, time aggregation and the construction of price indexes. *Journal of Econometrics*, 161(1), 24-35.
- Rao, D. S. P., and Hajargasht, G. (2016). Stochastic approach to computation of purchasing power parities in the International Comparison Program (ICP). *Journal of Econometrics*, 191(2), 414-425.

# Graphical approximation of Best Linear Unbiased Estimators for Extreme Value Distribution Parameters

## *Approssimazione grafica degli stimatori BLUE dei parametri delle distribuzioni dei valori estremi.*

Antonio Lepore

**Abstract** Graphical estimation methods play a central role in today's software because they may allow for a more straightforward analysis of the data and interpretation of results also by non-statisticians. In this paper, the best unbiased graphical estimators of distribution parameters, which have recently appeared in the literature for location-scale distributions, are conveniently approximated for the special case of the extreme value distributions for minima and maxima. The mean square deviation and bias of the resulting parameter estimators are compared to concurrent ones through proper pivotal indices via Monte Carlo simulation. The proposed approximation involves and is shown to produce also adequate results for the first two moments of order statistics from the standard extreme value distributions.

**Abstract** *I metodi di stima grafica ricoprono un ruolo centrale nei moderni strumenti software, in quanto facilitano l'analisi dei dati e l'interpretazione dei risultati anche in contesti non statistici. In questo lavoro, gli stimatori grafici BLUE recentemente proposti in letteratura per la famiglia di distribuzioni di posizione e scala vengono approssimati nel caso particolare delle distribuzioni dei valori estremi e vengono confrontati, mediante simulazione Monte Carlo, con le corrispondenti alternative proposte in letteratura tramite l'uso di opportuni indici non parametrici. Inoltre, la soluzione proposta rappresenta una soddisfacente approssimazione dei primi due momenti delle statistiche ordinate per le distribuzioni standard dei valori estremi.*

**Key words:** graphics and data visualization, linear unbiased estimators, location-scale distributions, extreme value distribution, probability plot.

---

Antonio Lepore

Department of Industrial Engineering, University of Naples Federico II, P.le V. Tecchio 80, 80125 Naples, Italy, e-mail: antonio.lepore@unina.it

## 1 Introduction

In many applicative fields, practitioners are used to exploit software tools that adopt graphical techniques to visualize data and check the fit provided by the chosen model. Even if a variety of effective analytical methods is available, graphical techniques give deeper insight and visual understanding of statistical information. This is commonly achieved through probability plots, which report ordered observations of a random variable (i.e., experimental data) against the corresponding estimates  $\hat{F}_i$  of the parent cumulative distribution function (cdf) (i.e., the plotting position) on properly scaled axis in a linear fashion. The more the points lie on a straight line, the more suitable the chosen parent distribution. As is known, the latter is usually required to belong or to be related to the location-scale family. This assumption allows probability plots to estimate parent distribution parameters through the slope and the intercept of the line of best fit [11]. However, the choice of the regressand (i.e., response variable) for the distribution fitting methods and their corresponding relative accuracy are not always clear [10] and the dispute of determining a unique plotting position approach has given rise to recent contributions and a wide controversial discussion [1, 2, 4, 5, 6, 7, 11, 14, 15, 16, 17, 18]. In particular, Pirouzi Fard and Holmquist [18] consider simple approximations of variances and covariances for order statistics from the standard extreme value (EV) distribution, whereas Pirouzi Fard [17] provides a comparison between the ordinary least-squares (OLS) and the generalized least-squares (GLS) distribution fitting methods when the data set arises from the standard EV distribution for minima. Cook and Harris [2] find out in the case of the EV distribution for maxima that the classical Gringorten estimator [8] of the order statistic mean gives asymptotic values for infinite sample sizes, whereas they are most often improperly used for small sample sizes. Fuglem et al. [7] support previous work by Cunnane [3] and state that plotting position methods dependent on the anticipated parent distribution should be used. However, Makkonen et al. [15, 16] still support the classical distribution-free approach [9]. In this paper, the graphical best linear unbiased estimators (BLUEs) [6], which have recently appeared in the literature, are elaborated for the EV distributions through a convenient approximation of the first two moments of order standard statistics and compared to the most popular and effective ones.

## 2 Approximation of the BLUEs of Extreme Value Distribution Parameters via probability plots

The EV cdf for minima (referred to as extreme value distribution in [17, 18]) and maxima (referred to Gumbel as in [2, 10, 11]) are, respectively, as follows

$$F_m(x; a, b) = 1 - e^{-e^{-\frac{x-a}{b}}}, \quad F_M(x; a, b) = e^{-e^{-\frac{x-a}{b}}}; \quad b > 0. \quad (1)$$

As is known, EV standard cdf's can be obtained by setting  $a = 0$  and  $b = 1$  and have inverse functions which are infinitely differentiable. Performances of graphical approaches for EV distributions are influenced by the choice of the plotting position formula, the distribution fitting method, as well as the covariance matrix (or its approximation) especially if the sample size is small [10]. In general, the use of the ordered observations of a sample of size  $N$ ,  $x_{(1)}, \dots, x_{(i)}, \dots, x_{(N)}$ , as regressand and the mean of the standard order statistics,  $\mu_{(1)}, \dots, \mu_{(i)}, \dots, \mu_{(N)}$ , as regressors achieves the best results and is mandatory for GLS estimation, which explicitly requires the specification of the covariance,  $\sigma_{(i,j)}$ , between the  $i$ -th and  $j$ -th order statistics ( $1 \leq i \leq j \leq N$ ) and leads to BLUEs of distribution parameters. In this paper the approximations suggested in [6] for  $\mu_{(i)}$  and  $\sigma_{(i,j)}$

$$\begin{aligned} \tilde{\mu}_{(i)} = & G^{-1}(p_i) + \frac{1}{2}G^{-1(2)}(p_i) \frac{p_i(1-p_i)}{(N+2)^2} + \frac{1}{3}G^{-1(3)}(p_i) \frac{p_i(1-p_i)(1-2p_i)}{(N+2)^2} \\ & + \frac{1}{8}G^{-1(4)}(p_i) \frac{p_i^2(1-p_i)^2}{(N+2)^3} \end{aligned} \quad (2)$$

$$\begin{aligned} \tilde{\sigma}_{(i,j)} = & \frac{p_i(1-p_i)}{N+2}G^{-1}(p_i) + \frac{p_i(1-p_j)}{(N+2)^2} \left[ (1-2p_i)G^{-1(2)}(p_i)G^{-1}(p_j) \right. \\ & + (1-2p_j)G^{-1(2)}(p_j)G^{-1}(p_i) + \frac{1}{2}p_i(1-p_i)G^{-1(3)}(p_i)G^{-1}(p_j) \\ & \left. + \frac{1}{2}p_j(1-p_j)G^{-1(3)}(p_j)G^{-1}(p_i) + \frac{1}{2}p_i(1-p_j)G^{-1(2)}(p_i)G^{-1(2)}(p_j) \right] \end{aligned} \quad (3)$$

are conveniently elaborated for the EV distribution for minima (resp. maxima), where  $G(x) = F_m(x; 0, 1)$  (resp.  $G(x) = F_M(x; 0, 1)$ ),  $p_i = i/(N+1)$ , and  $G^{-1(k)}(x)$  is the  $k$ -th derivative of the inverse function  $G^{-1}(x)$ . However, simple closed forms are available for  $\mu_{(1)}$  and  $\sigma_{(1,1)}$ , in the case of the EV distribution for minima

$$\mu_{(1)} = -\gamma - \ln N, \quad \sigma_{(1,1)} = \pi^2/6 \quad (4)$$

and for  $\mu_{(N)}$  and  $\sigma_{(N,N)}$ , in the case of the EV distribution for maxima

$$\mu_{(N)} = \gamma + \ln N, \quad \sigma_{(N,N)} = \pi^2/6 \quad (5)$$

where  $\gamma$  is the Euler's constant. Therefore, expressions (4) and (5) can be more opportunely utilized in this area in place of the general approximation formulas (2) and (3). Then, the GLS regression of  $\tilde{\mu}_{(i)}$  on the sample observations through the covariance approximation  $\tilde{\sigma}_{(i,j)}$  lead to graphical estimators for  $a$  and  $b$ , that are not unbiased because of the approximations. Hence, based on the results drawn in [10], it can be of interest to compare the latter approach with the most effective ones among those mentioned in the introduction and summarized in the first two rows of Table 1, namely Pirouzi Fard (PF) [17] and Hong and Li (HL) [11]. In general, each approximation  $\tilde{\mu}_{(i)}$  is associated with a plotting position  $\hat{F}_i = G^{-1}(\tilde{\mu}_{(i)})$  and vice versa. The last two rows of Table 1 report the Cook and Harris (CH) [2] and

**Table 1** Summary of the analysed probability plots ( $1 \leq i \leq j \leq N$ ). The correction factors  $\gamma_{Nk}$  ( $k = 1, \dots, 5$ ) are defined as in [11].

	$\hat{F}_i$	$\hat{\sigma}_{(i,j)}$
PF	$\begin{cases} 1 - e^{-\frac{i-\gamma}{N}} & i = 1 \\ \frac{i-0.4866}{N+0.1840} & \text{elsewhere} \end{cases}$	$\begin{cases} \pi^2/6 & i = j = 1 \\ \frac{(i-0.469)(N+0.831-i)^{-1}(N+0.073)^{-1}}{\ln\left(\frac{N+0.779-i}{N+0.356}\right)\ln\left(\frac{N+0.831-i}{N+0.356}\right)} & \text{elsewhere} \end{cases}$
HL	$\begin{cases} e^{-\frac{i-\gamma}{N}} & i = N \\ \frac{(i-0.37+0.232/\sqrt{N})}{(N+0.144+0.232/\sqrt{N})} & \text{elsewhere} \end{cases}$	$\begin{cases} \pi^2/6 & i = j = N \\ \frac{N+1-j-\gamma_{N1}}{(N+2-\gamma_{N2})(j-\gamma_{N3})\left(\ln\left(\frac{i-\gamma_{N5}}{N+1-\gamma_{N4}}\right)\ln\left(\frac{i-\gamma_{N3}}{N+1-\gamma_{N4}}\right)\right)} & \text{elsewhere} \end{cases}$
CH	$\frac{(i-0.439+0.466/\ln(N))}{(N+0.113+0.466/\ln(N))}$	-
GU	$\frac{i}{N+1}$	-

the classical Gumbel (GU) [9] plotting positions that rely instead on the use of the OLS method. Note that PF only applies to the EV distribution for minima, whereas HL and CH only apply to that for maxima.

### 3 Simulation Study and Results

A simulation study is carried out by drawing  $M = 10^5$  pseudo-random samples from the EV distributions for minima and maxima at sample sizes  $N = 5$  and  $N = 30$  to compare

- (i) the goodness of the approximations used of  $\mu_{(i)}$  and (when applicable) of  $\sigma_{(i,j)}$ ,
- (ii) the bias and the efficiency of graphical estimators for  $a$  and  $b$ ,

corresponding to the different approaches reported in Table 1 and that proposed. Slightly differently from [10, 11, 18], the following root mean square error (*RMSE*) and maximum absolute deviation (*MAD*) indices are utilized to compare (i)

$$RMSE = \sqrt{\sum_{i=1}^N (\mu_{(i)} - \tilde{\mu}_{(i)})^2 / N}, \quad MAD = \max_{1 \leq i \leq j \leq N} |\sigma_{(i,j)} - \tilde{\sigma}_{(i,j)}|. \quad (6)$$

Note that they are not pivotal (parameter-free), then may vary according to the actual distribution parameters. Therefore, the latter are set to standard values without the *RMSE* being undetermined when any  $\mu_{(i)} = 0$  as in [10]. The exact evaluation of  $\mu_{(i)}$  and  $\sigma_{(i,j)}$  in (6) is obtained using numerical integration [13]. Moreover, the following indices, namely the pivotal root deviation (*PRD*) and the pivotal absolute bias (*PAB*) of estimators  $\hat{a}$  and  $\hat{b}$  are introduced

$$PRD(\hat{a}) = \sqrt{E\{(\hat{a} - a)^2\}}/b, \quad PRD(\hat{b}) = \sqrt{E\{(\hat{b} - b)^2\}}/b^2 \quad (7)$$

$$PAB(\hat{a}) = |E\{\hat{a}\} - a|/b, \quad PAB(\hat{b}) = |E\{\hat{b}\} - b|/b \quad (8)$$

in order to compare (ii). It is trivial to show that (7) and (8) are pivots (see, e.g., [6, 12]) and therefore, the obtained results stand for whatever parameter. The lower the *RMSE* and the *MAD*, the better the proposed approximation of  $\mu_{(i)}$  and  $\sigma_{(i)}$ , respectively. Table 2 reports *RMSE* and *MAD* achieved by the different approaches reported in Table 1 and the proposed one, whereas Table 3 reports *PRD* and *PAB* of the estimators  $\hat{a}$  and  $\hat{b}$ . As anticipated, note that CH and GU approaches do not involve the approximation of  $\sigma_{(i,j)}$ , thus do not apply for *MAD*.

### 4 Conclusions

Table 2 clearly shows that the proposed approximations for the mean and the covariances of the order statistics from the EV distributions achieve the best performances at each considered sample size ( $N = 5, 30$ ) both in terms of *RMSE* and *MAD*. Table 3 confirms that the corresponding graphical estimators of distribution parameters achieve the smallest bias (*PAB*) and the highest efficiency (i.e., the smallest *PRD*)

**Table 2** *RMSE* and *MAD* achieved by approaches reported in Table 1 and that proposed for EV distributions at different sample sizes – bold text highlights the smallest value of each column.

	EV distribution for minima				EV distribution for maxima			
	<i>RMSE</i>		<i>MAD</i>		<i>RMSE</i>		<i>MAD</i>	
	$N = 5$	$N = 30$	$N = 5$	$N = 30$	$N = 5$	$N = 30$	$N = 5$	$N = 30$
Proposed	<b>0.00355</b>	<b>0.00057</b>	<b>0.01453</b>	<b>0.00193</b>	<b>0.00355</b>	<b>0.00057</b>	<b>0.01453</b>	<b>0.00193</b>
PF	-	-	-	-	0.01564	0.00288	0.02531	0.00324
HL	0.00756	0.00282	0.24872	0.32144	-	-	-	-
CH	0.04349	0.00743	-	-	-	-	-	-
GU	0.23592	0.12330	-	-	0.23592	0.12332	-	-

**Table 3** *PRD* and *PAB* of  $\hat{a}$  and  $\hat{b}$  achieved by approaches reported in Table 1 and that proposed for EV distributions at different sample sizes – bold text highlights the smallest value of each column.

	EV distribution for minima				EV distribution for maxima			
	<i>PRD</i> ( $\hat{a}$ )	<i>PAB</i> ( $\hat{a}$ )	<i>PRD</i> ( $\hat{b}$ )	<i>PAB</i> ( $\hat{b}$ )	<i>PRD</i> ( $\hat{a}$ )	<i>PAB</i> ( $\hat{a}$ )	<i>PRD</i> ( $\hat{b}$ )	<i>PAB</i> ( $\hat{b}$ )
Proposed	0.48020	<b>0.00191</b>	<b>0.40820</b>	<b>0.00081</b>	<b>0.48020</b>	<b>0.00191</b>	0.40820	<b>0.00081</b>
PF	-	-	-	-	0.48176	0.01852	<b>0.40403</b>	0.01018
$N = 5$ HL	<b>0.48009</b>	0.00836	0.41496	0.00425	-	-	-	-
CH	0.48095	0.04568	0.46173	0.00569	-	-	-	-
GU	0.48358	0.00383	0.61444	0.24902	0.48358	0.00383	0.61444	0.24902
Proposed	<b>0.19266</b>	<b>0.00079</b>	<b>0.14686</b>	<b>0.00009</b>	<b>0.19266</b>	<b>0.00079</b>	<b>0.14686</b>	<b>0.00009</b>
PF	-	-	-	-	0.19273	0.00216	0.14714	0.00173
$N = 30$ HL	0.19371	0.00165	0.14949	0.00164	-	-	-	-
CH	0.19684	0.00462	0.18467	0.00148	-	-	-	-
GU	0.19481	0.00640	0.21633	0.08979	0.19481	0.00640	0.21633	0.08979

even when plugging in the proposed approximations for large sample sizes ( $N = 30$ ). However, as expected, some rather biased estimators can be slightly more efficient at small sample sizes ( $N = 5$ ), namely PF and HL. According to [10], note that graphical estimators of distribution parameters that rely on the OLS instead of the GLS estimation method, namely CH and GU, are always the least efficient. Hence, Makkonen's claims in the plotting position controversy mentioned in the introduction cannot be supported. In the view of these results, the proposed approximation for EV distributions allows practitioners not to drastically abandon classical graphical methods and opt out of more efficient analytical solutions.

## References

1. Cook, N.: Rebuttal of Problems in the extreme value analysis. *Struct. Saf.* (2012)
2. Cook, N.J., Harris, R.I.: The Gringorten estimator revisited. *Wind Struct. An Int. J.* **16**(4), 355–372 (2013). DOI 10.12989/was.2013.16.4.355
3. Cunnane, C.: Unbiased plotting positions – A review. *J. Hydrol.* **37**(3–4), 205–222 (1978). DOI 10.1016/0022-1694(78)90017-3
4. Erto, P., Lepore, A.: A Note on the Plotting Position Controversy and a New Distribution-free Formula. In: *Proceeding of the 45th Scientific Meeting of the Italian Statistical Society*, pp. 16–18 (2010)
5. Erto, P., Lepore, A.: New Distribution-Free Plotting Position Through an Approximation to the Beta Median. *Adv. Theor. Appl. Stat.* pp. 23–27 (2013). DOI 10.1007/978-3-642-35588-2\_3
6. Erto, P., Lepore, A.: Best unbiased graphical estimators of location-scale distribution parameters: application to the Pozzuoli's bradyseism earthquake data. *Environ. Ecol. Stat.* **23**(4), 605–621 (2016). DOI 10.1007/s10651-016-0356-9
7. Fuglem, M., Parr, G., Jordaan, I.: Plotting positions for fitting distributions and extreme value analysis. *Can. J. Civ. Eng.* **40**(2), 130–139 (2013). DOI 10.1139/cjce-2012-0427
8. Gringorten, I.I.: A plotting rule for extreme probability paper. *J. Geophys. Res.* **68**(3), 813–814 (1963). DOI 10.1029/JZ068i003p00813
9. Gumbel, E.J.: *Statistics of Extremes* (1958)
10. Hong, H.P.: Selection of regressand for fitting the extreme value distributions using the ordinary, weighted and generalized least-squares methods. *Reliab. Eng. Syst. Saf.* **118**, 71–80 (2013). DOI 10.1016/j.res.2013.04.003
11. Hong, H.P., Li, S.H.: Plotting positions and approximating first two moments of order statistics for Gumbel distribution: Estimating quantiles of wind speed. *Wind Struct. An Int. J.* **19**(4), 371–387 (2014). DOI 10.12989/was.2014.19.4.371
12. Lawless, J.: Confidence interval estimation for the Weibull and extreme value distributions. *Technometrics* **20**(4), 355–364 (1978)
13. Lieblein, J.: Efficient methods of extreme-value methodology. *Tech. rep.* (1974)
14. Makkonen, L.: Bringing Closure to the Plotting Position Controversy. *Commun. Stat. - Theory Methods* **37**(3), 460–467 (2008). DOI 10.1080/03610920701653094
15. Makkonen, L., Pajari, M., Tikanmäki, M.: Closure to Problems in the extreme value analysis (*Struct. Safety* 2008: 30: 405419). *Struct. Saf.* (2013)
16. Makkonen, L., Pajari, M., Tikanmäki, M.: Discussion on Plotting positions for fitting distributions and extreme value analysis 1. *Can. J. Civil. Eng.* **40**(9), 927–929 (2013)
17. Pirouzi Fard, M.N.: Probability plots and order statistics of the standard extreme value distribution. *Comput. Stat.* **25**(2), 257–267 (2010). DOI 10.1007/s00180-009-0174-8
18. Pirouzi Fard, M.N., Holmquist, B.: Approximations of Variances and Covariances for Order Statistics from the Standard Extreme Value Distribution. *Commun. Stat. - Simul. Comput.* **37**(8), 1500–1506 (2008). DOI 10.1080/03610910802244059



# Monitoring ship performance via multi-way partial least-squares analysis of functional data

## *Monitoraggio delle prestazioni di una nave mediante analisi multi-way partial least squares di dati funzionali*

Antonio Lepore, Biagio Palumbo and Christian Capezza

**Abstract** The multi-sensor systems installed on board of modern ships provide massive amounts of data that require opportune multivariate methods for continuous performance monitoring during voyages. In this paper, functional data are obtained from variables that describe operating conditions of a Ro-Pax cruise ship owned by the Grimaldi Group and are analysed via multi-way partial least-squares regression of the fuel consumption per hour. The proposed procedure is shown to well predict and monitor ship performance and to indicate if and when an anomaly may occur in ship operating conditions throughout each voyage.

**Abstract** *I sistemi di acquisizione dati installati a bordo delle moderne navi generano un'enorme quantità di dati che rende necessario lo sviluppo di opportuni metodi multivariati per il monitoraggio delle prestazioni durante la navigazione. Nel presente lavoro, viene effettuata un'analisi dei dati funzionali che descrivono le condizioni operative di viaggio di una nave da carico e passeggeri di proprietà della società armatoriale italiana Grimaldi Group. La procedura proposta in questo articolo consente, mediante regressione multi-way partial least squares, la previsione e il monitoraggio continuo delle prestazioni della nave ed è in grado di supportare l'individuazione delle anomalie e dell'istante in cui esse si presentano durante un viaggio.*

**Key words:** Functional data analysis, Multi-way partial least squares (MPLS), fuel consumption monitoring, multivariate control chart

---

Antonio Lepore  
Department of Industrial Engineering, University of Naples Federico II  
Biagio Palumbo  
Department of Industrial Engineering, University of Naples Federico II  
Christian Capezza  
Department of Industrial Engineering, University of Naples Federico II,  
e-mail: christian.capezza@unina.it

## 1 Introduction

Nowadays, thanks to real-time multi-sensor systems installed on board, modern ships are able to continuously measure and store operating data overload that requires opportune multivariate methods for monitoring ship performance. Monitoring fuel consumption of a ship usually has been limited to single measurements of variables for each voyage, or to continuously monitor a single variable throughout the entire voyage, generally the speed over ground (SOG). Even though functional data analysis is applied in several subject areas [1,2], it has never been implemented in the maritime field. In this paper, functional data are obtained from variables that describe ship operating conditions and are used to apply multi-way partial least-squares (MPLS) [3,4] for monitoring ship performance. Based on trajectories of different ship operating conditions, squared prediction error charts are used to monitor anomalies in ship operating conditions, whereas prediction error chart monitors the fuel consumption per hour (FCPH).

## 2 The procedure

Time is the functional domain used in most of the functional data analysis applications. However, on a given ship route, travel duration varies significantly from voyage to voyage. Therefore, a more appropriate domain needs to be chosen to allow comparing navigation variable measurements over different replications, i.e., different voyages. In this paper, percentage of total distance travelled at each voyage by the ship is suitably chosen as functional domain. At a given domain point, the ship is almost in the same position over different replications of a given route and its operating conditions are reasonably expected to be similar when no anomalies in ship performance occurred. Discrete measured values of a navigation variable at different domain points for each voyage are then converted to functional data.

The main underlying idea of the proposed procedure is to get a three-dimensional array  $\underline{\mathbf{X}}$  by evaluating ship operational reference (functional) data at given domain point with the following three dimensions: number of replications  $I$ , number of variables  $J$ , and number of evaluation points  $K$ . Thus, a MPLS regression model can be suitably built on the ship FCPH at each voyage as the scalar response variable, which is organized into the  $(I \times 1)$  vector  $\mathbf{y}$ . Functional data analysis is found useful to obtain instantaneous information not only about the SOG, which, as is known, represents the most significant variable determining FCPH [5], but also about its derivative, i.e., acceleration (that can be used as additional predictor). Furthermore, discrete operating conditions are generally available (e.g., departure/arrival operating conditions, route type) for each voyage and can be stored in a  $(I \times M)$  matrix  $\underline{\mathbf{Z}}$ . Data are mean centered and scaled prior to perform the analysis.

MPLS is thus applied by unfolding the array  $\underline{\mathbf{X}}$  into a large  $(I \times JK)$  matrix  $\tilde{\mathbf{X}}$  and considering the  $(I \times (JK + M))$  matrix  $\mathbf{X} = [\tilde{\mathbf{X}} \quad \mathbf{Z}]$ , which can be decomposed via the partial least-squares (PLS) method into a smaller number of  $R$  orthogonal score vectors  $\mathbf{t}_1, \dots, \mathbf{t}_R$ , arranged in a  $(I \times R)$  matrix  $\mathbf{T}$ . The latter can be eventually used as regressor for  $\mathbf{y}$  as follows

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}; \quad \mathbf{y} = \mathbf{Tq} + \mathbf{f}, \quad (1)$$

where  $\mathbf{P}$  is the  $((JK + M) \times R)$  matrix of the  $\mathbf{X}$ -loadings,  $\mathbf{q}$  is the  $(R \times 1)$  vector of  $\mathbf{y}$ -loadings, and  $\mathbf{E}$   $(I \times (JK + M))$  and  $\mathbf{f}$   $(I \times 1)$  are residual matrices. The matrix  $\mathbf{T}$  is given by [6,7]

$$\mathbf{T} = \mathbf{XW}(\mathbf{P}^T\mathbf{W})^{-1}, \quad (2)$$

where  $\mathbf{W}$  is the  $((JK + M) \times R)$  matrix of the  $\mathbf{X}$ -weights. Forthcoming voyages can then be monitored by specializing the squared prediction error statistic  $SPE_X$  [3] for residuals in the predictor variable space at each voyage to a single instantaneous evaluation point  $k$  (i.e., a given percentage of distance travelled) as  $SPE_k = \sum_{c=(k-1)J+1}^{kJ} \mathbf{e}(c)^2$ , where the  $(1 \times (JK + M))$  vector  $\mathbf{e}$  contains the corresponding  $\mathbf{X}$ -residuals.  $SPE_k$  represents the perpendicular distance of the instantaneous ship operating condition measurements from the reduced predictor variable space obtained based on the reference data. Control limits for both  $SPE_X$  and  $SPE_k$  statistics are given by [8]. Detailed information can be obtained about plausible causes of anomalies by interrogating the MPLS model. While  $SPE_k$  statistic is able to clearly detect problems at a specific point  $k$ , one can examine contribution of the  $j$ -th individual variable to the  $SPE_X$  statistic through  $SPE_{X,j} = \sum_{c=0}^{K-1} \mathbf{e}(j + cJ)^2$ .

If a forthcoming voyage shows no anomalies, i.e., the monitoring statistics do not exceed control limits, approximate prediction intervals for the future observation of FCPH  $y$  can be calculated through the limits  $\hat{y} \pm t_{I-R-1, \alpha/2} \hat{\sigma} \sqrt{1 + \mathbf{t}_{\text{new}}^T (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{t}_{\text{new}}}$  [3], where  $\mathbf{t}_{\text{new}}$  is obtained as  $\mathbf{t}_{\text{new}}^T = \mathbf{x}_{\text{new}}^T \mathbf{W}(\mathbf{P}^T \mathbf{W})^{-1}$  from the future  $x$ -observations  $\mathbf{x}_{\text{new}}$  and Equation (2),  $\hat{y} = \mathbf{t}_{\text{new}}^T \mathbf{q}$ ,  $\hat{\sigma} = \mathbf{f}^T \mathbf{f} / (I - 1)$  and  $t_{I-R-1, \alpha/2}$  is the  $100\alpha/2$  percentile of a Student's distribution with  $(I - R - 1)$  degrees of freedom.

### 3 Application

The proposed procedure has been applied to real operational data acquired on board of a Ro-Pax ship operating in the Mediterranean Sea, owned by the Grimaldi Group. For each voyage  $J = 7$  functional variables have been considered as predictors of the FCPH:

1. SOG [ $kn$ ];
2. acceleration [ $kn/s$ ];
3. power difference between port and starboard propeller shafts [ $kW$ ];
4. power difference between port and starboard shaft generators;
5. longitudinal wind [ $kn$ ];
6. side wind [ $kn$ ];
7. distance from the mean route [ $NM$ ].

The last variable takes into account the path differences between voyages of the same route type. Each functional variable has been evaluated in  $K = 100$  equally spaced domain points. The matrix  $\mathbf{Z}$  contains two indicator variables that distinguish the three route types that the ship sails. MPLS has been then applied to a set of  $I = 192$  reference voyages. A single run of 10-fold cross validation procedure based on PRESS statistic [9] has been carried out and selected  $R = 4$  latent variables. The coefficient of determination is equal to 0.93 and confirms the model is able to adequately predict the FCPH at each voyage. Ship performance has been then monitored on 51 successive voyages. Figure 1 shows the  $SPE_x$  at each voyage and highlights unusual operating conditions for voyages 22, 23, 26, 27, 30, 39 and 44. These voyages are further individually examined using the  $SPE_k$  control chart. As an example, the  $SPE_k$  statistic for voyage 14 reported in Figure 2a does not show unusual variations throughout the voyage. Whereas it is clearly above the 95% control limit for voyage 39, as shown in Figure 2b. The anomalous voyages can be checked against the reference model to determine the reason for their difference. This can be investigated by using the contribution plots.

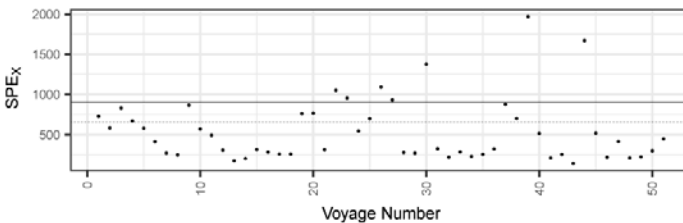


Figure 1: Monitoring charts for  $SPE_x$  statistics with 95% and 99% control limits (dashed and solid line) for 51 new voyages.

Figure 3 displays the contribution of each variable to the  $SPE_x$  statistic for voyage 39 and indicates the variable 6 (distance from the mean route) as the main variable responsible for the out-of-control observation. For voyages with  $SPE_x$  and  $SPE_k$  statistics in control, FCPH can be monitored through the prediction error chart illustrated in Figure 4. In this chart, voyages that fall outside the prediction limits require further investigation on those variables that have not been considered in the MPLS model.

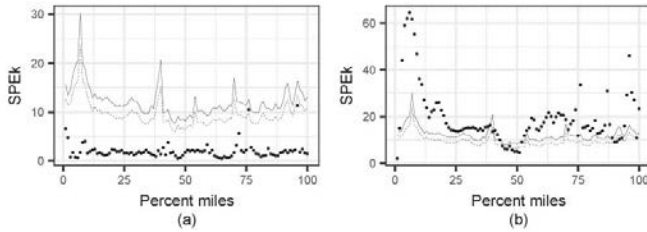


Figure 2: Monitoring charts for  $SPE_x$  statistics with 95% and 99% control limits (dashed and solid line) for a new regular voyage (voyage 14) with no anomalies (a) and for a new voyage (voyage 39) where problem is clearly identified (b).

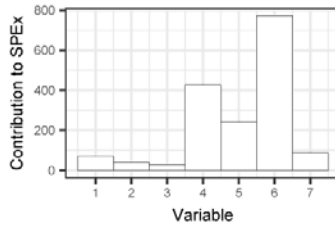


Figure 3: Contribution of variables to  $SPE_x$  statistic for voyage 39 in Figure 1b.

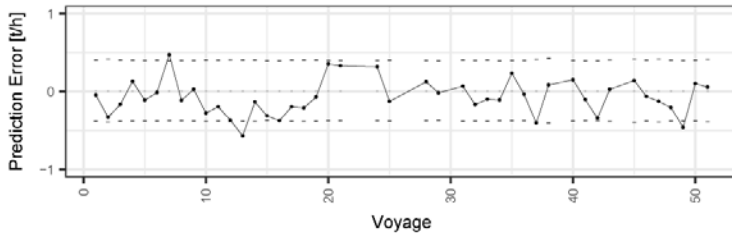


Figure 4: Monitoring FCPH for new voyages for which squared prediction error statistic shows no anomalies through prediction error chart, with 95% prediction limits.

## 4 Conclusion

In the shipping industry, there is a lack of methods that allow multivariate continuous monitoring during voyages. Functional data give instantaneous information on ship operating conditions and can be used to build linear models via multi-way partial least squares to monitor ship performance and predict fuel consumption per hour. The application illustrated in this paper shows that the proposed procedure is able to furnish adequate predictions and to indicate if and when anomalies occur. The squared prediction error statistic evaluated at a single domain point gives clear indications in this regard. This would have not been feasible through statistical models built using a single variable observation per each voyage. Functional data analysis could be exploited for numerous applications, such as the crucial theme of developing predictive maintenance techniques on ship engines and identifying types of faults in order to provide early warnings.

## References

1. Ramsay, J.O. and Silverman, B.W.: *Functional Data Analysis*. Second Edition Springer Series in Statistics, ISBN 978-0387-40080-8, (2005)
2. Ramsay, J., Hooker, G., and Graves, S.: *Functional data analysis with R and MATLAB*. (2009)
3. Nomikos, P. and MacGregor, J.F.: Multi-way partial least squares in monitoring batch processes. *Chemom. Intell. Lab. Syst.* **30**(1), 97–108, (1995) doi:10.1016/0169-7439(95)00043-7
4. Kourti, T., Nomikos, P., and MacGregor, J.E.: Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *J. Process Control* **5**(4), 277–284, (1995)
5. Bialystocki, N. and Konovessis, D.: On the estimation of ship's fuel consumption and speed curve: A statistical approach. *J. Ocean Eng. Sci.* **1**(2), 157–166, (2016) doi:10.1016/j.joes.2016.02.001
6. Helland, I.S.: On the structure of partial least squares regression. *Commun. Stat. - Simul. Comput.* **17**(2), 581–607, (1988) doi:10.1080/03610918808812681
7. Phatak, A. and Jong, S. De: The geometry of partial least squares. *J. Chemom.* **11**(December 1996), 311–338, (1997) doi:10.1002/(SICI)1099-128X(199707)11:4<311::AID-CEM478>3.0.CO;2-4
8. Nomikos, P. and MacGregor, J.F.: Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics* **37**(1), 41–59, (1995) doi:10.1016/0967-0661(95)00014-L
9. Geladi, P. and Kowalski, B.R.: Partial Least-Squares Regression - a Tutorial. *Anal. Chim. Acta* **185**, 1–17, (1986)

# Dynamic profiling of banking customers: a pseudo-panel study

## *Segmentazione dinamica di clienti bancari: uno studio basato su indagini ripetute*

Caterina Liberati, Lisa Crosato, Paolo Mariani and Biancamaria Zavanella

**Abstract** The analysis of the evolution of satisfaction in business context is usually based on pseudo panels studies, because they are less costly and easy to build with the available data. As in the cross-section case, detailed information about customers are collected at each time point, but the dynamic comparison generally involves few temporal lags (due to short life time of products and services). Accordingly, in our paper we apply the Dual Multiple Factor Analysis. Such a technique allows the synthesis of the multivariate tables and their visualization on a common space that sheds light on customers' trajectories of satisfaction. A real case study of an Italian bank is illustrated.

**Abstract** *L'analisi dell'evoluzione della soddisfazione nel contesto di business, è di solito basato su studi pseudo-panel, perchè sono meno costosi e facili da costruire con i dati disponibili. Come in casi cross-section, informazioni dettagliate sui clienti vengono raccolte in ogni istante temporale, ma il confronto dinamico avviene generalmente considerando soli pochi ritardi temporali (a causa del breve ciclo di vita di prodotti e servizi). Di conseguenza, nel nostro contributo proponiamo l'utilizzo dell'analisi fattoriale multipla duale. Tale tecnica permette la sintesi delle tabelle multivariate e la visualizzazione delle stesse su uno spazio comune che fa luce su traiettorie di soddisfazione dei clienti. I vantaggi della tecnica vengono illustrati in un caso di studio relativo ad una banca italiana.*

---

Caterina Liberati

Università di Milano-Bicocca, p.zza Ateneo Nuovo 1, e-mail: caterina.liberati@unimib.it

Lisa Crosato

Università di Milano-Bicocca, p.zza Ateneo Nuovo 1, e-mail: lisa.crosato@unimib.it

Paolo Mariani

Università di Milano-Bicocca, p.zza Ateneo Nuovo 1, e-mail: paolo.mariani@unimib.it

Biancamaria Zavanella

Università di Milano-Bicocca, p.zza Ateneo Nuovo 1, e-mail: biancamaria.zavanella@unimib.it

**Key words:** Customers Profiling, Customer Satisfaction Surveys, Dual Multiple Factor, Pseudo-Panels

## 1 Introduction

To date, the study of Customer Satisfaction has dominated marketing behavioural literature.[5] Despite the strong recognition that consumer behaviour should be viewed from a dynamic perspective, only a residual percentage of the studies published in marketing has addressed the problem in this manner.[12] [10] [9] [8] The dearth of panel studies appears to be largely a consequence of costs to the company and difficulty in obtaining longitudinal data sets and/or maintaining the sample over time[7], maybe due to little incentive to build databases of historical performance for products and services.

Given the deficiencies of cross-sectional data and the problems associated with collecting longitudinal panel data, one practical solution is to exploit, as much as possible, all of the information already available in various cross-sectional data sources.[3]

The econometric literature proposes a way to perform such matching: the collection of pseudo-panel data, that makes it possible to monitor gross change utilising a time series of cross-sectional data. At this regards, Ref. 1 introduced the use of cohorts to estimate a fixed effects model from repeated cross-sections. The benefits of such a procedure are several, from a decrease in attrition, to a drop in individual measurement errors. Although the econometric approach has provided a valuable contribution to studying pseudo panel data, such a strategy seems inadequate for the treatment of marketing surveys, where individual level changes must be monitored.

In an attempt to approach the Customer Satisfaction study from a dynamic perspective based on pseudo panel surveys, this work proposes the usage of Multiple Factor Analysis (MFA). MFA is an extension of the Principal Component Analysis (PCA), tailored to handle multiple data tables that measure sets of variables collected over the same observations, or, alternatively, (in the Dual MFA) multiple data tables where the same variables are measured over different sets of observations. The advantages of such a technique are several and ranging from full information employment (in terms of instances and variables), to synthesising the dimensionality of the tables and easily visualising points across time. Indeed, once all of the instances have been embedded into the common factorial plan, a post-hoc stratification can be performed to reduce the number of instances into a manageable number of profiles that are mutually exclusive and share well-defined characteristics.[4] [11]



## 2 Modeling Data Over Time

The analysis of several sets of individuals described by a same set of variables is a problem frequently encountered, not only in marketing. Dual Multiple Factor Analysis (DMFA) answers exactly such task.[2] The general idea behind DMFA is to normalize each of the datasets and then to combine these data tables into a common representation of the variables called the compromise map.[6] Let's denote with  $X$  a  $N \times K$  matrix composed by column-wise juxtaposition of  $L$  sub-matrixes, each of them collecting on the same set of variables but different observations. The mathematical formulation of DMFA can be described into two steps. In the first one a grand matrix  $\bar{X}$  is obtained, juxtapositioning the standardized  $X_{[q]}$  by column and weighting them with the first eigenvalue coming from separate PCAs on  $X_{[q]}$ . In the second step, we performed a Principal Component Analysis of the grand matrix.

$$(\bar{X}'D\bar{X}) = \Lambda\Gamma\Lambda' \quad (1)$$

where  $D$  is the  $(N \times N)$  diagonal matrix (metric) whose terms are the masses associated to the observations,  $\Gamma$  is a  $K \times K$  orthonormal basis and  $\Lambda$  is a  $K \times K$  diagonal matrix of eigenvalues.

The spectral decomposition theorem ensures the best reconstruction in terms of least squares of the weighted correlation matrix  $(\bar{X}'D\bar{X})$ ; the solution provides individuals' factor scores of the total matrix  $X$ , which represent a compromise of the  $K$  sub-matrixes:

$$F = X \Delta^2 U \quad (2)$$

The data analysed in this study aim to monitor several aspects related to customer satisfaction across three years (2010-2012). It collects clients' appreciations about:

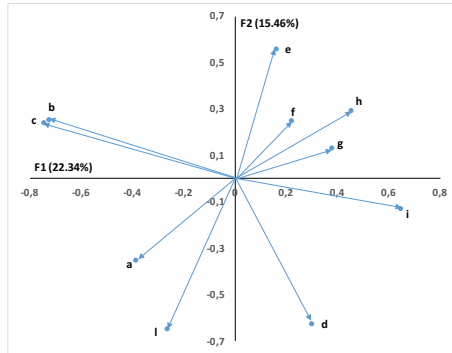
- Banking touch points (Personnel (a), ATM (b), Internet-banking (c))
- Imagine of the credit institute (Prestige (e), Innovation (f), Honesty (g), Trust (h))
- Proxy of customers engagement (Probability to recommend the bank to someone(d), Interest for the customers (i), Overall Satisfaction (l))

The total number of the collected instances was 6193, summarizing, respectively, the 2068 (2010), 2058 (2011), 2067 (2012) observations over three years.

Our data matrix  $X$  ( $6193 \times 10$ ) was obtained by summing up (by column) the three data tables  $X = [X_{2010}; X_{2011}; X_{2012}]$ . The two-step procedure was employed on the matrices under study: first,  $X_{2010}, X_{2011}, X_{2012}$  were centered and standardised and a separate PCA was run on each of the tables to obtain the weights to balance the within-groups inertia. Second, a PCA was performed on the grand matrix obtained.

The solution provided by the spectral decomposition uncovers a peculiar configuration of the variables onto the principal compromise plane (Fig. 1). Accordingly we named  $f_1$  as *standardized-customised banking service* and  $f_2$  as *imagined-experienced bank*.

Instances can also be projected on the compromise plane. Due to the large number of individuals composing the sample, it is very difficult to visualise the dynamic

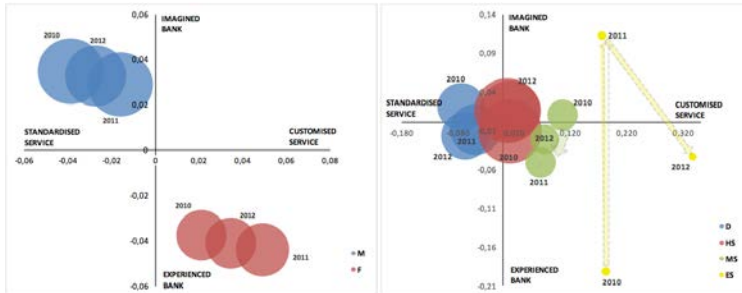


**Fig. 1** Variables projection onto the principal plane

paths of each subject; therefore, a segmentation was performed to profile the long-term behaviours of customers' profiles. We focus our attention on those groups that were selected according to the Italian bank's interests, but the analysis can easily be replicated for any other group.

We compare the trajectories of clients distinguished by socio-demographic characteristics, since such variables usually play an important role in the assessment formulation. Visual inspection of the graphical representations in figure 2, which depict average positions and inner variability of the profiles over time, reveals different evolutions of the monitored groups. In our case, female customers move along the positive side of the first axis, showing a high appreciation for a customised banking service (Fig. 2 left panel). On the contrary, males move exactly on the opposite direction, seeming more involved with a standardised assistance. Also the relationship with the bank has different connotations when studied by gender: females prefer to experience services provided by the financial institution while males take in high consideration the bank's image. Both tendency paths highlight an involution in the temporal trends, probably underlying a lack of specific actions/stimuli per gender.

A second comparison has been performed that contrasts customers' behaviours distinguished by different educational levels (Fig. 2 right panel). This time, trajectories appear different for shapes and lengths. Graduates, lying on the negative side of the first axis over the three years, show a high appreciation for remote touch points while middle school graduates have more fuzzy evaluations. On the contrary, poorly educated profiles (as middle school or elementary school graduates) seem to prefer customised assistance.



**Fig. 2** Dynamic behaviours: trajectories per gender (left panel); trajectories per educational levels (right panel)

### 3 Conclusion

This paper presents a new approach to Customer Satisfaction management. The idea originated from the necessity of a longitudinal perspective in the assessment of Customer Satisfaction, also expressed by several contributions in behavioral studies. Today, this task is even more important because information is available and inexpensive. The strength of our approach is that it is model-free, so it can be applied to every data table's comparison/visualisation. In reference to the real case study illustrated in the paper, the evidence found is quite interesting and can easily be interpreted in terms of management implications. The compromise factorial plane obtained allows to distinguish the profiles of the clients who prefer remote services from those who are involved with a customised assistance. It also uncovers different types of relationship between customers and the bank: such evidences, if monitored over time, can help the management to better respond to clients' demands.

### References

1. Deaton, A.: Panel data from time series of cross-sections. *Journal of Econometrics* **30**, 109–126 (1985)
2. Escofier, B. Pagés, J.: Multiple factor analysis. *Computational Statistics & Data Analysis* **18** 121–140 (1990)
3. Frethey-Bentham, C.: Pseudo panels as an alternative study design. *Australasian Marketing Journal*, **19**, 281–292 (2011)
4. Green, P., Krieger, A.: Alternative approaches to cluster based market segmentation. *Journal of the Market Research Society* **37**, 221–239 (1995)
5. Homburg, C. Koschate, N., Hoyer, W. D.: The role of cognition and affect in the formation of customer satisfaction: A dynamic perspective. *Journal of Marketing* **70**, 21–31 (2006)
6. Lê, S., Pagés, J.: Dmfa: Dual multiple factor analysis. *Communications in Statistics-Theory and Methods*. **39** 483–492 (2010)

7. Leonidou, L. C., Barnes, B. R., Spyropoulou, S. Katsikeas, C. S.: Assessing the contribution of leading mainstream marketing journals to the international marketing discipline. *International Marketing Review* **27**, 491–518 (2010)
8. Liberati, C., Mariani, P.: Banking customer satisfaction evaluation: a three-way factor perspective. *Advances in Data Analysis and Classification*. **6**, 323–336 (2012)
9. Masserini, L. Liberati, C., Mariani, P.: Quality service in banking: a longitudinal approach. *Quality and Quantity*. **51** 509–523 (2017)
10. Rindfleisch, A. , Malter, A. J., Ganesan, S. Moorman, C.: Cross-sectional versus longitudinal survey research: concepts, findings, and guide-lines. *Journal of Marketing Research*. **45**, 261–279 (2008)
11. Wedel, M., Kamakura, A.: *Market Segmentation: Conceptual and Methodological Foundations*. Kluwer Academic Publishers, Dordrecht (1998)
12. Williams, B. C., Plouffe, C. R.: Assessing the evolution of sales knowledge: a 20-year content analysis. *Industrial Marketing Management* **36**, 408–419 (2007)

# A comparison between seasonality indices deployed in evaluating unimodal and bimodal patterns

## *Un confronto tra indici di stagionalità utilizzati nella misura di pattern unimodali e bimodali*

Giovanni L. Lo Magno, Mauro Ferrante, Stefano De Cantis

**Abstract** This paper will discuss a recently proposed index for measuring seasonality, which is based on the solution of the well-known transportation problem. A specific characterization of the cost matrix will permit the taking into account of the cyclical structure of time periods, which characterizes the phenomenon under observation. Various features of the proposed index will be evaluated by comparing it with other indices which are commonly used in the measurement of seasonality, such as the Gini concentration index. Given the wide range of disciplines with an interest in the analysis of seasonal phenomena, the approach proposed may be of wide interest.

**Abstract** *Il presente lavoro discute un indice recentemente proposto per la misura della concentrazione stagionale e basato sulla soluzione del problema del trasporto. Una specifica caratterizzazione della matrice dei costi consente di tenere in considerazione la struttura ciclica dei periodi che caratterizza il fenomeno osservato. Diverse caratteristiche dell'indice proposto sono valutate confrontandolo con altri indici che sono comunemente utilizzati per la misura della stagionalità, come ad esempio l'indice di concentrazione di Gini. Considerata la varietà di ambiti interessati allo studio della stagionalità, l'approccio proposto può essere d'interesse da diverse prospettive.*

**Key words:** Seasonal amplitude, transportation problem, concentration index

---

Giovanni L. Lo Magno  
Department of Economics, Business and Statistics, University of Palermo, e-mail: lomagno.gl@virgilio.it

Mauro Ferrante  
Department of Culture and Society, University of Palermo, e-mail: mauro.ferrante@unipa.it

Stefano De Cantis  
Department of Economics, Business and Statistics, University of Palermo, e-mail: stefano.decantis@unipa.it

## 1 Introduction

The notion of seasonality is very simple and very complex: referring to the former, it is a feature of many natural and human phenomena. The differing intensity of solar rays determines numerous effects on the environment, such as varying levels of humidity and temperature, but also on animal behaviour and human habits [2]. With reference to the latter, the notion of seasonality is an extremely complex concept, whose measurement and analysis is a challenge. Indeed, having reviewed the main indices used in different study contexts, we observed a lack of appropriate measures relating to seasonality, which are capable of taking the cyclical structure of time periods into account. That is, the majority of indices currently used in measuring seasonality (e.g. the Gini concentration index, the coefficient of seasonal variation or the Theil index) do not take into account the natural ordering of time periods (e.g. months). Subsequently, given a pattern in which the total amount of the phenomenon of interest is concentrated in two consecutive months (e.g. January and February) and another pattern in which the phenomenon is concentrated in two very distant months (e.g. January and August), the currently-used indices would evaluate the two patterns as the same. One of the aims of the seasonality index discussed in this paper is to overcome this issue. Thus, after a discussion of a recently-proposed index for measuring seasonality, we will evaluate some of its properties in relation to how it behaves in evaluating unimodal and bimodal distributions.

## 2 A new index for measuring seasonality and its main properties

The approach we would like to propose for measuring seasonality is based on the solution of an appropriately defined transportation problem [1]. The transportation problem is a well-known, linear, minimization problem in which the goal is to minimize the cost of transferring units from a set of warehouses to a set of customers, satisfying the constraints given by the available resources and the requested demands.

We can define a *seasonal pattern* as the vector  $\mathbf{P} = (y_1, y_2, \dots, y_T)$ , with observations for time periods from 1 to  $T$ , such that  $y_t$ , for  $t = 1, 2, \dots, T$ , is non negative. The total amount of the observed phenomenon is  $Y = \sum_{t=1}^T y_t$  and the average value is  $Y/T$ . Furthermore, let  $\mathcal{A} = \{t : y_t > Y/T\}$  be the set of time periods for which the observed value is over the average; each of these time periods has a surplus  $a_t = y_t - Y/T$ . Similarly, let  $\mathcal{B} = \{t : y_t < Y/T\}$  be the set of time periods with observed values under the average, each with deficit  $b_t = Y/T - y_t$ .

In order to eliminate seasonality, the  $T$ -dimensional pattern  $(\frac{Y}{T}, \frac{Y}{T}, \dots, \frac{Y}{T})$  can be obtained by transferring units from time periods in  $\mathcal{A}$  to time periods in  $\mathcal{B}$ . The amount which is transferred from time period  $i \in \mathcal{A}$  to time period  $j \in \mathcal{B}$  is  $x_{ij}$ . We can suppose that transferring one unit from  $i$  to  $j$  has a cost  $c_{ij}$ , thus the cost of all the transfers is  $c = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{B}} c_{ij} x_{ij}$ . The minimum cost  $c^*$  which is

required to eliminate seasonality, through the aforementioned transfers, is the cost corresponding to the optimal solution to the following transportation problem:

$$\begin{aligned}
 \min c &= \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{B}} c_{ij} x_{ij} \\
 \text{s.t.} & \\
 \sum_{j \in \mathcal{B}} x_{ij} &= a_i, \quad \forall i \in \mathcal{A} \\
 \sum_{i \in \mathcal{A}} x_{ij} &= b_j, \quad \forall j \in \mathcal{B} \\
 x_{ij} &\geq 0, \quad \forall i \in \mathcal{A}, j \in \mathcal{B}
 \end{aligned} \tag{1}$$

where the first set of constraints ensures that the amount  $a_i$  is transferred from each time period  $i \in \mathcal{A}$ ; the second set of constraints ensures that each time period  $j \in \mathcal{B}$  receives the amount  $b_j$ ; and the third set of constraints ensures that each transfer is non-negative. The minimum cost  $c^*$  corresponding to the optimal solution to the transportation problem (1) is our absolute measure for seasonality, namely

$$S(\mathbf{P}) = c^* \tag{2}$$

In [3] we demonstrated that the maximum value of  $S(\mathbf{P})$ , holding  $Y$  constant, is:

$$S_{\max}(\mathbf{P}) = \frac{Y}{T} \max_{t \in M} \left\{ \sum_{j=1}^T c_{ij} \right\} \tag{3}$$

This result permits us to construct the relative seasonality index  $S_R$  which is bounded in the interval  $[0, 1]$ :

$$S_R(\mathbf{P}) = \frac{S(\mathbf{P})}{S_{\max}(\mathbf{P})} \tag{4}$$

We recognize that the relation between time periods is cyclical. Time periods can be thought of as collocated in a circumference: the shorter arc with  $i$  and  $j$  as extremes is what is termed *cyclical distance*, and this is the unitary cost we consider for a transfer from  $i$  to  $j$ , namely:

$$c_{ij} = \begin{cases} |i-j| & \text{if } |i-j| \leq \frac{T}{2} \\ T - |i-j| & \text{otherwise} \end{cases} \tag{5}$$

Different unitary costs may be employed in the defining our seasonality index; however, when those costs defined in (5) are adopted, the  $S_R$  index has three important properties, which we deem desirable for measuring seasonality:

- *Scale invariance*:  $S_R(\mathbf{P}) = S_R(\lambda \mathbf{P})$ , with  $\lambda > 0$ .
- *Rotation invariance*:  $S_R(y_1, y_2, \dots, y_T) = S_R(y_T, y_1, y_2, \dots, y_{T-1})$
- *Sensitivity to permutations*: generally, and excluding rotations (see the ‘‘rotation invariance’’ property), the permutations of a pattern are evaluated in a different way (the Gini index is permutation-invariant, thus it can not capture, for example, an increase in seasonality which is consequent on a reduction in the distance between two relevant modes).

### 3 Comparing seasonality indices

In this section we will compare: the Gini index  $G_C$  (normalized in the  $[0, 1]$  interval), the precipitation concentration index  $PCD$  [5] and the relative seasonality index  $S_R$ , when evaluating unimodal and bimodal patterns. Unimodal and bimodal patterns, obtained by adapting von Mises distributions to the discrete case, were used in these comparisons.

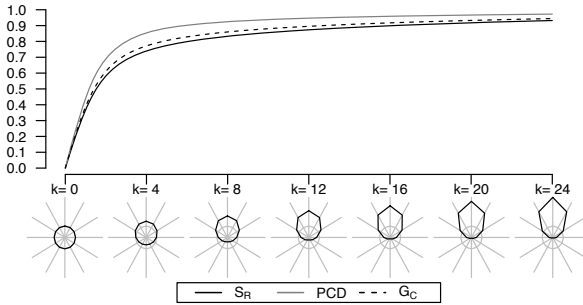
The von Mises distribution [4] is a circular, continuous, unimodal and symmetric distribution, with support in the  $[0, 2\pi]$  interval. Its shape depends on two parameters: the first is the expected value  $\mu$ , which corresponds to the mode; the second is  $\kappa \geq 0$  and it affects the degree of concentration. The higher is  $\kappa$ , the greater is the concentration around the mode; if  $\kappa = 0$ , then the von Mises distribution coincides with the circular uniform distribution.

In order to construct unimodal patterns of size  $T$ , we discretized von Mises distributions by partitioning their supports into  $T$  equal intervals of size  $2\pi/T$  and then calculating the probability related to each interval. Thus, the value  $y_i$  in each resulting pattern is the probability which is related to the  $i$ -th interval of the von Mises' support and, consequently,  $\sum y_i = 1$ . In this study 241 von Mises distributions were considered for the unimodal patterns; these distributions have the same  $\mu$ , but  $k = 0, 0.1, 0.2, \dots, 24$ . The parameter  $\mu$  was set to  $\pi/T$ , namely the midpoint of the first interval. Fig. 1 shows the results of all the indices which were applied to the 241 unimodal distributions under consideration. In order to clarify how the shape of the distribution changes as  $\kappa$  increases, radar charts for the distributions with  $\kappa = 0, 4, 8, 12, 16, 20, 24$  have been included in the lower part of Fig. 1. It can be observed that all the indices increase as  $\kappa$  increases, thus they correctly capture the concentration parameter  $\kappa$ . Furthermore, the relation between the three indices is  $S_R < G_C < PCD$ , with the only exception for the distribution with  $\kappa = 0$ , for which all the indices have the value 0.

In order to observe the behaviour of the indices when they evaluate bimodal patterns, size 12 bimodal patterns were constructed as discrete versions of a 50% mixture of two von Mises distributions, both with the same  $\kappa$  parameter value. The distance  $d$  between the two modes varied between 0 (the unimodal case) and 6 (the maximum possible distance between two time periods). 1 and 5 were considered as two extreme values for  $\kappa$ , which corresponded to a low and a high concentration case around the two modes respectively. The indices values for the patterns when  $\kappa = 1$  are plotted on Fig. 2 and those for the patterns when  $\kappa = 5$  are displayed in Fig. 3.

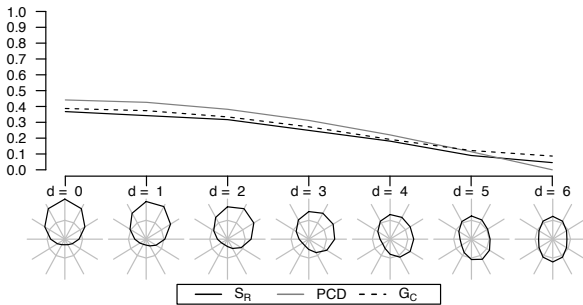
Fig. 2 demonstrates that, for bimodal patterns where  $\kappa = 1$ , the indices decrease in value while  $d$  increases; furthermore they exhibit similar values. However, Fig. 3 also shows that the indices decrease in value while  $d$  increases, but they decrease at a quite different rate. To understand why this difference, it has to be noted that if  $\kappa$  is very high then there is a high concentration around the two modes, thus patterns where  $d > 2$  can be approximately expressed as  $\mathbf{P} = (y_1 = 0.5, 0, \dots, 0, y_{1+d} = 0.5, 0, \dots, 0)$ . Thus, for  $d > 2$ , all the patterns can be roughly considered as permutations. As the Gini index is relatively insensitive to permutations, its value is quite



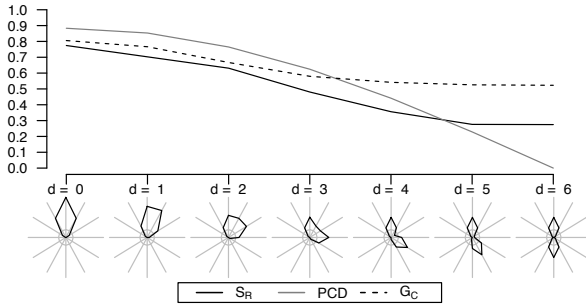


**Fig. 1** Relative seasonality index  $S_R$ , corrected Gini index  $G_C$  and precipitation concentration degree index  $PCD$ , all calculated for discretized von Mises distributions with various values of  $\kappa$  and  $T = 12$ . Radar charts are reported for the distributions with  $\kappa = 0, 4, 8, 12, 16, 20, 24$ . The regular dodecagon in each radar chart corresponds to the uniform pattern.

stable for patterns where  $d > 2$ . This is not a desirable behaviour because a seasonality index should significantly decrease when the distance between two highly concentrated modes increases. In contrast, the  $PCD$  index clearly reveals decreasing values but it is zero when  $d = 6$ . This occurs because  $PCD$  equals zero for patterns where the first half of the pattern is the same as the second, like the pattern (1, 3, 5, 1, 3, 5). A seasonality index should not return zero for patterns like these. Thus, the only index which shows a desirable behaviour is  $S_R$ , which decreases at an appreciable rate and does not reach zero when  $d = 6$ .



**Fig. 2** Relative seasonality index  $S_R$ , corrected Gini index  $G_C$  and precipitation concentration degree index  $PCD$ , all calculated for discretized, bimodal von Mises distributions with  $T = 12$ ,  $\kappa = 1$  and values of the distance between the two modes  $d = 0, 1, \dots, 6$ . Radar charts are reported for the distributions with  $\kappa = 0, 4, 8, 12, 16, 20, 24$ . The regular dodecagon in each radar chart corresponds to the uniform pattern.



**Fig. 3** Relative seasonality index  $S_R$ , corrected Gini index  $G_C$  and precipitation concentration degree index  $PCD$ , all calculated for discretized, bimodal von Mises distributions with  $T = 12$ ,  $\kappa = 5$  and values of the distance between the two modes  $d = 0, 1, \dots, 6$ . Radar charts are reported for the distributions with  $\kappa = 0, 4, 8, 12, 16, 20, 24$ . The regular dodecagon in each radar chart corresponds to the uniform pattern.

## 4 Concluding remarks

The authors of this paper contend that the challenge of measuring seasonality has not received adequate attention in the literature. For example, the Gini index does not take into account the natural ordering of time periods. In addition to permitting a specification of the cost matrix which takes into account for the cyclical structure of time periods, our seasonality index performed well when applied to unimodal and bimodal patterns. Although a set of desirable properties has been highlighted in this paper, we hope that new related issues will come to light as a result of this research. Given these challenges and considering the interdisciplinary nature of seasonal phenomena, the topic of measuring seasonality merits greater attention and from a wider range of points of view.

## References

1. Dantzig, G. B.: Linear Programming and Extensions. Princeton University Press, Princeton (1963)
2. De Cantis, S., Ferrante, M.: Seasonality and tourism. The Sage Encyclopedia of Travel and Tourism. Sage (2016)
3. Lo Magno, G.L., Ferrante, M., De Cantis, S.: A new index for measuring seasonality: a transportation cost approach. (*Submitted to Mathematical Social Sciences*).
4. Jammalamadaka, S.R., SenGupta, A.: Topics in circular statistics. Series on multivariate analysis; vol.5, World Scientific Publishing, Singapore (2001)
5. Zhang, L., Qian, Y.: Annual distribution features of precipitation in China and their interannual variations. *Acta Meteorologica Sinica* 17 (2), 146–163 (2003)

# Three-way Correspondence Analysis for Ordinal-Nominal Variables

## *L'Analisi delle Corrispondenze a Tre-vie per Variabili Ordinali-Nominali*

Rosaria Lombardo and Eric J Beh

**Abstract** This paper presents some variants of three-way polynomial correspondence analysis to analyse associations in three-way contingency tables that are constructed from ordinal and nominal variables. Historically, three-way correspondence analysis has been used for this purpose without regard to the ordinal structure of the variables. Recently, Lombardo et al.(2017) proposed an alternate orthogonal basis of Emerson's polynomials for modelling interactions in three-way contingency tables. Here, we propose the *hybrid* decomposition for modelling cases where not all variables are ordered.

**Abstract** *In questo articolo si propone lo studio di alcune varianti dell'analisi delle corrispondenze a tre-vie con polinomi ortogonali, per analizzare le interazioni tra variabili ordinali e nominali. In letteratura, l'analisi delle corrispondenze a tre-vie è utilizzata per lo studio della dipendenza tra le variabili qualitative senza tener conto della natura ordinale delle categorie delle variabili. Recentemente, per tener conto di tale caratteristica, Lombardo et al. (2017) hanno proposto di considerare i polinomi ortogonali di Emerson come una base ortonormale alternativa per analizzare le interazioni nelle tabelle di contingenza a tre-vie. In presenza di variabili categoriche miste (nominali-ordinali), modelli ibridi di decomposizione saranno considerati.*

**Key words:** Three-way Correspondence Analysis, Ordinal and Nominal Categorical Variables, Tucker3 Decomposition, Trivariate Moment Decomposition, Hybrid Decomposition.

---

R. Lombardo  
Economics Department, University of Campania, via Gran Priorato di Malta, Capua (CE), Italy  
Tel.: +390810601382  
Fax: +390823622984  
e-mail: rosaria.lombardo@unina2.it

E.J. Beh  
School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, 2308, NSW, Australia e-mail: eric.beh@newcastle.edu.au

## 1 Introduction

In this paper we aim to present some variants of ordered three-way correspondence analysis recently proposed in the literature (Lombardo et al. 2017) for analysing nominal-ordered variables. The objective of ordered three-way correspondence analysis is to gain insight into the symmetric association among the three ordered variables that make up a contingency table (Lombardo et al. 2017). This insight is in addition to what the classical approaches reveal when using methods of generalized singular value decomposition such as the Tucker3 or the PARAFAC decomposition (Kroonenberg 2008). Our proposal involves the trivariate moment decomposition of the data (Lombardo et al. 2017) and the visualization of category trends using polynomial biplots, both of which procedures will be explained below. Indeed, from a geometrical point of view, modelling the association using a new dimensional space based on Emerson's polynomial components for the ordered variables, can be seen as the major purpose of ordered three-way correspondence analysis. Here we consider the case where not all the three variables have ordered categories. As a consequence, we merge the various features of the Tucker3 decomposition with the trivariate moment decomposition using Emerson's polynomials and Tucker3 components, via *hybrid* decompositions. Modelling this association using Emerson's polynomials and Tucker3 components deserves special attention as does the construction of the graphical representations.

The paper is organised as follows. In Section 2 we briefly present classic three-way correspondence analysis. Section 3 describes its ordinal variant based on orthogonal polynomials and Section 4 presents some possible *hybrid* decompositions. In Section 5, we will illustrate in brief the proposed analysis using data from the results of a survey of the Dutch Central Bureau of Statistics (Israëls 1987).

## 2 Three-way correspondence analysis for unordered variables

Three-way correspondence analysis can be viewed as a generalization of two-way correspondence analysis for analysing the three-way chi-squared statistic  $X_{IJK}^2$  or its analog  $X_{IJK}^2/n$ . This index can be partitioned into four orthogonal terms (see, for example, Lancaster 1951) where the first three terms are the pairwise chi-squared statistics obtained by aggregating across the categories of each variable and the last term represents the trivariate interaction among the variables. That is, the sum of these four terms gives Pearson's mean squared three-way contingency coefficient.

$$X_{IJK}^2/n = X_{IJ}^2/n + X_{IK}^2/n + X_{JK}^2/n + X_{int}^2/n. \quad (1)$$

To decompose  $X_{IJK}^2/n$ , we can consider a three-way generalisation of the singular value decomposition, i.e. the Tucker3 model decomposition (Tucker 1966; Kroonenberg 2008, p. 54; Beh and Lombardo 2014, Section 11.6) which computes the components for each of the spaces of the three categorical variables, and in addition,

a three-way core array containing the elements that reflect the strength of the links among these components. For a detailed discussion of three-way correspondence analysis for nominal variables, see Carlier and Kroonenberg (1996), Kroonenberg (2008, Chap. 17), Beh and Lombardo (2014, Chapter 11). Let  $\mathbf{P} = (p_{ijk})$  be a general three-way table of joint relative frequencies from the cross-classification of  $n$  units according to three variables, called row, column and tube variables, respectively. Define  $\mathbf{D}_I, \mathbf{D}_J$  and  $\mathbf{D}_K$  as  $I \times I, J \times J$  and  $K \times K$  diagonal matrices whose general elements are the row, column and tube marginal proportions,  $p_{i\bullet\bullet}, p_{\bullet j\bullet}$  and  $p_{\bullet\bullet k}$ , respectively. Furthermore, let  $\mathbf{\Pi} = \left( \frac{p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}} - 1 = \pi_{ijk} \right)$  be the array of the deviations from the three-way independence model.

Using the Tucker3 model, the general form of the  $X^2_{IJK}/n$  decomposition can be written as

$$\mathbf{\Pi} = \mathbf{AG}(\mathbf{B} \otimes \mathbf{C})' + \mathbf{e} = \hat{\mathbf{\Pi}} + \mathbf{e}, \tag{2}$$

where  $\mathbf{\Pi}$  is the flattened table of the deviations from the three-way independence model of dimension  $I \times JK$ ;  $\mathbf{G}$  is the flattened core array of dimension  $P \times QR$ , and  $\mathbf{A}, \mathbf{B}$  and  $\mathbf{C}$  are the component matrices associated with the row, column and tube variables, of dimension  $I \times P, J \times Q$  and  $K \times R$  (with  $P \leq I, Q \leq J$  and  $R \leq K$ ), respectively, and  $\mathbf{e}$  represents the error of approximation between the observed  $\pi_{ijk}$  and their predicted values  $\hat{\pi}_{ijk}$ . When  $P = I, Q = J, R = K$ , Equation (2) yields an exact decomposition. The component matrices are orthonormal with respect to  $\mathbf{D}_I, \mathbf{D}_J$  and  $\mathbf{D}_K$ . The general elements of the core array can be interpreted as the generalized, or three-way analogue, of the two-way singular values. For the sake of brevity, we do not provide a comprehensive description of the bivariate interaction terms of  $X^2_{IJK}/n$  which can also be modelled.

### 3 Three-way correspondence analysis for ordered variables

When the categories of the variables are ordered, Lombardo et al. (2017) recently proposed to incorporate such a design feature into the decomposition model by replacing the Tucker3 components with the orthogonal polynomials of Emerson (1968), thereby defining the *trivariate moment decomposition*. To form the polynomial basis we can generate as many orthogonal polynomials as there are ordered categories. In general, the first polynomial for each ordered variable of the three-way table, represents the *zeroth-order* polynomial and consists of values that are all constant. The second polynomial is the *first-order* polynomial that describes the variation in the location of the categories. The third polynomial is the *second-order* orthogonal polynomial that reflects the variation in the dispersion of the categories. Higher-order polynomials represent higher-order moments of the ordered categories. Emerson considered a computationally efficient way of calculating these orthogonal polynomials by using a three-term recurrence relation; see for example Beh and Lombardo (2014, p. 94) and Lombardo et al. (2016).

The matrices of the row, column and tube orthogonal polynomials are denoted by  $\mathcal{A} = \{\alpha_{iu}\}$ , (for  $i = 1, \dots, I$  and  $u = 0, \dots, I - 1$ )  $\mathcal{B} = \{\beta_{jv}\}$  (for  $j = 1, \dots, J$  and  $v = 0, \dots, J - 1$ ) and  $\mathcal{C} = \{\gamma_{kw}\}$  (for  $k = 1, \dots, K$  and  $w = 0, \dots, K - 1$ ), respectively. Like the Tucker3 components, the row polynomials are orthogonal with respect to the marginal proportions  $p_{i\bullet\bullet}$ . The column and tube polynomials are orthogonal with respect to the marginal proportions  $p_{\bullet j\bullet}$  and  $p_{\bullet\bullet k}$ , respectively. As mentioned above, the zeroth-order orthogonal polynomial is a constant, the first polynomial is linear (respecting the ordinality of categories), the second polynomial is quadratic (describing the variation of the category dispersion), and so on. The *trivariate moment decomposition* also calculates the generalized correlations that replace the links between components in the core array; they are also referred to as the *trivariate generalized correlation* between the  $u$ -th-order polynomial component of the first variable, the  $v$ -th-order polynomial component of the second variable and the  $w$ -th-order polynomial component of the third variable. Like the core array, the generalized correlation table of dimension  $U \times V \times W$  is not super-diagonal. Its flattened form is given by  $\mathbf{Z} = \mathcal{A}' \mathbf{D}_I \mathbf{\Pi} (\mathbf{D}_J \mathcal{B} \otimes \mathbf{D}_K \mathcal{C})$ . By replacing  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  in Equation (2) with their orthogonal polynomial equivalents,  $\mathcal{A}$ ,  $\mathcal{B}$  and  $\mathcal{C}$ , Lombardo, et al. (2017) demonstrated that the decomposition of the three-way Pearson's coefficient,  $X_{IJK}^2/n$  can be expressed in the similar manner as in Equation (2) such that

$$\mathbf{\Pi} = \mathcal{A} \mathbf{Z} (\mathcal{B} \otimes \mathcal{C})' + \mathbf{e}. \quad (3)$$

The dimension of the  $\mathbf{Z}$  table is  $U \times VW$  (with  $U \leq I, V \leq J$  and  $W \leq K$ ) where  $U, V$  and  $W$  represent the number of columns in the polynomial component matrices  $\mathcal{A}$ ,  $\mathcal{B}$  and  $\mathcal{C}$  for the first, second and third way, respectively, and  $\mathbf{e}$  is an error term that is equal to zero when  $U = I, V = J$  and  $W = K$ . As for the Tucker3 model, the bivariate and trivariate interaction terms of the global association can also be modelled; see Equation (1).

#### 4 Three-way correspondence analysis for mixed variables

When the three-way contingency table consists of nominal and ordinal variables, the approximation of the global dependence, or total inertia, in  $\mathbf{\Pi}$  involves computing the Tucker3 components for the nominal variables and Emerson's polynomial for the ordered variables. Two cases involving such a structure can arise: 1) one nominal and two ordered variables; 2) two nominal variables and only one ordered. For three ordered variables refer to Lombardo et al. (2017). After computing the polynomial for an ordered variable, say the column variable, and the Tucker3 components for the row and tube variables, the *hybrid* decomposition takes on the form

$$\mathbf{\Pi} = \mathbf{A} \mathbf{Z} (\mathcal{B} \otimes \mathcal{C})' + \mathbf{e} \quad (4)$$

where  $\mathbf{e}$  is the error term and the components  $\mathbf{A}$ ,  $\mathbf{C}$  and  $\mathcal{B}$  for the row, tube and column variables are computed using an iterative *hybrid* algorithm. At the first stage of the algorithm, the components  $\mathbf{A}$  and  $\mathbf{C}$  are derived using the singular vectors of the flattened tables  $\mathbf{\Pi}_{I \times JK}$  and  $\mathbf{\Pi}_{K \times IJ}$ , respectively. The components of the third ordered variable are computed by Emerson's polynomials related to  $\mathbf{\Pi}_{J \times IK}$ . A full rank decomposition of the association in a three-way contingency table using the *hybrid* models may be achieved by choosing  $(P = I, Q = J, R = K, U = I, V = J, W = K)$  provided that the products of  $PQ \geq R, UV \geq W, PR \geq Q, UW \geq V$ , and  $QR \geq P, VW \geq U$  (Kroonenberg, 2008, p. 66), in this case the convergence of the algorithm is quickly reached. While the number of polynomials should be always equal to the number of categories in a variable (Lombardo et al. 2017), the number of Tucker3 components can be smaller. In this situation, the convergence of the *hybrid* algorithm will be reached only after a small number of iterative steps. A full decomposition is always used when all the three variables are ordered, as it is for model (3), but is seldom used in practice when the variables are not all ordered. In all cases, given the orthogonality of the components  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$ , and of the polynomials  $\mathcal{A}, \mathcal{B}, \mathcal{C}$  with respect to the marginal matrices  $\mathbf{D}_I, \mathbf{D}_J$  and  $\mathbf{D}_K$ , respectively, the inertia or Pearson's mean squared contingency coefficient can be written also in terms of generalized correlations, such that  $\frac{\chi^2_{JK}}{n} = \|\mathbf{\Pi}\|^2 = \|\mathbf{G}\|^2 = \|\mathbf{Z}\|^2$ .

### 4.1 Polynomial biplots

To graphically depict the association structure in a three-way contingency table where at least one categorical variable is ordinal, we shall focus here on the single-variable polynomial biplot. This is only one of a variety of different types of biplots discussed in the literature; see for example, Kroonenberg (2008), Carlier and Kroonenberg (1996), Gower et al. (2016) and Lombardo et al. (2017). For the sake of brevity and consistency with the example illustrated in the next Section, here we define the coordinates of the single-variable polynomial biplot when only the column variable is ordinal. The principal polynomial coordinates, or reference-mode coordinates, are defined as

$$\mathbf{F}_{(I \times V R)} = \mathbf{AZ}_{(P \times V R)} \quad \{ = f_{i,vr} = \sum_{p=1}^P a_{ip} z_{pvr} \}$$

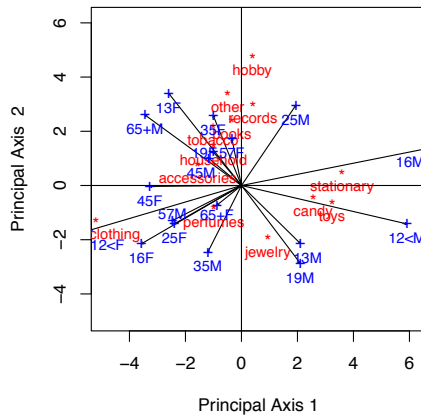
and are the row Tucker3 components weighted by the generalized correlations. The column-tube interactive coordinates are standard polynomial coordinates and are defined as

$$\mathbf{H}_{(JK \times V R)} = (\mathcal{B} \otimes \mathbf{C}) \quad \{ = h_{jk,vr} = \beta_{jv} c_{kr} \}. \tag{5}$$

For this biplot, the coordinates for both the row and column-tube categories are displayed in the space defined by the column $\times$ tube interactive components. We get as many interactive polynomial axes as the product number of the  $V$  column polynomials times the  $R$  tube Tucker3 components.

## 5 Example: Shoplifting data

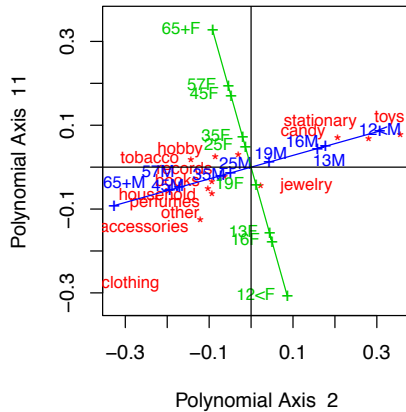
To illustrate the applicability of three-way, ordered and nominal, correspondence analysis, we consider the contingency table from the a survey undertaken by the Dutch Central Bureau of Statistics (Israëls 1987). The data concerns the number of men and women suspected of shoplifting in 1977 and 1978, both in Dutch general stores and in big textile stores. The row categories consists of 13 items stolen: *clothing, clothing accessory, tobacco and/or provisions, stationary, books, records, household goods, candy, toys, jewelry, perfume, hobby and/or tools* and *other items*. The column categories consist of 9 age groups (in years) of the perpetrators: less than 12, 12 to 14, 15 to 17, 18 to 20, 21 to 29, 30 to 39, 40 to 49, 50 to 64, 65 and over, and the tube categories are *male* and *female*. The  $X^2 = 22317.9$  indicates that there is a strong significant association among the three variables. The chosen number of dimensions of the *hybrid* model is  $P = 4, V = 9, R = 2$  where the algorithm converges after 10 iterations and reconstructs about 92% of the total inertia. The main purpose of our study of these data is to describe the shoplifting behaviour



**Fig. 1** Shoplifting data: three-way CA

of items as a function of age and gender. We want to investigate how does the age of the perpetrator and gender influence the types of items that are stolen. What sources





**Fig. 2** Shoplifting data: Polynomial biplot, ordered three-way CA-linear trend

of variation of the age groups for males or females can help to describe this association? For the sake of brevity, here we only look at the linear polynomial of age that can allow to see when there is a linear growth or decline over time for males and females. To highlight the difference of the hybrid analysis with respect to the classic analysis, we first display in Figure 1 the results of the classic three-way correspondence analysis. Figure 1 shows the *nested-mode* biplot classically proposed for three-way correspondence analysis (see Kroonenberg 2008, p.443), where the age and sex categories are combined to reflect the interactive nature of the variables along each axis. Here the origin of the plot indicates no association. It shows that males in the young age groups have a propensity to steal *toys*, *candy* and *stationary* and males in the older age groups together with the young female 13F principally stole *household, goods* and *tobacco*. Further it also shows that the female young age group, *less than 12*, mainly stole *clothing*. However it does not illustrate a clear trend of the age groups or a prevalence of the males with respect to the female perpetrators, unlike the polynomial biplot of Figure 2. To portray the association among the three categorical variables where only the column variable is ordered we use the polynomial biplot described in Section 4.1. Since we have only the columns that are ordered, we obtain nine ordered column polynomials (0th, 1st, 2nd, etc.) and two Tucker3 components for the gender variable. In total we compute 18 interactive polynomial axes- the first is the combination of the constant column polynomial and the first Tucker3 component of the gender variable, the second interactive axis is defined by the combination of the linear column polynomial and the first Tucker3 component of the gender variable and so on. This axis reflects changes in the linearity of columns (ages) given the tube categories (gender). In Figure 2, we consider the 2nd and 11th axis, since the eleventh axis represents the combination of the lin-

ear column polynomial and the second Tucker3 component of the tube variable *-it again represents the linearity of ages given the gender*. The origin of the axis represents the mean age of the perpetrators. Figure 2 shows a clear linear trend of age groups for the males and females. The male linear trend is associated with the first axis while the female linear trend is associated with the second axis. It is more evident that males stole a great variety of items than females (indeed there were 22597 stolen items by males against 10504 stolen items by females). On the right-hand side of the first axis, we observe that males in the young age groups had a propensity to steal *toys, candy* and *stationary*. The left-hand side of the first axis shows that males in the older age groups principally stole *records, books, household, perfumes, other* and *accessories*. Further it also shows that females in the young age groups stole mainly *clothing* and *jewelry*, while female in the older age groups mainly stole *hobby* and *tobacco* items.

## 6 Conclusion

In this paper we have discussed in a unified framework three-way correspondence analysis consisting of nominal and ordered categorical variables. We have done this by considering the use of the Tucker3 orthogonal base with the orthogonal polynomial base (Emerson 1968) defining a three-way hybrid decomposition. We have proposed that one way to visually summarize the association is to consider the *polynomial biplot*; see Figure 2. By observing the proximity and position of the points in this display, we can obtain a more detailed summary of the associations among the variables in terms of category trends.

## References

1. Beh EJ, Lombardo R (2014) Correspondence Analysis, Theory, Practice and New Strategies. Wiley, Chichester
2. Carlier A, Kroonenberg PM (1996) Biplots and decompositions in two-way and three-way correspondence analysis. *Psychometrika* 61: 355–373
3. Emerson PL (1968) Numerical construction of orthogonal polynomials from general recurrence formula. *Biometrics* 24: 696–701
4. Gower JC, Le Roux NJ, Sugnet GL (2016) Biplots: Qualitative Data. *WIREs Computational Statistics* 8: 82–111
5. Israëls (1987) Eigenvalue Techniques for Qualitative Data. Leiden: DSWO Press.
6. Kroonenberg PM (2008) Applied Multiway Data Analysis. Wiley, Hoboken, NJ
7. Lancaster O H (1951) Complex contingency tables treated by the partition of the chi-square. *Journal of Royal Statistical Society, Series B* 13: 242–249
8. Lombardo R, Beh EJ, Kroonenberg PM (2016) Modelling trends in ordered correspondence analysis using orthogonal polynomials. *Psychometrika* 81 (2): 325–349
9. Lombardo R, Kroonenberg PM, Beh EJ (2017) Three-way polynomial correspondence analysis for ordered contingency tables. Submitted
10. Tucker LR (1966) Some mathematical notes on three mode factor analysis. *Psychometrika* 31: 279–311

# Log-mean linear models for causal inference

## *Modelli log-mean linear per inferenza causale*

Monia Lupparelli and Alessandra Mattei

**Abstract** We discuss a log-mean linear regression approach to deal with causal inference when the interest is in assessing the effect of treatment on a set of multiple (binary) outcomes which might be not independent. We explore how the effect of treatment on joint outcomes can be decomposed considering the effect on single outcomes and the effect on their joint distribution. The method is illustrated through a randomized experiment concerning the effect of honey on nocturnal cough associated with childhood upper respiratory tract infections.

**Abstract** *Un approccio basato su regressioni log-mean linear viene discusso per fare inferenza causale quando si è interessati a valutare l'effetto di un trattamento su variabili risposta (binarie) multiple che potrebbero non essere indipendenti. Si esplora come l'effetto del trattamento su variabili congiunte possa essere decomposto considerando l'effetto sulle singole variabili e sulla loro distribuzione congiunta. Il metodo viene illustrato attraverso uno studio randomizzato relativo all'effetto del miele sulla tosse notturna durante l'infanzia in presenza di infezioni del tratto respiratorio.*

**Key words:** Causal Relative Risks; Intrinsic and Extrinsic causal effects; Product potential outcomes

## 1 Introduction

We focus on assessing causal effects on multiple binary outcomes using the potential outcome approach to causal inference (see [1] for a comprehensive review). In

---

Monia Lupparelli

Department of Statistical Sciences, University of Bologna, e-mail: monia.lupparelli@unibo.it

Alessandra Mattei

Department of Statistica, Informatica, Applicazioni, University of Florence,

e-mail: mattei@disia.unifi.it

particular, we define an enlarged set of potential outcomes of interest and suitable relative risk causal estimands for these outcomes. We also show that casual effects on joint outcomes can be decomposed into two components accounting respectively for the effects on single outcomes and the effect of their joint dependence structure. For inference, we propose the log-mean linear regression approach for multiple binary outcomes recently developed by [2].

We apply our framework to a real analysis of a double-blinded randomization study, that we name the *Honey study*, aimed to evaluate the effects of a single nocturnal dose of buckwheat honey versus no treatment on nocturnal cough and sleep difficulty associated with childhood upper respiratory tract infections. From September 2005 through March 2006, 72 children aged between 2 and 18 years with cough attributes characterized by the presence of rhinorrhea and cough for 7 or fewer days duration were recruited on presentation for an acute care visit from a single university-affiliated pediatric practice in Hershey, Pennsylvania (USA). Subjective parental assessments about their child cough symptoms were assessed both previous and after the treatment administration (see [4] for further details about the study). Here we explore the effect of honey on three attributes of nocturnal cough: *Cough Bothersome*, *Cough Frequency* and *Cough Severity*.

## 2 The model

### 2.1 Preliminaries and background

Given the finite set  $V = \{1, \dots, p\}$ , let  $Y_V = (Y_v)_{v \in V}$  be the vector of binary outcomes of interest. For every  $D \subseteq V$ , let  $Y_D = 1_D$  be the event of joint success given that  $Y_v = 1$ , for each  $v \in D$ , where  $1_D$  is a vector of 1s of size  $|D|$ . Then, for every non-empty subset  $D$  of  $V$ , we define the product outcome

$$Y^D = \prod_{v \in D} Y_v \tag{1}$$

which is a binary variable taking value 1 in case of joint success and 0 otherwise. The interest in our analysis is exploring the effect of a treatment both on joint successes, that is, the effect on each product outcomes, as well as on single outcomes. Therefore, let  $Y^V = (Y^D)_{D \subseteq V, D \neq \emptyset}$  be the augmented vector of outcomes of interest.

In order to formally define the causal effects of interest, we first introduce the potential outcome approach to causal inference. We take a super-population perspective, considering the observed group of units as a random sample from an infinite super-population, and we focus on causal effects of a binary treatment, so that, each unit in the sample can potentially be assigned to an active treatment group ( $w = 1$ ) or to a control group ( $w = 0$ ). Under the Stable Unit Treatment Value Assumption (SUTVA, [6]), we can define for each outcome variable,  $Y_v$ ,  $v \in V$ , two potential outcomes for each unit. Let  $Y_v(0)$  denote the value of  $Y_v$  under treatment  $w = 0$ , and

let  $Y_v(1)$  denote the value of  $Y_v$  under treatment  $w = 1$ . Let  $Y_V(w) = (Y_v(w))_{v \in V}$  be the random vector including potential outcomes for every variable under treatment level  $w$ ,  $w = 0, 1$ . Then, for every non-empty subset  $D$  of  $V$ , let

$$Y^D(w) = \prod_{v \in D} Y_v(w), \quad w = 0, 1 \tag{2}$$

be the potential product outcome and let  $Y^V(w) = (Y^D(w))_{D \subseteq V, D \neq \emptyset}$  be the vector of all product potential outcomes, for  $w = 0, 1$ . Let  $P(Y^D(w) = 1)$  denote the probability of the joint success  $Y_D = 1_D$  under treatment  $w = 0, 1$ , for any non-empty subset  $D$  of  $V$ .

### 2.2 Causal estimands

In the potential outcome approach causal effects are defined as comparison of potential outcomes,  $Y_v(1)$  versus  $Y_v(0)$ , for a common set of units, for every  $v \in V$ . We focus on causal relative risks, that is  $RR_v = \frac{P(Y_v(1)=1)}{P(Y_v(0)=1)}$ , for each single outcome. Then, we define the causal relative risk for any product outcome  $Y^D$ :

$$RR_D = \frac{P(Y^D(1) = 1)}{P(Y^D(0) = 1)}, \quad D \subseteq V. \tag{3}$$

We adopt the convention,  $RR_\emptyset = 1$ .

For any product outcome  $Y^D \in Y^V$ , we expect that the causal effect in (3) combines the effect of the treatment on “nested” product outcome  $Y^{D'}$ , for any non-empty subset  $D'$  strictly included in  $D$ , with the effect of the treatment on the joint distribution of  $Y_D$  with  $D \subseteq V$ . In particular, for any  $Y^D$ , we define the *intrinsic causal effect (ICE)* and the *extrinsic causal effect (ECE)*:

$$ICE_D = g(RR_{D'})_{D' \subset D}, \quad D \subseteq V \tag{4}$$

$$ECE_D = h[P(Y_D(0), Y_D(1))], \quad D \subseteq V, \tag{5}$$

for two suitable functions  $g(\cdot)$  and  $h(\cdot)$ . Basically, for any  $Y^D \in Y^V$ , the *ICE* accounts for the effect of treatment deriving from the product structure of  $Y^D$ , whereas the *ECE* accounts for the effect of treatment of the joint (dependence) structure of  $Y_D$ . We expect that for any product outcome  $Y^D \in Y^V$ , the causal effect of treatment defined in (3) is a suitable combination of the intrinsic and the extrinsic components in (4) and in (5), respectively.

### 2.3 Observed data and assignment mechanism

Unfortunately, we cannot directly observe both  $Y_V(0)$  and  $Y_V(1)$  for any subject. For each unit let  $W$  denote the actual treatment assignment:  $W = 0$  for units assigned to the control group, and  $W = 1$  for units assigned to the treatment group. We observe  $Y_V^{obs} \equiv Y_V(W) = W \cdot Y_V(1) + (1 - W) \cdot Y_V(0)$ , but the other potential outcomes,  $Y_V^{mis} \equiv Y_V(1 - W) = (1 - W) \cdot Y_V(1) + W \cdot Y_V(0)$ , are missing. Therefore, in order to learn about the causal effects of interest it is crucial to posit an assignment mechanism. In what follows, we will maintain the following assumption:

**Assumption 1** *Random treatment assignment:*  $P(W|Y_V(0), Y_V(1), X) = P(W)$  where  $X$  is a vector of observed covariates

Under the randomization assumption, which holds by design in randomized experiments, we propose a model-based approach to causal inference. Specifically, we propose a log-mean linear model for potential outcomes, and we derive maximum likelihood estimators of the causal parameters of interest. This approach appears to be particularly appealing because the causal effects of interest are directly related to model parameters.

### 2.4 Log-mean linear model

We assume that each random binary vector  $Y_V(w)$  with  $w = 0, 1$  follows a Multivariate Bernoulli distribution with mean parameter vector  $\mu_V(w) = (\mu_D(w))_{D \subseteq V}$ , with

$$\mu_D(w) = P(Y_D = 1_D), \quad w = 0, 1, \quad D \subseteq V. \quad (6)$$

Therefore, any product potential outcome  $Y^D(w)$  is a Bernoulli variable with probability parameter  $\mu_D$ , with  $D \subseteq V$ . From [5], let  $\gamma_W(w) = (\gamma_D(w))_{D \subseteq V}$  be the log-mean linear parameter vector of the joint distribution of  $Y_V(w)$ , with

$$\gamma_D(w) = \sum_{D' \subseteq D} (-1)^{|D \setminus D'|} \log \mu_{D'}(w), \quad w = 0, 1, \quad D \subseteq V. \quad (7)$$

Following [2] we propose a log-mean linear regression for modelling the distribution of the multivariate potential outcome  $Y_V(w)$ . The resulting model is given by a sequence of joint regressions

$$\gamma_D(w) = \alpha_D + \beta_D w, \quad w = 0, 1, \quad D \subseteq V. \quad (8)$$

[2] proved that  $\log RR_D = \sum_{D' \subseteq D} \beta_{D'}$  and, therefore, the treatment effect in Equation (3) is

$$RR_D = \prod_{D' \subseteq D} \exp(\beta_{D'}), \quad D \subseteq V, \quad (9)$$

for any product outcome  $Y^D \in Y^V$ . Furthermore, for any  $D \subseteq V$ , we define

**Table 1** Honey Study: Estimates of the causal effects (standard errors in parentheses)

Relative Risk	Extrinsic		Intrinsic		
	Estimate	Causal Effect	Estimate	Causal Effect	
$RR_B$	1.578 (0.344)				
$RR_F$	1.923 (0.511)				
$RR_S$	1.537 (0.412)				
$RR_{\{B,F\}}$	1.914 (0.554)	$ECE_{\{B,F\}}$	0.631 (0.128)	$ICE_{\{B,F\}}$	3.034 (1.342)
$RR_{\{B,S\}}$	1.581 (0.453)	$ECE_{\{B,S\}}$	0.652 (0.134)	$ICE_{\{B,S\}}$	2.425 (1.086)
$RR_{\{F,S\}}$	1.725 (0.516)	$ECE_{\{F,S\}}$	0.583 (0.140)	$ICE_{\{F,S\}}$	2.956 (1.495)
$RR_{\{B,F,S\}}$	1.632 (0.499)	$ECE_{\{B,F,S\}}$	1.459 (0.288)	$ICE_{\{B,F,S\}}$	1.119 (0.304)

$$ICE_D = \prod_{D' \subset D} \left[ RR_{D'}^{(-1)^{|D \setminus D'|}} \right]^{-1}, \tag{10}$$

$$ECE_D = \exp(\beta_D). \tag{11}$$

Next proposition shows that the decomposition of the causal effect for each product outcome  $Y^D$  into the intrinsic and the extrinsic component naturally follows.

**Proposition 1.** *Under the log-mean linear regression approach in Equation (8), for any product outcome  $Y^D$ ,*

$$RR_D = ICE_D \times ECE_D, \quad D \subseteq V. \tag{12}$$

*Proof.* The proof derives from Lemma 3 and Proposition 1 in [3] marginalizing over the set of covariates. In particular, the result follows because  $ICE_D = \prod_{D' \subset D} \exp(\beta_{D'})$ .

Notice that in case of a single outcome  $Y_v$ ,  $ICE_v = 1$  and  $RR_v = ECE_v = \beta_v$ , for any  $v \in V$ . Furthermore, in case  $Y_D$  should be a subset of independent outcomes, from [5] we have that  $\beta_D = 0$  and, therefore,  $RR_D = ICE_D$ , for any  $D \subseteq V$ . Finally, see [3] for an in-depth discussion about intrinsic and extrinsic effects in a wider context including also a set of covariates, which allows us to assess the heterogeneity of the causal effects.

### 3 The Honey study

Given the finite set  $V = \{b, f, s\}$ , let  $Y_V = (Y_b, Y_f, Y_s)$  define the vector of the three binary outcomes associated respectively to the three cough attributes, *Bothersome*, *Frequency* and *Severity*. These outcomes have been properly dichotomized such that they take level 1 in case the attribute is absent or low and the level 0 otherwise (see [3] for details). We are interested in exploring the effect of honey on achieving joint successes, that is, the effect on each single outcome  $Y_b, Y_f, Y_s$  and on the four resulting product outcomes, shortly denoted as  $Y^{bf}, Y^{bs}, Y^{fs}$  and  $Y^{bfs}$ . Therefore,

let  $Y^V = (Y^D)_{D \subseteq V, D \neq \emptyset} = (Y_b, Y_f, Y_s, Y^{bf}, Y^{bs}, Y^{fs}, Y^{bfs})$  be the augmented vector of outcomes of interest.

We specified a log-mean linear regression model for the joint distribution of the three outcomes  $Y_V = (Y_b, Y_f, Y_s)$  related to the cough attributes in the honey data set. The estimates of all causal effects with their standard errors are collected in Table 1.

We get positive estimates of the honey causal effects on each single outcome, and in particular the honey shows the strongest positive effect in reducing the frequency of nocturnal cough. Moreover the treatment has a positive effect on any product outcome and this means that the honey improves conditions of children with more than one critical cough attribute. In particular the honey treatment is more effective for patterns including the frequency attribute. We also notice that for product outcomes, the intrinsic effect is always stronger than the extrinsic one, except for the greatest pattern including all the attributes. These results show that the effect of treatment on the outcome dependent structure is not negligible and a multivariate approach for causal inference is definitively more suitable than an univariate one.

**Acknowledgements** We are grateful to Professor Ian Michael Paul (Penn State University College of Medicine, USA) and Professor Tonya Sharp King (Penn State University College of Medicine, USA) for providing us the data about the honey study.

## References

1. Imbens, G. W., Rubin, D. B.: Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction. Cambridge University Press, New York (2015)
2. Lupporelli, M., Roverato, A.: Log-mean linear regression models for binary responses with an application to multimorbidity. *Journal of the Royal Society, Series C.* (to appear)
3. Lupporelli, M., Mattei, A.: Causal inference for binary non-independent outcomes. (submitted)
4. Paul, I. M., Beiler, J., McMonagle, A., Shaffer, M. L., Duda, L., Berlin, C. M.: Effect of honey, dextromethorphan, and no treatment on nocturnal cough and sleep quality for coughing children and their parents. *Archives of Pediatrics and Adolescent Medicine.* **161**, 1140–1146 (2007)
5. Roverato, A., Lupporelli, M., La Rocca, L.: Log-mean linear models for binary data. *Biometrika.* **100**, 485–494 (2013)
6. Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* 25, 279–292



# Research on the Risk Factors accountable for the occurrence of degenerative complications of type 2 diabetes in Morocco: a prospective study

## *Ricerca di fattori di rischio legati all'occorrenza di complicanze degenerative del diabete di tipo 2 in Marocco: uno studio prospettico*

Badiaa Lyoussi<sup>2</sup>, Zineb Selih<sup>1,2</sup>, Mohamed Berraho<sup>1</sup>, Karima El Rhazi<sup>1</sup>, Youness El Achhab<sup>1</sup>, Adiba El Marrakchi<sup>3</sup>, Chakib Nejjari<sup>1,4</sup>

### Abstract

*Aims:* Our study aims to determine associated risk factors with complications of diabetes in patients with type 2 diabetes followed in primary care centers in Morocco.

*Methods:* We conducted a nested case – control study. Cases were type 2 diabetic's patients who suffered from degenerative complication after diabetes diagnosis; controls were type 2 diabetic's patients with no complications of diabetes at the time of inclusion in the cohort. The analysis was performed separately for women and men in order to determine the specificity of each sex factor.

*Results:* 732 patients with or without complications were identified. Retinopathy is the most frequent (41.2%) followed by diabetic neuropathy (28.4%) and cardiovascular complications (26.2%). For women, low economic level (OR<sub>adj</sub> = 11.36, 95% CI 5.59 - 23.25), forget the treatment (OR<sub>adj</sub> = 3.42, 95% CI 1.29 - 9.09), urban environment (OR<sub>adj</sub> = 3.97, 95% CI 0.04 - 0.17), very high level of stress (OR<sub>adj</sub> = 2.94, 95% CI 1.00 - 8.63), and overweight (OR<sub>adj</sub> = 2.50, 95% CI 1.12 - 5.53), remained significant with the risk of degenerative complications after adjustment.

However, in unadjusted analysis for men, the low socioeconomic level and the patients without professional activities increased the degenerative complication risk. The patients with overweight [5.96 (95% CI: 1.61 – 22.10)], with dyslipidemia [3.09 (95% CI: 1.51- 6.33)] and patients treated by a general physician [4.57 (95% CI: 1.24 – 16.82)] were a higher risk for degenerative complication.

*Conclusion:* These findings suggest that some risk factors of degenerative complication of type 2 diabetes are strongly linked with the Moroccan context. This study highlighted important

---

<sup>1</sup> Laboratory of Epidemiology, Clinical Research and Community Health, Faculty of Medicine and Pharmacy, Fez, Morocco.

<sup>2</sup> Laboratory of Physiology, Pharmacology and Environmental Health, Faculty of Sciences Dhar El Mehraz, Fez, Morocco.

<sup>3</sup> Reference Centre for diabetes care, Fez, Morocco.

<sup>4</sup> National School of Public Health in Rabat, Morocco.

areas for health care intervention and provided a reminder for vigilance when known risk factors for complications are present.

**Key words:** Type 2 diabetes, Risk factor, Degenerative complication

## **Introduction**

Morocco is currently experiencing a demographic, nutritional and epidemiological transition, with a proliferation of cases of diabetes similar with current trends on a global scale. Its degenerative complications represent a heavy burden in terms of morbidity, mortality, but also in terms of impact and socio-economic cost. Many of the complications of diabetes can be avoided or delayed by preventative measures and programs to manage this disease. Indeed, actions and preventive measures targeting the determinants of complications of diabetes require their knowledge and identification. Besides, real work missed about the complications of diabetes and these determinants.

This work was included in the study of the degenerative complications of diabetes in the Moroccan diabetic in terms of descriptive and analytical epidemiology by studying the main factors associated with degenerative complications. Use this like a template.

Instead of simply listing Section headings of different levels we recommend to let every heading be followed by at least a short passage of text. Please note that the first line of text that follows a Section heading is not indented, whereas the first lines of all subsequent paragraphs are.

## **Methodology**

### **1. Study subjects**

The present study was conducted as a nested case–control study in a cohort of type 2 diabetes patients. The recruitment of cases and controls is made from diabetic patients in the EpiDiaM cohort (Epidemiology Diabetes Morocco).

### **2. Case/Control study:**

We recruited 366 cases and 366 controls.

#### The case:

All diabetic patients in our cohort with one or more complications of diabetes (Macrovascular, nephropathy, neuropathy and retinopathy) have been our target population of cases. We excluded patients who had complications before diagnosis of diabetes and patients with inability to determine the dates for the diagnosis of complication.

#### The control:

All diabetic patients in our cohort with no complications represented our target population controls. Were defined as controls for this study; diabetic with no

complications of diabetes at the time of inclusion in the cohort. Controls were matched to cases on age ( $\pm 5$  years), sex and diabetes duration ( $\pm 5$  years). We excluded patients with complications indeterminate situation (Figure 1).

**Figure 1:** Flow chart of eligibility for the study and inclusion in the analysis



### 3. Analytic Plan

The analysis was adjusted for all potential confounders. Sex is known as a modifier factor for association between risk factors and cardiovascular complications [1 - 2]. In addition, some factors are specific to women (such as contraception) and other men especially in the Moroccan context (tobacco, alcohol ...). For these reasons the analysis was performed separately for women and men in order to determine the specificity of each sex factor. P = 0.05 was the level of statistical significance. All analyses were performed using SPSS version 20.

## Results

To meet our objective, in the study of factors related to degenerative complications of type 2 diabetes, we showed that among 732 patients (366 cases with complication (s)

and 366 controls) eligible. We first identified a total of 1,196 diabetic patients from the EPIDIAM cohort study. The majority of the population was female (77.7%). Mean age was  $57.5 \pm 10.4$  years and mean diabetes duration was  $8 \pm 6.60$  years. 41.2% had retinopathy and 28.4 % had diabetic neuropathy, while 26.2% had cardiovascular complications. In the multivariate analysis, a higher risk of degenerative complications was observed among women with low economic status (OR, adj = 11.36, 95% CI 5.59 to 23.25), women having forgotten their treatment (Adj. OR = 3.42, 95% CI 1.29-9.09), urban women (adj. OR = 3.97, 95% CI 0.04 - 0.17), Women with a very high level of stress (adj. OR 2.94, 95% CI 1.00 to 8.63), and overweight women (adj. OR = 2.50, 95% CI 1.12 to 5.53 ). However, for men, in the unadjusted analysis, low socio-economic and occupational inactivity were increased the risk of degenerative complication. Overweight men [5.96 (95% CI: 1.61-22.10)], with dyslipidemia [3.09 (95% CI: 1.51-6.33)] and patients treated with a general practitioner [4.57 (95% CI: 1.24 to 16.82)] had a higher risk of degenerative complications. Our results confirm that the risk factors for degenerative complication of type 2 diabetes are strongly related to the Moroccan context.

## Discussion

We conducted a case-control study nested in a cohort to determine the factors associated with degenerative complications of type 2 diabetes in Morocco.

For women, we observed that the low socioeconomic level was associated with a significant increasing on the risk of degenerative complications among women. Although its mechanism has not been completely clarified, the low socioeconomic level can be a responsible of development of complications of diabetes with type 2 through different and complex processes [3]. In Moroccan context, poor women, like those in other studies among low-income women [4-5], often put their families' needs and preferences before their own.

However, higher levels of psychosocial stress may affect a person's socioeconomic status, use of medical services and overall health [6-7]. Surwit et al. showed that stress management training for one year was associated with a reduction significance of HbA1c. However, very anxious patients didn't obtain a reduction in HbA1c level [8]. Additionally, obesity is associated with increased risks for complications [9-10]. This connection is maintained after adjustment with other risk factors, with risk multiplied by 2.5 in women. Obesity is an important modifiable risk factor for type 2 diabetes [11], cardiovascular disease [12-13] and renal failure [14]. In some regions of Morocco, especially the South, the weight of the woman is even seen as a competitive advantage increasing her chances of finding a husband [15]. However, our analysis did not indicate a significant association between complication risk and physical activity.

While the observed association between the area and complication risks was demonstrated in previous epidemiologic studies [16-17], the risk of degenerative complications is multiplied by 3.9 in women who reside in urban areas. This may be

due to the sedentary lifestyle as well as a reduced access to a healthy food in urban areas.

For men, we have not been able to do a multivariate analysis. Due to lack of statistical power (low number of diabetic patients) conditions for statistical modeling was not satisfied.

To our knowledge, in Morocco our case-control study is the first to directly analyze the relationship between degenerative complication and all risk factors that may lead to the occurrence of these complications among Moroccan diabetics. Therefore, our study added an important knowledge of risk factors responsible for the occurrence of degenerative complications resulting in hospitalization or death. This study highlighted important areas for health care intervention and provided a reminder for vigilance when known risk factors for complications are present.

### ***Author Disclosures***

The authors have no financial interests to disclose directly or indirectly related to the research in the manuscript.

### **References**

1. Collier A, Ghosh S, Hair M, Waugh N. Impact of socioeconomic status and gender on glycaemic control, cardiovascular risk factors and diabetes complications in type 1 and 2 diabetes: A population based analysis from a Scottish region. *Diabetes & Metabolism* 2015; 41:145–151.
2. Raphael D, Anstice S, Raine K, *et al*. The social determinants of the incidence and management of type 2 diabetes mellitus: are we prepared to rethink our questions and redirect our research activities? *Leadership Health Serv* 2003; 16:10–20.
3. Huxley R, Barzi F, Woodward M. Excess risk of fatal coronary heart disease associated with diabetes in men and women: Meta-analysis of 37 prospective cohort studies. *BMJ* 2006; 332 (7533):73–78.
4. Yu VL, Raphael D. Identifying and addressing the social determinants of the incidence and successful management of type 2 diabetes mellitus in Canada. *Can J Public Health* 2004; 95:366–368.
5. Hepworth J. Gender and the capacity of women with NIDDM to implement medical advice. *Scandinavian J Public Health* 1999; 27:260–266.
6. Lin EH, Katon W, Von Korff M, Rutter C, Simon GE, Oliver M, *et al*. Relationship of depression and diabetes self-care, medication adherence, and preventive care. *Diabetes Care* 2004; 27(9):2154–2160.
7. Spangler JG, Summerso JH, Bell RA, Konen JC. Smoking status and psychosocial variables in type 1 diabetes mellitus. *Addict Behav* 2001; 26 (1):21–29.
8. Surwit RS, van Tilburg MA, Zucker N, *et al*. Stress management improves long-term glycemic control in type 2 diabetes. *Diabetes Care* 2002; 25(1):30-34.
9. Albu J, Konnarides C, Pi-Sunyer FX. The weight control: Metabolic and Cardiovascular Effects. *Diabetes Journal* 1995; 3:335-347.
10. Nguyen NT, Nguyen XM, Lane J, Wang P. Relationship Between Obesity and diabetes in adult population A United States: Results from the National Survey Nutrition Health Review, 1999-2006. *Obesity Surgery* 2011; 21: 351-355.
11. Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. *Nature* 2001; 414:782–787.

12. Zalesin KC, Franklin BA, Miller WM, Peterson ED, McCullough PA. Impact of obesity on cardiovascular disease. *Med Clin North Am* 2011; 95:919–937.
13. Tirosh A, Shai I, Afek A, Dubnov-Raz G, Ayalon N, Gordon B, *et al*. Adolescent BMI trajectory and risk of diabetes versus coronary disease. *N Engl J Med* 2011; 364:1315–1325.
14. Hsu CY, McCulloch CE, Iribarren C, Darbinian J, Go AS. Body mass index and risk of end stage renal disease. *Ann Intern Med* 2006; 144:21–28.
15. Rguibi M, Belahsen R. Body size preferences and sociocultural influences on attitudes towards obesity among Moroccan Sahraoui women. *Public Health Nutrition* 2007; 9(6):722–727.
16. Dahiru T, Ejembi CL. Clustering of cardiovascular disease risk-factors in semi-urban population in Northern Nigeria. *Niger J Clin Pract* 2013; 16:511–516.
17. Ramachandran A, Mary S, Yamuna A, Murugesan N, Snehalatha C. High prevalence of diabetes and cardiovascular risk factors associated with urbanization in India. *Diabet Care* 2008; 31 (5): 893–898.

# Bootstrap group penalty for high-dimensional regression models

## *Una procedura di penalizzazione bootstrap a gruppi per modelli di regressione ad alta dimensionalità*

Valentina Marneli, Debora Slanzi and Irene Poli

**Abstract** The paper presents a new penalization procedure for variable selection in regression models. We propose the Bootstrap Group Penalty (BGP) that extends the bootstrap version of the LASSO method by taking into account the grouping structure which may be present or introduced in a model. Based on a simulation study we demonstrate that the new procedure outperforms some existing group penalization methods in terms of both prediction accuracy and variable selection quality.

**Abstract** Il presente lavoro propone una nuova procedura per la selezione delle variabili in modelli di regressione penalizzati, chiamata penalizzazione bootstrap a gruppi (BGP), che estende la versione bootstrap del metodo LASSO tenendo conto della struttura di raggruppamento fra i predittori che può essere presente o può essere introdotta in un modello. Uno studio di simulazione rivela che la nuova procedura fornisce risultati migliori rispetto ad alcuni metodi di penalizzazione a gruppo esistenti. La bontà di questi risultati è misurata sia in termini di accuratezza di previsione sia in termini di qualità nella selezione delle variabili.

**Key words:** Bi-level selection, high-dimensionality, regression models

## 1 Introduction

One of the most challenging problems arising in many scientific contexts is modelling data characterised by a huge number of variables interacting with each other in some complex and unknown pattern. Often, the sample size considered in the analysis is small compared to the number of variables. The development of new

---

Valentina Marneli, Debora Slanzi, Irene Poli  
Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, via Torino 155, Mestre (IT)  
European Centre for Living Technology, S. Marco 2940, Venice (IT),  
e-mail: valentina.marneli@unive.it, debora.slanzi@unive.it, irenepoli@unive.it

statistical tools tailored to analyse these data is therefore crucial in contemporary statistical learning. Many proposals are available in literature with the main aim of reducing the dimensionality and selecting the most relevant variables for the problem under study [Fan and Lv, 2010]. An important line of research is concerned with penalized regression procedures; see for example [Fan and Lv, 2010, Breheny and Huang, 2009, Tibshirani, 1996, Zhang, 2010]. The sparsity condition is widely considered under this scenario assuming that only a subset of predictors is associated with the response variable. The penalized procedures with the sparsity assumption are designed with the aim of both selecting the most relevant predictors and estimating the parameters of the model. The penalties in regression models can be subdivided into three wide classes which relate to individual variable selection, group variable selection and bi-level variable selection. In this paper we propose a new penalization procedure which we call the Bootstrap Group Penalty (BGP) which is obtained by coupling the properties of group variable and bi-level selection methods with the bootstrap re-sampling methods. The approach extends the work of Bach (2008) where a bootstrap version of LASSO was proposed. The paper proceeds as follows. In Section 2 we review penalized procedures and we introduce the Bootstrap Group Penalty procedure. The Section 3 investigates the performance of our method through a simulation study. In Section 4 we present some concluding remarks on the results of the simulation study.

## 2 The Bootstrap Group Penalty procedure

### 2.1 Model set-up

We consider the multiple linear regression model

$$y_i = X_i \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

where  $X_i = (x_{i1}, \dots, x_{ip})^T$  is a  $p$ -dimensional vector of predictors,  $y_i$  is the response variable,  $\varepsilon_i$  is the error term and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$  is the regression vector of  $p$  unknown parameters which must be estimated from the data. It is known that when the number of covariates (dimension of the system) considerably exceeds the number of observations ( $p \gg n$ ) the model is not identifiable from a statistical perspective and therefore not estimable with standard classical statistical procedures. Penalized regression procedures, also known as regularized regression methods, are important methods frequently used in high dimensional problems as they are able to estimate reliable models and to improve on prediction capabilities, also when the number of predictors is much larger than the number of observations. In order to estimate the vector of regression coefficients  $\boldsymbol{\beta}$  we minimize the so-called penalized least squares function, defined as



$$Q(\boldsymbol{\beta}) = \frac{1}{2n}(y - X\boldsymbol{\beta})^T(y - X\boldsymbol{\beta}) + P(\boldsymbol{\beta}|\lambda),$$

where  $y = (y_1, \dots, y_n)$  and  $X = (X_1, \dots, X_n)^T$  is the  $n \times p$  design matrix. The function  $P(\cdot)$  is a penalty on the regression coefficient parameters  $\boldsymbol{\beta}$  which controls the complexity of the model. The parameter  $\lambda$  is a tuning parameter which can be selected using cross validation or information criteria like the Akaike and the Bayesian information criteria [Akaike, 1974, Schwarz, 1978]. Various penalties have been proposed in the literature, see for example [Fan and Lv, 2010] and [Huang *et al.*, 2012]. These penalties can be subdivided into three big groups which relate to individual variable selection, group variable selection and bi-level variable selection. There is a large literature on penalized regression procedures for individual variable selection; the least absolute shrinkage selection operator (LASSO) proposed by Tibshirani (1996) is surely the most used and famous procedure. Recently, there has been a large number of works extending these approaches to grouped predictors; indeed it is possible to take account of a grouping structure among the predictors in order to improve model prediction capabilities ([Ogutu and Piepho, 2014]). When a grouping structure is introduced into a model, interest may rely entirely on selecting relevant groups and not individual variables, but when both individual variables and groups are important we will consider bi-level selection procedures to select both the important groups and variables within these groups. Under this setup, among others, we can cite the following variable selection procedures, the group LASSO ([Yuan and Lin, 2006]), the smoothly clipped absolute deviation penalty ([Fan and Li, 2001]), the minimax concave penalty method ([Zhang, 2010]), the composite MCP ([Breheny and Huang, 2009]), the group bridge penalty ([Fu, 1998]) and the group exponential LASSO ([Breheny, 2015]). These selection procedures have been introduced to overcome some limitations of the LASSO estimator and present a number of appealing properties in terms of both estimation accuracy as well as variable selection properties. In order to improve the performances of these procedures in high-dimensional settings, when the number of observations is very small, it may be desirable to use a bootstrap sampling technique which is able to perturb an initial dataset to gain information from the multiple datasets resulting from the bootstrap procedure. This approach was suggested for the LASSO estimator by Bach (2008).

## 2.2 The Bootstrap Group Penalty (BGP)

We introduce a new penalized procedure which is obtained by coupling the properties of group variable and bi-level selection methods with the bootstrap re-sampling methods.

If the training data consists of  $n$  observations  $(X_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ ,  $i = 1, \dots, n$ , we consider  $B$  bootstrap replications of the  $n$  pairs  $(X_i, y_i)$ . More specifically, for  $b = 1, \dots, B$ , we consider  $(X_{bi}, y_{bi}) \in \mathbb{R}^p \times \mathbb{R}$  sampled uniformly at random with

replacement from the original training pairs  $(X_i, y_i)$ ,  $i = 1, \dots, n$ . Then at each bootstrap iteration we estimate the regression parameters  $\beta_j$ , for  $j = 1, \dots, p$ , by using a penalized group or bi-level group selection procedure. At each bootstrap step we select the subset of covariates with non-zero coefficients selected by the penalized procedure specified by the set of covariates indices:

$$J_b = \{j | \hat{\beta}_j^b \neq 0, j = 1, \dots, p\}, \quad b = 1, \dots, B.$$

Among all the  $B$  sets  $J_b$ , we select only the predictors which have high frequency out of the  $B$  bootstrap replications. The proportion of predictors to be taken into account depends strongly on the penalization method. The set of the variables selected will be then used to estimate the model using a penalized group procedure. The procedure is synthesized in algorithm 1.

---

#### Algorithm 1 BGP procedure

---

- 1: **Input**  $n$  sample size,  $B$  number of bootstrap replicates,  $(X_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ ,  $\tau$  threshold value to select relevant covariates in the bootstrap replications.
  - 2: **for**  $b = 1$  to  $B$  **do**
  - 3:     **repeat**
  - 4:         Generate  $B$  bootstrap samples  $(X_{bi}, y_{bi}) \in \mathbb{R}^p \times \mathbb{R}$
  - 5:         Compute the estimates of the regression parameters by using a penalized procedure  $\hat{\beta}_j^b$  for  $j = 1, \dots, p$ .
  - 6:         Compute  $J_b = \{j | \hat{\beta}_j^b \neq 0, j = 1, \dots, p\}$ .
  - 7:     **until**
  - 8:     **end for**
  - 9: Compute  $J = \{j | \text{which are present in at least } \tau\% \text{ of the } J_b\}$ .
  - 10: Compute  $\hat{\beta}$  by using a penalized procedure on the restricted set  $J$  of covariates.
- 

### 3 A simulation study

In this section, we perform a simulation study to evaluate some group penalization methods and their corresponding bootstrap counterparts. Among the several penalties that can be used here we focus on the group bridge (gBridge), the composite MCP (cMCP), the maximum concave penalty (MCP), the group exponential LASSO (gel) and the group LASSO (gLASSO), and on their bootstrap counterparts, which will be referred as BgBridge, BcMCP, BMCP, Bgel and BgLASSO, respectively. The simulation is based on the linear regression model  $y_i = X_i \beta + \varepsilon_i$ ,  $i = 1, \dots, n$ , where  $\varepsilon_i \sim N(0, \sigma^2)$  as introduced in Equation 1. The standard deviation  $\sigma$  is assumed to be 3 and covariates were generated from the normal distribution as in the simulation study proposed by Breheny (2015). We consider the following setup: 10 groups with 20 variables in each group, then the total number of predictors is  $p = 200$ . The number of non zero coefficients is 4, and all the non-zero coefficients belongs to the same group. We randomly split the data into training and testing datasets where the training set is assumed approximately 50% of the full data

set. The data set size considered in this simulation is  $n = 100$ . Regarding the bootstrap samples, we used  $B = 500$ . To evaluate the performance of the various group penalization methods with respect to their bootstrap counterpart we calculate some measures of prediction accuracy and variable selection efficiency. In particular we repeat the simulation 1000 times and we compute the predictive mean square error, the sensitivity (the ratio between the number of selected important variables and the number of important variables), and the specificity (the ratio between the number of removed unimportant variables and the number of unimportant variables) as defined in Geng *et al.* (2015). The results of this simulation are presented in Table 1.

**Table 1** Simulation results over 1000 replicates: Predictive Mean Square Error (PMSE), Sensitivity, Specificity. The statistical significant differences between the original approach and the bootstrap counterpart are reported in bold.

Method	PMSE	Sensitivity	Specificity
gBridge	0.885 (0.176)	0.934 (0.122)	0.843 (0.025)
BgBridge	<b>0.835</b> (0.293)	<b>1.000</b> (0.000)	<b>0.907</b> (0.031)
cMCP	0.936 (0.207)	0.797 (0.190)	0.849 (0.014)
BcMCP	<b>0.839</b> (0.150)	<b>0.910</b> (0.136)	<b>0.913</b> (0.014)
MCP	1.281 (0.257)	1.000 (0.000)	0.489 (0.077)
BMCP	<b>1.038</b> (0.911)	1.000 (0.000)	<b>0.899</b> (0.040)
gel	0.949 (0.185)	1.000 (0.000)	0.462 (0.083)
Bgel	<b>0.755</b> (0.608)	0.9998(0.008)	<b>0.937</b> (0.044)
gLASSO	<b>0.924</b> (0.221)	0.554 (0.497)	<b>0.916</b> (0.158)
BgLASSO	1.293 (0.544)	<b>1.000</b> (0.000)	0.652 (0.104)

From this Table we can notice that the Bootstrap Group Penalty procedure is able to improve the PMSE and the Specificity in almost all the compared approaches as highlighted in bold in Table 1 (see for example, the very good performance in prediction of Bgel with respect to all the other approaches using different penalization methods, confirmed also for Specificity index). Moreover, when considering Sensitivity, we can notice that all the BGP procedures are able to increase or at least to confirm very good performances achieving values of Sensitivity very close to 1.

## 4 Concluding remarks

Our results suggest that combining penalized group and bi-level variable selection methods with re-sampling methods is a promising approach to handle high-dimensional problems. The method could be easily adapted to handle other penalization procedures. Further works will be devoted to investigate the method in a real case study and in other simulations set-up.

## References

- [Akaike, 1974] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In: B.N. Petrov and F. Csaki (eds.) 2nd International Symposium on Information Theory: 267–81 Budapest: Akademiai Kiado.
- [Bach, 2008] Bach, F.R. (2008). Bolasso: Model Consistent Lasso Estimation through the Bootstrap. Proceedings of the 25-th International Conference on Machine Learning, Helsinki, Finland, 2008.
- [Breheny and Huang, 2009] Breheny, P., Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, **2(3)**, 369–380.
- [Breheny, 2015] Breheny, P. (2015). The Group Exponential Lasso for Bi-Level Variable Selection. *Biometrics*, **71**, 731–740.
- [Fan and Li, 2001] Fan, J. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.
- [Fan and Lv, 2010] Fan, J., Lv, J. (2010) A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica*, **20**, 101–148.
- [Fu, 1998] Fu, W. J. (1998). Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, **7(3)**, 397–416.
- [Geng et al., 2015] Geng, Z., Wang, S., Yu, M., Monahan, P.O., Champion, V., Wahba, G. (2015). Group variable selection via convex log-exp-sum penalty with application to a breast cancer survivor study. *Biometrics*, **71(1)**, 53–62.
- [Huang et al., 2009] Huang, J., Ma, S., Xie, H., and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika*, **9**, 339–355.
- [Huang et al., 2012] Huang, J., Breheny, P., Ma, S. (2012) A Selective Review of Group Selection in High-Dimensional Models. *Statistical Sciences*, **27(4)**, 481–499.
- [Lee et al., 2012] Lee, Y.K., Lee, E.R., Park, B.U. (2012). Principal component analysis in very high-dimensional spaces. *Statistica Sinica* **22**, 933–956.
- [Liu et al., 2015] Liu, J., Wang, F., Gao, X., Zhang, H., Wan, X., Yang, C. (2015) A Penalized Regression Approach for Integrative Analysis in Genome-Wide Association Studies. *Journal of Biometrics and Biostatistics*, **6(228)**, 7 pages.
- [Ogutu and Piepho, 2014] Ogutu, J.O., Piepho, H.-P. (2014) Regularized group regression methods for genomic prediction: Bridge, MCP, SCAD, group bridge, group lasso, sparse group lasso, group MCP and group SCAD. *BMC Proceedings*, **8(Suppl. 5)**, S7.
- [Park et al., 2007] Park, M. Y., Hastie, T., Tibshirani R. Averaged gene expressions for regression. *Biostatistics*, 212–227.
- [Tibshirani, 1996] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58(1)**, 267–288.
- [Sharma et al., 2014] Sharma, D.B., Bondell H.D., Zhang, H.H. (2013). Consistent Group Identification and Variable Selection in Regression with Correlated Predictors. *Journal of Computational and Graphical Statistics*, **22(2)**, 319–340.
- [Schwarz, 1978] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464.
- [Yuan and Lin, 2006] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, **68**, 49–67.
- [Zhang, 2010] Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, **38(2)**, 894–942.

# Improving small area estimates of households' share of food consumption expenditure in Italy by means of Twitter data

## *Migliorare la precisione delle stime per piccola area della quota di spesa dei generi alimentari tramite i dati del social network Twitter*

Marchetti S., Pratesi M., Giusti C.

**Abstract** In this work we use emotional data coming from Twitter as auxiliary variable in a small area model to estimate Italian households' share of food consumption expenditure (the proportion of food consumption expenditure on the total consumption expenditure) at the provincial level. We show that the use of Twitter data has a potential in predicting our target variable, reducing the estimated mean squared error with respect to what obtained by the same working model without the Twitter data.

**Abstract** *In questo lavoro si mostra come l'uso di dati ricavati da Twitter possa migliorare in termini di efficienza le stime per piccole aree della quota di spesa per generi alimentari a livello provinciale in Italia.*

**Key words:** Big Data, Area level model, Emotional data

## 1 Introduction

Recently, an increasing number of researchers have investigated the value of using big data (huge amounts of digital information about human activities) in socio-economic studies, see for example Eagle et al (2010); Blumenstock et al (2015); Decuyper et al (2014). Marchetti et al (2015) suggested three approaches to use big data in synergy with small area estimation methods. Another approach to use big data in small area estimation was suggested by Porter et al (2014).

In this paper we focus on the use of data coming from the social network Twitter to investigate their potential in predicting the share of food consumption expenditure of Italian households at the province level. The paper has the following structure: the description of the data used in the analysis is in section 2; the small area estimation

---

Stefano Marchetti, Monica Pratesi, Caterina Giusti  
University of Pisa, Via Ridolfi, 10, 56124 Pisa (PI), e-mail: stefano.marchetti@unipi.it

model is presented in section 3; the results of the application are detailed in section 4. Finally, we draw some concluding remarks in section 5.

## 2 Data used in the application

The primary source of data on households' expenditure in Italy is the Household Budget Survey (HBS) carried out annually by ISTAT. In 2012 the sample of the HBS was composed by approximately 28000 households. Data were collected on the basis of a two-stage sample design where the first stage were the municipalities and the second stage were the households. The regions (NUTS 2 level according to Eurostat) are the finest geographical level for which direct estimates of the target indicators are reliable. However, the knowledge of measures able to assess households' living conditions and well-being at a more detailed geographical level is often crucial, since this knowledge can for example enable policy makers in planning local policies aiming at reducing poverty and social exclusion (Giusti et al, 2016).

The households' consumption expenditure can be classified into food (and beverages) and non food expenditure. The share of total expenditure that an household dedicate to food items is an important indicator of the household living conditions: at risk of poverty households usually spend an higher share of their total expenditure on food with respect to the other households, with a lower impact of the share of expenditure dedicated to other resources and commodities.

To estimate the target at the province level we resort to model-based area-level small area methods, since direct estimates are unreliable. As possible sources of auxiliary variables – needed in model-based estimation – we use data coming from the Population and Housing Census 2011 and from the Survey<sup>1</sup> on Social Actions and Services on Single and Associates Municipalities 2012.

From the Population Census we collected information at provincial level such as the number of households, the average households' size, the tenure status, the female-headed households quota. As the target variable of our analysis can be considered as a proxy of the households' living conditions, we also considered as valuable source of auxiliary information the expenditure that Italian municipalities made in 2012 for interventions of social protection. These interventions includes the costs information on local welfare policies, such as services, benefits and transfers directed to households with children, old-age persons, poor and social excluded persons, immigrants.

Besides these sources of official statistics, we also considered as a potential source of auxiliary information big data from Twitter. In particular, we considered here as potential covariate for our small area working models the iHappy indicator referring to the year 2012. The iHappy indicator is made available every year since 2012 for all the 110 Italian provinces on the Opinion Analytics platform *Voices from the Blogs*. The iHappy indicator referring to the year 2012 was computed by

<sup>1</sup> This survey is a census survey, although some nonresponses can occur. Here we ignore the non-responses and we use these data as census data.

collecting and coding more than 43 millions of tweets posted on a daily basis in all the Italian provinces. The words and emoticons of the tweets were classified using a training set in two categories: “happy” and “unhappy”, together with a residual class “other”. Then, Curini et al (2015) derived the frequency distribution of the happy and unhappy tweets in the entire population. The iHappy indicator was then computed for each Italian province as the percentage ratio of the number of happy tweets to the sum of happy and unhappy tweets. The overall average of the iHappy indicator in 2012 was equal to 44.5%, with a minimum value of 35.1% for Oristano and a maximum value of 56.6% for Sassari, both provinces of the Sardinia region. Indeed, the spatial variability of the iHappy values was rather high, as it is evident from the “emotional map” of Figure 1 (right).

### 3 Short review of the Fay-Herriot model for small area estimation

Data obtained from surveys are often used to estimate characteristics for subsets of the survey population. If the sample from a subset is small, then a traditional design-based survey estimator can have unacceptably large variance. These subsets has been defined as *small areas* (Rao and Molina, 2015).

In this study the available data allow us to rely only on area-level models (relate small area direct estimates to area-specific auxiliary variables). In addition, we do not have time-series data and the spatial correlation of the target direct estimates is low. So our choice falls on the Fay and Herriot (1979) estimator (FH). In what follows a summary description of the method is given.

Let  $m$  be the number of small areas and  $\theta_i$  be the target parameter of the area  $i$  (mean or proportion). A survey provides a direct estimator  $\hat{\theta}_i^{dir}$  of  $\theta_i$ ,  $E[\hat{\theta}_i^{dir}] = \theta_i$  under the sampling design. A  $p$ -vector  $\mathbf{X}_i$  contains the auxiliary data sources – exactly known – of population characteristics for area  $i$ . The FH model is as follows:

$$\hat{\theta}_i^{dir} = \mathbf{X}_i^T \beta + u_i + e_i \quad i = 1, \dots, m, \quad (1)$$

where  $u_i \stackrel{iid}{\sim} N(0, \sigma_u^2)$ ,  $i = 1, \dots, m$  are the model errors and  $e_i \stackrel{ind}{\sim} N(0, \psi_i^2)$ ,  $i = 1, \dots, m$  are the design errors, with  $e_i$  independent from  $u_j$  for all  $i$  and  $j$ . It is assumed that the quantity of interest in area  $i$  is  $\theta_i = \mathbf{X}_i^T \beta + u_i$ .

Under the assumption of normality of both the errors (model and sampling design), the best linear unbiased predictor (BLUP) of  $\theta_i$  is  $\tilde{\theta}_i^{FH} = \gamma_i \hat{\theta}_i^{dir} + (1 - \gamma_i) \mathbf{X}_i^T \tilde{\beta}$ ,  $\gamma_i = \sigma_u^2 / (\sigma_u^2 + \psi_i^2)$ , where  $\tilde{\beta}$  is the Best Linear Unbiased Estimator of  $\beta$ . According to the theory of small area estimation (Rao and Molina, 2015), the parameters  $\beta$  and  $\sigma_u^2$  are unknown and must be estimated, while  $\psi_i^2$  is assumed to be known.

Estimators of  $\beta$  and  $\sigma_u^2$  can be obtained using the restricted maximum likelihood from the marginal distribution  $\hat{\theta}_i^{dir} \sim N(\mathbf{X}_i^T \beta, \sigma_u^2 + \psi_i^2)$ . By plugging in the estimates of  $\beta$  and  $\sigma_u^2$  into the BLUP we obtain the empirical best linear unbiased predictor

$$\hat{\theta}_i^{FH} = \hat{\gamma}_i \hat{\theta}_i^{dir} + (1 - \hat{\gamma}_i) \mathbf{X}_i^T \hat{\beta}, \quad \hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i^2}. \quad (2)$$

The estimator (2) has the following  $MSE(\hat{\theta}_i^{FH}) = \gamma_i \psi_i^2 + (1 - \gamma_i)^2 \mathbf{X}_i^T V(\hat{\beta}) \mathbf{X}_i + \psi_i^4 (\psi_i^2 + \sigma_u^2)^{-3} V(\hat{\sigma}_u^2) = g_{1i} + g_{2i} + g_{3i}$ , where  $g_{2i}$  is the contribution to the MSE from estimating  $\beta$  and  $g_{3i}$  is the contribution to the MSE from estimating  $\sigma_u^2$ ;  $V(\hat{\beta})$  and  $V(\hat{\sigma}_u^2)$  are the asymptotic variances of an estimator  $\hat{\beta}$  of  $\beta$  and an estimator  $\hat{\sigma}_u^2$  of  $\sigma_u^2$ , respectively. An estimator of the MSE is as follows

$$mse(\hat{\theta}_i^{FH}) = \hat{g}_{1i} + \hat{g}_{2i} + 2\hat{g}_{3i}, \quad (3)$$

where  $\hat{g}_{1i} = \hat{\gamma}_i \psi_i^2$ ,  $\hat{g}_{2i} = (1 - \hat{\gamma}_i)^2 \mathbf{X}_i^T [\sum_{i=1}^m \mathbf{X}_i \mathbf{X}_i^T / (\psi_i^2 + \hat{\sigma}_u^2)]^{-1} \mathbf{X}_i$ ,  $\hat{g}_{3i} = \psi_i^4 (\psi_i^2 + \hat{\sigma}_u^2)^{-3} 2 [\sum_{i=1}^m 1 / (\hat{\sigma}_i^2 + \psi_i^2)]^{-1}$ .

#### 4 Area-level small area model *with* and *without* Twitter data to estimate the share of food consumption expenditure in the Italian provinces

In this section we show that the use of Twitter data can improve the precision of the Share of Food Consumption Expenditure (SFCE) estimates in the Italian provinces, obtained using small area methods.

First, we estimated the SFCE at provincial level using the FH model (1) selecting the more predictive variables among the data described in section 2 without considering the iHappy variable, the one computed using Twitter 2012 data. In this way we obtained a reduction in MSE in all the provinces. Second, we added the iHappy variable to the other auxiliary variables and we estimated the SFCE again. If the iHappy variable is linearly correlated with the SFCE and this relation is not yet explained by the other auxiliary variables, then we expect a better performance in terms of MSE when using iHappy. We will show that the results obtained support this expectation.

The target variable, the SFCE, was obtained from the HBS 2012 survey as the ratio between the consumption expenditure for food (including beverages) and the total consumption expenditure. Its direct estimate at provincial level was obtained using the Horvitz and Thompson (1952) expansion estimator,  $\hat{\theta}_i^{dir}$ .

In 2012 the Italian provinces were 110 in total. However, in 2012 no HBS sample data were available for the province of Enna (Sicily) therefore it was not possible to obtain a direct estimate for this province, so we computed a synthetic estimator given that we know the auxiliary data for this province.

The selected auxiliary variables for the model without the iHappy variable are: the share of owners of the house  $x_1$ , the share of households lead by a female  $x_2$ , the per-household local government expenses to support several categories of citizens, households with children ( $x_3$ ), old-aged persons ( $x_4$ ), immigrants ( $x_5$ ), at risk of poverty persons ( $x_6$ ), services to families ( $x_7$ ). So let  $\mathbf{X}_i = [1, x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}]^T$  be the design  $p$ -vector for model (1) for the area  $i$ ,



where  $x_{ki}$ ,  $k = 0, \dots, p = 7$ ,  $i = 1, \dots, m$ , is the value of the  $k$ th auxiliary variable in area  $i$  (with  $x_{0i} = 1$ ).

The FH model without the iHappy variable is then  $\hat{\theta}_i^{dir} = \mathbf{X}_i^T \beta + u_i + e_i$ . Estimates of  $\beta$  and  $\sigma_u^2$  were obtained under the Normality assumptions made in section 3 using the restricted maximum likelihood (REML). From the analysis of  $\hat{u}_i = \hat{\gamma}(\hat{\theta}_i^{Dir} - \mathbf{X}_i^T \hat{\beta})$ , the Normality assumption seems reasonable. Indeed, the Shapiro and Wilk (1965) Normality test is equal to 0.978 with a  $p$ -value of 0.063.

To check the hypothesis that big data can help to increase the precision of the small area estimates - if used as auxiliary variables - we added the iHappy variable ( $x_8$ ), obtained from the analysis of Twitter data as explained in section 2, to the set of the selected auxiliary variables ( $x_1, x_2, \dots, x_7$ ). Let  $\mathbf{Z}_i = [\mathbf{X}_i, x_{8i}]^T$ , where  $x_{8i}$  is the iHappy value for area  $i$ . The FH model is  $\hat{\theta}_i^{dir} = \mathbf{Z}_i^T \beta^{BD} + u_i^{BD} + e_i^{BD}$ , where the superscript  $^{BD}$  refers to parameters under the model that makes use of big data (the Twitter data). Point and  $mse$  estimates are then obtained according to the methodology described in section 3 (replacing  $\mathbf{X}_i$  by  $\mathbf{Z}_i$ ).

In both the models - with and without iHappy variable - we selected the auxiliary variables using a step-wise procedure based on AIC (Hastie and Pregibon, 1992). The selected variables show a negative linear correlation with the target that range from  $-0.130$  to  $-0.509$ . The negative correlations were expected for all the variable, but the share of households lead by a female. In general, in Italy, households lead by a female are positively correlated with poverty indexes and deprivation variables. However, we can suppose that the households lead by a female are associated with a reduction of the household size, so the expenses in food and beverages decreases so that to increase the SFCE. This hypothesis is supported by a linear correlation between the share of the households lead by a female and the household size equal to  $-0.857$ . As done for the model without iHappy variable, we estimated  $\beta^{BD}$  and  $\sigma_u^{BD}$  under the Normality assumptions made in section 3 using the REML. The Shapiro and Wilk (1965) Normality test for  $\hat{u}_i^{BD}$ s is equal to 0.980 with a  $p$ -value of 0.107.

The regression parameters estimated for both the models - with and without iHappy - are showed in table 1. The  $\beta$ s obtained under the two models are similar, the introduction of the iHappy variable in the FH model does not change significantly the model, it just add predictive power to it. The parameter  $\sigma_u$  is estimated

Table 1: Regression parameters of the FH model with/without the iHappy variable.

	$\hat{\beta}^{BD}$	p-value <sup>BD</sup>	$\hat{\beta}$	p-value
Intercept	0.7165	0.0000	0.6446	0.0000
iHappy2012	-0.0019	0.0067	-	-
Share of owners of the house	-0.0038	0.0000	-0.0039	0.0000
Share of household lead by female	-0.3164	0.0009	-0.3222	0.0012
Expenses for household with children	-0.0001	0.2121	-0.0002	0.0513
Expenses for old-aged persons	-0.0001	0.0123	-0.0001	0.0280
Expenses for immigrants	-0.0013	0.0003	-0.0013	0.0009
Expenses for at risk of poverty persons	0.0006	0.0009	0.0007	0.0006
Expenses for services to families	-0.0005	0.0460	-0.0006	0.0246

equal to 0.020 for the model without iHappy and to 0.019 for the model with iHappy. To verify the null hypothesis that  $\sigma_u^2 = 0$ , we used the test proposed by Datta et al (2011) and we reject the null hypothesis  $\sigma_u^2 = 0$  for both the models.

It is important to highlight that the iHappy indicator is based on self-selected data, the Twitter data. However, in this application we are not able to treat the self-selection bias due to lack of information. Thus, we assume that the self-selection is negligible. Moreover, the iHappy indicator can be affected by measurement error, since not any happy tweet corresponds to a happy person. In our application the MSE of the iHappy is very small, due to the very large sample size (43 millions of tweets), therefore model that account the measurement error, such as the one proposed by Ybarra and Lohr (2008), approximately corresponds to the traditional FH model.

Results on the SFCE estimates are summarized in table 2. Using the FH estimator (2) with the set of auxiliary variables  $\mathbf{X}_i$ s the *rmse* is reduced in all the provinces. The average reduction is about 30% with a 25% of provinces where the reduction is at least about 40%(table 2). Moreover, using also the iHappy variable the reduction of the *rmse* results in an average gain of 2%. A clearer picture of the gain in precision due to the introduction of the iHappy variable in the FH model can be see in the last line of table 2, which shows the efficiency of  $\hat{\theta}_i^{FH,BD}$  against  $\hat{\theta}_i^{FH}$ . There is a gain in all the areas, but one where we observe a loss of 0.5%. The gain goes from about 2% up to about 7%. Given that the small area estimates obtained without the use of the iHappy variable show a remarkable gain in terms of reduction of *mse*, the further reduction of the *mse* due to the introduction of the iHappy variable in the model is a very good result. This is particularly important also because the iHappy variable can be computed every year, while updated census information on the population is not always available.

Table 2: Summary of point estimates of SFCE and their efficiency.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\hat{\theta}_i^{Dir}(\%)$	15.38	19.44	21.34	22.45	25.56	35.42
$\hat{\theta}_i^{FH}(\%)$	15.37	19.70	21.60	22.19	24.47	29.91
$\hat{\theta}_i^{FH,BD}(\%)$	15.44	19.68	21.64	22.17	24.65	29.55
$rmse(\hat{\theta}_i^{FH})/rmse(\hat{\theta}_i^{Dir})(\%)$	19.79	60.66	74.90	70.38	82.16	99.39
$rmse(\hat{\theta}_i^{FH,BD})/rmse(\hat{\theta}_i^{Dir})(\%)$	18.44	58.29	72.37	68.49	80.35	99.43
$rmse(\hat{\theta}_i^{FH,BD})/rmse(\hat{\theta}_i^{FH})(\%)$	93.18	95.73	97.33	97.02	98.22	100.50

In order to obtain a clearer picture of the estimates across the country, we mapped them out in figure 1. In the same figure we contrast our estimates with the map of the iHappy variable to show the relationship between the two variables. The SFCE point estimate for the out of sample province of Enna has been computed using the regression synthetic estimator (see Rao and Molina, 2015). In particular, the estimated SFCE for the province of Enna is 25.29% with an rmse of 1.98%. These results seem plausible according to the estimates obtained for the neighbors provinces.

In Italy the SFCE is 22.2% at national level, showing that in average the consumption of food does not represent a large amount on total expenses for con-

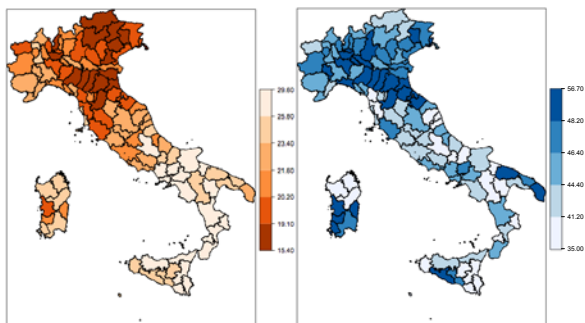


Fig. 1: Map of the FH estimates of the SFCE (left) and map of the iHappy variable for 110 provinces in Italy (right).

sumption. At provincial level (table 2) the SFCE varies between 15.44% (Ravenna, central Italy) and 29.55% (Caserta, southern Italy), so there is evidence of spatial heterogeneity. About a quarter of the provinces have an SFCE  $\geq 25\%$ . All these provinces are in the southern part of Italy. Nine provinces have an estimated SFCE that is below 18%, five of these provinces are in the central part and the other four are in the northern part of Italy, confirming the well known Italian north-south divide.

For a more detailed description of this application see Marchetti et al (2016)

## 5 Conclusions

In this paper we focused on the iHappy indicator obtained from the analysis of Twitter data. The data consist of all the geo-referenced tweets posted in 2012 in the Italian provinces, classified by Curini et al (2015) as the percentage of happy tweets to the total of tweets at provincial level.

In our analysis the iHappy indicator resulted a good additional covariate to predict households' SFCE, given the net influence of other covariates characterizing the provinces, such as the tenure status of the house, the gender of the head of the households, the level of the expenses of the local government to support vulnerable groups.

In Italy the SFCE shows a territorial variability that mimics that of many socio-economic indicators: in 2014 the north-eastern and north-western part of Italy had the lowest level of SFCE (respectively 15.7% and 15.5%) while the southern part (islands included) had the highest (21%). This north-south divide is evident also from the territorial distribution of the iHappy indicator, with few exceptions (some provinces of Sardinia, Puglia and Sicily).

Concluding, the iHappy indicator on happiness can provide useful covariates on yearly bases, free of charge and broken by provinces. It comes affected by self-

selection bias and measurement error. In this application we assumed that the self-selection is negligible and that the measurement error appears to be a minor issue.

## References

- Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. *Science* 350:1073–1076
- Curini L, Iacus S, Canova L (2015) Measuring idiosyncratic happiness through the analysis of twitter: An application to the italian case. *Social Indicators Research* 121(2):525–542
- Datta G, Hall P, Mandal A (2011) Model selection and testing for the presence of small area effects, and application to area-level data. *Journal of the American Statistical Association* 106:362–374
- Decuyper A, Rutherford A, Wadhwa A, Bauer J, Krings G, Gutierrez T, Blondel V, Luengo-Oroz M (2014) Estimating food consumption and poverty indices with mobile phone data. Tech. rep., UNITED NATIONS GLOBAL PULSE
- Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. *Science* 328:1029–1031
- Fay R, Herriot R (1979) Estimation of income from small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association* 74:269–77
- Giusti C, Masserini L, Pratesi M (2016) Local comparisons of small area estimates of poverty: an application within the tuscan region in italy. *Social Indicators Research*
- Hastie T, Pregibon D (1992) Generalized linear models, Wadsworth and Brooks/Cole, chap 6
- Horvitz D, Thompson D (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47:663–85
- Marchetti S, Giusti C, Pratesi M, Salvati N, Giannotti F, Pedreschi D, Rinzivillo S, Pappalardo L, Gabrielli L (2015) Small area model-based estimators using big data sources. *Journal of Official Statistics* 31:263–281
- Marchetti S, Giusti C, Pratesi M (2016) The use of twitter data to improve small area estimates of households' share of food consumption expenditure in italy. *ASTA Wirtsch Sozialstat Arch* 10(79)
- Porter A, Holan S, Wikle C, Cressie N (2014) Spatial fay-herriot models for small area estimation with functional covariates. *Spatial Statistics* 10:27–42
- Rao J, Molina I (2015) Small Area Estimation. *Wiley Series in Survey Methodology*, Wiley, URL [https://books.google.it/books?id=i1B\\_BwAAQBAJ](https://books.google.it/books?id=i1B_BwAAQBAJ)
- Shapiro S, Wilk M (1965) An analysis of variance test for normality (complete samples). *Biometrika* 67:215–216
- Ybarra L, Lohr S (2008) Small area estimation when auxiliary information is measured with error. *Biometrika* (95):919–931

# Gross Annual Salary of a new graduate: is it a question of profile?

## *Retribuzione Annuale Lorda di un neo-laureato: dipende tutto dal profilo?*

Paolo Mariani, Andrea Marletta and Mariangela Zenga

**Abstract** The paper aims to identify an ideal profile for the new graduates in the recruitment process. Moreover, the distribution of their gross annual salary is analyzed in relation to a selected profile. The analysis is based on the Education-for-Labour Elicitation from Companies' Attitudes towards University Studies Project using the methodology of a Conjoint Analysis. The data refers to 471 enterprises operating in Lombardy in different economic sectors with particular focus on the tertiary sector.

**Abstract** *Il lavoro si propone di individuare un profilo ideale dei neo laureati nel processo di selezione. Viene inoltre presa in considerazione la distribuzione della retribuzione annua lorda del neo assunto in relazione ad uno specifico profilo selezionato. L'analisi si basa sulla ricerca Education-for-Labour Elicitation from Companies' Attitudes towards University Studies utilizzando la metodologia della Conjoint Analysis. I risultati sono relativi a 471 imprese operanti in Lombardia, secondo diversi settori economici, con particolare attenzione al settore del terziario.*

**Key words:** New graduates, Conjoint Analysis, Utility Score, Gross Annual Salary.

---

Paolo Mariani

Department of Economics Management and Statistics, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, Milano, Italy e-mail: paolo.mariani@unimib.it

Andrea Marletta

Department of Economics Management and Statistics, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, Milano, Italy e-mail: andrea.marletta@unimib.it

Mariangela Zenga

Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, Milano, Italy e-mail: mariangela.zenga@unimib.it

## 1 Introduction

This work concerns the comprehension policies about relationships between the enterprises and universities, with reference to the labour market for the new graduates. In particular, the study is based on the multi-centre research, Education-for-Labour Elicitation from Companies' Attitudes towards University Studies (Electus) [3], a research project involving 6 Italian universities. The aims of Electus are various. Firstly, it focus on the identification of an ideal graduate profile for several job positions. Secondly, it works toward some across-the-board skills, universally recognized as 'best practices' for a graduate. Finally, the analysis allows to achieve differences and valuations between wage and competencies for new graduates.

The paper is organized as follows. The first section contains the presentation of the statistical method (Conjoint Analysis). The Mariani-Mussini coefficient of economic valuation is introduced in the second section. The results of Electus survey is showed in the third section. Finally, conclusions and main remarks are discussed in the last part of the paper.

## 2 Methodology: the Conjoint Analysis and the coefficient of Economic Variation

Conjoint analysis (CA) is a technique widely used to investigate consumer choice behaviour [4]. In particular, in this study CA refers to the stated preference model used to obtain part-worth utilities. The aim of this model consists in estimating a utility function (UF) for the characteristics describing several profiles. The UF is defined as follow:

$$U_k = \sum_{s=0}^n \beta_s x_{sk} \quad (1)$$

where  $x_{0k}$  is equal to 1 and  $n$  is the number of all level of attributes which define the combination of a given profile,  $x_{sk}$  is the dummy variable that refers to the specific attribute level. As a result, the utility associated with  $k$  alternatives ( $U_k$ ) is obtained by summing the terms  $\beta_s x_{sk}$  over all attribute levels, where  $\beta_s$  is the partial change in  $U_k$  for the presence of the attribute level  $s$ , holding all other variable constants. Usually when CA is performed, all respondents answer to every possible profile. In this experiment the possible profiles obtained from combining every level in a full factorial fashion were so numerous, so it was necessary to apply an ad-hoc fractional factorial design. According to several criteria [3], an individual random sample of four profiles was administered to each respondent which had to mark them on a scale of 1 to 10. This experimental final design results both orthogonal and balanced.

Part-worth utilities of levels obtained from non-standard CA represents the starting point to re-evaluate the proposed Gross Annual Salary of the job vacancies. Secondly, economic re-evaluation will be carried out through relative importance of

attributes in non-standard CA using Mariani-Mussini coefficient of economic valuation [5]. The general formulation of the coefficient is:

$$MI_{ij} = \frac{U_i - U_b}{U_b} * I_j \quad (2)$$

where  $U_i$  is the sum of part-worth utility scores associated with the profile  $i$ ,  $U_b$  the sum of utility scores associated with a baseline profile and  $I_j$  is the relative importance for the attribute  $j$ .

The coefficient  $MI_{ij}$  could be also used for estimating the variation in terms of the salary associated to profile  $i$  compared to the baseline one. Given the salary associated with the baseline profile  $\pi$ , the coefficient of economic re-evaluation can be expressed as:

$$V_{ij} = MI_{ij} * \pi \quad (3)$$

Variations  $V_{ij}$  change in proportion of the  $I_j$ , this entails two basic considerations. Firstly, when an attribute has a very high value of importance,  $V_{ij}$  assumes higher variations. Secondly,  $V_{ij}$  concern attribute variations one at a time, that is to say profile comparisons are possible only varying an attribute, holding fixed all others. Moreover, if the baseline profile is the best/worst one, all coefficients  $MI_{ij}$  and all variations  $V_{ij}$  will be negative/positive.

### 3 Application

The survey was conducted in 2015 using CAWI technique. Data were collected using a software program called Sawtooth [6]. Data manipulation and Conjoint Analysis were obtained using *R* software and *Conjoint* package [1].

The questionnaire contained two sections: conjoint experiment for the five job positions and general information about the company (demographic questions). Regarding the five job positions for the new graduates, Administration clerk, HR assistant, Marketing assistant, ICT professional and CRM assistant were considered. To specify the candidates' profile, six attributes were used:

- *Field of Study* with 10 levels (Philosophy and literature, Educational sciences, Political science/ Sociology, Economics, Law, Statistics, Industrial engineering, Mathematics/ Computer sciences, Psychology, Foreign languages),
- *Degree Mark* with 3 levels (Low, Medium, High),
- *Degree Level* with 2 levels (Bachelor, Master),
- *English Knowledge* with 2 levels (Suitable for communication with foreigners, Inadequate for communication with foreigners),
- *Relevant work experience* with 4 levels (No experience at all, Internship during or after completion of university studies, Discontinuous or occasional work during university studies, One year or more of regular work)

- *Willingness to Travel on Business* with 3 levels (Unwilling to travel on business, Willing to travel on business only for short periods, Willing to travel on business even for long periods).

After having rated the selected profile and chosen the best one, the entrepreneurs had to propose a Gross Annual Salary for the chosen profile in order to measure the so-called 'willingness to pay' [2].

As far as the Milano-Bicocca research unit is concerned, interviewees were representatives of companies registered on the Portal of Almalaurea for recruitment and linkage, limited to the university site. Final respondents were 471. Companies profile shows that they were in prevalence (52%) small sized (15-49 employees), followed (25.6%) by medium sized, ranging from 50 to 249 employees and (22.4%) by the large companies with 250 employees or more. The most represented activity sectors were services to the industry (62.1%), services to the person or the family (16.2%) and manufacturing (14.9%). The majority of companies (89.4%) operated-fully or partially within the domestic market. Moreover, they were mainly under the management of the entrepreneur (63%).

Five CAs are achieved corresponding to the different job positions in order to measure entrepreneurs' preferences. Results for path-worth utilities are similar for all the attributes, except for the attribute *Field of Study*. This means that all other competencies have some levels that are universally identified as 'best practice' for a graduate.

The attributes named *Relevant work experience* and *English Knowledge* show always the same level as best for each vacancy. After all, it is easy to imagine that companies prefer to employ a candidate with one year or more of regular work and suitable for communication with foreigners. Variables *Degree Mark* and *Willingness to travel on business* are competencies where best two levels are always preferred, so a medium-high marked degree and the willing to travel on business for short or long periods are preferable among candidates.

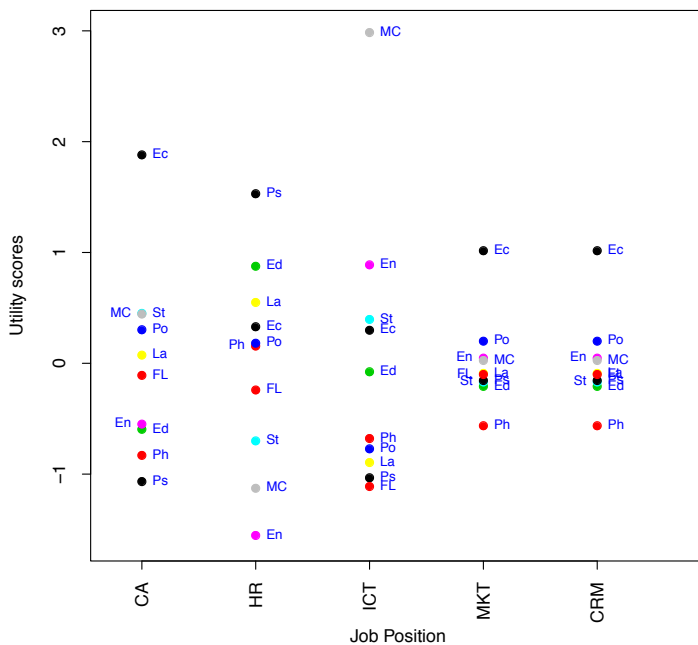
Utility scores for variable *Degree level* are very close to 0 for each position, this means that there is no substantial difference for a bachelor or a master degree for a graduate.

The attribute *Field of Study* is more complex to analyse since it is less cross and a degree in a field could result the best for a position and the worst for an other one. For this reason, in this paper only variations about Field of Study are taking into account. This allows to make a comparison of the coefficient of economic evaluation  $MI_{ij}$  and its associated variation  $V_{ij}$  through different job positions.

In Fig. 1 part-worth utilities for *Field of Study* attribute from 5 non-standard CA are shown for each job position. Economics studies represents the best profile considering 3 job positions, while a degree in Psychology and Mathematics/ Computer sciences optimizes utility respectively for HR Assistant and ICT Professional.

The contribution of these part-worth utilities is very relevant for total utility  $U_i$ , since variable *Field of Study* has the higher relative importance of attributes  $I_j$  for each job vacancies. The quota of explained  $I_j$  ranges from a minimum of 42.98% for a position in customer relationship management (CRM) to a maximum of 80.54%





**Fig. 1** Part-worth utilities for job position and field of study

for a vacancy in ICT technical positions. Importance for other attributes is always under 20% for each position.

In this application, baseline profile is the best profile which optimizes the total utility, so  $U_b$  is the sum of the highest part-worth utilities (plus an intercept) for each attribute  $j$ . This means that all  $MI_{ij}$  coefficients and all variations  $V_{ij}$  are negative.

Table 1 shows  $MI_{ij}$  coefficients of economic re-evaluation for Field of Study, as expected each  $MI_{ij} \leq 0$  and  $MI_{ij} = 0$  only in correspondence of the best Field of Study for vacancy. Comparing the job positions, ICT professionals displays higher coefficients. This is due to the fact the  $I_j$  is very high and a degree in Mathematics/Computer sciences is fully specialized for this position. The biggest coefficient is for a degree in Foreign languages for ICT professionals, in comparison with a graduate in Mathematics/Computer sciences they earn an halved Gross Annual Salary.

**Table 1**  $M_{ij}$  coefficients for Field of Study

Field of Study\ Job position	AC	HR	ICT	MKT	CRM
Philosophy and literature	-18.27%	-10.97%	-45.96%	-12.35%	-10.84%
Educational sciences	-16.68%	-5.23%	-38.39%	-13.24%	-8.41%
Political science/ Sociology	-10.63%	-10.76%	-47.12%	-11.03%	-5.61%
Economics	---	-9.57%	-33.70%	---	---
Law	-12.17%	-7.83%	-48.68%	-15.71%	-7.60%
Statistics	-9.63%	-17.79%	-32.48%	-11.41%	-8.13%
Industrial engineering	-16.37%	-24.60%	-26.29%	-14.70%	-6.65%
Mathematics/ Computer sciences	-9.68%	-21.21%	---	-14.82%	-6.80%
Psychology	-19.86%	---	-50.40%	-10.47%	-8.04%
Foreign languages	-13.40%	-14.13%	-51.39%	-9.24%	-7.67%

## 4 Conclusion and future research

The article proposes the use of a non-standard Conjoint Analysis in detecting best profiles for graduates using data from the Electus project. Moreover, a new evaluation of the Gross Annual Salary is proposed using the Mariani-Mussini index derived from utility scores. Analysis deriving from 5 different job positions show how all graduates' competencies are across-the-board, except for *Field of Study*. English knowledge, medium or high level for degree mark, relevant work experience and willingness to travel are the most important required attributes for a graduate. About *Field of Study*, a degree in Economics seems to be the most attractive for entrepreneurs, except for very specialized vacancies as ICT professionals, where other degree courses exhibit an halved salary respect to Computer Sciences graduates.

Future research will focus the attention on results coming from stratified CA based on socio-demographic features of companies responding in the Electus project.

## References

1. Bak A. and Bartlomowicz T. (2012), Conjoint analysis method and its implementation in conjoint R package, In: Pociucha J., Decker R. (Eds.), *Data analysis methods and its applications*, C.H. Beck, p. 239-248
2. Breidert C., *Estimation of Willingness-to-Pay. Theory, Measurement, Application*. Deutscher Universitts-Verlag, Wiesbaden (2006)
3. Fabbri L., Scioni M.: The ideal candidate to a job through a conjoint study. In: Meerman, A., Kliewe, T. (eds.) *Academic Proceedings of the 2015 University-Industry Interaction Conference: Challenges and Solutions for Fostering Entrepreneurial Universities and Collaborative Innovation*, Amsterdam (2015) .
4. Green P.E., Srinivasan V.: Conjoint analysis and Consumer research: Issues and Outlook. *Journal of Consumer Research*. **5**(2), 103-123 (1978)
5. Mariani P., Mussini M.: A new coefficient of Economic evaluation based on Utility Scores. *Argumenta Oeconomica*. **1**(30), 33-46 (2013)
6. Sawtooth Software <http://www.sawtoothsoftware.com>

# Dynamic random coefficient based drop-out models for longitudinal responses

## *Modelli a coefficienti casuali dinamici per risposte longitudinali affette da drop-out non-ignorabile*

Maria Francesca Marino and Marco Alfó

**Abstract** We propose a dynamic random coefficient based drop-out model for the analysis of longitudinal data subject to potentially non-ignorable drop-out. The presence of a non-ignorable missingness may severely bias inference on the observed data. In this framework, random coefficient based drop-out models represent an flexible approach to jointly model both longitudinal responses and missingness. We extend such an approach by allowing the random parameters in the longitudinal data process to evolve over time according to a non-homogeneous hidden Markov chain. The resulting model offers great flexibility and allows us to efficiently describe both between-outcome and within-outcome dependence.

**Abstract** *Gli studi longitudinali sono spesso caratterizzati dalla presenza di dati mancanti dovuti ad alcuni individui che lasciano lo studio anticipatamente. Quando il meccanismo che conduce al dato mancante è non ignorabile, è possibile giungere a conclusioni inferenziali valide solo modellando congiuntamente due outcome: il processo longitudinale ed il processo generatore del dato mancante stesso. A questo scopo, si propone un modello di regressione per dati longitudinali soggetti ad a drop-out potenzialmente non ignorabile in cui coefficienti casuali tempo-costanti e tempo-variabili vengono congiuntamente presi in considerazione. Questo permette di modellare in maniera opportuna sia la dipendenza esistente tra le misurazioni di ripetute di uno stesso outcome per una stessa unit statistica, sia la dipendenza esistente tra outcome diversi.*

**Key words:** Hidden Markov models, nonparametric maximum likelihood, random effects, missingness,

---

Maria Francesca Marino

Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, e-mail: mariafrancesca.marino@unifi.it

Marco Alfó

Dipartimento di Scienze Statistiche, "Sapienza" Università di Roma, e-mail: marco.alfó@uniroma1.it

## 1 Introduction

Longitudinal studies are frequently affected by drop-out. If the selection of individual staying in the study still depends on (future) unobserved responses once conditioning on the observed data, the missingness mechanism is said to be non-ignorable [9]. In this respect, to obtain valid inference, missingness should be taken in explicit account.

Different alternatives are available in the literature to deal with non-ignorable drop-outs [8]. Among them, random coefficient based drop-out model [RCBDM - 7] represent an interesting approach. They allow for the presence of two different sets of individual-specific random parameters for the longitudinal and the missing data process, respectively. These capture the dependence between repeated measurements from the same individual (within-individual dependence). The corresponding joint distribution provides instead a measure of dependence between the longitudinal and the missingness process (between-outcomes dependence).

When dealing with longitudinal data, the assumption of time-constant, individual-specific, sources of unobserved heterogeneity may be too restrictive [1]. Starting from the proposal by [10], we introduce a dynamic random coefficient based drop-out model, where time-varying random parameters are considered to model the longitudinal outcome. To explain our proposal, we assume that the dependence between the longitudinal and the missing data process is captured by an individual-specific *upper-level* mixture. Also, to describe the dependence within profiles, we consider two further sets of random parameters. For the longitudinal outcome, individual-specific, time-varying, random parameters that evolve over time according to a non-homogeneous hidden Markov chain are exploited. On the other hand, for the missing data outcome, we consider individual-specific, time-constant, random parameters identifying non-homogeneous propensities to stay into the study.

The proposed model is applied to the Leiden 85+ dataset where the effect of demographic and genetic factors on the evolution of cognitive functioning in elder people is of main interest [3]. Due to poor health conditions or death, individuals enrolled in the study may present incomplete sequences. We show how the proposed model specification may be fruitfully exploited to derive valid inference on the parameters of interest.

## 2 Motivating example: the Leiden 85+ study

The Leiden 85+ study is a longitudinal study conducted by the Leiden University Medical Centre in the Netherlands, with the aim at analysing the evolution of cognitive functioning in the elderly. The study entails Leiden inhabitants who turned 85 years old between September 1997 and September 1999. The sample is made by 541 elderly who were followed for six consecutive yearly visits until they reached 90 years of age. Patient conditions were assessed via the Mini Mental Status Examination [MMSE, 6] index taking values between 0 and 30 with higher values corre-

sponding to better cognitive skills. The aim of the study is that of identifying demographic and genetic factors influencing the dynamics of cognitive functioning and healthy aging. To this purpose, the following covariates were measured: age, *gender*, *educational status*, and *APOE genotype*. The latter identifies the Apolipoprotein E genotype of the patient; in particular,  $\epsilon_4$  allele is known to be linked to the risk of dementia. Due to the design of the study, a number participants present incomplete responses (i.e. drop-out), because of poor health conditions or death.

Preliminary analysis show that MMSE values generally reduce with time but such a trend is more evident for subjects dropping out prematurely. Such a finding poses the question on whether the process leading to missing data may be ignored. In the next section, we will introduce a dynamic RCBDM to account for both the potential dependence between the longitudinal data process and the drop-out mechanism and the within-profile dependence.

### 3 The dynamic RCBDM

Let us suppose a longitudinal study is designed to collect measures for a response variable  $Y_{it}, i = 1, \dots, n, t = 1, \dots, T$ , on a sample of  $n$  individuals at  $T$  time occasions and let  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT})$  denote the vector of individual response sequences. As it is frequent when dealing with longitudinal studies, some individuals in the sample may drop-out prematurely and, thus, may present incomplete sequences. In this framework, let  $\mathbf{R}_i = (R_{i1}, \dots, R_{iT_i^*})'$  indicate the  $T_i^*$ -dimensional missing data vector, where  $T_i^* = \min(T_i + 1, T)$  and  $T_i$  denotes the number of available measurements for the  $i$ -individual.  $R_{it}$  is defined as a binary variable with  $R_{it} = 0$  if the  $i$ -th individual drops-out from the study between occasion  $t - 1$  and  $t$  and  $R_{it} = 1$  otherwise.

Furthermore, let  $Z_{it} \in \{1, \dots, G\}$  and  $U_i \in \{1, \dots, K\}$  be two individual-specific, discrete, latent variables influencing the longitudinal and the missing data process, respectively. As it is clear, while the latter variable is assumed to depend on the individual  $i$  only, the former variable is individual- and time-specific. This allows us to capture sources of unobserved dynamics that influence  $Y_{it}$  and that would be barely captured by a time-constant latent variable.

We assume that the longitudinal outcome  $Y_{it}$  only depends on the corresponding latent variable  $Z_{it}$  and, conditional on the vector  $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iT})$ , the elements of  $\mathbf{Y}_i$  are independent, with joint (conditional) density given by

$$f(\mathbf{y}_i | \mathbf{Z}_i = \mathbf{z}_i) = \prod_{t=1}^T f(y_{it} | Z_{it} = z_{it}).$$

Similarly, we assume that conditional on the latent variable  $U_i$ , missingness indicators are independent and the corresponding joint (conditional) density is

$$f(\mathbf{r}_i | U_i = u_i) = \prod_{t=1}^{T_i^*} f(r_{it} | U_i = u_i).$$

To describe the effect of the observed covariates on the outcomes  $(Y_{it}, R_{it})$ , the following regression models are also defined:

$$\begin{cases} g[\mathbb{E}(Y_{it} | Z_{it} = g)] = \zeta_g + \mathbf{x}'_{it}\beta, \\ \text{logit}[\text{Pr}(R_{it} = 0 | U_i = k)] = \xi_k + \mathbf{w}'_{it}\gamma. \end{cases}$$

In the expressions above,  $g(\cdot)$  represents an appropriate link function, while the parameters  $\beta$  and  $\gamma$  describe the effects of observed covariates,  $\mathbf{x}_{it}$  and  $\mathbf{w}_{it}$ , on  $Y_{it}$  and  $R_{it}$ , respectively. Also,  $\zeta_g, g = 1, \dots, G$ , denotes the value of the random intercept in the longitudinal data model when  $Z_{it} = g$ . To simplify the interpretation of such parameters, we introduce the following ordinal constraint:

$$\zeta_1 \leq \zeta_2 \leq \dots \leq \zeta_G, \quad (1)$$

so that lower values of  $Z_{it}$  correspond to lower values for the longitudinal response. Last,  $\xi_k, k = 1, \dots, K$ , denotes the discrete random intercept associated to the missing data process when  $U_i = k$ .

Following an approach similar to that suggested by [10], we model the dependence between  $Z_i$  and  $U_i$  and, therefore, between the longitudinal and the missing data process, by considering a discrete *upper-level* latent variable,  $V_i$ , defined on the support set  $\{1, \dots, H\}$ , with  $\tau_h = \text{Pr}(V_i = h), h = 1, \dots, H$ . In particular, we assume that, conditional on  $V_i = h$ , the latent variables  $Z_i$  and  $U_i$  are independent with joint distribution described by the following (association) model:

$$f(\mathbf{Z}_i, U_i) = \sum_{h=1}^H \tau_h [\text{Pr}(\mathbf{Z}_i = \mathbf{z}_i | V_i = h) \text{Pr}(U_i = u_i | V_i = h)].$$

With the aim of accounting for time-varying sources of unobserved heterogeneity influencing the longitudinal data process, we assume that, conditional on the  $h$ -th component of the upper-level mixture, that is conditional on  $V_i = h$ , the latent variables  $Z_{it}$  evolve over time according to a first order hidden Markov chain with initial probability vector  $\delta_h$  and transition probability matrix  $\mathbf{Q}_h$ , with  $h = 1, \dots, H$ .

### 3.1 Reducing model complexity

As it can be noticed, the adopted parameterization is quite complex. This could lead to numerical difficulties when deriving the corresponding maximum likelihood estimates. In order to reduce the number of parameters, we follow an approach similar to that by [4] and specify  $\delta_h$  and  $\mathbf{Q}_h$  via a global logit parameterization. This choice is motivated by the constraints specified in equation (1) which, in turn, lead to considering the latent variable  $Z_{it}$  having as ordinal. In this framework, initial

probabilities of the hidden Markov chain are defined according to the model

$$\log \frac{\delta_{g|h} + \dots + \delta_{G|h}}{\delta_{1|h} + \dots + \delta_{g-1|h}} = \alpha_{0g} + \psi_{0h}, \tag{2}$$

with  $h = 1, \dots, H$  and  $g = 2, \dots, G$ . For identifiability purposes, we set  $\psi_{01} = 0$ , so that the number of parameters to be estimated reduces to  $(G - 1) + (H - 1)$ . On the other hand, transition probabilities are modelled according to the following ordinal logit:

$$\log \frac{q_{gg'|h} + \dots + q_{Gg'|h}}{q_{1g'|h} + \dots + q_{g-1g'|h}} = \alpha_{1gg'} + \psi_{1h}, \tag{3}$$

with  $h = 1, \dots, H$ ,  $g = 1, \dots, G$ , and  $g' = 2, \dots, G$ . As above, to ensure parameter identifiability, we set  $\psi_{11} = 0$ , so that  $G(G - 1) + (H - 1)$  parameters need to be estimated.

### 4 Model inference

Let  $\theta$  denote the vector of all model parameters. Estimation of such parameters can be carried out via a maximum likelihood approach. Due to the local independence assumption within and between the longitudinal and the missing data responses,  $\mathbf{Y}_i$  and  $\mathbf{R}_i$ , inference may be based on the following observed data likelihood:

$$L(\theta) = \prod_{i=1}^n \sum_h^H \tau_h \left\{ \sum_{Z_{i1} \dots Z_{iT_i}} \left[ \prod_{t=1}^{T_i} f(y_{it} | Z_{it} = z_{it}) \delta_{z_{it}|h} \prod_{t=2}^T q_{z_{it-1}z_{it}|h} \right] \times \right. \\ \left. \times \left[ \prod_{t=1}^{T_i^*} \sum_{u_i} f(r_{it} | U_i = u_i) \pi_{u_i|h} \right] \right\},$$

To avoid multiple summations over all possible realisations of the hidden chain,  $Z_{i1}, \dots, Z_{iT_i}$ , we may rely on the EM algorithm [5].

In this framework, two separated steps need to be alternated. In the E-step, we need to compute expected value of the complete data log-likelihood, conditional on the observed data and the current value of parameter estimates. Such a computation can be consistently simplified by extending the standard forward-backward algorithm [2] which is typically used in the hidden Markov model framework. In the M-step, we need to maximize the expected value of the complete data log-likelihood with respect to model parameters. The E- and the M-steps are iterated until convergence. As it is frequent when dealing with discrete latent variables, to avoid local maxima or spurious solutions, we may consider a multi-start strategy based on both deterministic and random solutions. Also, the number of upper- and lower-level components/states is treated as fixed and known. The algorithm is run

for varying choices of  $(G, K, H)$  and the best model is chosen via standard model selection techniques.

## References

- [1] F. Bartolucci and A. Farcomeni. A discrete time event-history approach to informative drop-out in mixed latent markov models with covariates. *Biometrics*, 71:80–89, 2015.
- [2] L. E Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.
- [3] A. Bootsma-Van Der Wiel, E. Van Exel, A.J.M. De Craen, J. Gussekloo, A.M. Lagaay, D.L. Knook, and R.G.J. Westendorp. A high response is not essential to prevent selection bias: results from the leiden 85-plus study. *Journal of clinical epidemiology*, 55:1119–1125, 2002.
- [4] R. Colombi and A. Forcina. Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, pages 1007–1019, 2001.
- [5] A.P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38, 1977.
- [6] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. “mini-mental state”: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12:189–198, 1975.
- [7] Nisha C. Gottfredson, Daniel J. Bauer, and Scott A. Baldwin. Modeling change in the presence of nonrandomly missing data: Evaluating a shared parameter mixture model. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2):196–209, 2014.
- [8] R. J.A. Little and D. B. Rubin. *Statistical analysis with missing data*. Wiley, 2002.
- [9] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
- [10] Alessandra Spagnoli and Marco Alfò. Random coefficient based dropout models: a finite mixture approach. In *46th Scientific Meeting of the Italian Statistical Society*, 2012.



# Hidden Markov models: dimensionality reduction, atypical observations and algorithms

## *Modelli Markoviani latenti: riduzione dimensionale, dati anomali e algoritmi*

Antonello Maruotti and Jan Bulla

**Abstract** We develop a new class of parsimonious models to perform time-varying clustering and dimensionality reduction in a time-series setting, while accounting for atypical observations. The problem of similarity search in time-series data is addressed by specifying a hidden Markov model. For the maximum likelihood estimation of the model parameters, we outline an ad-hoc Alternating Expected Conditional Maximization (AECM) algorithm. As the inclusion of covariates in the model might alter the formation of the latent states and that parameter estimation could become infeasible with large numbers of time points and covariates, we firstly construct the observed model, then obtain the latent state classifications, and subsequently study the relationship between covariates and latent state memberships.

**Abstract** *In questo lavoro, introduciamo una nuova classe di modelli parsimoniosi per classificare e ridurre la dimensione dello spazio delle variabili in un contesto di serie storiche, tenendo conto nel processo di stima di dati anomali. Un modello Markoviano latente specificato per classificare le serie storiche. Per la stima di massima verosimiglianza dei parametri del modello, definiamo un algoritmo ad hoc di tipo AECM. L'inclusione di covariate nel modello potrebbe alterare la formazione degli stati latenti. La stima dei parametri potrebbe diventare computazionalmente complessa. Abbiamo, perciò, costruito un modello per la parte osservata e, successivamente, ottenuto la classificazione delle osservazioni nei vari stati latenti; infine studiamo la relazione tra covariate e stati latenti.*

**Key words:** Factor model, AECM, Three-step algorithm, Contaminated Gaussian distribution

---

Antonello Maruotti

Libera Università Maria Ss. Assunta, Via Pompeo Magno 22 - 00192 Roma, e-mail: a.maruotti@lumsa.it

Jan Bulla

University of Bergen, Realfagbygget, Allgt. 41, Bergen e-mail: Jan.Bulla@uib.no

## 1 Introduction

Hidden Markov models (HMMs) are state of the art in the analysis of time-dependent data. These models have been applied in time series analysis for more than four decades [1]. Being dependent mixture models, HMMs allow to unambiguously recover of the structure of the data by rigorously defining homogeneous latent subgroups; simultaneously, they provide meaningful interpretation of the inferred partition. Nowadays, Gaussian HMMs are commonly used for clustering continuous data (see, e.g., [2]), although some robust (conditional) distributions have been recently proposed in the literature [4, 8].

Real data are often contaminated by outliers, spurious points or noise (collectively called *atypical observations* herein, that may affect both the parameters estimates and the ability to recover the latent structure. Despite the wide literature on robust estimation of mixture models, there are not many papers dealing with robustness issues in HMMs [3, 5].

The challenge of modeling multivariate time-series and their interactions is fairly common to all analyzes of high dimensional data with many variables of interests. Dimensionality-related aspects present a challenge, because these time-series could be potentially highly correlated. Therefore, estimation and interpretation of the parameters of interest may become non-trivial. In order to examine the interrelationships between time-series and to perform dimensionality reduction in the variable space simultaneously (allowing for an easy interpretation of model parameters), we propose the use of a latent factor model. Accordingly, we define a general class of parsimonious HMMs by imposing a factor decomposition on state-specific covariance matrices. The loadings and noise terms of the covariance matrix may be constrained to be equal or unequal across latent states. In addition, the noise term may be subject to further restrictions, resulting in a set of eight parsimonious covariance structures [6, 7].

At last, in order to characterize transitions between hidden states along with estimating the effects of observed covariates on the transitions, we use a multinomial logistic regression model, which is capable of revealing the heterogeneity in the transition process.

## 2 Methodology

### 2.1 Notation and assumptions

Let  $\{\mathbf{Y}_t, t = 1, \dots, T\}$  denote sequences of multivariate longitudinal observations recorded on  $T$  times, where  $\mathbf{Y}_t = (Y_{t1}, \dots, Y_{tP})' \in R^P$ , and let  $\{S_t; i=1, \dots, I, t = 1, \dots, T\}$  be a first-order Markov chain defined on the state space  $\{1, \dots, k, \dots, K\}$ . A HMM is a particular kind of dependent mixture. It is a stochastic process consisting of two parts: the underlying unobserved process  $\{S_t\}$ , fulfilling the Markov

property, i.e.

$$\Pr(S_t = s_t \mid S_1 = s_1, S_2 = s_2, \dots, S_{t-1} = s_{t-1}) = \Pr(S_t = s_t \mid S_{t-1} = s_{t-1}),$$

and the state-dependent observation process  $\{\mathbf{Y}_t\}$  for which the conditional independence property holds, i.e.

$$f(\mathbf{Y}_t = \mathbf{y}_t \mid \mathbf{y}_1 = \mathbf{y}_1, \dots, \mathbf{Y}_t = \mathbf{y}_t, S_1 = s_1, \dots, S_t = s_t) = f(\mathbf{Y}_t = \mathbf{y}_t \mid S_t = s_t),$$

where  $f(\cdot)$  is a generic probability density function.

The hidden Markov chain has  $K$  states with initial probabilities

$$\pi_{ik} = \Pr(S_1 = k), \quad k = 1, \dots, K,$$

and transition probabilities

$${}_{t-1}\pi_{k|j} = \Pr(S_t = k \mid S_{t-1} = j), t = 2, \dots, T; j, k = 1, \dots, K. \tag{1}$$

In (1),  $k$  refers to the current state, whereas  $j$  refers to the one previously visited; this convention will be used throughout the paper. Assuming that the hidden process follows a first-order Markov chain is equivalent to the assumption that any latent variable  $S_t$  given  $S_{t-1}$  is conditionally independent of  $S_1, S_2, \dots, S_{t-2}$ . This dependence structure is seldom considered restrictive, and, due to its easy interpretation usually preferred to more complex structures of the latent variables.

The hidden Markov chain has  $K$  states, labeled from 1 to  $K$ , with initial probabilities

$$\pi_k = \Pr(S_1 = k), \quad k = 1, \dots, K,$$

and transition probabilities

$${}_t\pi_{k|j} = \Pr(S_t = k \mid S_{t-1} = j), \quad t = 2, \dots, T; j, k = 1, \dots, K.$$

Note that  $k$  refers to the current state in the above definitions, whereas  $j$  refers to the one previously visited; this convention will be used throughout the paper. Moreover, the initial probabilities are collected in the  $K$ -dimensional vector  $\boldsymbol{\pi}$ , whereas the  $K \times K$  transition probability matrix  ${}_t\Pi$  contains the time-varying transition probabilities. The simplest model in this framework is the homogeneous HMM, which assumes time-homogeneous transition probabilities, i.e. independence of  $t$  and thus  ${}_t\Pi = \Pi$ . This specification fails to take into account how atmospheric observed conditions affect the evolution of unobserved exposure states and, in general, time heterogeneity of the transition probability matrix. In order to overcome this potential drawback, the transitions probabilities may be parametrized as a function of  $\tilde{P}$  exogenous covariates  $\mathbf{x}_t = \{x_{t1}, \dots, x_{t\tilde{P}}\}$  by

$${}_t\pi_{k|j} = \frac{\exp(\mathbf{x}'_t \boldsymbol{\gamma}_{jk} + \gamma_{jk0})}{1 + \sum_{h=1}^K \exp(\mathbf{x}'_t \boldsymbol{\gamma}_{jh} + \gamma_{jh0})} \tag{2}$$

where  $\gamma_{jk} = \{\gamma_{jk1}, \dots, \gamma_{jk\bar{P}}\}$  represents a vector of fixed regressors and  $\gamma_{jk0}$  is an intercept term. To ensure identifiability, we impose  $\gamma_{jj} = 0$  and  $\gamma_{j0} = 0$  for  $j = 1, \dots, K$ . Accordingly, the probability of no transition at time  $t$  is given by

$${}^t\pi_{j|j} = \frac{1}{1 + \sum_{h=1}^K \exp(\mathbf{x}'_t \gamma_{jh} + \gamma_{j0})}.$$

This model specification permits to investigate the dynamics of the hidden state sequence over time, and allows for a potential impact of covariates on its evolution.

### 2.2 The contaminated factor HMM

As concerns  $\mathbf{Y}_t | S_t = k$ , in the fashion of [8], we assume a contaminated Gaussian distribution

$$f_{CN}(\mathbf{Y}_t | S_t = k; \mu_k, \Sigma_k, \alpha_k, \eta_k) = \alpha_k \phi(\mathbf{Y}_t | S_t = k; \mu_k, \Sigma_k) + (1 - \alpha_k) \phi(\mathbf{Y}_t | S_t = k; \mu_k, \eta_k \Sigma_k), \tag{3}$$

where  $\phi(\cdot; \mu_k, \Sigma_k)$  denotes a  $P$ -variate Gaussian distribution with mean  $\mu_k$  and covariance matrix  $\Sigma_k$ ,  $\alpha_k \in (0, 1)$  is the proportion of good points in state  $k$ , and  $\eta_k > 1$  is an inflation parameter denoting the degree of bad points contamination in state  $k$ . In symbols,  $\mathbf{Y}_t | S_t = k \sim CN_P(\mu_k, \Sigma_k, \alpha_k, \eta_k)$ .

In order to allow for dimension reduction and parsimony, we further assume a contaminated Gaussian factor analyzers model for  $\mathbf{Y}_t | S_t = k$ . Such a model postulates

$$\mathbf{Y}_t | (S_t = k) = \mu_k + \Lambda_k \mathbf{U}_{itk} + \mathbf{e}_{itk}, \tag{4}$$

where  $\mathbf{U}_{itk}$  is a  $Q$ -dimensional ( $Q \ll P$ ) vector of latent factors,  $\Lambda_k$  is a  $P \times Q$  matrix of factor loadings, and  $\mathbf{e}_{itk}$  is the error term. The contaminated Gaussian factor analyzers model generalizes the corresponding Gaussian model by assuming

$$\begin{pmatrix} \mathbf{Y}_t | S_t = k \\ \mathbf{U}_{itk} \end{pmatrix} \sim CN_{P+Q}(\mu_k^*, \Sigma_k^*, \alpha_k, \eta_k), \tag{5}$$

where

$$\mu_k^* = \begin{pmatrix} \mu_k \\ \mathbf{0}_Q \end{pmatrix} \quad \text{and} \quad \Sigma_k^* = \begin{pmatrix} \Lambda_k \Lambda_k' + \Psi_k & \Lambda_k \\ \Lambda_k' & \mathbf{I}_Q \end{pmatrix}. \tag{6}$$

To further analyze the model it is useful to introduce the dichotomous variable  $V_{itk}$  assuming value 1, with probability  $\alpha_k$ , if observation  $i$  at time  $t$  in state  $k$  is a good point and zero, with probability  $1 - \alpha_k$ , if it is a bad point. Thus, for good and bad points we have

$$\begin{pmatrix} \mathbf{Y}_t | S_t = k \\ \mathbf{U}_{itk} \end{pmatrix} \Big|_{V_{itk} = 1} \sim N_{P+Q}(\mu_k^*, \Sigma_k^*) \quad \text{and} \quad \begin{pmatrix} \mathbf{Y}_t | S_t = k \\ \mathbf{U}_{itk} \end{pmatrix} \Big|_{V_{itk} = 0} \sim N_{P+Q}(\mu_k^*, \eta_k \Sigma_k^*),$$

respectively, where  $N_B(\mu, \Sigma)$  denotes a  $B$ -variate Gaussian distribution with mean vector  $\mu$  and covariance matrix  $\Sigma$ . Thus,

$$\begin{aligned} \mathbf{Y}_t | S_t = k, V_{itk} = 1 &\sim N_P(\mu_k, \Lambda_k \Lambda_k' + \Psi_k), \\ \mathbf{Y}_t | S_t = k, V_{itk} = 0 &\sim N_P[\mu_k, \eta_k (\Lambda_k \Lambda_k' + \Psi_k)], \\ \mathbf{U}_{itk} | V_{itk} = 1 &\sim N_Q(\mathbf{0}_Q, \mathbf{I}_Q), \\ \mathbf{U}_{itk} | V_{itk} = 0 &\sim N_Q(\mathbf{0}_Q, \eta_k \mathbf{I}_Q), \\ \mathbf{e}_{itk} | V_{itk} = 1 &\sim N_P(\mathbf{0}_P, \Psi_k), \\ \mathbf{e}_{itk} | V_{itk} = 0 &\sim N_P(\mathbf{0}_P, \eta_k \Psi_k), \end{aligned}$$

so that

$$\begin{aligned} \mathbf{Y}_t | S_t = k &\sim CN_P(\mu_k, \Lambda_k \Lambda_k' + \Psi_k, \alpha_k, \eta_k), \\ \mathbf{U}_{itk} &\sim CN_Q(\mathbf{0}_Q, \mathbf{I}_Q, \alpha_k, \eta_k), \\ \mathbf{e}_{itk} &\sim CN_P(\mathbf{0}_P, \Psi_k, \alpha_k, \eta_k), \end{aligned}$$

where  $\Psi_k = \text{diag}(\psi_{k1}, \dots, \psi_{kp}, \dots, \psi_{kP})$ .

### 2.3 Parsimonious HMCNFA family

In this section, in the fashion of [6], we extend the hidden Markov of contaminated Gaussian factor analyzers model by allowing additional constraints across states on the  $\Lambda_k$  and  $\Psi_k$  matrices and on whether or not  $\Psi_k = \psi_k \mathbf{I}_P$  (isotropic constraint). The full range of possible constraints provides a family of eight different parsimonious hidden Markov of contaminated Gaussian factor analyzers models, which are given in Table 1.

**Table 1** Parsimonious structures derived from the hidden Markov of contaminated Gaussian factor analyzers model.

Identifier	$\Lambda_1, \dots, \Lambda_K$	$\Psi_1, \dots, \Psi_K$	Isotropy on $\Psi_k$	$\Sigma_k$	# of free parameters for $\Sigma_1, \dots, \Sigma_K$
UUU	Unconstrained	Unconstrained	Unconstrained	$\Sigma_k = \Lambda_k \Lambda_k' + \Psi_k$	$K[PQ - Q(Q-1)/2] + KP$
UUC	Unconstrained	Unconstrained	Constrained	$\Sigma_k = \Lambda_k \Lambda_k' + \psi_k \mathbf{I}_P$	$K[PQ - Q(Q-1)/2] + K$
UCU	Unconstrained	Constrained	Unconstrained	$\Sigma_k = \Lambda_k \Lambda_k' + \Psi$	$K[PQ - Q(Q-1)/2] + P$
UCC	Unconstrained	Constrained	Constrained	$\Sigma_k = \Lambda_k \Lambda_k' + \psi \mathbf{I}_P$	$K[PQ - Q(Q-1)/2] + 1$
CUU	Constrained	Unconstrained	Unconstrained	$\Sigma_k = \Lambda \Lambda' + \Psi_k$	$[PQ - Q(Q-1)/2] + KP$
CUC	Constrained	Unconstrained	Constrained	$\Sigma_k = \Lambda \Lambda' + \psi_k \mathbf{I}_P$	$[PQ - Q(Q-1)/2] + K$
CCU	Constrained	Constrained	Unconstrained	$\Sigma_k = \Lambda \Lambda' + \Psi$	$[PQ - Q(Q-1)/2] + P$
CCC	Constrained	Constrained	Constrained	$\Sigma_k = \Lambda \Lambda' + \psi \mathbf{I}_P$	$[PQ - Q(Q-1)/2] + 1$

### 3 Maximum likelihood estimation

Even in this relatively general framework, the parameters of the proposed parsimonious HMMs can be estimated using the method of maximum-likelihood. In order to perform maximum likelihood estimation of the above model on the basis of the multivariate response  $\mathbf{y}_t = \{y_{t1}, \dots, y_{tp}\}$ , computation of the likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{\pi}' \mathbf{f}(\mathbf{y}_1) {}_2\Pi \mathbf{f}(\mathbf{y}_2) {}_3\Pi \dots \mathbf{f}(\mathbf{y}_{T-1}) {}_T\Pi \mathbf{f}(\mathbf{y}_T) \mathbf{1} \quad (7)$$

is necessary. Here,  $\boldsymbol{\theta}_k = \{\mu_k, \Lambda_k, \Psi_k, \gamma_{jk}, \pi_k, k = 1, 2, \dots, K\}$  is the set of all model parameters,  $\mathbf{f}(\mathbf{y}_t)$  denotes a diagonal matrix with conditional probability densities  $f(\mathbf{Y}_t = \mathbf{y}_t | S_t = k; \mu_k, \Lambda_k, \Psi_k)$  on the main diagonal and  $\mathbf{1}$  represents a unit vector of size  $K$ .

To maximize (7) with respect to  $\boldsymbol{\theta}$ , we introduce a three-step AECM based on the following steps:

- Step 1. Fit a homogeneous HMM, i.e. without covariates, for the multivariate continuous outcomes. Maximum likelihood estimation is performed by maximizing (7) under the constraint  ${}_t\Pi = \Pi$  using an AECM algorithm. The motivation beyond the use of the AECM algorithm lies in its ability to break the model into smaller models. On the basis of this preliminary fitting, we obtain the final estimates of the conditional distribution parameters.
- Step 2. For each time  $t = 1, \dots, T$ , we obtain the posterior expected values of state membership on the basis of the first step.
- Step 3. Maximize the component of the (complete-data log-) likelihood involving the hidden structure parameters.

After an initial estimate of the latent parameters, the second and the third steps are iterated until convergence, while keeping fixed the estimates of the conditional distribution parameters from the first step.

At the first step of the algorithm, we partition the set of unknown parameters  $\boldsymbol{\theta}$  in two disjoint subsets  $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ :  $\boldsymbol{\theta}_1$  contains the hidden chain parameters  $\boldsymbol{\pi}$  and  $\Pi$  and the elements of the state-specific means  $\mu_k$ , while  $\boldsymbol{\theta}_2$  consists of  $\Lambda_k$  and  $\Psi_k$ . Then, the following steps are alternated until convergence in order to carry out the AECM algorithm:

- First stage
  - E-step: compute the conditional expectation of the complete-data log-likelihood, given the observed data and the current estimate of the parameter vector  $(\boldsymbol{\mu}, \boldsymbol{\pi}, \Pi)$ , while keeping  $(\Lambda, \Psi)$  fixed at their values resulting from the previous iteration.
  - M-step: maximize the preceding expected complete-data log-likelihood function with respect to  $(\boldsymbol{\mu}, \boldsymbol{\pi}, \Pi)$ .
- Second stage

E-step: compute the conditional expectation of the complete-data log-likelihood, in this step conditional on  $(\Lambda, \Psi)$ , while considering  $(\mu, \pi, \Pi)$  fixed as given by the calculations in the first stage of the AECM algorithm.

CM-step: maximize the preceding expected complete-data log-likelihood function with respect to  $(\Lambda, \Psi)$ . The (conditional) maximization step depends on the imposed model restrictions.

Once the AECM achieves convergence at the first step, we obtain the posterior probabilities of belonging to a state and use these to get estimates of  ${}_t\pi_{k|j}$ . The estimated parameters for the hidden process are the solutions of weighted sums of  $K$  multinomial regressions. We then update the posterior probabilities and iterate Step 2 and Step 3, plugging in the estimated transition probabilities into the log-likelihood function, till further convergence.

## References

1. Baum, L.E. and Petrie, T.: Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, **37**, 1554–1563 (1966).
2. Bartolucci, F. and Farcomeni, A.: A note on the mixture transitions distribution and hidden Markov models. *Journal of Time Series Analysis*, **31**, 132–138 (2010).
3. Bulla, J.: Hidden Markov models with  $t$  components. Increased persistence and other aspects. *Quantitative Finance*, **11**, 459–475 (2011)
4. Farcomeni, A. and Greco, L.: S-estimation of hidden Markov models. *Computational Statistics*, **30**, 57–80 (2015).
5. Maruotti, A.: Robust fitting of hidden Markov regression models under a longitudinal setting. *Journal of Statistical Computation and Simulation*, **84**, 1728–1747 (2014)
6. Maruotti, A., Bulla, J., Lagona, F., Picone, M. and Martella, F.: Dynamic mixtures of factor analyzers to characterize multivariate air pollutant exposures. *The Annals of Applied Statistics*, to appear (2017)
7. McNicholas, P.D. and Murphy, T.B.: Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**, 285–296 (2008).
8. Punzo, A. and Maruotti, A.: Clustering multivariate longitudinal observations: The contaminated Gaussian hidden Markov model. *Journal of Computational and Graphical Statistics*, **25**, 1097–1116 (2016).





# A flexible analysis of PISA 2015 data across countries, by means of multilevel trees and boosting.

## *Analisi comparativa dei risultati di PISA2015: un'applicazione di alberi a effetti misti e Boosting.*

Chiara Masci, Geraint Johnes and Tommaso Agasisti

**Abstract** The aim of this work is to analyze and compare PISA2015 results in mathematics in nine world countries, finding out which are student and school levels characteristics related to students' performances. Based on the fact that education systems are different across countries, the main methodological issue is to use flexible methods that do not force any functional relationships between the variables. We therefore apply tree-based methods in a two-stage procedure: in the first stage, random effect regression trees are used in order to relate student performances to students' characteristics and to estimate school-value added; while in the second stage, school value-added is related to school level characteristics by means of regression trees and boosting. Results show that three-based methods well fit the problem, being able to explain a good part of variability and identifying different significant features across countries.

**Abstract** *L'obiettivo di questo lavoro è analizzare e paragonare i risultati del test PISA 2015 in matematica in nove paesi del mondo, individuando quali sono le caratteristiche a livello studente e scuola maggiormente legate alle performance degli studenti. Viste le differenze nei vari sistemi scolastici del mondo, l'obiettivo metodologico dell'analisi è trovare un metodo abbastanza flessibile, da non forzare nessuna relazione tra le variabili. Appliciamo quindi metodi basati sugli alberi di regressione in una procedura a due stadi: nel primo stadio, applichiamo alberi di regressione multilivello per identificare le variabili a livello studente legate ai risultati degli studenti e per stimare l'effetto scuola; nel secondo, identifichiamo le variabili a livello scuola legate all'effetto scuola, usando alberi di regressione e boosting. I risultati mostrano che tecniche basate sugli alberi sono adatte ad analizzare questo tipo di dati e rivelano come le caratteristiche legate alle performance degli studenti siano diverse nei vari paesi.*

---

Chiara Masci  
Politecnico di Milano, via Bonardi 9, Milano 20133, e-mail: chiara.masci@polimi.it

Geraint Johnes  
Lancaster University, Lancaster, LA14YX e-mail: g.johnes@lancaster.ac.uk

**Key words:** Random effect regression trees, boosting, student achievements, school value-added.

## 1 Introduction

The educational system is a complex and unknown process that varies across and within countries. The determinants that play a role in this process are various and arising from different levels of the scholastic system and of students life. Indeed, the learning process of students is not only influenced by students' own characteristics, but also by the family, the peers, the context in which they live, their class/school-mates, and by the characteristics of the school that they are attending. When trying to analyze the educational process, it is worth but difficult to take into account all these aspects, especially, because their marginal impact on student achievements is unknown and the interactions between them further complicates the process itself.

Our aim is to identify which are the student and school levels characteristics that are related to student achievements, to investigate their impacts on the outcome and how these impacts interact among them, within nine world countries<sup>1</sup>. In particular, our research questions are:

- Which student level characteristics are related to student achievements?
- How much of the total variability between student achievements can be explained by the difference between schools and how can we estimate the school value-added?
- Which school level characteristics are related to school value-added and in which way?
- How the important variables interact among them in influencing the outcome variable?
- How these relationships vary across countries?

In order to address these issues, we develop a two stage-analysis: (i) in the first stage, we apply random effects tree-based estimation methods, called RE-EM tree (see *Sela and Simonoff (2012)*) in which we consider students (level 1) nested within schools (level 2) - by means of this model we can both analyze which are the student level variables that are related to student achievements and estimate the school value-added -; (ii) in the second stage, we apply regression trees (see *Gareth et al. (2013)*) and boosting (see *Elith et al. (2013)*) to identify which are the school level characteristics related to school value-added (estimated at stage (i)), how they are related to the outcome and how they interact among eachother.

---

<sup>1</sup> The 9 selected countries are: Australia, Canada, France, Germany, Italy, Japan, Spain, UK, USA.

## 2 The Dataset

The Programme for International Student Assessment (PISA) is a triennial international survey (started in 2000) which aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students (see *Pena-Lopez (2016)*). Students are assessed in various disciplines and a set of student and school levels characteristics are available, thanks to questionnaires that students and school principals had to fill out. In our analysis, we use PISA data of 2015 of nine world countries: Australia, Canada, France, Germany, Italy, Japan, Spain, UK and USA. Regarding the student level, we consider information about his/her gender, immigrant status, socio-economical index, the time he/she spends studying, some built indicator about his/her approach to the school and to the subject (anxiety, effort, collaboration, perception of school climate..) and information about his/her family (home resources, support..). While at school level, we consider information about the school body composition (school size, percentages of disadvantaged students..), school resources (computers, number of teachers, materials..), “management” (principal characteristics, funds..), school climate (students truancy, teacher absenteeism..) and participation of students’ families.

## 3 Methodology

There are three main points to be taken into account when modeling this kind of educational data: data levels of grouping (students within schools), realistic assumptions on the relationships across variables and interactions.

This is why we decide to move to a machine learning approach, applying a two-stage procedure. In the first stage, we apply a (two-level) random effect regression tree (RE-EM tree), with random intercept. The response variable is the student (level 1) PISA test score in maths, that is regressed against a set of student level characteristics (fixed effects) and where students are nested within schools (level 2). By means of this model, we can estimate the fixed effects of student level predictors on the outcome and we can also estimate the school value-added. In the second stage, we regress the estimated school value-added against a set of school level characteristics, by means of regression trees and boosting.

### 3.1 First stage: RE-EM trees

RE-EM trees work basically as random effects linear models (see *Snijders (2011)*), but relaxing the linearity assumptions of the fixed covariates with the response. Instead, a regression tree is built for the fixed part. If we consider students (level 1) nested within schools (level 2), the model takes the form:

$$y_{ij} = f(x_{ij1}, \dots, x_{ijp}) + b_j + \varepsilon_{ij} \quad (1)$$

with

$$b \sim N(0, \sigma_b^2), \quad (2)$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2) \quad (3)$$

where  $f(x_{ij1}, \dots, x_{ijp})$  involves a partition of the predictor space and

$y_{ij}$  is the maths PISA test score of student  $i$  within school  $j$ ;

$x_{ij1}, \dots, x_{ijp}$  are the  $p$ -predictors at student level;

$b_j$  is the random effect of school  $j$ ;

$\varepsilon_{ij}$  is the error.

One of the advantages of multilevel models is that we can compute the Percentage of Variability explained by Random Effects (PVRE):

$$PVRE = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\varepsilon^2}. \quad (4)$$

### 3.2 Second stage: Boosting

Regression trees have a series of advantages: they do not force any functional relationship between the answer variable and the covariates; they can be displayed graphically and are easily interpretable; they can handle qualitative predictors and they allow interactions between the variables. Nevertheless, they suffer from high variance and they are very sensitive to outliers. For this reasons, there are methods that reduce variance and increase predictive power, like *Bagging*, *Random Forest* and *Boosting* (see *Gareth et al. (2013)*).

*Boosting* (see *Elith et al. (2013)*) is a method for improving model accuracy, based on the idea that it is easier to find and average many rough rules of thumb, than to find a single, highly accurate prediction rule. Related techniques - including bagging, stacking and model averaging - also build, then merge results from multiple models, but boosting is unique because it is sequential: it is a forward, stagewise procedure. In boosting, models (e.g. regression trees) are fitted iteratively to the training data, using appropriate methods gradually to increase emphasis on observations modeled poorly by the existing collection of trees.

## 4 Results

Table 1 shows RE-EM trees results. For each country, we obtain the portion of explained variability ( $R^2$ ) by the RE-EM tree model, the PVRE, the tree of fixed

effects and the estimated school values-added.  $R^2$ s are relatively high in almost all the countries, suggesting that the model is able to catch a good part of variability in the data. The PVREs are quite different across countries, meaning that there are countries (e.g. France or Japan) where differences across schools are quite big, and others (e.g. Spain or Australia) where the impact of attending certain schools respect than others is small.

Country	$\sigma_\varepsilon^2$	$\sigma_b^2$	PVRE	$R^2$
Australia	0.690	0.125	15.41%	33.59%
Canada	0.724	0.143	16.49%	29.93%
France	0.464	0.419	47.47%	55.28%
Germany	0.525	0.437	45.44%	50.17%
Italy	0.568	0.395	41.04%	45.57%
Japan	0.510	0.437	46.13%	50.32%
Spain	0.706	0.068	0.08%	30.11%
UK	0.695	0.162	18.97%	32.51%
USA	0.689	0.132	16.15%	33.45%

**Table 1** RE-EM trees results in the nine selected countries.

Results of second stage regression tree boosting may be summarized, in each country, in (i) variables importance ranking (boosting gives an idea of how much each school level variable is “important” in explaining the school value-added); (ii) single and joint partial plot (partial plot gives a graphical representation of the marginal effect of each predictor on the response variable, after “averaging-out” the effects of the other predictors. Joint plots represent the joint impact of two predictors on the response); (iii) percentage of explained variability by the model.

In order to give an example, in Australia, we are able to explain about 40.2% of the total variability and the four most important variables result to be: percentage of disadvantaged students, percentage of funds given by the government, student truancy and percentage of students with special needs. Figure 1 reports the partial plot of these 4 most important variables and Figure 2 reports an example of 2 joint plots.

## 5 Conclusions

This paper analyzes PISA2015 test scores in mathematics in nine world countries, by means of a flexible way able to fit different education systems and to identify different patterns within the data. Methodology takes into account the hierarchical structure of data, it does not force any functional relationships between variables and it allows for interactions between them. Results show the high predictive power of tree-based methods in this context, identifying the most significant variables in affecting students’ performances and highlighting heterogeneities across countries.

## References

1. Sela, Rebecca J., and Jeffrey S. Simonoff: Re-em trees: a data mining approach for longitudinal and clustered data. *Machine learning*, 86(2):169-207 (2012).
2. James, Gareth, et al.: *An introduction to statistical learning*. Vol. 6. New York: springer (2013).
3. Elith, Jane, John R. Leathwick, and Trevor Hastie: A working guide to boosted regression trees. *Journal of Animal Ecology* 77.4 (2008): 802-813.
4. Snijders, Tom AB.: *Multilevel analysis*. Springer Berlin Heidelberg (2011).
5. Pea-Lpez, Ismael et al.: *Pisa 2015 results (volume i). excellence and equity in education* (2016).

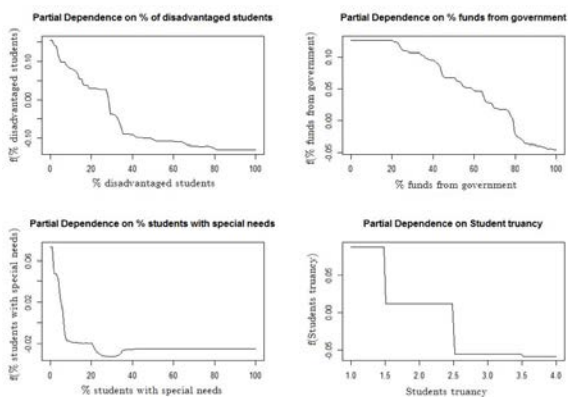


Fig. 1 Partial plots of the four most important variables in Australia.

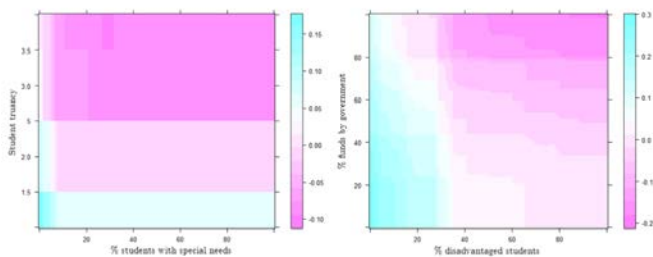


Fig. 2 Joint Partial plots in Australia. Color identifies the values of school value-added.

# Impact of the 2008 and 2012 financial crises on the unemployment rate in Italy: an interrupted time series approach

## *Impatto delle crisi finanziarie del 2008 e del 2012 sul tasso di disoccupazione in Italia: approccio di analisi basato sulle serie temporali interrotte*

Lucio Masserini and Matilde Bini

### **Abstract**

One of the most widely recognized indicators of a recession is a rising unemployment rate. In Italy, from the late nineties this indicator continuously decreased over time until 2007. The aim of this paper is to study the immediate impact and persistence of the 2008 global financial crisis and the 2012 European sovereign debt crisis on the Italian unemployment rate by using a segmented regression analysis approach of interrupted time series. Quarterly data were collected from the website of the Italian National Institute of Statistics. In particular, the impact of the financial crises was evaluated across some subpopulations of interest by stratifying unemployment rate for age groups, in order to highlight the effects on youth unemployment, gender and macro-regions. Finally, to provide a more in-depth analysis, some information on the effects of the two economic recessions was also given about the people not engaged in Education, Employment or Training.

**Abstract** *Uno degli indicatori di recessione più utilizzati è il tasso di disoccupazione. In Italia, dalla fine degli anni novanta tale indicatore è costantemente diminuito fino al 2007. Lo scopo di questo lavoro è quello di studiare*

---

<sup>1</sup> Lucio Masserini, Statistical Observatory – University of Pisa; email: [lucio.masserini@unipi.it](mailto:lucio.masserini@unipi.it)  
Matilde Bini, European University of Rome; email: [matilde.bini@uniroma2.it](mailto:matilde.bini@uniroma2.it)

*l'impatto immediato e la persistenza della crisi finanziaria globale del 2008 e la crisi del debito sovrano europeo 2012 sul tasso di disoccupazione italiano, utilizzando un'analisi di regressione di serie temporali interrotte. I dati sono stati raccolti sul sito dell'Istituto Nazionale Italiano di Statistica. In particolare, l'impatto delle crisi finanziarie è stato valutato per alcune sottopopolazioni di interesse stratificando il tasso di disoccupazione per età, al fine di evidenziare gli effetti sulla disoccupazione giovanile, per genere e per macro-regioni. Infine, per fornire una descrizione più approfondita del fenomeno, l'analisi è stata estesa anche ai giovani non occupati e non in istruzione e formazione.*

**Key words:** unemployment rate, interrupted time series analysis, segmented regression

## 1 Introduction

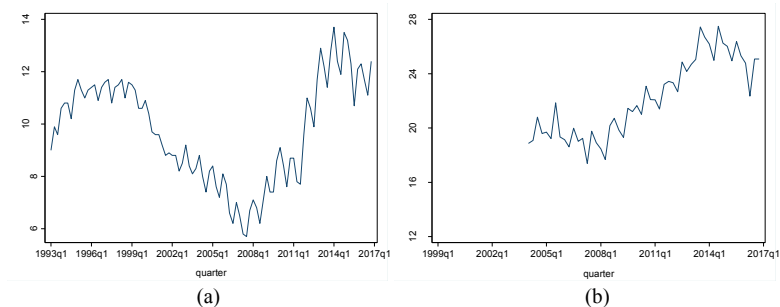
During this past decade two economic crises had a severe impact in all countries around the world. More specifically, after the economic decline observed in world markets during the late 2000s and early 2010s, which generated the Great Recession defined by the International Monetary Fund as the worst global recession since the Great Depression of the 1930s, a sovereign debt crisis faced to European countries in 2009, resulting in a second economic recession in the years after (2011–2015). These crises produced negative effects on GDP growth, on economic performance, on the labour productivity and on labour markets. The International Labour Organization (ILO, 2001) revealed that due to the global economic crisis, in 2009 about 22 million people were unemployed worldwide in particular in developed economies and in the European Union. During this period the unemployment rate continued towards a dramatically increase with high and persistent levels of unemployment in young people. In Italy the unemployment problem is worrying since it affects particular segments of the labour market, such as the younger generations and some macro regions.

The aim of this study is to assess and measure whether and how much the aforementioned financial crises have changed the level and trend in the UR and in the young people who are neither employed nor in education or training (NEET), immediately and over time, and to see if these changes are short-or long-term. Whereas UR is a widely recognized indicator of a recession, the NEET has been considered since it provides a measure of disengagement from the labour market and perhaps, more generally, quantifies also how many people are sliding towards the borders of the active society. A segmented regression approach of Interrupted Time Series (ITS) analysis is used by analysing quarterly data collected from the Italian National Institute of Statistics (ISTAT).



## 2 Data and empirical strategy

Data were collected from I.Stat, the warehouse of statistics currently produced by the Italian National Institute of Statistics (ISTAT) which provides an archive of about 1,500 time series (<http://dati.istat.it/>). Quarterly data on two different kind of indicators were downloaded from the theme ‘Labour and wages’: UR for the period 1993–2016, overall and stratified by gender, age groups and macro-regions; and the percentage of NEET for the period 2004–2016, overall and stratified by gender. Such data are derived from the official estimates obtained in the Labour force survey, carried out on a quarterly basis interviewing a sample of nearly 77,000 households representing 175,000 individuals. According to the Eurostat definition (Eurostat, 2017), UR is given by the number of people unemployed as a percentage of the labour force. The youth unemployment rate (YUR) is the number of unemployed 15–24 year-olds expressed as a percentage of the youth labour force, and the NEET refers to the percentage of people aged between 15 and 29 years who currently do not have a job, are not enrolled in training or are not classified as a student. Figure 1a illustrates the trend in the overall UR in Italy from 1993 to 2016. The choice of such a long period allows for a more accurate estimate of the secular trend, and this will be useful for the subsequent analysis. As shown, UR rises since the mid-nineties until 1998, when it reaches 11.6%. Thereafter, it steadily declines until the third quarter of 2007 (2007q3), falling to 5.7%, which represents the minimum value observed throughout the period. Starting from the fourth quarter of 2007 (2007q4), period in which the effects of the financial crisis following the bankruptcy of Lehman Brothers begin to appear, UR undergoes a first shock. As a result, it shows a trend reversal, although with some obvious fluctuations, and its value rises in the subsequent two-years period, known as Great Recession, oscillating between 7% and 9% in 2010–2011, when it reaches roughly the same level of a decade before. Afterwards, the European sovereign-debt crisis which occurred in the late 2011 (2011q4) causes a second shock, and UR increases even more dramatically up to 13.5% at the end of 2014 (2014q4), after a six quarter recession for the euro area economy. After this peak, UR seems to show a slight trend reversal, although it is perhaps still too early to consider this as a possible structural change.



**Figure 1:** Total UR (a) and percentage of NEET (b) in Italy over the observed period

On the other hand, Figure 1b illustrates the trend of the overall percentage of NEET in Italy from 2004 to 2016. For this indicator, the series is shorter because data from previous years are not available. However, the trend can still be detected and it seems, at least partly, similar to that of UR. Indeed, after a slight decrease in the period before the onset of the global financial crisis (2007q3), the percentage of NEET starts a steeply and steady growth that continues unchanged also after the occurrence of the sovereign debt crisis (2011q3). And here too, a trend reversal seems to occur starting from the end of 2014 (2014q4). In the light of the previous considerations, the analysis period was divided into the following four sub-periods, arising from the two successive financial shocks and by a slight trend reversal observed in the last two years: the period before the so-called 2008 global financial crisis (until 2007q3); the subsequent three-year period known as the Great Recession aftermath of the financial crisis, characterised by a general economic decline observed in world markets (2007q4–2011q3); the period following the European sovereign debt crisis, which resulted in a second economic recession (2011q4–2014q4); and finally, the last two years (2015q1–2016q4), during which it seems to glimpse a slight decrease in both the UR and the percentage of NEET. Consequently, the three breaks in the series were set in 2007q4, 2011q4 and 2015q4. Moreover, as regard the UR, since the historical trend has changed substantially in the late nineties (see Figure 1a), data prior to 1999 were removed from the analysis in order to obtain a more accurate estimate of the underlying trend before the first interruption of the series. Therefore, the analysis period is limited to the years 1999q1–2016q4 for the UR and to 2004q1–2016q4 for the percentage of NEET. The interruptions allow to highlight the severity of the two financial crises, respectively, and continuation of their effects in the subsequent years of recession, as shown by the sharp change observed in UR at the beginning of each period and the successive trend.

### 3 Interrupted time series analysis

In this study, a segmented regression approach of interrupted time series (ITS) analysis was carried out in order to assess and measure, in statistical terms, whether and how much the two financial crises have changed the level and trend in the outcome variables, immediately and over time, and to see if these changes are short- or long-term (Wegner, Soumerai and Zhang, 2002).

ITS analysis (Shadish, Cook and Campbell, 2002) is a simple but powerful tool used in quasi-experimental designs for estimating the impact of population-level or large scale interventions on an outcome variable observed at regular intervals before and after the intervention. In such circumstances, ITS allows to examine any change on the outcome variable in the post-intervention period given the trend in the pre-intervention period (Bernal, Cummins and Gasparrini, 2016). In this respect, the underlying secular trend in the outcome before the intervention is determined and used to estimate the counterfactual scenario, which represents what would have happened if the intervention had not taken place and serves as the basis for

comparison. For the purposes of our study the interventions are given by two unplanned and real-world events, the aforementioned and well recognized financial crises. In segmented regression of ITS, each sub-period of the series is allowed to exhibit its own level and trend, which can be represented by the intercept and slope of a regression model, respectively. The intercept indicates the value of the series at the beginning of an observation sub-period; and the slope is the rate of change during a segment (or sub-period). Therefore, by following this approach it is possible to compare the pre-crisis level and trend with their post-crisis counterpart, in order to estimate the magnitude and statistical significance of any differences.

The ITS regression model with a single group under study (here, the Italian population), two interventions, which in this study are given by the two economic recessions in 2007q4 and 2011q4, and a possible UR trend reversal in 2015q1, can be represented as it follows (Linden and Adams, 2011; Bernal, Cummins and Gasparrini, 2016):

$$y_t = \beta_0 + \beta_1 T_t + \beta_2 x_{t2007q4} + \beta_3 T_{t2007q4} x_{t2007q4} + \beta_4 x_{t2011q4} + \beta_5 T_{t2011q4} x_{t2011q4} + \beta_6 x_{t2015q1} + \beta_7 T_{t2015q1} x_{t2015q1} + \varepsilon_t.$$

In particular,  $y_t$  is the aggregated outcome variable at each equally-spaced time-points  $t$ , here represented by quarters;  $T_t$  is the time elapsed since the start of the study, where  $t$  varies between 1999q1 to 2016q4 for UR and between 2004q1 to 2016q4 for NEET, respectively;  $x_{t2007q4}$  is a dummy variable indicating the onset of the global financial crisis in fourth quarter of 2007, coded as 0 (pre-crisis period) and 1 (post-crisis period);  $T_{t2007q4} x_{t2007q4}$  is the interaction term between time and the 2007q4 global financial crisis;  $x_{t2011q4}$  is a dummy variable indicating the onset of the 2011q4 European sovereign debt crisis, coded as 0 (pre-crisis period) and 1 (post-crisis period); and  $T_{t2011q4} x_{t2011q4}$  is the interaction term between time and 2011q4 European sovereign debt crisis. Finally,  $x_{t2015q1}$  is a dummy variable indicating the time in which a trend reversal might have occurred, coded as 0 (before the trend reversal) and 1 (after the trend reversal); and  $T_{t2015q1} x_{t2015q1}$  is the usual interaction term. Accordingly,  $\beta_0$  is the intercept and represents the starting level of the outcome variable at  $T = 1999q1$  for UR and  $T = 2004q1$  for NEET, respectively;  $\beta_1$  is the slope and represents the trajectory (or secular trend) of the outcome variable until the 2007q4 global financial crisis;  $\beta_2$  is the level change that occurs immediately following the 2007q4 global financial crisis (compared to the counterfactual);  $\beta_3$  is the difference between the slope pre and post the global financial crisis;  $\beta_4$  is the level change that occurs immediately following the 2011q4 European sovereign debt crisis;  $\beta_5$  is the difference between the slope pre and post the European sovereign debt crisis;  $\beta_6$  is the level change that occurs immediately following the 2014q4 (compared to the counterfactual);  $\beta_7$  is the difference between the slope pre and post the trend reversal; and  $\varepsilon_t$  represents the random error term which is assumed to follow a first auto-regressive (AR1) process. The regression coefficients are estimated by using Ordinary least-squares (OLS) method with the Newey-West (1987) standard errors.



**Table 1:** Estimates of the impact of the 2007q4 and 2011q4 financial crises on the UR in Italy

...	Base rate (1999)	Trend 1999q1- 2007q3	Rate change 2007q4	Trend change 2007q4	Rate change 2011q4	Trend change 2011q4	Rate change 2015q1	Trend change 2015q1
Overall	11.020***	-0.138***	0.788**	0.255***	1.583***	0.137	-0.492	-0.425***
Males	8.352***	-0.095***	0.602*	0.250***	1.459***	0.112	-1.303	-0.428***
Females	14.989***	-0.205***	1.061***	0.267***	1.653***	0.190***	-1.638**	-0.280**
15-24	26.909***	-0.183***	0.970	0.716***	3.696**	0.345	-3.350	-1.611***
25-34	11.142***	-0.068***	0.127	0.290***	1.566**	0.246*	-1.873	-0.588***
35-44	7.863***	-0.095***	0.997***	0.177***	1.245**	0.166***	-1.333***	-0.245**
45-54	6.505***	-0.099***	0.725***	0.196***	1.111***	0.109*	-0.748	-0.296***
55-64	8.076***	-0.166***	0.847***	0.219***	1.350***	-0.029	-0.167	-0.029
North	5.927***	-0.067***	0.826**	0.222***	0.888**	0.014	-0.895	-0.307***
Northwest	4.640***	-0.038***	0.244	0.171***	0.842	0.011	-0.637	-0.306***
Center	8.736***	-0.098***	1.061***	0.193***	1.319**	0.140**	-1.022	-0.356**
South	20.683***	-0.269***	0.992**	0.371**	2.634***	0.360***	-2.445**	-0.478***

**Table 2:** Estimates of the impact of the 2007q4 and 2011q4 financial crises on the percentage of NEET in Italy

...	Base rate (2004)	Trend 2004q1- 2007q3	Rate change 2007q4	Trend change 2007q4	Rate change 2011q4	Trend change 2011q4	Rate change 2015q1	Trend change 2015q1
Overall	19.973***	-0.061	-0.358	0.352***	0.006	0.032	-1.673**	-0.542***
Males	15.148***	0.013	-0.410	0.361***	0.309	0.004	-1.507**	-0.744***
Females	24.713***	-0.127*	-0.346	0.332***	-0.302	0.062	-1.841**	-0.331**
Northwest	12.639***	-0.078	0.607	0.400***	-1.161	0.067	-1.734*	-0.634***
North East	10.521***	-0.008	0.521	0.377***	0.113	-0.144	-0.791	-0.653***
Center	15.479***	-0.089***	-1.243	0.477***	0.157	-0.055	-1.808**	-0.538***
South	29.884***	-0.074	0.385	0.297***	0.549	0.132**	-2.001**	-0.445**

\* p < 0.10; \*\* p < 0.05; \*\*\* p < 0.01

After this second financial shock, the UR seems to further accelerate its increase only in some sub-groups while at national level no significant trend difference was observed. In particular, such acceleration was particularly higher for the South macro-regions (+0.360;  $p < 0.001$ ), age group 25–34 (+0.246;  $p < 0.05$ ) and females (+0.190;  $p < 0.001$ ) while no significant further rate increase was detected for YUR. However, it should be emphasized here that this further increase, although lower than the one highlighted during the Great Recession, where present has to be added to that already existing, thus making particularly critical the situation. As regards the percentage of NEET, a considerable heterogeneity was found in the 2004 base rate, which was 19.973 at national level. Its value was higher for the South macro-regions (29.884) but lower for the North East (10.521) and North West (12.639) macro-regions; moreover, it was higher for females (24.713) than males (15.148). On the other hand, its trend prior to the 2008 global financial crisis (2004q1–2007q3) was basically constant at national level, with the only exception for the macro-regions of Center, which showed a slightly descending trajectory (-0.089;  $p < 0.001$ ). The onset of the of the global financial crisis (2007q4) did not cause an immediate impact on the percentage of NEET, overall and in any of the considered sub-groups. However, a significant trend change was found both at national level (+0.352;  $p < 0.001$ ) and for all the analysed sub-groups; such change was particularly higher only for the macro-regions of Center (+0.477;  $p < 0.001$ ). The European sovereign debt crisis (2011q4) does not seem to alter this situation, neither for the rate change nor for the trend change. Therefore, this means that after this second financial crisis the rise of the percentage of NEET remains steady and equal to the previous period without showing any jump.

## References

1. Cumby, R. E., and Huizinga, J.: Testing the autocorrelation structure of disturbances in ordinary least squares and instrumental variables regressions. *Econometrica*, 60, 185–195 (1992).
2. Linden, A., and Adams, J. L.: Applying a propensity-score based weighting model to interrupted time series data: Improving causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, 17, 1231–1238 (2011).
3. Lopez Bernal, J., Cummins, S., and Gasparrini, A.: Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int. J. Epidemiol.*, 1–8 (2016).
4. Newey, W. K., and West, K. D.: A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703–708 (1987).
5. Shadish, W. R., Cook, T. D., and Campbell, D. T.: *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin (2002).
6. Wagner, A. K., Soumerai, S. B., Zhang, F., and Ross-Degnan, D.: Segmented regression analysis of interrupted time series studies in medication use research. *J. Clin. Pharm. Ther.*, 27(4), 299–309 (2002).
7. Eurostat: *Glossary*. Luxembourg: Eurostat (2017).
8. ILO: *World of work report 2011: Making markets work for jobs*. Geneva: International Labour Office (2011).

# An R Package for Cluster-Weighted Models

## *Un Pacchetto R per i Modelli Cluster-Weighted*

Angelo Mazza and Antonio Punzo and Salvatore Ingrassia

**Abstract** Cluster-weighted models (CWMs) are mixtures of regression models with random covariates. However, besides having recently become rather popular in statistics and data mining, there is still a lack of support for CWMs within the most popular statistical suites. In this paper, we introduce **flexCWM**, an R package specifically conceived for fitting CWMs. The package supports modeling the conditioned response variable by means of the most common distributions of the exponential family and by the  $t$  distribution. Covariates are allowed to be of a mixed-type and parsimonious modeling of multivariate normal covariates, based on the eigenvalue decomposition of the component covariance matrices, is supported. Furthermore, either the response or the covariates distributions can be omitted, yielding to mixtures of distributions and mixtures of regression models with fixed covariates, respectively. The expectation-maximization (EM) algorithm is used to obtain maximum-likelihood estimates of the parameters and likelihood-based information criteria are adopted to select the number of groups and/or the parsimonious model. For the component regression coefficients, standard errors and significance tests are also provided. Parallel computation can be used on multicore PCs and computer clusters, when several models have to be fitted.

**Abstract** *I modelli cluster-weighted (CWMs) sono misture di regressioni con covariate random divenuti piuttosto popolari negli ultimi anni. Nonostante*

---

Angelo Mazza

Department of Economics and Business, University of Catania e-mail: a.mazza@unict.it

Antonio Punzo

Department of Economics and Business, University of Catania e-mail: antonio.punzo@unict.it

Salvatore Ingrassia

Department of Economics and Business, University of Catania e-mail: s.ingrassia@unict.it

*ciò, i software statistici più comuni non offrono supporto per tali modelli. Per ridurre tale gap, in questo lavoro introduciamo il pacchetto **R** denominato **flexCWM** che permette di fittare un'ampia gamma di modelli cluster-weighted. In particolare, il pacchetto supporta le più comuni distribuzioni della famiglia esponenziale, nonché la distribuzione  $t$ , per quanto riguarda la variabile risposta. Le covariate possono essere di tipo misto; per quelle distribuite secondo una normale multivariata, è possibile considerare CWMs parsimoniosi attraverso una ben nota scomposizione spettrale delle matrici di covarianze. Inoltre, sia la variabile risposta che la distribuzione delle covariate possono essere omesse andando a definire, rispettivamente, misture di distribuzioni e misture di regressioni con covariate fisse. L'algoritmo EM è utilizzato per ottenere le stime di massima verosimiglianza dei parametri, mentre criteri di scelta del modello basati sulla verosimiglianza sono utilizzati per scegliere il numero di componenti della mistura e/o la configurazione parsimoniosa ottimale. Per quanto riguarda le stime dei coefficienti di regressione, vengono calcolati gli standard errors e i comuni test di significatività. Infine, il pacchetto permette di fittare più modelli simultaneamente utilizzando parallelizzazione dei processi.*

**Key words:** cluster-weighted models, EM algorithm, mixture models, model-based clustering, random covariates

## 1 Introduction

When data at hand are composed by a response variable  $Y$  and by a set of  $d$  covariates  $\mathbf{X}$ , say  $(\mathbf{X}, Y)$ , and there is a latent source of heterogeneity, mixtures of regression models with fixed covariates (see, e.g., DeSarbo and Cron, 1988 and Frühwirth-Schnatter, 2006) constitute a reference framework of analysis. However, by assuming fixed covariates, modeling for  $\mathbf{X}$  is not considered; furthermore, the assignment of the data points  $(\mathbf{x}, y)$  to the clusters is required to be independent from the covariates distribution, as noted by Hennig (2000). This *assignment independence* assumption is generally not true in observational studies, and makes mixtures of regression models with fixed covariates inadequate in many real data applications. Mixtures of regression models with random covariates overcome this problem by allowing for *assignment dependence*: the component distributions for  $\mathbf{X}$  can also be distinct and they can affect the assignment of the data points to the clusters. Therefore, they are often to be preferred in real data analyses (Hennig, 2000). For a comparison between the two approaches, see also Ingrassia *et al.* (2012) and Ingrassia and Punzo (2016).

A member of the class of mixtures of regression models with random covariates is the cluster-weighted model (CWM; Gershensfeld, 1997). The CWM assumes a (parametric) functional relation for the local expectation of



$Y|\mathbf{X} = \mathbf{x}$ , and factorizes the local joint distribution  $p(\mathbf{x}, y)$  into the product between the conditional distribution of  $Y|\mathbf{x}$  and the marginal distribution of  $\mathbf{X}$ . Some recent developments in CWMs can be found in Subedi *et al.* (2013, 2015), Punzo (2014), Ingrassia *et al.* (2014, 2015), Berta *et al.* (2016), Punzo and Ingrassia (2016), Punzo and McNicholas (2017), and Dang *et al.* (2017).

## 2 The proposal

In this contribution, we introduce the R (R Core Team, 2013) package **flexCWM**, available from CRAN at <http://cran.r-project.org/web/packages/flexCWM/index.html>, specifically conceived for fitting CWMs. The package supports modeling of the conditioned response variable by means of the most common distributions of the exponential family and by the  $t$  distribution. Covariates may be of mixed-type; supported distributions are multivariate Gaussian, multinomial, binomial, and Poisson. Following Banfield and Raftery (1993) and Celeux and Govaert (1995), parsimonious modeling for multivariate normal covariates, based on the eigenvalue decomposition of the component covariance matrix, is supported (see Punzo and Ingrassia, 2015). The expectation-maximization (EM) algorithm is used to obtain maximum-likelihood estimates of the parameters and several likelihood-based information criteria are adopted to select the number of groups and/or the parsimonious model. For the local regression coefficients, standard errors and significance tests are also provided.

## 3 Conclusions

Several CRAN packages, supporting modeling by mixtures of regressions, are available. A list of them may be found in the task view “Cluster Analysis & Finite Mixture Models” of Leisch and Grün (2012), in the section entitled “Cluster-wise Regression”. **flexmix** is one of the most widely used packages for mixtures of regression models (Leisch, 2004) and mixtures of regression models with concomitant variables (Grün and Leisch, 2008); it implements an user-extensible framework for estimation, via the EM algorithm. Other packages for mixtures of regression models, include: **fpc** for mixtures of linear regression models and fixed point clusters for linear regression (Hennig, 2013), **mixreg** for mixtures of one-variable regression models (Turner, 2011), and **mixtools**, which provides a set of functions for analyzing a variety of finite mixture models, including mixtures of regression models with fixed covariates (see Benaglia *et al.*, 2009, Section 5). Within this context, the **flexCWM** package aims at giving support for cluster-weighted modeling, providing also an alternative for estimating other classical mixture models.

## References

- Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**(3), 803–821.
- Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (2009). **mixtools**: An R package for analyzing mixture models. *Journal of Statistical Software*, **32**(6), 1–28.
- Berta, P., Ingrassia, S., Punzo, A., and Vittadini, G. (2016). Cluster-weighted multilevel models for the evaluation of hospitals. *METRON*, **74**(3), 275–292.
- Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**(5), 781–793.
- Dang, U. J., Punzo, A., McNicholas, P. D., Ingrassia, S., and Browne, R. P. (2017). Multivariate response and parsimony for Gaussian cluster-weighted models. *Journal of Classification*, **34**(1), 4–34.
- DeSarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, **5**(2), 249–282.
- Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer-Verlag, New York.
- Gershenfeld, N. (1997). Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, **808**(1), 18–24.
- Grün, B. and Leisch, F. (2008). **FlexMix** version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, **28**(4), 1–35.
- Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, **17**(2), 273–296.
- Hennig, C. (2013). **fpc**: *Flexible Procedures for Clustering*. Version 2.1-6.
- Ingrassia, S. and Punzo, A. (2016). Decision boundaries for mixtures of regressions. *Journal of the Korean Statistical Society*, **45**(2), 295–306.
- Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local statistical modeling via the cluster-weighted approach with elliptical distributions. *Journal of Classification*, **29**(3), 363–401.
- Ingrassia, S., Minotti, S. C., and Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics and Data Analysis*, **71**, 159–182.
- Ingrassia, S., Punzo, A., Vittadini, G., and Minotti, S. C. (2015). The generalized linear mixed cluster-weighted model. *Journal of Classification*, **32**(1), 85–113.
- Leisch, F. (2004). **FlexMix**: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, **11**(8), 1–18.
- Leisch, F. and Grün, B. (2012). *CRAN Task View: Cluster Analysis & Finite Mixture Models*. Version 2012-03-22.
- Punzo, A. (2014). Flexible mixture modelling with the polynomial Gaussian cluster-weighted model. *Statistical Modelling*, **14**(3), 257–291.

- Punzo, A. and Ingrassia, S. (2015). Parsimonious generalized linear Gaussian cluster-weighted models. In I. Morlini, T. Minerva, and M. Vichi, editors, *Advances in Statistical Models for Data Analysis*, Studies in Classification, Data Analysis and Knowledge Organization, pages 201–209, Switzerland. Springer International Publishing.
- Punzo, A. and Ingrassia, S. (2016). Clustering bivariate mixed-type data via the cluster-weighted model. *Computational Statistics*, **31**(3), 989–1013.
- Punzo, A. and McNicholas, P. D. (2017). Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. *Journal of Classification*, **34**(2).
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, **7**(1), 5–40.
- Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. D. (2015). Cluster-weighted  $t$ -factor analyzers for robust model-based clustering and dimension reduction. *Statistical Methods & Applications*, **24**(4), 623–649.
- Turner, R. (2011). *mixreg: Functions to Fit Mixtures of Regressions*. Version 0.0-4.



# Methods and applications for the treatment of Big Data in strategic fields

## *Metodi ed Applicazioni per il trattamento di Big Data in domini strategici*

Antonino Mazzeo and Flora Amato

**Abstract** Nowadays, in a broad range of application areas, the daily data production has reached unprecedented levels. These data origin from multiple sources, such as documental sources, social media posts, digital pictures and videos and so on. The technical and scientific issues related to the data booming have been designated as the “Big Data” challenges. To deal with big data analysis, innovative algorithms and data mining tools are needed in order to extract information and discover knowledge from the continuous and increasing data growing. In most of data mining methods, the data volume and variety directly affect computational load.

In this paper, we consider a strategic field like the e-Government one. We illustrate the strategies and the methodologies for big data processing and document management.

**Abstract** *Oggi, in una vasta gamma di domini, la produzione dei dati ha raggiunto livelli senza precedenti. Questi dati derivano da più fonti, come sorgenti documentali, messaggi di social media, immagini digitali, video e così via.*

*Le questioni tecniche e scientifiche inerenti la gestione di grosse moli di dati sono state designate come “Big Data”. Per riuscire a processare con profitto grandi moli dei dati, è necessario il ricorso ad algoritmi innovativi e strumenti di data mining*

---

<sup>1</sup> Antonino Mazzeo, Flora Amato  
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'informazione  
Università degli Studi di Napoli Federico II  
Via Claudio,21 - 80125 - Napoli  
{mazzeo,flora.amato}@unina.it

*per estrarre informazioni e scoprire le conoscenze dal crescente incremento dei dati.*

*Nella maggior parte dei metodi di data mining, il volume dei dati e la varietà influiscono direttamente sul carico computazionale.*

*In questo documento consideriamo un settore strategico come quello dell'e-government. Illustriamo le strategie e le metodologie per l'elaborazione di grandi moli di dati e la gestione dei documenti.*

**Key words:** Big Data, e-Government, Data Mining

## **1 Big Data Processing**

In many application areas the daily data production has reached unprecedented levels. According to recently published statistics, in 2012 every day 2.5 EB (Exabyte) were created, with 90% of the data created in the last two years [1].

This data originates from multiple sources: sensors used to gather climate information, social networks, digital pictures and video streaming, and so on. Moreover, the size of this data is growing exponentially due to not expensive media (smartphones and sensors), and to the introduction of big Cloud Datacentres.

The technical and scientific issues related to the data booming have been designated as the “Big Data” challenges and have been identified as highly strategic by major research agencies. Most definitions of big data refer to the so-called three V s: volume, variety and velocity, referring respectively to the size of data storage, to the variety of source and to the frequency of the data generation and delivery[2,3].

To deal with big data analysis, innovative approaches for data mining and processing are required in order to enable process optimization and enhance decision making tasks. To achieve this, an increment on computational power is needed and dedicated hardware can be adopted.

## **2 The strategic field of e-Government in Italy**

E-Government, or electronic management of public services (or e-Gov), or processes of democratic governance, concerns the reorganization of the bureaucratic processes

in both central and local Public Administrations. In this context, one of main goal of e-Gov is that of providing a strong computerized management of electronic documents in order to optimize the work of the governmental offices and offer the users (citizens and businesses) both faster and more effective services and new ways of accessing such services.

From a general point of view, the theme of e-Government can be traced back to the overlap between two worlds that are apparently different and distant from each other; in particular, it can be considered as the application of Information and Communication Technology (ICT) to problems that are typical both of the Public Administration and the legal domain.

The use of ICT in the public administrations is not new, being introduced some decades ago with a series of specific projects, which were often the evolution of pre-existent legacy applications, conceived to automate single parts of the information and bureaucratic system and devoid of a systemic and global vision[4,5].

Many initiatives, often supported by facilitated finances, were introduced in the eighties within the Community in order to deeply introduce ICT into public administrations and realize strong and flexible information systems, flexible to changing and with the objective of supporting the principal bureaucratic processes within specific domains (Ministries, Local Bodies, Regions, etc.).

In the nineties and until the beginning of the present decade, with the spread of the Internet and the related technologies, the focus has been moved towards the opening of such systems to the web, in order to carry out initiatives of e-Gov and define a first level of interconnectivity shared among the administrations belonging to different domains, principally in the national environment, but also in an international one[6,7].

Nowadays, the process of combining the effectiveness of the services and their transparency within Public Administration context, goes through a strong automation of the internal processes and in addition through the capacity of using open systems, able to cooperate at application levels, following federate models: in this way, it is possible to ensure the observance of legal and organizational binding forces established by the autonomy of the various governmental Entities and the achievement of automatic and inter-domain bureaucratic processes.

Note that such technologies are not always directly and easily suitable to the specificities of the Italian bureaucratic applications (of e-Government) because of the binding forces of the specific regulations.

Generally speaking, the strategic plans provided for by all the actions of e-Government have the aims of establishing cooperation and coordination among the

different subjects of Public Administration. In the last decade and more, Public Administration in Italy has been changing its own organizational structure to enable the development of its own information systems with respect to the new application requirements, by opening and reorganizing itself, enacting new regulations, implementing its own standards and using European and international ones, resorting to solutions that often realize real “technological leaps” in the automation solutions applied.

Why revolutionize a bureaucratic organization existing since more than a century and based on paper documents and mechanical processes? Why change?

A first simple answer is given in the following. Looking at the Italian system, the need of change is principally due to the strong necessity of a de-bureaucratization and a simplification of the processes in order to: i) provide the public and private administrative acts with transparencies; ii) to increase in the quality of the offered services; iii) to decrease the costs of the organization, thus increasing its efficiency.

Looking, instead, from a wider point of view, we conclude that there is a great need to arrange, for a national system, convenient instruments able to ensure its growing, development and competitiveness.

The system of a Nation can't compete in the International and Community environment without a modern and suitable bureaucratic system, based on the use of the new technologies, operating in the Internet, and able to grant to the administrative actions continuity, definite times, quality, safety and privacy.

It is necessary to pass from systems based on computerized procedures, which are often centralized and supporting organizations based on paper documents and manual processes, to information systems focused on processes, which are often so totally automated and completely based on electronic documents that are able to optimize and rationalize the use of the human resources involved.

The incentives to change are above all represented by the spread usage of electronic documents and the related processes of dematerialization, by the implementation in full cooperation and interoperability of inter-intra domain processes, by the availability of qualifying and low-cost technology; by the evolution of the communication networks both in terms of available band and capillarity, by the safety of the various levels of the system, by effective systems of access control and profiling of the users.



The main instruments achieved, but still in evolution, concern electronic signature for documents legal validity, temporal mark-up for providing temporal evidence, digital protocol, long term preservation of electronic documents according to the regulations, the service of certified electronic mails to give evidence to the posting and receipt of documents.

In Italy the CNIPA has regulated a model of reference for the interoperability and the applicatory cooperation for the Public Administration named “Architecture of the Public System of Connectivity and Cooperation (PSC)”; the Public System of Cooperation (PSCoop) is a set of technological standards and infrastructural services whose objective is enabling the interoperability and the cooperation of the information systems for the fulfillment of administrative actions; the services offered aim at creating a groundwork to which all the Regions can connect in order to use and distribute services through standard protocols, with rules of safety and access that are shared and with a prearranged and monitored quality of the service.

Many Regions and Local Bodies have been equipping themselves to take advantage of the offered services and many initiatives promoted by the Ministry of Innovation are leading to the sharing of the models and the solutions adopted in order to achieve in a short-term period a real solution of interoperability.

### **3 Documents and Data Processing**

Note that all the e-Gov applications so far described have dematerialization activities as a common and fundamental factor: information, previously stored using graphic marks on material (paper) supports, is made immaterial using a codified electronic representation, and can be nowadays stored on several digital supports such as memories, magnetic or optical disks, tapes or other mature technologies nowadays in use.

Dematerialization is not only a normative and technological challenge but also an organizational matter involving various human resources. The transformation of a bureaucratic organization based on paper into one based on electronic documents is not easily achievable according to general models that are exportable among the organizations themselves.

So far, we have described the main characteristic of the e-Gov system, in particular, we note that e-Gov processes are usually characterized by a huge quantity of paper documents that need to be properly managed, stored and distributed. In order to reduce the huge amount of hard papers for optimizing information communication in

terms of consumed time and resources, it is widely agreed that a semantic-based dematerialization process will greatly enhance e- Government systems and application procedures.

The dematerialization process implies the application of syntactic-semantic methodologies in order to automatically transform the unstructured or sometimes semi-structured document into a formally structured, machine readable records.

The core aspect related to a novel and efficient dematerialization process is the idea standing beyond the common document concept, that can be defined as the representation of acts, facts and figures directly made or by means of electronic processing, and stored on a intelligible support. In other words, a document consists of objects such as text, images, drawings, structured data, operational codes, programs and movies, that, according to their relative position on the support, determine the shape and, consequently the structure of the document itself through the relationships between them. During the various and different e- Government processing phases, that are really different from an application domain to another, a document is processed and eventually stored on various kinds of media, properly defined in order to archive and preserve papers, photographic films and microfilms, VHS cassettes, Magnetic Tapes, DVD disks, and more.

## References

1. F. Amato, A. Mazzeo, A. Penta, A. Picariello, "Knowledge Representation and Management for E-Government Documents", Book Chapter of E-Government Ict Professionalism and Competences Service Science, pp.31-40, Springer Boston, 2008.
2. W. I. Grosky (1997), "Managing Multimedia Information in Database Systems", *Comm. Of ACM*, vol.40, n.12.
3. D. A. Adjero, and K. C. Nwosu (1997), "Multimedia Database Management - Requirements and Issues", *IEEE Transaction Multimedia*, pp. 24-33, July-September.
4. M. S. Lew, N. Sebe, D. Djeraba, and J. Rain, (2006) "Content-based multimedia information retrieval: State of the art and challenges", *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, n.1, pp. 1-19.
5. G. Boccignone, A. Chianese, V. Moscato, and A. Picariello, (2008) "Context-sensitive queries for imageretrieval in digital libraries.", *Journal of Intelligent Information Systems*, vol. 31, Issue 1, pp. 53-84.
6. F. Amato, A. Mazzeo, A. Penta, A. Picariello, "An information system for the extraction of relevant information for e-government activity", *sebd*, pp.366-373,2008 Sixteenth Italian Symposium on Advanced Database Systems, 2008.

# Happy parents' tweets

## *Twitter, genitori e felicità*

Letizia Mencarini, Viviana Patti, Mirko Lai, and Emilio Sulis

**Abstract** This article explores opinions and semantic orientation around fertility and parenthood by scrutinizing filtered Italian Twitter data. We propose a novel methodological framework relying on Natural Language Processing techniques for text analysis and social media corpora development, which is aimed at extracting sentiments from texts. A multi-layered manual annotation for exploring sentiment and attitudes to fertility and parenthood was applied to Twitter data. The corpus was analysed through sentiment and emotion lexicons in order to highlight how affective language is used in this domain. It emerges that parents express a generally positive attitude towards children, while children are more critical towards parents. The corpus constitutes a first step to improve our understanding of attitudes towards fertility and parenthood in this kind of contents.

**Abstract** *L'articolo esplora le opinioni e l'orientamento semantico intorno ai temi della fecondità e della genitorialità a partire da un'analisi di dati Twitter italiani. Viene proposto un nuovo quadro metodologico basato su tecniche di Natural Language Processing per l'analisi del testo e lo sviluppo di corpora linguistici da social media, finalizzato a estrarre sentimenti da testi. Un'annotazione manuale a più livelli è stata applicata ai dati Twitter per esplorare il sentiment e gli atteggiamenti degli utenti nei confronti della fecondità e della genitorialità. Il corpus è stato analizzato mediante risorse lessicali di emozioni e sentiment, per evidenziare come il linguaggio affettivo viene utilizzato in questo dominio. Dall'analisi emerge che i genitori esprimono un atteggiamento generalmente positivo nei confronti dei figli, mentre i figli sono più critici. Il corpus costituisce un primo passo verso la comprensione degli atteggiamenti verso fecondità e genitorialità espresse in forma spontanea in questo tipo di testi.*

**Key words:** sentiment analysis, social media, fertility, subjective well-being, linguistic corpora

---

<sup>1</sup> Letizia Mencarini, Dondena Centre for Research on Social Dynamics and Public Policy & Dept. of Management and Technology, Bocconi University, Italy; email: letizia.mencarini@unibocconi.it.

Viviana Patti, Mirko Lai, Emilio Sulis, Dipartimento di Informatica, University of Turin, Italy, email: {patti,lai,sulis}@di.unito.it.

## Introduction

The proliferation of sensors, together with the increasing popularity of social media leaves traces. This massive dissemination of information heralds a new era in social studies, bringing about new research challenges and opportunities (King, 2011; Lazer et al., 2009; Aggarwal, 2013). Several studies have exploited online social media (i.e., Facebook, Instagram, Twitter). In particular, Twitter analysis has been used to distinguish cultural traits (Golder and Macy et al., 2011), as well as a multitude of aspects, ranging from political polarization (Conover et al., 2011) and polls (O'Connor et al., 2010) to finance (Bollen et al., 2011). Tweets have also proven useful in the analysis of sentiment (Pang and Lee, 2008), as well as in distinguishing emotions (Mohammad et al., 2013) or different kinds of irony (Sulis et al., 2016; Hernandez-Farias et al., 2016). These kinds of digital traces have already been used to study human behaviour. For example, web searches have been used to predict the spread of infectious diseases (Ginsberg et al., 2010); email has been used to track migration (Zagheni and Weber, 2012), and mobile phones for daily life patterns (Gonzalez et al., 2008), as well as for economic development (Eagle et al., 2010). We, instead, focus here on the nexus between fertility and subjective wellbeing (SWB) by using filtered Twitter data in Italian. In particular, we investigate opinions and semantic orientation for fertility and parenthood.

There has been a recent increase in studies on subjective wellbeing and fertility (Clark et al. 2008; Kohler et al. 2005; Myrskylä & Margolis 2014). While these studies provide important information on the dynamics that link subjective wellbeing and childbearing and childrearing, they can only provide limited insights into the substantive role SWB plays in terms of individual fertility behaviour. Therefore, it can be difficult to explain fertility change without greater insight into the nature of SWB, and how it is discussed in relation to fertility. In this context, we want to understand whether social media content, and in particular Twitter data, can be exploited for investigating the opinions and semantic orientation around fertility and parenthood. This approach may provide new insights into the SWB-fertility nexus.

Using Twitter data, SWB can be read indirectly. In particular, we propose a novel methodological framework relying on Natural Language Processing (NLP) techniques for text analysis and social media corpora development, which is aimed at extracting sentiments or moods, which in turn can be used to construct indirect SWB measures. This is, of course, different from survey questionnaires, where respondents typically report their wellbeing on a grading scale; and where skewed distribution is the norm, with few people reporting very low levels of SWB. With Twitter individuals' opinions are posted spontaneously and often as a reaction to some emotionally-driven observation. Moreover, using Twitter we can incorporate, into our analysis, additional measures of attitudes towards children and parenthood. This offers wider geographical coverage than is found in normal survey information. As a reference dataset, we adopted all the tweets posted in Italian in 2014 from the TWITA collection (Basile and Nissim, 2013). A multi-step methodology was established in order to filter and select the relevant tweets concerning fertility and

parenthood. Then, in order to enable a deeper and more finely-grained analysis of sentiment-related phenomena for fertility and parenthood, a multi-layered manual annotation was applied to a random sample of the selected data. Here sentiment and irony on parenthood-related topics were annotated. One of the novelties of the semantic annotation scheme we created is that it allowed us to mark up information not only for sentiment polarity, but also for the specific semantic areas/sub-topics that may be the target of sentiment in the analysis of the link between SWB, parenthood, and fertility. This is a necessary first step in enabling further analysis of this kind of content.

The corpus was also analysed with sentiment and emotion lexicons in order to highlight relationships between the use of affective language and specific sub-topics. This analysis is useful per se, but it is also functional in addressing the automatic sentiment classification task. The annotated corpus is available to the research community. Its development constitutes only a first step and is a precondition for further analysis. Further analysis would involve extracting from the corpus, which includes semantically enriched data, measures of SWB constructed in an indirect way, which might improve our understanding of attitudes to fertility and parenthood.

## **TW-SWELLFER: Dataset and Annotation Methodology**

As a reference dataset, we adopted all the tweets posted in Italian language in 2014, which were retrieved through the Twitter Streaming API and applying the Italian filter proposed within the TWITA project (Basile and Nissim, 2013). The dataset includes 259,893,081 tweets (4,766,342 geotagged). We applied a multi-step methodology in order to filter and select those relevant tweets concerning fertility and parenthood. We could not rely on the exploitation of one or few hashtags or other elements that allow identifying posts on fertility and parenthood. In fact, these topics are somehow spread in the dataset and messages may contain relevant information on such subjects even if the main topic of the post is different. We are facing a situation where, on the one hand, the set of the data that are potentially relevant for our specific analysis is wider than usual; on the other hand, it is more difficult to identify the presence of information related to the topics we are interested in. In a first step, eleven hashtags<sup>1</sup> and other nineteen keywords have been chosen for selecting tweets of interest. This list is the result of a combination of a manual content analysis and a linguistic analysis on synonyms. We obtain a total amount of 3.9 million tweets. A second filtering step consisted in removing noisy tweets from corpus. Tweets posted by companies/institutions/newspapers accounts have been deleted: they are messages not concerning individual expressions. Finally, duplicated tweets not marked as RT were deleted.

---

<sup>1</sup> #papa, #mamma, #babbo, #incinta, #primofiglio, #secondofiglio, #futuremamma, #maternita, #paternita, #allattamento, #gravidanza.

## 1.1 *Annotation scheme and annotation process*

We developed and applied to our dataset an annotation model aimed at studying two aspects: the polarity of sentiment expressed in the tweets, but also specific parenthood-related topics discussed in Twitter that are the target of the sentiment.

**Sentiment polarity.** To build our annotation model, we relied on a standard annotation scheme on sentiment polarity (POLARITY), by exploiting the same labels POS, NEG, NONE and MIXED provided the organizers of the shared task for sentiment analysis in Twitter for Italian (Basile et al., 2014). Also the presence/absence of irony has been marked in order to be able to reason on sentiment polarity also in case of use of figurative devices. In order to mark irony, we introduced two polarized ironic labels: HUMNEG, for ironic tweets with negative polarity, and HUMPOS for ironic tweets with positive polarity.

**Parenthood-related semantic areas.** A set of labels marks the specific semantic areas (or SUBTOPICS) of the tweets related to the parenthood domain. This part of the annotation scheme is very important since somehow provides us with a semantic grid in order to analyse which are the aspects of parenthood that are discussed on Twitter. We considered 7 labels, suggested by a group three experts on the subjective well-being and fertility domain, after a manual analysis of a subset of the tweets: TOBEPA - Being parents (to mark when the user generically comments about his status of parent; TOBESO - Being sons/daughters (to mark the when the user is a son/daughters that comments on the parent-son/daughters relationship; DAILYLIFE - Daily life (to mark messages commenting on recurring situation in everyday life in the relationship between parents and children); JUDGOTHERPA - Judgment over other parents behaviour (to mark comments on educations of children, e.g., comments of behaviours which does not seems to be appropriated for the parent role; FUTURE - Children' future (to mark tweets where parents do express sentiments about the future of children; BECOMPA - To become parents (to mark tweets where users speak about the prospect or fear of being parents; POL - Political side (to mark tweets talking about laws having impact on being parents.

Two additional tags (IN-TOPIC/OFFTOPIC) have been added to allow annotators to mark if the tweet is relevant. The addition of this tag was necessary because of the noise still present in the dataset. Furthermore, the manual annotation will produce also data to be used in order to create a supervised topic classifier from the whole TW-SWELLFER corpus.

A random sample of 5,566 tweets from TW-SWELLFER has been collected. On this sample we applied crowdsourcing for manual annotation via the Crowdfunder platform<sup>1</sup>. We relied on CrowdFlower controls to exclude unreliable annotators and spammers based on hidden tests created by developing a set of gold-standard test questions equipped with gold reasons. The annotator's task was, first, to mark if the post is IN- or OFF-TOPIC (or unintelligible), and then to mark for IN-TOPIC posts, on the one hand, the polarity and presence of irony, on the other hand, the subtopics. Precise guidelines were provided to the annotators.

---

<sup>1</sup> <https://www.crowdfunder.com/>

Overall, for each tweet at least three independent annotations were collected. We used majority voting to select the true label. We obtained the following results.

**In-topic vs off-topic.** Manual annotation on this aspect resulted in 2,355 in-topic tweets (42.3%) and 3,136 off-topic (56.3%); the remaining 75 tweets were discarded (cases of disagreement). Thanks to the preliminary filtering steps, the proportion of in-topic tweets is pretty high compared to common results from different Twitter based content and opinion analysis (Ceron et al., 2014).

**Polarity, irony, sub-topics (in-topic tweets).** We obtained 1,545 tweets labeled with the same tags for all the layers (POLARITY, IRONY and SUBTOPICS). We call it the TW-SWELLFER-GOLD corpus.

## Analysis

Regarding IN-TOPIC tweets (2,355 posts), the 26.4% has been labeled as positive and 22.3% as negative, giving us a guidance on what might be the general feeling in Twitter about the research topics on happiness and parenthood. The irony issue is limited to a 15.7% of all the messages and negative irony prevails (10.1% of negative ironic tweets and 5.6% of positive ironic tweets), while neutral tweets are just the 8.3%. The amount of mixed tweets is limited to 1.2% (remaining 26% are labelled as NULL because annotators didn't agree on polarity, irony and subtopics labels). Overall, it seems that positive and negative feelings towards family, parenthood and fertility appear more or less equally spread through Twitter Italy. Even if the positive posts are a little bit more than the negative ones, ironic tweets must be considered: most of them are negative ironic posts (i.e., insulting/damaging the target) balancing the slight difference between pure positive and negative tweets. Furthermore, this particular topic, combined with the nature of communication in Twitter via short direct message, discourages people to stand in the grey (neutral) area, as could happens in other cases: about the 90% of the tweets shows an explicit polarity, meaning that people take a side and express their opinions.

**Which are these opinions and about what?** Going further with the analysis and looking also at the contents, so taking into consideration the "topic specification attribute and its values (Fig. 1), the largest category refers to sons tweets (TOBESO, 40.3%), in which children are discussing and posting about being children and/or about relating themselves with parents. Parents tag (TOBEPAPA) settles on 15% and becoming tag (BECOMEPA) on 10%. Remaining categories have minor impact, all being in between 1% and 6% (e.g., JUDGOTHERPA, 6,5%; DAILYLIFE: 5,6%).

### 1.2 *Sentiment and emotion analysis*

We performed a lexical analysis on the annotated corpus which concerns different aspects of affect: sentiment and emotions. As we will see, the distribution of terms in each group of messages reveals interesting patterns.

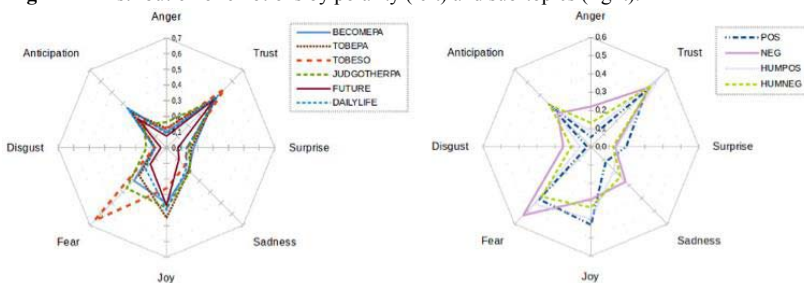
The whole polarity of messages has been computed by exploiting four existing sentiment lexical resources (Nissim and Patti, 2016) and summing positive and negative terms. A normalization is finally performed, i.e. dividing the polarity value by the number of terms in each group. In particular, the four lexica considered (LIWC, HuLiu, EmoLex and Afinn<sup>1</sup>) count more positive terms in positive messages. Similarly, negative terms are more frequent in negative messages. Ironic messages reveal a similar pattern, even if smoothed. Table 1 presents some results.

**Table 1:** Polarity values according to different lexicons in tweets tagged with different labels.

Tag	polarityLIWC	polarityHuLiu	polarityEmoLex	polarityAfinn
<b>POS</b>	1.062	0.220	0.621	3.512
<b>NEG</b>	-1.609	0.037	0.122	0.390
<b>HUMPOS</b>	0.194	0.122	0.225	2.293
<b>HUMNEG</b>	-0.336	0.078	0.637	0.610
<b>BECOMEPA</b>	1.502	0.732	0.182	-1.643
<b>TOBESO</b>	1.969	0.876	0.018	1.561
<b>FUTURE</b>	0.931	0.079	0.174	-2.058
<b>TOBEPa</b>	1.939	1.379	0.178	5.036
<b>JUDGOTHERPA</b>	1.883	0.896	0.118	-1.110

The emotion lexicon indicates also larger frequency of terms related to anger, sadness, fear and disgust in negative messages than in positive ones (Fig. 2, left). Instead, messages contain more terms related to joy, anticipation and surprise. Some suggestions can be derived in the comparison of polarity categories and the corresponding ironic ones. For instance, terms related to joy are more frequent in ironic negative messages than in negative ones. It is an insight of the polarity reversal phenomena, where a shift is produced by the adoption of a seemingly positive statement, to reflect a negative one (Sulis et al., 2016).

**Figure 1:** Distribution of emotions by polarity (left) and sub-topics (right).



The analysis of sub-topic specifications reveals a positive polarity for messages concerning TOBEPa, while BECOMEPA has a more negative polarity (Table 1). Focusing on the emotion lexicon, TOBEPa has an higher incidence of Joy words

<sup>1</sup> LIWC (<http://liwc.wpengine.com/>); Hu&Liu (Hu and Liu, 2004); AFINN (Nielsen, 2011); EmoLex (Mohammad et al., 2013).



(Fig. 2, right). Messages concerning educations of children (JUDGOTHERPA) contain a high frequency of anger and disgust term. The category TOBESO is more controversial, having the higher frequency of negative terms as fear, but also trust, as well as having the lower frequency of Joy terms. Coherently, anticipation is more frequent in the BECOMEPA group of messages. Overall, it seems that daughter and sons are more critics toward parents, whereas, parents seem to express a more positive attitude towards their daughters and sons.

## Conclusions

The contribution of this paper is the exploration of opinions and semantic orientations related to fertility and parenthood as found in about three million Italian tweets. To this end, we developed a Twitter corpus of social media contents. This corpus was, then, annotated with a novel semantic annotation scheme not only for sentiment polarity, but also for the specific semantic areas/sub-topics which were the target of sentiment in the fertility-SWB domain. The corpus was further analysed by using sentiment and emotion lexicons in order to highlight the relationships between the use of affective language and specific sub-topics in the fertility-SWB domain.

In addition, this work brings Italy into the debate on the nexus between subjective wellbeing and fertility. Italy, in fact, has been excluded from ongoing research on the topic because of a lack of suitable longitudinal data (Frey and Stutzer 2000, Kohler et al. 2005; Clark et al. 2008; Myrskylä and Margolis 2014). More must be done in order to enable a fruitful exploitation of these data, for demographic purposes. It would be particularly important to extract the information about the educational and socio-demographic traits of users in the dataset. Investigations into the relationship between social media data and official statistics is also a promising direction. By using the geocodes associated with tweets, research can link major – positive and negative – signals stemming from the sentiment analysis of the resident population in a given area (Italian provinces or NUTS-3 level) with the socio-economic characteristics of that area and the presence of childcare services. In addition, further investigations might exploit the information about the specific semantic areas considered in the present study. Aggregating geo-referenced messages into administrative areas, other interesting correlations can be detected. This analysis might shed light on the use of social media content in predicting demographic variables.

## Acknowledgments

The authors gratefully acknowledge financial support from the European Research Council under the European ERC Grant Agreement n. StG- 313617 (SWELL-FER: Subjective Well-being and Fertility, P.I. Letizia Mencarini).

## References

1. Aggarwal, Charu C. and Abdelzaher, Tarek F. Social Sensing. In *Managing and Mining Sensor Data*, Chapter 9, 237-297 (2013)
2. Basile, V, Bolioli, A., Nissim, M. Patti, V. and Rosso, P.. Overview of the Evalita 2014 SENTiment POLarity Classification Task. In Proc. of EVALITA '14, 50–57, Pisa, Italy. (2014)
3. Basile V. and Nissim, M.. Sentiment analysis on Italian tweets. In Proc. of the 4th Workshop on Comp. Approaches to Subjectivity, Sentiment and Social Media Analysis, 100–107, ACL (2013)
4. Bollen, J., Mao, H., & Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8 (2011)
5. Castells, M. *The Rise of the Network Society* (2nd ed.). Blackwell Publishers, Inc., Cambridge, MA, USA (2000)
6. Clark, R., Ogawa, N., Lee, S.-H., and Matsukura, R. Older Workers and National Productivity in Japan. *Population and Development Review* 34(Supplement): 257-274 (2008)
7. Clark, A.E., Oswald A.J. A simple statistical method for measuring how life events affect happiness. *Int J Epidemiol* 31 (6): 1139-1144. (2002). doi: 10.1093/ije/31.6.1139
8. Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. Political polarization on twitter. *ICWSM*, 133, 89-96 (2011)
9. Eagle, N., Macy, M., and Claxton, R. Network diversity and economic development. *Science*, 328(5981):1029–1031 (2010)
10. Frey, B. S and Stutzer, A. Happiness, economy and institutions. *The Economic Journal*, 110 (466), 918-938 (2000)
11. Golder, S. A. and Macy, M. W. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881. (2011)
12. Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782. (2008)
13. Hernández Fariás, D. I., Patti, V. and Rosso, P. Irony detection in Twitter: The role of affective content. *ACM Transaction of Internet Technology*, 16(3):19:1–19:24. (2016).
14. Hu, M. and Liu, B.. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04*, pages 168–177, Seattle, WA, USA. ACM (2004)
15. King, G.. Ensuring the Data-rich Future of the Social Sciences. *Science*, 331 (6018), 719-721, (2011)
16. Kohler, H. P., Behrman, J. R., and Skytthe, A. Partner + Children = Happiness? The Effects of Partnerships and Fertility on Well- Being. *Population and development review*, 31(3), 407-445. (2005)
17. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Alstynne, M. V. *Social science: Computational social science. Science*, 323(5915):721–723 (2009)
18. Mohammad, S. M. and Turney, P. D.. Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465. (2013)
19. Myrskylä, M. and Margolis, R. Happiness: Before and after the kids. *Demography*, 51(5), 1843-1866. (2014)
20. Nielsen, F. A. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the Workshop on ‘Making Sense of Microposts’: Big things come in small packages*, volume 718 of CEUR WP, pages 93–98. CEUR-WS.org. (2011).
21. Nissim, M. and Patti, V.: Semantic aspects in sentiment analysis. In: *Sentiment Analysis in Social Networks*, chap. 3, pp. 31–48. Elsevier. (2017)
22. O'Connor, B., Balasubramanyam, R., Routledge, B. R., & Smith, N. A. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129), 1-2. (2010)
23. Pang, B. and Lee, L. *Opinion Mining and Sentiment Analysis*. Found. Trends Inf. Retr. 2. (2008)
24. Sulis, E., Hernández Fariás, D. I., Rosso, P., Patti, V., and Ruffo, G. Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132 – 143. Elsevier (2016)
25. Zaghenni, E. and Weber, I. You are where you E-mail: Using E-mail Data to Estimate International Migration Rates. *Proceedings of ACM Web Science* (2012)

# Space-Time Analysis of Movements in Basketball using Sensor Data

## *Un'analisi dei movimenti spazio-temporali nella Pallacanestro con l'utilizzo di dati provenienti da sensori*

Rodolfo Metulini and Marica Manisera and Paola Zuccolotto

**Abstract** Global Positioning Systems (GPS) are nowadays intensively used in Sport Science as they permit to capture the space-time trajectories of players, with the aim to infer useful information to coaches in addition to traditional statistics. In our application to basketball, we used Cluster Analysis in order to split the match in a number of separate time-periods, each identifying homogeneous spatial relations among players in the court. Results allowed us to identify differences in spacing among players, distinguish defensive or offensive actions, analyze transition probabilities from a certain group to another one.

**Abstract** *I sistemi di posizionamento globali (GPS) sono ampiamente utilizzati in campo sportivo in quanto ci permettono di rilevare in diversi istanti temporali il posizionamento dei giocatori in campo, allo scopo di fornire indicazioni utili in aggiunta alle statistiche tradizionali. Con un'applicazione sulla pallacanestro, utilizziamo una Cluster Analysis allo scopo di suddividere la partita in gruppi omogenei in termini di relazioni spaziali tra giocatori. Identifichiamo inoltre se ciascun gruppo corrisponde ad azioni di attacco o di difesa, e stimiamo le matrici di transizione che quantificano la probabilità di passaggio da un gruppo ad un'altro.*

**Key words:** Sport Science; Big Data; Basket; GPS; Trajectories; Data Mining

---

Rodolfo Metulini

Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 50, 25122 Brescia, e-mail: rodolfo.metulini@unibs.it

Marica Manisera

Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 50, 25122 Brescia, e-mail: marica.manisera@unibs.it

Paola Zuccolotto

Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 50, 25122 Brescia, e-mail: paola.zuccolotto@unibs.it

## 1 Introduction

Studying the interaction between players in the court, in relation to team performance, is one of the most important issue in Sport Science. In recent years, thanks to the advent of Information Technology Systems (ITS), it became possible to collect a large amount of different types of spatio-temporal data, which are, basically, of two kinds. On the one hand, play-by-play data report a sequence of relevant events that occur during a match. Events can be broadly categorized as player events such as passes and shots; and technical events, for example fouls and time-outs. Carpita et al. [1, 2] used cluster analysis and principal component analysis in order to identify the drivers that affect the probability to win a football match. Social network analysis has also been used to capture the interactions between players [3]; Passos et al. [4] used centrality measures with the aim of identifying central players in water polo. On the other hand, object trajectories capture the movement of players or the ball. Trajectories are captured using optical- or device-tracking and processing systems. Optical systems use cameras, the images are then processed to compute the trajectories [5], and commercially supplied to professional teams or leagues [6, 7]. Device systems rely on devices that infer location based on Global Positioning Systems (GPS) and are attached to the players' clothing [8]. The adoption of this technology and the availability of data is driven by various factors, particularly commercial and technical. Even once trajectories data become available, explaining movement patterns remains a complex task, as the trajectory of a single player depends on a large amount of factors. The trajectory of a player depends on the trajectories of all other players in the court, both teammates and rivals. Because of these interdependencies a player action causes a reaction. A promising niche of Sport Science literature, borrowing from the concept of Physical Psychology [9], expresses players in the court as agents that face with external factors [10, 11]. In addition, typically, there are certain role definitions in a sports team that influence movement. Predefined plays are used in many team sports to achieve specific objectives; moreover, teammates who are familiar with each other's playing style may develop productive interactions that are used repeatedly. Experts want to explain why, when and how specific movement behavior is expressed because of tactical behavior and to retrieve explanations of observed cooperative movement patterns. A common method to approach with this complexity in team sport analysis consists on segmenting a match into phases, as it facilitates the retrieval of significant moments of the game. For example, Perin et al. [12] developed a system for visual exploration of phases in football.

The aim of this paper is to study the spatial pattern of the players in the court and contribute, with our results, to the literature of data-mining methods for trajectories analysis in team sports, with the final objective of suggesting new useful strategies to improve the team's performance. Using a basketball case study, and having available the spatio-temporal trajectories extracted from GPS tracking systems, we applied a cluster analysis in order to identify different game phases allowing us to characterize the spatial pattern of the players in the court. Each cluster defines a game phase, because it groups all the moments being homogenous in terms of spacings among players. First, we characterize each cluster in terms of players' position in the court.

Then, we define whether each cluster corresponds to defensive or offensive actions and compute the transition matrices in order to examine the probability of switching to another group from time  $t$  to time  $t + 1$ .

## 2 Data and Methods

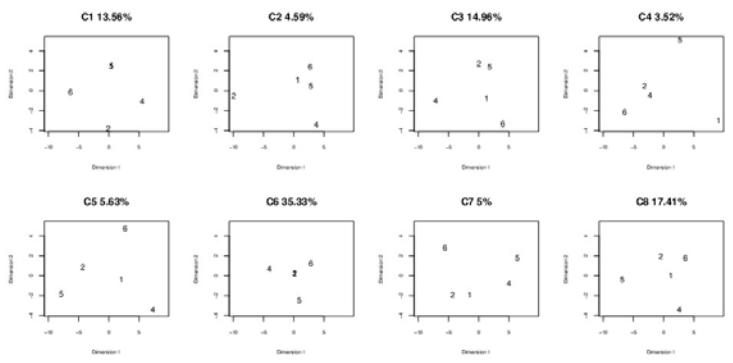
Basketball is a sport generally played by two teams of five players each on a rectangular court ( $28m \times 15m$ ). The match, according to International Basketball Federation (FIBA) rules, lasts 40 minutes, and is divided in four periods of 10 minutes each. The objective is to shoot a ball through a hoop  $46cm$  in diameter and mounted at a height of  $3.05m$  to backboards at each end of the court. The data we used in the analyses that follow refers to a friendly match played on March 22th, 2016 by a team based in the city of Pavia (Italy). This team played the 2015-2016 season in the C-gold league, the fourth league in Italy. Totally, six players took part to the friendly match. All those players wore a microchip in their clothing. The microchip collects the position (in pixels of  $1 m^2$ ) in both the  $x$ -axis and the  $y$ -axis, as well as in the  $z$ -axis (i.e. how much the player jumps). The positioning of the players has been detected at millisecond level. Considering all the six players, the system recorded a total of 133,662 space-time observations ordered in time. In average, the system collects positions about 37 times every second. Considering that six players are in the court at the same time, the position of each single player is collected, in average, every 162 milliseconds.  $x$ -axis (length) and  $y$ -axis (width) coordinates have been filtered with a Kalman approach. The Kalman filtering is an algorithm used to predict the future state of a system based on the previous ones, in order to produce more precise estimates. We cleaned the dataset by dropping the pre-match, the half-time break and the post-match periods. We completed the dataset by replacing all the missing coordinates, referred to the milliseconds that were not detected, with the value of the coordinates of the first previous instant with non-missing values. We then reshaped the dataset in order to obtain a data matrix with rows uniquely identified by the millisecond and columns devoted to the players' variables. The final dataset counts for 3,485,147 total rows. We applied a  $k$ -means Cluster Analysis in order to group a set of objects. Cluster analysis is a method of grouping a set of objects in such a way the objects in the same group (clusters) are more similar to each other than to those in other groups. In our case, the objects are represented by the time instants, expressed in milliseconds, while the similarity is expressed in terms of players' distance<sup>1</sup>. Based on the value of the between deviance (BD) / total deviance (TD) ratio and the increments of this value by increasing the number of clusters by one, we chose  $k=8$  (BD/TD=50% and relatively low increments for increasing  $k$ , for  $k \geq 8$ )

---

<sup>1</sup> In the analyses that follows, for the sake of simplicity, we only consider the period where player 3 was in the bench.

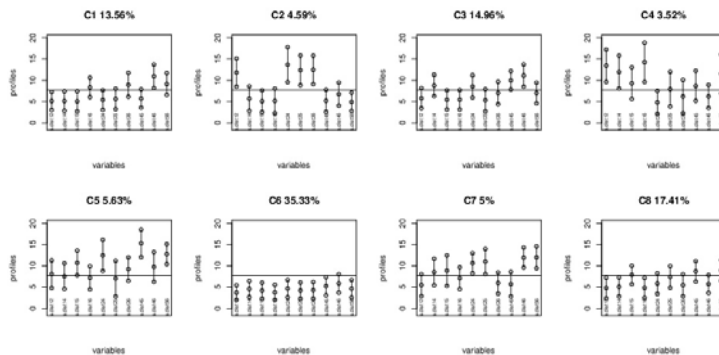
### 3 Results

The first cluster (C1) embeds 13.56% of the observations (i.e. 13.56% of the total game time). The other clusters, named C2, ..., C8, have size of 4.59%, 14.96%, 3.52%, 5.63%, 35.33%, 5.00% and 17.41% of the total sample size, respectively. We used Multidimensional Scaling (MDS) in order to plot the differences between the groups in terms of positioning in the court. With MDS algorithm we aim to place each player in  $N$ -dimensional space such that the between-player average distances are preserved as well as possible. Each player is then assigned coordinates in each of the  $N$  dimensions. We choose  $N=2$  and we draw the related scatterplots as in Figure 1. We observe strong differences between the positioning pattern among groups. In C1 and C5 players are equally spaced along the court. C6 also highlights an equally spaced structure, but the five players are more closed by. In other clusters we can see a spatial concentration: for example in C2 players 1, 5 and 6 are closed by while in C8 this is the case of players 1, 2 and 6. Figure 2 reports cluster profile plots and



**Fig. 1** Map representing, for each of the 8 clusters, the average position in the  $x - y$  axes of the five players, using MDS.

helps us to better interpret the spacing structure in Figure 1, characterizing groups in terms of average distances among players. Profile plot for C6 confirms that players are more close by, in fact, all the distances are smaller than the average distance. At the same way, C2 presents distances among players 1, 5 and 6 smaller than the average. After having defined whether each moment corresponds to an offensive or a defensive action looking to the average coordinate of the five players in the court, we also found that some clusters represent offensive actions rather than defensive. More precisely, we found that clusters C1, C2, C3 and C4 mainly correspond to offensive actions (respectively, for the 85.88%, 85.91%, 73.93% and 84.62% of the times in each cluster) and C6 strongly corresponds to defensive actions (85.07%). Figure 3 shows the transition matrix, which reports the relative frequency in which



**Fig. 2** Profile plots representing, for each of the 8 clusters, the average distance among each pair of players.

subsequent moments in time report a switch from a cluster to a different one. It emerges that for the 31,54% of the times C1 switches to a new cluster, it switches to C3, another offensive cluster. C2 switches to C3 for the 42.85% of the times. When the defensive cluster (C6) switches to a new cluster, it switches to C8 for the 56.25% of times.

NA	1	2	3	4	5	6	7	8
1	0	10.71	23.53	47.83	0	20.83	31.25	20.23
2	0.77	0	9.15	0	1.85	2.08	8.33	2.89
3	31.54	42.86	0	8.7	44.44	20.83	18.75	20.23
4	6.15	3.57	1.96	0	7.41	0	10.42	1.16
5	0.77	3.57	16.99	17.39	0	0	16.67	8.09
6	27.69	7.14	18.95	0	1.85	0	0	43.93
7	15.38	21.43	3.92	4.35	18.52	0	0	3.47
8	17.69	10.71	25.49	21.74	25.93	56.25	14.58	0

**Fig. 3** Transition matrix reporting the relative frequency subsequent moments ( $t, t + 1$ ) report a switch from a group to a different one.

## 4 Conclusions and future research

In recent years, the availability of "big data" in Sport Science increased the possibility to extract insights from the games that are useful for coaches, as they are interested to improve their team's performances. In particular, with the advent of Information Technology Systems, the availability of players' trajectories permits to analyze the space-time patterns with a variety of approaches: Metulini [13], for example, adopted motion charts as a visual tool in order to facilitate interpretation of results. Among the existing variety of methods, in this paper we used a cluster analy-

sis approach based on trajectories' data in order to identify specific pattern of movements. We segmented the game into phases of play and we characterized each phase in terms of spacing structure among players, relative distances and whether they represent an offensive or a defensive action, finding substantial differences among different phases. These results shed light on the potentiality of data-mining methods for trajectories analysis in team sports, so in future research we aim to i) extend the analysis to multiple matches, ii) match the play-by-play data with trajectories in order to extract insights on the relationship between particular spatial patterns and the team's performance.

**Acknowledgements** Research carried out in collaboration with the Big&Open Data Innovation Laboratory (BODaI-Lab), University of Brescia (project nr. 03-2016, title Big Data Analytics in Sports, [www.bodai.unibs.it/BDSports/](http://www.bodai.unibs.it/BDSports/)), granted by Fondazione Cariplo and Regione Lombardia. Authors would like to thank MYagonism (<https://www.myagonism.com/>) for having provided the data.

## References

1. Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P.: Football mining with R. *Data Mining Applications with R* (2013)
2. Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P.: Discovering the Drivers of Football Match Outcomes with Data Mining. *Quality Technology & Quantitative Management* 12.4, 561-577 (2015)
3. Wasserman, S., Katherine F.: *Social network analysis: Methods and applications*, Vol. 8. Cambridge university press (1994)
4. Passos, P., Davids, K., Araujo, D., Paz, N., Minguens, J., Mendes, J.: Networks as a novel tool for studying team ball sports as complex social systems. *Journal of Science and Medicine in Sport* 14.2, 170-176 (2011)
5. Bradley, P., O'Donoghue, P., Wooster, B., Tordoff, P.: The reliability of ProZone MatchViewer: a video-based technical performance analysis system. *International Journal of Performance Analysis in Sport* 7.3, 117-129 (2007)
6. Corporation. Tracab Player Tracking System (2015). URL <http://chyronego.com/sports-data/player-tracking>
7. AG. Impire AG (2015). URL <http://www.bundesliga-datenbank.de/en/products/>
8. Sports Ltd. Catapult USA - Wearable Technology for Elite Sports (2015). URL <http://www.catapultsports.com/>
9. Turvey, M. T., Shaw R.E.: *Toward an ecological physics and a physical psychology. The science of the mind: 2001 and beyond*, 144-169 (1995)
10. Travassos, B., Davids, K., Araujo, D., Esteves, P. T.: Performance analysis in team sports: Advances from an Ecological Dynamics approach. *International Journal of Performance Analysis in Sport* 13.1, 83-95 (2013)
11. Passos, P., Araujo, D., Volossovitch, A.: *Performance Analysis in Team Sports*. Routledge (2016)
12. Perin, C., Vuillemot, R., Fekete, J. D.: SoccerStories: A kick-off for visual soccer analysis. *IEEE transactions on visualization and computer graphics* 19.12, 2506-2515 (2013)
13. Metulini, R.: *Spatio-Temporal Movements in Team Sports: A Visualization approach using Motion Charts*. arXiv preprint arXiv:1611.09158 (2016)



# An ordinal Latent Markov model for the evaluation of health care services

## *Un modello Latent Markov ordinale per la valutazione di servizi assistenziali*

Montanari Giorgio E., Doretti Marco and Bartolucci Francesco

**Abstract** This work studies the dynamic behavior of the health status of some elderly hosted in different nursing homes. Specifically, we consider a dataset gathered from the Long Term Care Facilities (LTCF) Programme, a longitudinal study carried on in Umbria (Italy). The final goal of our analysis is to understand whether the evolution of elderly' health conditions significantly change across different nursing homes. To this end, an ordinal Latent Markov model accounting for both dropout and intermittent missing data patterns is proposed. Then, some performance measures are computed on a standardized elderly population in order to rule out the effect of patient case-mix.

**Abstract** *Questo lavoro analizza il comportamento dinamico dello stato di salute di anziani ricoverati in varie case di cura. A questo scopo si analizzano dati longitudinali provenienti dal Programma Long Term Care Facilities (LTCF), realizzato in Umbria (Italia). L'obiettivo finale è valutare se l'evoluzione delle condizioni di salute dei pazienti varia significativamente tra le case di cura. A tal fine, si propone l'utilizzo di un modello Latent Markov ordinale che consideri la presenza di dati mancanti per abbandono dello studio o altri motivi. Nell'applicazione, per ciascuna casa si propongono alcuni indicatori di risultato corretti per il case-mix.*

**Key words:** Health care services; Latent Markov models; Longitudinal data.

## 1 Introduction

Health care is one of the most relevant concerns for regional governments in Italy. In Umbria, a region of central Italy, public programs exist which aim to take care of specific population segments such as elderly or disabled people. Clearly, it is of im-

---

Montanari G.E. - University of Perugia, e-mail: [giorgio.montanari@unipg.it](mailto:giorgio.montanari@unipg.it)

Doretti M. - University of Perugia, e-mail: [marco.doretti@unipg.it](mailto:marco.doretti@unipg.it)

Bartolucci F. - University of Perugia, e-mail: [francesco.bartolucci@unipg.it](mailto:francesco.bartolucci@unipg.it)

The authors acknowledge the financial support of the Umbria region for this research.

portance for policy makers to have a deep understanding of the general health status of these people and of the effectiveness of health care provided. To this end, we focus on the *Long Term Care Facilities* (LTCF) protocol, a program mainly addressed to elderly people hosted in regional nursing homes (NHs). Through the administration of specifically designed questionnaires, some information is periodically collected on these people to monitor their physical and psychological conditions as well as the care services provided by NHs. In this work, using data coming from the LTCF dataset, we propose an ordinal Latent Markov model accounting for both dropout and intermittent missing data patterns aimed at understanding whether the evolution of elderly' health conditions significantly change across different NHs.

## 2 The LTCF data

LTCF data are collected through a questionnaire routinely administered approximately every six months. The questionnaire is formed by several sections dealing with different aspects of the health status such as cognitive conditions, humour and behavioral disorders or problems with Activities of Daily Living (ADL). Therefore, when the entire questionnaire is considered, the underlying trait is multidimensional. However, as a first step of a more complex data analysis to be developed, in this work we consider a single section of the questionnaire, namely the ADL section. The latter includes ten ordinal items that we suppose to be the outcome of an ordinal latent trait. Indeed, they report the level of difficulty patients experience in taking simple actions like getting dressed, walking or stooping, using the bathroom or eating by themselves.

The sample we consider covers the years 2012 and 2013 and includes  $n = 1292$  patients hosted in 41 NHs whose number of patients ranges from 5 to 96. Ideally, there should be  $T = 4$  measurement occasions for each patient. However, this is not always the case as dropout due to death or discharge might occur. Furthermore, intermittent missingness is also present (missing occasions between valid occasions). As is common in longitudinal studies, we need to carefully evaluate the missingness mechanism in the context we deal with. In this work, we treat intermittent missing data as well as dropout due to discharge as missing at random [3]. This choice seems a reasonable one as the former have an unknown cause, while motivations for discharge are various. On the other hand, dropout due to death is clearly non-ignorable as death is associated to a worsening of health status. This non-ignorable mechanism must be accounted for somehow in the model.

## 3 The model

Latent Markov (LM) models are of use when some categorical outcome variables are measured at a number of time occasions. These outcomes (i.e., questionnaire

items) are assumed to be probabilistically influenced by an unobserved process (i.e., health status), which is modeled like a first order discrete-time Markovian process with a finite number of states; see [1]. Three sets of parameters characterize this structure: conditional response probabilities (probabilities of specific outcome categories given the latent state), initial probabilities (probabilities of latent states at the first measurement occasion) and transition probabilities (probabilities of latent states at following occasions given previous latent state).

Assume we have data on  $n$  independent units, indexed by  $i = 1, \dots, n$ , and that unit  $i$  is observed at  $T_i$  measurement occasions, with  $T_i \leq T = 4$ .  $\mathbf{Y}_i^{(t)}$  denotes the response vector of unit  $i$  at occasion  $t$  ( $t = 1, \dots, T_i$ ). Such a vector includes  $J$  univariate categorical responses, that is  $\mathbf{Y}_i^{(t)} = (Y_{i1}^{(t)}, \dots, Y_{iJ}^{(t)})$ . In principle, each indicator  $Y_{ij}^{(t)}$  might have a generic number of response categories indexed from 1 to  $c_j$ . Similarly, the Markovian latent process is denoted by  $\mathbf{V}_i = (V_i^{(1)}, \dots, V_i^{(T_i)})$ , where, at each time occasion  $t$ ,  $V_i^{(t)}$  is a categorical variable with  $k$  levels. Each  $Y_{ij}^{(t)}$  is assumed to be independent of any other variable in the model conditionally on  $V_i^{(t)}$ . Therefore, the parameters of interest are the conditional response probabilities

$$\phi_{jy_j,v} = P(Y_{ij}^{(t)} = y_j | V_i^{(t)} = v), \quad j = 1, \dots, J, \quad y_j = 1, \dots, c_j, \quad v = 1, \dots, k,$$

which are constant with respect to time. As discussed in Section 2,  $V_i^{(t)}$  can be thought of as an ordinal variable. As a consequence, a global logit parametrization

$$\log \frac{\phi_{jm+1,v} + \dots + \phi_{jc_j,v}}{\phi_{j1,v} + \dots + \phi_{jm,v}} = \tau_{jm} + \delta_v, \tag{1}$$

( $j = 1, \dots, J; m = 1, \dots, c_j - 1; v = 1, \dots, k$ ) can be imposed. In Equation (1), it is assumed that  $\tau_{j1} > \dots > \tau_{jc_j-1}$  for  $j = 1, \dots, J$  and that  $\delta_1 < \dots < \delta_k$ , with  $\delta_1 = 0$ . This parametrization fixes the direction of the association between the responses and the latent variable. In so doing, label switching - a well-known problem in this class of models - is ruled out. In this case, such association is positive so that higher latent states correspond to increasing difficulties in the ADL.

We also allow initial and transition probabilities to depend on individual covariates denoted by  $\mathbf{X}_i = (\mathbf{X}_i^{(1)}, \dots, \mathbf{X}_i^{(T_i)})$ . These include personal characteristics as age and gender as well as binary indicators for NH membership. The latter allow to model the nursing home effect on the initial and transition probabilities for evaluation purposes. Specifically, we set

$$\pi_i^{(1)}(v) = P(V_i^{(1)} = v | \mathbf{X}_i^{(1)} = \mathbf{x}_i^{(1)})$$

( $v = 1, \dots, k$ ) and

$$\pi_i^{(t)}(v|\bar{v}) = P(V_i^{(t)} = v | V_i^{(t-1)} = \bar{v}, \mathbf{X}_i^{(t)} = \mathbf{x}_i^{(t)})$$

$(\bar{v}, v = 1, \dots, k; t = 2, \dots, T_i)$ . Relying on the ordinal nature of the latent process  $\mathbf{V}_i$ , the models for the conditional initial and transition probabilities can be respectively expressed by the regression equations

$$\log \frac{\pi_i^{(1)}(v+1) + \dots + \pi_i^{(1)}(k)}{\pi_i^{(1)}(1) + \dots + \pi_i^{(1)}(v)} = \xi_v + \mathbf{x}_i^{(1)} \boldsymbol{\beta} \quad (2)$$

$(v = 1, \dots, k-1)$ , and

$$\log \frac{\pi_i^{(t)}(v+1|\bar{v}) + \dots + \pi_i^{(t)}(k|\bar{v})}{\pi_i^{(t)}(1|\bar{v}) + \dots + \pi_i^{(t)}(v|\bar{v})} = \psi_{\bar{v}} + \omega_v + \mathbf{x}_i^{(t)} \boldsymbol{\gamma} \quad (3)$$

$(\bar{v} = 1, \dots, k; v = 1, \dots, k-1; t = 2, \dots, T_i)$ . In (2) and (3) a global logit parametrization like in (1) is assumed. Under this parametrization, the covariate effects, represented by the column vectors  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$ , are constant across the logit equations while the sequences of thresholds  $\xi_1 > \dots > \xi_{k-1}$  and  $\omega_1 > \dots > \omega_{k-1}$  must be decreasing to ensure that the cumulative sums of probabilities along the ordered categories of the latent variables are increasing. Finally, notice that for identification purposes  $\psi_1$  is always set to 0.

We extend the model described so far to account for the missingness mechanism occurring in our application. First, we add an extra response category  $c_j + 1$  in a way such that, in the case of death after the  $t$ -th occasion ( $t < T$ ), we complete the data by setting  $Y_{ij}^{(u)} = c_j + 1$  for  $u = t + 1, \dots, T$ . Moreover, we define an extra absorbing latent state  $k + 1$  corresponding to the death. Consequently, some related additional probabilities have to be properly constrained. Specifically, for  $j = 1, \dots, J$ ,  $v = 1, \dots, k$  and  $t = 2, \dots, T$ :

- $\pi_i^{(1)}(k+1) = 0$ : no one can be in the extra latent state at the first occasion;
- $\pi_i^{(t)}(k+1|k+1) = 1$ : no one can revert to other states from death;
- $\phi_{jc_j+1,v} = 0$ : the extra category cannot be observed if one is not dead;
- $\phi_{jc+1,k+1} = 1$ : only the extra category can be observed if one is dead.

We also deal with intermittent missing data patterns by extending the set of covariates explaining transition probabilities (i.e., Equation (3)). Specifically, we add a variable measuring the time interval - in days - from the previous occasion. Although intermittent missingness is assumed to occur at random (see Section 2), the introduction of such a covariate seems necessary to take into account the interval between non missing occasions and get correctly interpretable results. Indeed, it allows the estimation of six-month ahead transition probabilities, with six months being the time interval between measurement waves originally designed in the LTCF study. Notice that the extension we propose to deal with missing data involves only two additional free parameters to estimate: the additional threshold  $\omega_k$  and the regression parameter in  $\boldsymbol{\gamma}$  associated to the aforementioned additional covariate. The other thresholds have to be set to  $-\infty$  to satisfy the constraints above. Parameter estimates are obtained by means of the Expectation-Maximization algorithm [2], which is a standard estimation tool for LM models.

### 4 Results

An important part of the model selection process in latent variable models is the choice of the number of latent states  $k$ . A commonly adopted strategy considers both formal criteria like the Bayesian Information Criterion (BIC) [4] and interpretability of results. For this application, we have fitted models with  $k$  ranging from 2 to 10. After some comparisons, the model with  $k = 5$  has been selected. This choice represents a compromise between the BIC criterion, interpretability of the resulting latent states, avoiding latent states with a very small number of patients.

A summary of the estimated conditional response probabilities  $\hat{\phi}_{jv}$  can be provided by the standardized item score

$$\hat{s}_{jv} = \frac{1}{c_j - 1} \sum_{y_j=1}^{c_j} (y_j - 1) \hat{\phi}_{jv}, \quad v = 1, \dots, k, \quad j = 1, \dots, J.$$

Specifically,  $\hat{s}_{jv}$  indicates on a 0-1 scale the difficulty of a patient in latent state  $v$  in taking the action described by item  $j$ . Table 1 reports the standardized item scores for the five-state model as well as their averages across latent states, denoted by  $\bar{\hat{s}}_j$ . According to Table 1, the first difficulties are experienced in taking a bath or a shower, dressing the lower part of the body, and maintain personal hygiene (respectively, items 1, 4 and 2), while the last difficulties are related to bed mobility and eating (item 9 and 10).

v	item j									
	1	2	3	4	5	6	7	8	9	10
1	0.236	0.124	0.064	0.101	0.007	0.005	0.012	0.024	0.002	0.000
2	0.477	0.421	0.340	0.420	0.129	0.107	0.187	0.272	0.052	0.007
3	0.667	0.622	0.571	0.648	0.444	0.410	0.508	0.585	0.287	0.063
4	0.876	0.835	0.810	0.871	0.788	0.767	0.806	0.851	0.621	0.304
5	0.992	0.988	0.985	0.992	0.986	0.984	0.986	0.991	0.949	0.848
$\bar{\hat{s}}_j$	0.650	0.598	0.554	0.606	0.471	0.455	0.500	0.545	0.382	0.244

**Table 1** Standardized item scores and their averages

Table 2 reports the estimated initial and transition probabilities for the latent process. These probabilities are averaged across the distribution of individual covariates. Note that patients are rather uniformly distributed across latent states at first occasion.

Looking at the transition probabilities, the probability of persistence (i.e., probability of remaining in the same latent state) decreases for higher states as the probability of migrating towards worse states or the extra state  $d$ , death, increases.

As regards the effect of the NH membership on the latent process, here we focus on the NHs with the higher and lower effects on the probabilities of transition

initial probabilities					$\bar{v} \setminus v$	transition probabilities					
1	2	3	4	5		1	2	3	4	5	d
0.118	0.128	0.171	0.241	0.342	1	0.877	0.121	0.002	0.000	0.000	0.000
					2	0.049	0.686	0.251	0.013	0.001	0.000
					3	0.001	0.081	0.613	0.274	0.029	0.002
					4	0.000	0.004	0.097	0.504	0.347	0.048
					5	0.000	0.000	0.010	0.128	0.539	0.322

**Table 2** Estimated average initial and transition probabilities

towards worse states. The estimated difference between these effects in the linear predictor is equal to 2.337, with a p-value lower than  $10^{-5}$ . In Table 3 we report the corresponding estimated average six-month ahead transition probabilities computed on the same set of elderly. These are comparable transition matrices as they have been standardized over the same distribution of other covariates (age and gender) to vanish the patient case-mix effect. From Table 3 we conclude that there is a large difference between NHs in their ability to maintain the elderly in good health. Further investigations are needed to explain the reasons of these differences.

$\bar{v} \setminus v$	lower effect NH						higher effect NH					
	1	2	3	4	5	d	1	2	3	4	5	d
1	0.959	0.040	0.001	0.000	0.000	0.000	0.566	0.423	0.011	0.000	0.000	0.000
2	0.118	0.778	0.100	0.004	0.000	0.000	0.007	0.315	0.612	0.061	0.004	0.000
3	0.003	0.186	0.685	0.117	0.008	0.001	0.000	0.012	0.265	0.578	0.133	0.011
4	0.000	0.009	0.218	0.592	0.167	0.014	0.000	0.000	0.015	0.185	0.590	0.209
5	0.000	0.001	0.025	0.270	0.568	0.136	0.000	0.000	0.001	0.021	0.238	0.739

**Table 3** Estimated six-month ahead transition probabilities for higher and lower nursing home effects

## References

- [1] F. Bartolucci, A. Farcomeni, and F. Pennoni. *Latent Markov Models for Longitudinal Data*. Statistics in the Social and Behavioural Sciences. Chapman & Hall/CRC, 2013.
- [2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, pages 1–38, 1977.
- [3] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, second edition, 2002.
- [4] G. Schwarz. Estimating the dimension of a model. *Ann Stat.*, 6(2):461–464, 1978.

# New fuzzy composite indicators for dyslexia

## *Indicatori sintetici fuzzy per la diagnosi precoce della dislessia*

Isabella Morlini and Maristella Scorza

**Abstract** Composite indicators should ideally identify multidimensional concepts that cannot be captured by a single variable. In this paper, we suggest a method based on fuzzy set theory for the construction of fuzzy synthetic indexes of dyslexia, using the set of manifest variables measured by means of reading tests. A few criteria for assigning values to the membership function are discussed, as well as criteria for defining the weights of the variables. An application regarding the diagnosis of dyslexia in primary and middle school in Italy is presented. In this application, the fuzzy approach is compared with the crisp approach actually used in Italy for detecting dyslexic children in compulsory school.

**Abstract** *La diagnosi precoce della dislessia nei bambini di età scolastica è di fondamentale importanza per poter garantire una didattica mirata e prevenire possibili effetti negativi sullo sviluppo della personalità. Attualmente in Italia la diagnosi della dislessia viene effettuata analizzando la velocità o l'accuratezza della lettura in test normativi. Poiché la dislessia è un fenomeno complesso che può essere misurato solo dall'analisi congiunta dei diversi aspetti inerenti le performances di lettura e non segue la suddivisione rigida fra "patologia presente" e "patologia assente" ma può manifestarsi con diversi livelli di gravità, in questo lavoro vengono proposti nuovi indici composti fuzzy per misurare il grado della patologia nei bambini frequentanti la scuola primaria e secondaria di primo grado.*

**Key words:** composite indicator, dyslexia, fuzzy index, membership function, weighting criteria

---

<sup>1</sup> Isabella Morlini, Department of Economics "Marco Biagi", University of Modena & Reggio Emilia, Viale J. Berengario 51, 41121 Modena, Italy; email: isabella.morlini@unimore.it

Maristella Scorza, Department of Social Sciences, University of Modena & Reggio Emilia, Viale Allegri, 42.121 Reggio Emilia, Italy; email: maristella.scorza@unimore.it

## 1 Introduction

Decoding ability in primary school in Italy and in countries with transparent orthography is currently assessed with the aid of standardized tests requiring the students to read aloud a selected list of words and non-words or a text. The most widely used standardized tests in Italy have been introduced by Sartori *et al.* (2007). Recently, a new screening procedure for identifying impaired decoders in elementary grades has been proposed by Morlini *et al.* (2014). What is important in the use of tests and screening procedures is the way the results are interpreted. One of the defining characteristic of a skilled decoder is that he or she not only is able to spell written words (or non-words) accurately, but also does so rapidly and automatically. An individual who spells accurately but very slowly cannot be considered a skilled decoder. Slow rate of word reading is then characteristic of impaired decoding as well as low accuracy, especially in transparent languages (World Health Organization (2008)). In Italy, decoding ability is assessed without taking into account both aspects and an individual can be classified as impaired because he or she is able to read words (or non-words) very rapidly, even though he or she misspells a fairly large number of words (or non-words). Individuals with weak decoding skills who are able to read a large number of words, provided they are given ample time, can be erroneously classified as adequate decoders. Many authors have outlined the necessity of considering both speed and accuracy for a valid assessment of decoding skills and a new challenge in learning disability research is to develop composite indicators that incorporate measures of speed as well as of accuracy (Morlini *et al.* (2015)). Since dyslexia is a vague concept and the rigid partition between impaired and not impaired readers does not always reflect reality, fuzzy theory should be used in defining these new indicators.

In Section 2, we deal with the general problem of obtaining a synthetic fuzzy measure of a latent phenomenon like dyslexia from a set of metric variables. We present two criteria to transform the values of a variable into fuzzy numbers. In Section 3, we discuss the problem of weighting the variables and aggregating them into composite indicators. Clearly, the weights should reflect the contribution of each variable to the latent phenomenon. In Section 4, we focus on the specific application of measuring dyslexia in compulsory schools. The gradual transition from skilled to impaired readers can be captured by fuzzy indexes, as well as the risk for dyslexia. We apply the method to a sample of 3932 students attending elementary and middle schools in Italy. The fuzzy indicators of dyslexia allow us to obtain membership functions that can be compared with the result of the currently used diagnostic procedure, which strictly identifies a student as being “dyslexic” or “not dyslexic”.

## 2 The fuzzy approach

Let  $X$  be a set of elements  $x \in X$ . A fuzzy subset  $A$  of  $X$  is a set of ordered pairs:

$$[x, \mu_A(x)] \quad \forall x \in X$$



where  $\mu_A(x)$  is the membership function (*m.f.*) of  $x$  to  $A$  in the closed interval  $[0,1]$ . If  $\mu_A(x) = 0$ , then  $x$  does not belong to  $A$ , while if  $\mu_A(x) = 1$ , then  $x$  completely belongs to  $A$ . If  $0 < \mu_A(x) < 1$ , then  $x$  partially belongs to  $A$  and its membership to  $A$  increases according to the values of  $\mu_A(x)$ . Let us assume that the subset  $A$  defines the position of each element with reference to achievement of the latent concept, e.g. dyslexia. In this case,  $\mu_A(x) = 1$  identifies a situation of full achievement of the disease, whereas  $\mu_A(x) = 0$  denotes a person not sharing the disease (a very skilled decoder). Consider a set of  $n$  individuals and  $p$  metric variables  $X_s$  ( $s = 1, 2, \dots, p$ ) reflecting the latent phenomenon. In case of dyslexia, these variables are measures of reading performances like the time of reading in seconds, the number of misspelled words or the number of syllables read in a second. Without loss of generality, let us assume that each variable is positively related with that phenomenon, i.e. it satisfies the property “the larger the more impaired”. If a variable  $X_s$  shows a negative correlation (like the number of syllables read in a second) we substitute it with the simple decreasing function transformation  $f(x_{si}) = \max(x_{si}) - x_{si}$ .

In order to define the *m.f.* for each variable it is necessary to identify the extreme situations such that  $\mu_A(x) = 0$  (non membership) and  $\mu_A(x) = 1$  (full membership) and to define a criterion for assigning the *m.f.* to the intermediate values. Many criteria has been proposed in literature, especially in the field of social sciences, for measuring latent concepts like well-being, satisfaction and poverty (see e.g. Zani *et al.* (2012) and (2013)). For the specific purpose of measuring dyslexia we will consider only two specifications.

Let us assume that  $X_s$  is a metric variable. For simplicity of notation in the following we will omit index  $s$ . We choose an inferior (lower) threshold  $l$  and a superior (upper) threshold  $u$ , with  $l$  and  $u$  finite, and we define the *m.f.* as follows:

$$\begin{cases} \mu_A(x_i) = 0 & x_i \leq l \\ \mu_A(x_i) = \frac{x_i - l}{u - l} & l < x_i < u \\ \mu_A(x_i) = 1 & x_i \geq u \end{cases} \quad (1)$$

In (1) the *m.f.* is a linear function between the values of the two thresholds. Alternatively, we define the *m.f.* as  $\mu_A(x_i) = 1/(1 + d(x_i))$ , where  $d(x_i)$  is the distance between the value  $x_i$  and dyslexia, measuring the degree of impairment and indicating the level of the achievement of dyslexia. If  $d(x_i) = 0$ , there is full membership to  $A$  and  $\mu_A(x) = 1$ . If  $d(x_i) > 0$ , then  $\mu_A(x) < 1$ . In general, the relationship between physical measures and perception takes an exponential form (Zimmerman (1993), Balamoune-Lutz (2004)). If we assume that the relationship between physical measures and decoding impairment takes the same form, then the distance  $d(x_i)$  can be expressed as  $d(x_i) = e^{-a(x_i-b)}$  and the *m.f.* is then defined as:

$$\mu_A(x_i) = \frac{1}{1 + e^{-a(x_i-b)}} \quad (2)$$

The parameter  $a$  represents the extent of uncertainty and  $b$  may be viewed as the point in which the performance of the subject changes from “bad” to “pathological”.

### 3 The fuzzy composite indicators

The most general aggregation function of variables for obtaining a composite

indicator is the weighted generalized mean  $\mu_A(i) = \sum_{s=1}^p [\mu_A(x_{si})]^\alpha w_s^{1/\alpha}$ , where  $w_s > 0$

is the normalized weight that expresses the relative importance of the variable  $X_s$ , with  $\sum_{s=1}^p w_s = 1$ . For the sake of simplicity, we consider  $\alpha=1$ , that is the weighted arithmetic mean. Furthermore, we consider different weights reflecting the importance of each variable in diagnosing dyslexia. Since previous studies (Morlini *et al.* (2014, 2015)) have shown that the first component of a principal component analysis (PCA) accounts for a high percentage of the total variance of the measures of speed and accuracy in psychometric reading test, we consider weights proportional to the correlation of each variable with the first component of a PCA.

Alternatively, in order to attach to each variable a weight sensitive to the fuzzy membership, we consider the fuzzy proportion of each variable to the achievement of dyslexia  $g(X_s) = \frac{1}{n} \sum_{i=1}^n \mu(x_{is})$  and define the normalized weights as follows:

$$w_s = \ln\left(\frac{1}{g(X_s)}\right) / \sum_{s=1}^p \ln\left(\frac{1}{g(X_s)}\right). \quad (3)$$

### 4 Fuzzy indicators for dyslexia: an application

We administer the standardized tests *Batteries for the Diagnosis of Reading and Spelling Disabilities* (Sartori *et al.* (2007)) to 3932 students attending elementary (from grade II) and middle school in Lombardia and Emilia Romagna regions (Northern Italy). Table 1 reports the frequency distribution of students in each grade. In these tests the metric variables measuring performances of decoding are:

$X_1$ : time (in seconds) in reading the list of words

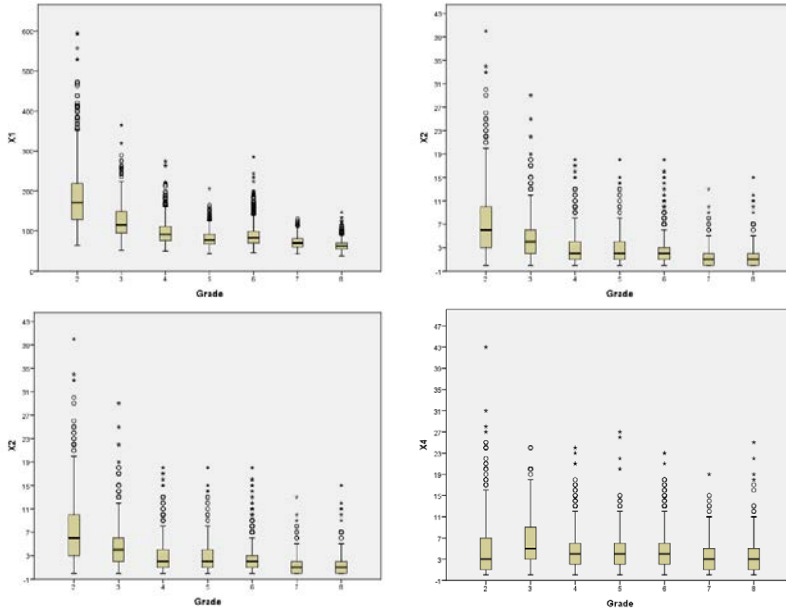
$X_2$ : number of words mispronounced in reading the list of words

$X_3$ : time (in seconds) in reading the list of non-words

$X_4$ : number of incorrect pronunciations in reading the list of non-words

**Table 1:** Frequency distributions of students in each grade

Number of students	Grade						
	Elementary school				Middle school		
	II	III	IV	V	VI	VII	VIII
	715	472	621	519	922	311	372



**Figure 1:** Distribution of each variable in each grade

Figure 1 shows the distribution of each variable in each grade. We perform a PCA on the correlation matrix. The first component accounts for 66.5% of the total variance, is highly correlated with all variables and is the only one with eigenvalue greater than one. We construct the following fuzzy indicators:

F<sub>11</sub>: using *m.f.* (1) with  $l = x_{5\%}$  (the fifth percentile) and  $u = x_{95\%}$  (the 95<sup>th</sup> percentile) in each grade and weights proportional to the factor loadings of the first PCA.

F<sub>12</sub>: using *m.f.* (1) with  $l = x_{5\%}$  and  $u = x_{95\%}$  in each grade and weights (3).

F<sub>21</sub>: using *m.f.* (2) with  $a = 0.5$  and  $b = x_{90\%}$  (the 90<sup>th</sup> percentile) in each grade and weights proportional to the factor loadings of the first PCA.

F<sub>21</sub>: using *m.f.* (2) with  $a = 0.5$  and  $b = x_{90\%}$  in each grade and weights (3).

Table 2 reports the frequency distribution of the values of the fuzzy indices. We may note that the differences in F<sub>11</sub> and F<sub>12</sub> and in F<sub>21</sub> and F<sub>22</sub> are negligible and thus the weighting system do not substantially change the fuzzy indicator. On the other hand, the choice of the membership function influence the results. Applying the diagnostic criterion actually used in Italy for which a student is classified as impaired if he or she shows a value above normative cut-off in two or more variables, 4.8% of the

students is classified as dyslexic. The fuzzy indicators give more insight into this percentage. According to *m.f.* (1), about 2% is definitely dyslexic, while should be considered at high risk for impairment the 2.9%. Another approximately 4% may be viewed as being at medium risk. According to *m.f.* (2), about 1% of the students are definitely dyslexic, while 1% is at high risk and approximately 1.6% at medium risk. We may also identify the prevalence of very skilled readers (64% according to *m.f.* (1) and 89% according to *m.f.* (2)) and the percentages of normal readers (given by the frequencies of values ranging from 0.4 to 0.7).

In conclusion, this paper presents a methodology to build fuzzy composite indicators with the aim of considering both speed and accuracy of reading in the early diagnosis of dyslexia and with the aim of going beyond the rigid unrealistic partition between “dyslexic” and “not dyslexic” student. The limit between a “bad” and a “pathological” performance in psychometric reading tests is somehow fuzzy. The application shows that the proposed indices work well in identify the level of impairment of the students and the results are in agreement with the percentages of dyslexic identified with the traditional diagnostic criterion but give more insights.

**Table 2:** Frequency distributions of the values of the fuzzy indices

	F <sub>11</sub>	F <sub>12</sub>	F <sub>21</sub>	F <sub>22</sub>
0.0 - 0.4	0.648	0.643	0.895	0.894
0.4 - 0.6	0.208	0.206	0.041	0.047
0.6 - 0.7	0.059	0.062	0.028	0.024
0.7 - 0.8	0.038	0.039	0.018	0.016
0.8 - 0.9	0.029	0.029	0.009	0.009
0.9 - 1.0	0.019	0.021	0.009	0.010
Tot	1.000	1.000	1.000	1.000

## References

1. Balamoune-Lutz, M.: On the Measurement of Human Well-Being: Fuzzy Set Theory and Sen's Capability Approach, UNU-WIDER, Helsinki (2004).
2. Morlini I., Stella G. and Scorza M.: A new procedure to measure children's reading speed and accuracy in Italian, *Dyslexia*, **20**, 1, 54–73 (2014)
3. Morlini I., Stella G. and Scorza M.: Assessing decoding ability: the role of speed and accuracy and a new composite indicator to measure decoding skill in elementary grades, *Journal of Learning Disabilities*, **48**, 2, 176–195 (2015)
4. Sartori, G., Job, R., and Tressoldi, P. E.: DDE-2 Batteria per la valutazione della dislessia e della disortografia evolutiva-2 [DDE-2 battery for the evaluation of dyslexia and disorthographya]. Giunti, Florence, Italy (2007)
5. World Health Organization: ICD-10: International statistical classification of diseases and related health problems. Geneva (2008)
6. Zani, S., Milioli, M.A., and Morlini I.: Fuzzy Methods and Satisfaction Indices. In Kenett, R.S. and Salini, S. (eds) *Modern Analysis of Customer Surveys*. Wiley, UK, 439–455 (2012)
7. Zani S., Milioli, M.A., and Morlini, I.: Fuzzy composite indicators: an application for measuring customer satisfaction. In Torelli N., Pesarin, F., and Bar-Hen, Avner (eds) *Advances in Theoretical and Applied Statistics*. Springer, Berlin, 241–251 (2013)
8. Zimmermann, H.J.: *Fuzzy Sets Theory and its Applications*, 2nd ed., Kluwer, Boston (1993).

# Big Textual Data: Lessons and Challenges for Statistics

## *Grandi dati testuali: le lezioni e le sfide per le statistiche*

Fionn Murtagh

**Abstract** At issue are a few early stage case studies relating to: research publishing and research impact; literature, narrative and foundational emotional tracking; and social media, here Twitter, with a social science orientation. Central relevance and importance will be associated with the following aspects of analytical methodology: context, leading to availing of semantics; focus, motivating homology between fields of analytical orientation; resolution scale, which can incorporate a concept hierarchy and aggregation in general; and acknowledging all that is implied by this expression: correlation is not causation. Application areas are: research publishing and qualitative assessment, narrative analysis and assessing impact, and baselining and contextualizing, statistically and in related aspects such as visualization.

**Key words:** mapping narrative, emotion tracking, significance of style, Correspondence Analysis, chronological hierarchical clustering

## 1 Underlying Themes in Methodology, Introduction

Clearly, through integration of analytical methodology and domain of application, the choice of methodology or even its development is dependent on the specific requirements. However the following general aspects of contemporary analytics, including textual data analytics, are useful to be noted.

An interview with Peter Norvig, Google, in C. Anderson [1] contained the following controversial perspectives: “Petabytes allow us to say: ‘Correlation is enough’. We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where

---

Fionn Murtagh

Institute of Mathematics and Data Science, University of Huddersfield, UK e-mail: fmurtagh@acm.org

science cannot.” “Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.”

To counteract this automation of all analytical reasoning, and to accept the need for inductive reasoning, there is: (1) Importance of: context (for our analytics); integration of data and domain; leading to the following. (2) Semantic analytics, and analytical synthesis in, and from, data and information. (3) Qualitative as well as quantitative evaluation and related analytics. All in all, this is leading to the Correspondence Analysis platform as an inductive reasoning framework for other analytical methodologies also.

Interestingly, the focus on regions of interest in information space is stressed by [21]. An article about the Internet of Things and Big Data by John Thornhill in the newspaper, Financial Times, on 9 January 2017 had this comment: “Sir Nigel Shadbolt, co-founder of the Open Data Institute ... The next impending revolution, he argues, will be about giving consumers control over their data.”

Ethical consequences of Big Data mining and analysis may be associated with the following, from [10]: “Rehabilitation of individuals. The context model is always formulated at the individual level, being opposed therefore to modelling at an aggregate level for which the individuals are only an ‘error term’ of the model.”

In [6], “There is the potential for big data to evaluate or calibrate survey findings ... to help to validate cohort studies”. Examples are discussed of “how data ... tracks well with the official”, far larger, repository or holdings. It is well pointed out how one case study discussed “shows the value of using ‘big data to conduct research on surveys (as distinct from survey research)”. Limitations though are clear: “Although randomization in some form is very beneficial, it is by no means a panacea. Trial participants are commonly very different from the external pool, in part because of self-selection, ...”. This is due to, “One type of selection bias is self-selection (which is our focus)”. Important points towards addressing these contemporary issues include the following. “When informing policy, inference to identified reference populations is key”: This is part of the bridge which is needed, between data analytics technology and deployment of outcomes.

Furthermore there is this: “In all situations, modelling is needed to accommodate non-response, dropouts and other forms of missing data. While “Representativity should be avoided”, here is an essential way to address in a fundamental way, what we need to address: “Assessment of external validity, i.e. generalization to the population from which the study subjects originated or to other populations, will in principle proceed via formulation of abstract laws of nature similar to physical laws”.

Hence our motivation for the following framework for analytical processes: Euclidean geometry for semantics of information; hierarchical topology for other aspects of semantics, and in particular how a hierarchy expresses anomaly or change. A further useful case is when the hierarchy respects chronological or other sequence information.

## **2 Towards: Qualitative as well as Quantitative Research Effectiveness and Impact**

For analysis of research funding, of publishing, and of commercial outcomes, account needs to be taken of measures of esteem. Also account is taken of research impact, through impact of research products: (1) research results, (2) organisation of science (journal editing, running conferences), (3) knowledge transfer, supervision, (4) technology innovations.

Correspondence Analysis when based on part of an ontology or concept hierarchy can be considered as “information focusing”. Correspondence Analysis provides simultaneous representation of observations and attributes. We project other observations or attributes into the factor space: these are supplementary or contextual observations or attributes. A 2-dimensional or planar view is an approximation of the full cloud of observations or of attributes. Therefore there can be benefit in the following: define a small number of aggregates of either observations or attributes, and carry out the analysis on them. Then project the full set of observations and attributes into the factor space.

In support of “The Leiden Manifesto for research metrics”, DORA (San Francisco Declaration on Research Assessment), Metrics Tide Report (HEFCE, Higher Education Funding Council England, 2015), qualitative judgement is primary. Research results may be assessed through first determining a taxonomic rank by mapping to a taxonomy of the domain (a manual action). There there will be unsupervised aggregation of criteria for stratification.

Research impact should be evaluated, first of all, based on qualitative considerations. Evaluation of research, especially at the level of teams or individuals can be organized by, firstly, developing and maintaining a taxonomy of the relevant subdomains and, secondly, a system for mapping research results to those subdomains that have been created or significantly transformed because of these research results. Of course, developing and/or incorporating systems for other elements of research impact, viz., knowledge transfer, industrial applications, social interactions, etc., are to be taken into account also.

See [19] for such work. Generally also see [5]. The latter maps out evolving vocabulary and associates this also with influential published articles.

## **3 Qualitative Style in Narrative for Analysis and Synthesis of Narrative**

For [11], the composition of the movie, Casablanca, is “virtually perfect”. Text is the “sensory surface” of the underlying semantics.

Here there is consideration as to how permutation testing and evaluation can be very relevant for qualitative appraisal. Considering the Casablanca movie, shot by Warner Brothers between May and August 1942, and also some early episodes of

the CSI Las Vegas, Crime Scene Investigation, television drama series, from the year 2000, the attributes used were as follows, [15].

All is based on the following: Euclidean geometry for semantics of information; hierarchical topology for other aspects of semantics, and in particular how a hierarchy expresses anomaly or change. The hierarchy respects chronological or other sequence information. Chronological hierarchical clustering, also termed contiguity constrained hierarchical clustering, is based on the complete link agglomerative clustering criterion [12, 2, 7].

1. Attributes 1 and 2: The relative movement, given by the mean squared distance from one scene to the next. We take the mean and the variance of these relative movements. Attributes 1 and 2 are based on the (full-dimensionality) factor space embedding of the scenes.
2. Attributes 3 and 4: The changes in direction, given by the squared difference in correlation from one scene to the next. We take the mean and variance of these changes in direction. Attributes 3 and 4 are based on the (full-dimensionality) correlations with factors.
3. Attribute 5 is mean absolute tempo. Tempo is given by difference in scene length from one scene to the next. Attribute 6 is the mean of the ups and downs of tempo.
4. Attributes 7 and 8 are, respectively, the mean and variance of rhythm given by the sums of squared deviations from one scene length to the next.
5. Finally, attribute 9 is the mean of the rhythm taking up or down into account.

For permutation testing, assessment was carried out relative to uniformly randomized sequences of scenes or sub-scenes.

#### **4 Statistical Significance of Impact**

Underlying [18] is the testing of social media with the aim of designing interventions, associated with statistical assessment of impact. The application here is to environmental communication initiatives. Measuring impact of public engagement theory, in the sense of the eminent political scientist, Jürgen Habermas, involves public engagement centred on communicative theory; by implication therefore, discourse as a possible route to social learning and environmental citizenship.

The case study here, was directed towards:

1. Qualitative data analysis of Twitter.
2. Nearly 1000 tweets in October, November 2012.
3. Evaluation of tweet interventions.
4. Eight separate twitter campaigns carried out.

Mediated by the latent semantic mapping of the discourse, semantic distance measures were developed between deliberative actions and the aggregate social effect. We let the data speak in regard to influence, impact and reach.



Impact was algorithmically specified in this way: semantic distance between the initiating action, and the net aggregate outcome. This can be statistically tested through the modelling of semantic distances. It can be further visualized and evaluated.

A fundamental aspect of the Twitter analysis was how a tweet, considered as a “campaign initiating tweet”, differed from an aggregate set of tweets. The latter was the mean tweet, where the tweets were first mapped into a semantic space. The semantic space is provided by the factor space, which is endowed with a Euclidean metric. For very high dimensions, we find “data piling” or concentration. That is, the cloud of points becomes concentrated in a point. Now that could be of benefit to us, when we are seeking a mean (hence, aggregate) point in a very high dimensional space. A further aspect is when it is shown that the cloud piling or concentration is very much related to the marginal distributions.

Here we show how we can test the statistical significance of effectiveness.

The campaign 7 case, with the distance between the tweet initiating campaign 7, and the mean campaign 7 outcome, in the full, 338-dimensional factor space is equal to 3.670904.

Compare that to all pairwise distances of non-initiating tweets. We verified that these distances are normal distributed, with a small number of large distances. By the central limit theorem, for very large numbers of such distances, they will be normal distributed. Denote the mean by  $\mu$ , and the standard deviation by  $\sigma$ . Mean and standard deviation are defined from distances between all non-initiating tweets, in the full dimensionality semantic (or factor) space. We find  $\mu = 12.64907$ ,  $\mu - \sigma = 8.508712$ , and  $\mu - 2\sigma = 4.368352$ .

We find the distance between initiating tweet and mean outcome, for campaign 7, in terms of the mean and standard deviation of tweet distances to be:  $\mu - 2.168451\sigma$ . Therefore for  $z = -2.16$ , the campaign 7 effectiveness is significant at the 1.5% level (i.e.  $z = -2.16$ , in the two-sided case, has 98.5% of the normal distribution greater than it in value).

In the case of campaigns 1, 4, 5, 6, their distances between initiating tweet and mean outcome are less than 90% of all tweet distances. Therefore the effectiveness of these campaigns is in the top 10% which is not greatly effective (compared to campaign 7).

In the case of campaigns 3 and 8, we find their distances to be less than 80% of all tweet distances. So their effectiveness is in the top 20%.

Finally, campaign 2 is the least good fit, relative to initiating tweet and outcome.

## 5 Tracking Emotion

This relates to determining and tracking emotion in an unsupervised way. This is as opposed to machine learning, like in sentiment analysis, which is supervised. Emotion is understood as a manifestation of the unconscious. Social activity causes

emotion to be expressed or manifested. This can lead to later discussion of psychoanalyst, Matte Blanco. See [14].

The foundation of this tracking of emotion, and determining the depth of emotion, is using the methodology of metric space mapping and hierarchical topology. The former here maps the textual data into a Euclidean metric endowed factor space, and the latter may be chronologically constrained hierarchical clustering.

The examples to follow are based on: in the Casablanca movie, dialogue (and dialogue only) between main characters Ilsa and Rick, having selected this dialogue from the scenes with both of these protagonists (scenes 22, 26, 28, 30, 31, 43, 58, 59, 70, 75 and last scene, 77); and chapters 9, 10, 11, 12 of Gustave Flaubert's 19th century novel, *Madame Bovary*. This concerns the three-way relationship between Emma Bovary, her husband Charles, and her lover Rodolphe Boulanger.

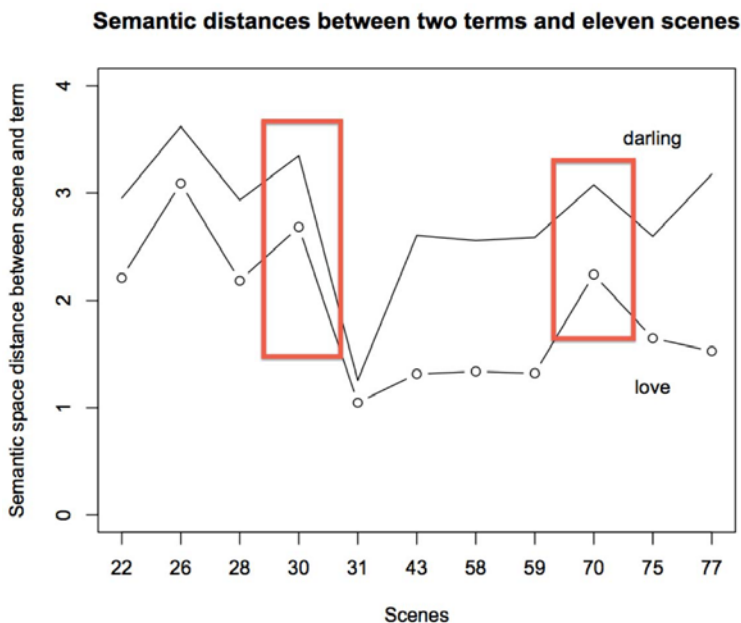
Following [16], in Figure 1 in the full dimensionality factor space, based on all interrelationships of scenes and words, the distance between the word "darling" in this space, was determined with each of the 11 scenes in this space. The same was done for the word "love". The semantic locations of these two words, relative to the semantic locations of scenes 30 and 70 are highlighted with boxes.

Then in Figure 2, hierarchical clustering, that is sequence constrained, is carried out on the scenes used, i.e. scenes 22, 26, 28, 30, 31, 43, 58, 59 70, 75, 77 (using the dialogue, between Ilsa and Rick). See how the big changes in scenes 30 and 70 are indicated in the previous figure.

Now there is consideration of the novel *Madame Bovary*, with the 3-way interrelationships of Emma Bovary, her husband Charles, and her lover, Rodolphe.

Figure 3 presents an interesting perspective that can be considered relative to the original text. Rodolphe is emotionally scoring over Charles in text segment 1, then again in 3, 4, 5, 6. In text segment 7, Emma is accosted by Captain Binet, giving her qualms of conscience. Charles regains emotional ground with Emma through Emma's father's letter in text segment 10, and Emma's attachment to her daughter, Berthe. Initially the surgery on Hippolyte in text segment 11 draws Emma close to Charles. By text segment 14 Emma is walking out on Charles following the botched surgery. Emma has total disdain for Charles in text segment 15. In text segment 16 Emma is buying gifts for Rodolphe in spite of potentially making Charles indebted. In text segments 17 and 18, Charles' mother is there, with a difficult mother-in-law relationship for Emma. Plans for running away ensue, with pangs of conscience for Emma, and in the final text segment there is Rodolphe refusing to himself to leave with Emma.

In Figure 4, there is display of the evolution of sentiment, expressed by (or proxied by) the terms "kiss", "tenderness", and "happiness". We see that some text segments are more expressive of emotion than are other text segments.

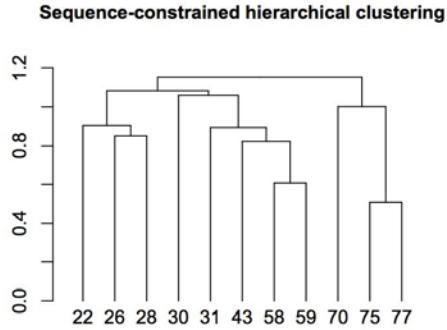


**Fig. 1** In the full dimensionality factor space, based on all interrelationships of scenes and words, we determined the distance between the word “darling” in this space, with each of the 11 scenes in this space. We did the same for the word “love”. The semantic locations of these two words, relative to the semantic locations of scenes 30 and 70 are highlighted with boxes.

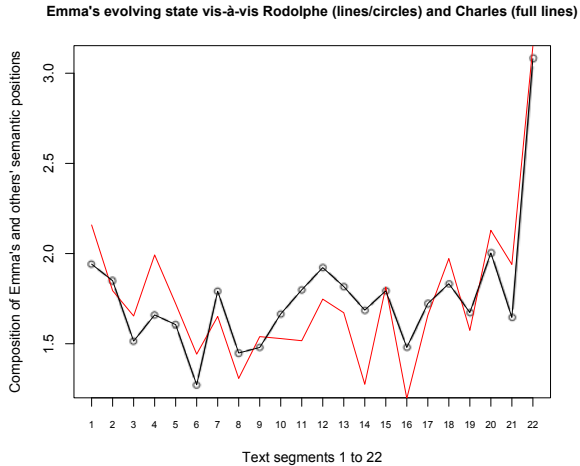
## 6 Analyses of Mapping of Behavioural or Activity Patterns or Trends

This concerns semantic mapping of Twitter data relating to music, film, theatre, etc. festivals. 75 languages were found to be in use, including Japanese, Arabic and so on, with the majority in Roman script. As indicative association to language, because the labelled language may be partially used or not in fact used, we take the following: English, Spanish, French, Japanese, Portuguese. We consider the days 2015-05-11 to 2016-08-02, with two days removed, due to lack of tweets. The numbers of tweets for these languages were as follows (carried out on 11 August 2016): en, 37681771; es, 9984507; fr, 4503113; ja, 2977159; pt, 3270839

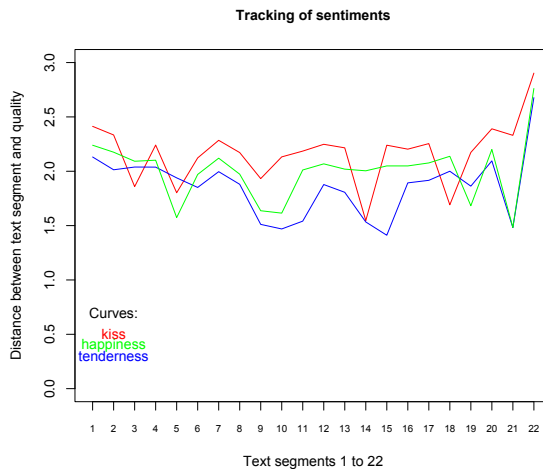
The tweeters and the festivals are as follows. Tweets characterized as French, 4913781 tweets. (For user, date and tweet content, the file size was: 667 MB.) The following were sought in the tweets: Cannes, cannes, CANNES, Avignon, avignon, AVIGNON. Upper and lower case were retained in order to verify semantic prox-



**Fig. 2** Hierarchical clustering, that is sequence constrained, of the 11 scenes used, i.e. scenes 22, 26, 28, 30, 31, 43, 58, 59, 70, 75, 77 (all with dialogue, and only dialogue, between Ilsa and Rick). Rather than projections on factors, here the correlations (or cosines of angles with factors) are used to directly capture orientation.



**Fig. 3** The relationship of Emma to Rodolphe (lines/circles, black) and to Charles (full line, red) are mapped out. The text segments encapsulate narrative chronology, that maps approximately into a time axis. Low or small values can be viewed as emotional attachment.



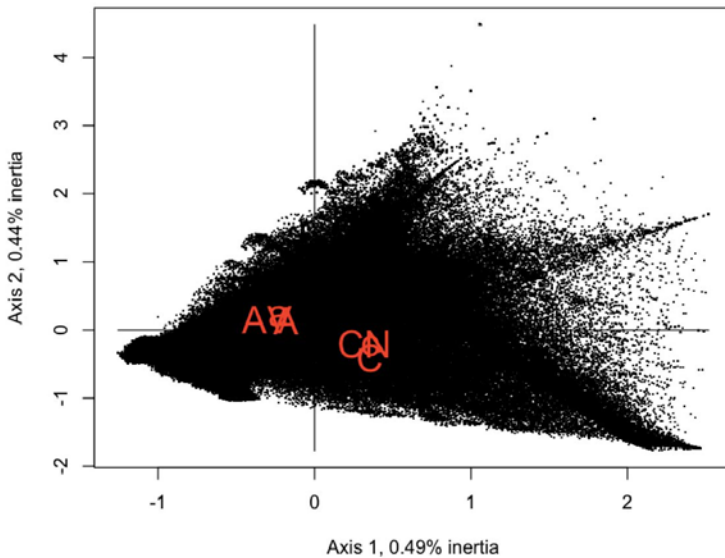
**Fig. 4** A low value of the emotion, expressed by the words “kiss”, “happiness” and “tenderness”, implies small distance to the text segment. These curves, “kiss”, “happiness” and “tenderness” start on the upper left on the top, the middle, and the bottom, respectively. The chronology of sentiment tracks the closeness of these different sentimental terms relative to the narrative, represented by the text segment. Terms and text segments are vectors in the semantic, factorial space, and the full dimensionality of this space is used.

imity of these variants. These related to the Cannes Film Festival, and the Avignon Theatre Festival. The following total numbers of occurrences of these words were found, and the maximum number of occurrences by a user, i.e. by a tweeter: Cannes, 1230559 and 3388; cannes, 145939 and 4024; CANNES, 57763 and 829; Avignon, 272812 and 4238; avignon, 39323 and 2909; AVIGNON, 14647 and 900.

The total number of tweeters, also called users here: 880664; total number of days retained, from 11 May 2015 to 11 Sept. 2016, 481. Cross-tabulated are: 880664 users by 481 days. There are 1230559 retained and recorded tweets. The non-sparsity of this matrix is just: 0.79%

In Figure 6, mapped are: C, c, CA (Cannes, cannes, CANNES) and A, a, AV (Avignon, avignon, AVIGNON). They are supplementary variables in the Correspondence Analysis principal factor plane. Semantically they are clustered. They are against the background of the Big Data, here the 880664 tweeters, represented by dots.

Current considerations, relating to approximately 55 million tweets per year (from May 2015), are as follows. Determine some other, related or otherwise, behavioural patterns that are accessible in the latent semantic, factor space. Retain selected terms from the tweets, and, as supplementary elements, see how they provide more information on patterns and trends. Carry out year by year trend analysis.



**Fig. 5** 880664 Twitter tweets projected on the principal factor, i.e. principal axis plane. Attributes are projected.

For further analyses and description of the data, see [4] and [17].

### **6.1 Baselineing or Contextualizing Analysis**

The following is in regard to such baselineing, i.e. contextualizing, against healthy reference subjects, from a case study in [9]. This repeats some of the description in [13], in regard to testing through statistically baselineing or contextualizing in a multivariate manner.

In [20], there is an important methodological development, concerning statistical inference in Geometric Data Analysis, i.e. based on MCA, Multiple Correspondence Analysis. At issue is statistical “typicality of a subcloud with respect to the overall cloud of individuals”. Following an excellent review of permutation tests, the data is introduced: 6 numerical variables relating to gait, body movement, related to the

following; a reference group of 45 healthy subjects; and a group of 15 Parkinsons illness patients, each before and after drug treatment. [8] (section 11.1) relates to this analysis, of the, in total,  $45 + 15 + 15$  observation vectors, of subjects between the ages of 60 and 92, of average age 74.

First there is correlation analysis carried out, so that when PCA of standardized variables is carried out, it is the case that the first two axes explain 97% of the variance. Axis 1 is characterized as “performance”, and axis 2 is characterized as “style”. Then the two sets of, before treatment, and after treatment, 15 Parkinsons patients are input into the analysis as supplementary individuals. [20] is directly addressing statistically the question of effect of treatment. Just as in [8], the healthy subjects are the main individuals, and the treated patients, before and after treatment, are the supplementary individuals. This allows to discuss the subclouds of the before, and of the after treatment individuals, relative to the first, performance, axis, and the second, style, axis. The test statistic, that assesses statistically the effect of medical treatment here, is a permutation-based distributional evaluation of the following statistic. The subcloud’s deviations relative to samples of the reference cloud are at issue. The Mahalanobis distance based on covariance structure of the reference cloud is used. The test statistic is the Mahalanobis norm of deviations between subcloud points and the mean point of the reference cloud.

In summary, this exemplifies in a most important way, how supplementary elements and the principal elements are selected and used in practice. The medical treatment context is so very clear in regard to such baselining, i.e. contextualizing, against healthy reference subjects.

## 7 Conclusion

Much that is at issue here is close to what is under discussion in [3]. The integral association of methodology and application domain will, of course, have shared and common methodological perspectives. However the application of statistical models, and other analytical stages such as feature selection, data aggregation with the various implications of this, and what is often termed data cleaning or data cleansing, all of these issues require analytical focus, and account to be taken of the analytical context. The latter may well include baselining, or benchmarking in an operational manner. In a sense, we might state that combinatorial inference is so paramount because of its applicability.

A good deal of the case studies reported on here made use of preliminary functionality, to be part of the R package, `XplorText`. This package makes use of these R packages, and add greatly to their functionality: `tm`, `FactoMineR`.

The software system, `SPAD`, is also extending greatly into support for text processing.

## References

1. Anderson, C.: “The end of theory: the data deluge makes the scientific method obsolete”, *Wired Magazine* (16 July 2008),
2. Bécue-Bertaut, M., Kostov, B., Morin, A., Naro, G.: “Rhetorical strategy in forensic speeches: Multidimensional statistics-based methodology”, *Journal of Classification*, 31, 85–106 (2014)
3. Gelman, A., Hennig, C.: “Beyond subjective and objective in statistics”, *Journal of the Royal Statistical Society Series A*, 180, Part 4, 1–31 (2017).
4. Goeriot, L., Mothe, J., Mulhem, P., Murtagh, F., SanJuan, E.: “Overview of the CLEF 2016 Cultural Micro-blog Contextualization Workshop”. In: Editors: N. Fuhr, P. Quaresma, T. Goncalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 7th International Conference of the CLEF Association, CLEF 2016, vora, Portugal, September 5-8, 2016, Proceedings, Lecture Notes in Computer Science, volume 9822, pp. 371–378 (2016)
5. Hernández, D.M., Bécue-Bertaut, M., Barahona, I.: “How scientific literature has been evolving over the time? A novel statistical approach using tracking verbal-based methods”, *JSM Proceedings, 2014, Section on Statistical Learning and Data Mining*, American Statistical Association, 1121–1132 (2014)
6. Keiding, N., Louis, T.A.: “Perils and potentials of self-selected entry to epidemiological studies and surveys”, *Journal of the Royal Statistical Society A*, 179, Part 2, 319–376 (2016)
7. Legendre, P., Legendre, L.: *Numerical Ecology*, 3rd edn., Elsevier, Amsterdam (2012)
8. Le Roux, B.: *Analyse Géométrique des Données Multidimensionnelles*, Dunod, Paris (2014)
9. Le Roux, R., Rouanet, H.: *Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis*, Kluwer, Dordrecht (2004)
10. Le Roux, B., Lebaron, F.: “Idées-clefs de l’analyse géométrique des données” (Key ideas in the geometric analysis of data). In F. Lebaron and B. Le Roux, editors, *La Méthodologie de Pierre Bourdieu en Action: Espace Culturel, Espace Social et Analyse des Données*, pages 3–20. Dunod, Paris (2015)
11. McKee, R.: *Story: Substance, Structure, Style, and the Principles of Screenwriting*, Methuen (1999)
12. Murtagh, F.: *Multidimensional Clustering Algorithms*, Physica-Verlag, Würzburg (1985)
13. Murtagh, F.: “Contextualizing Geometric Data Analysis and related data analytics: A virtual microscope for Big Data analytics”, *JIMIS*, submitted (2017)
14. Murtagh, F.: *Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics*, Chapman and Hall, CRC Press (2017)
15. Murtagh, F., Ganz, A., McKie, S.: “The structure of narrative: the case of film scripts”, *Pattern Recognition*, 42, 302–312 (2009)
16. Murtagh, F., Ganz, A.: “Pattern recognition in narrative: Tracking emotional expression in context”, *Journal of Data Mining and Digital Humanities*, vol. 2015 (published May 26, 2015).
17. Murtagh, F.: “Semantic mapping: towards contextual and trend analysis of behaviours and practices”. In: K. Balog, L. Cappellato, N. Ferro, C. MacDonald, Eds., *Working Notes of CLEF 2016 – Conference and Labs of the Evaluation Forum*, Évora, Portugal, 5-8 September, 2016, pp. 1207–1225 (2016). <http://ceur-ws.org/Vol-1609/16091207.pdf>
18. Murtagh, F., Pianosi, M., Bull, R.: “Semantic mapping of discourse and activity, using Habermas’s Theory of Communicative Action to analyze process”, *Quality and Quantity*, 50(4), 1675–1694 (2016)
19. Murtagh, F., Orlov, M., Mirkin, B.: “Qualitative judgement of research impact: Domain taxonomy as a fundamental framework for judgement of the quality of research”, *Journal of Classification* (in press, 2017). Preprint: <https://arxiv.org/abs/1607.03200>
20. Bienaise, S., Le Roux, B.: “Combinatorial typicality test in Geometric typicality test in geometric data analysis”, preprint (2016)
21. Wessel, M.: “You dont need Big Data – You need the right data”. *Harvard Business Review* (3 Nov. 2016). <https://hbr.org/2016/11/you-dont-need-big-data-you-need-the-right-data>.



# Workers' skills and wage inequality: A time-space comparison across European Mediterranean countries

## *Competenze e disuguaglianze salariali: un'analisi spazio-temporale per i paesi dell'Europa mediterranea*

Gaetano Musella and Gennaro Punzo

**Abstract** The work aims at exploring how the changes in the demand for skills in the labour market affect wage inequality comparatively for four countries of Southern Europe (Greece, Italy, Portugal, and Spain). Through the Recentered Influence Function (RIF) regression of Gini on EU-SILC data, Italy is compared to each other country concerned in order to assess the evolution of spatial inequality divides during the Great Recession (2005-2013). Gini gaps are then decomposed into the composition effect (employees' endowments) and wage structure (how employees' skills are rewarded). Based on our results, Italy appears to be a less unequal country as part of the Mediterranean Europe. A clearer employment structure may slow country's inequality growth and reduce spatial gaps.

**Abstract** *Il lavoro intende analizzare come i recenti cambiamenti nella domanda di competenze lavorative influenzino la disuguaglianza salariale nei principali paesi dell'Europa Mediterranea (Grecia, Italia, Portogallo e Spagna). Usando dati EU-SILC, il lavoro propone la stima di modelli di regressione RIF su Gini al fine di valutare le dinamiche spazio-temporali della disuguaglianza dei salari negli anni della Grande Recessione. I divari spaziali di disuguaglianza sono decomposti con il duplice obiettivo di separare la quota di gap totale attribuibile alla distribuzione territoriale delle dotazioni dei lavoratori dalla quota legata alla differente capacità dei mercati del lavoro nazionali di trasformare tali dotazioni in opportunità lavorative e guadagni. I risultati restituiscono un'immagine dell'Italia "meno diseguale" rispetto agli altri paesi del Mediterraneo ed evidenziano come strutture di mercato ben definite possano rallentare la disuguaglianza e ridurre i gap spaziali*

**Key words:** Employment structure, wage inequality, RIF-regression

---

<sup>1</sup> Gaetano Musella, University of Naples Parthenope, Department of Economics and Law Studies; e-mail: gaetano.musella@uniparthenope.it

Gennaro Punzo, University of Naples Parthenope, Department of Economics and Law Studies; e-mail: gennaro.punzo@uniparthenope.it

## 1. Background and aim of the work

The Great Recession is the biggest macroeconomic downturn since the 1930s. Its origins can be traced back to the 2007 global financial crisis that, in turn, was generated by the housing bubble and banking crisis in the United States. The economic recession reached its peak in Europe before the end of 2009 with the inevitable rising of unemployment and income inequality. Although the crisis hit all EU Member States hard, each country was affected with different proportion, timing and strength. However, nations with weaker economies suffered the most, and in particular, the four European Mediterranean countries – Portugal, Italy, Greece, and Spain – were damaged more extensively [7]. For instance, their total and youth unemployment rates were significantly greater than the Eurozone average (*Eurostat on-line database*).

The potential causes of rising unemployment and changing in the employment composition are widely discussed in literature, and for some years, many Authors have been focusing their attention on the shrinking of middle-skill jobs [1,2,3,5]. The progressive decrease in the demand for the middle-skilled workers has generated different structures of the labour market – such as job polarisation, upgrading and downgrading of occupations – of most developed economies. Specifically, job polarisation represents the case in which the demand for jobs requiring high and low skills grows simultaneously. Upgrading of occupations occurs if the growth involves the high-skill jobs exclusively, while the downgrading is the case in which low-skill occupations grow faster than other ones.

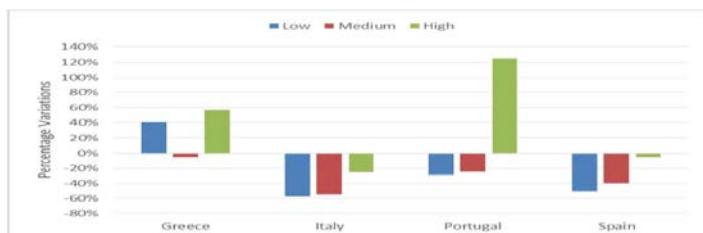
In this field, the goal of the present work is to explore the ways in which the decline of middle-skill jobs affects wage inequality comparatively for four countries of Southern Europe. At this purpose, the Recentered Influence Function (RIF) regression [4] allows achieving two objectives. First, evaluating the direction and intensity of the inequality spatial gap of Italy compared to each other country concerned (Greece, Portugal, and Spain). Second, decomposing the spatial inequality gaps into the two components of *composition effect*, which quantifies how much of the gap is due to the employees' endowments, and *wage structure*, which measures how much of the same gap is attributable to the capability of the country's labour market to reward the employees' characteristics.

## 2. Sketching the labour market structures of the Southern Europe

This Section focuses on the changes in employment composition by skill levels that have occurred between 2005 and 2013 in Greece, Italy, Portugal, and Spain. To do this, three groups of high-, middle- and low-skilled workers have been created using the average level of formal education as proxy of workers' skills [3]. Data are from the EU-SILC (European Union-Survey on Income and Living Condition) survey that categorizes jobs according to the International Standard Classification of Occupation (ISCO-08). The analysis focuses on employees, aged 16-64, defined as anyone who

works for a public or private employer and receives compensation in the form of a wage, salary, fee, payment by result or in kind.

Concisely, our results show that the shrinking of middle-skill jobs is a common element within the Mediterranean countries. More precisely, it is worth noting how the drop of middle-skill jobs may generate different patterns of the employment structure in each country's labour market (Figure 1).



**Figure 1:** Employment shares by skill levels. Percentage changes 2005–2013.

The relative growth of both low- (+40%) and high-skill jobs (+57%) allows classifying the Greek labour market as polarised in 2005–2013. Instead, in Portugal, the sharp rise in the demand of high-skilled employees (+125%), combined with the fall in the demand of low-skilled employees (-28%), provides evidence of upgrading of occupations. Italy and Spain share more hybrid patterns because neither of the two phenomena of job polarisation or upgrading clearly prevails. The drop of job opportunities for each differently skilled group of employees makes it difficult to draw clear conclusions about the structure of the Italian and Spanish labour markets.

### 3. Methodology: RIF of Gini on log-wage

The Recentered Influence Function regression [4] of Gini on log-wage will be performed to evaluate Gini differentials between Italy and each other country concerned. Once estimated RIF regression by country, the twofold decomposition (composition effect and wage structure) is carried out.

The observed wage ( $Y_i$ ) can be written without imposing a specific functional form considering the wage determination function of observed components  $X_i$  and some unobserved components  $\varepsilon_i$ :

$$Y_{gi} = f_g(X_i, \varepsilon_i), \quad \text{for } g = A, B \quad (1)$$

where  $g = A$  for employees belonging to the country of reference (in this work, Italy) and  $g = B$  for employees from each other country (alternatively, Greece, Portugal, and Spain).

The RIF-regression replaces the log-wage as dependent variable with the recentered influence function of Gini coefficient  $v(F)$ . Let  $v$  be the generic distributional statistic to study and IF the influence function [6], the RIF-regression is:

$$(2) \quad RIF(Y; v) = IF(Y; v) + v$$

The above expression can be written as:

$$(3) \quad E[RIF(Y; v)|X] = X\beta^v$$

where  $\beta^v$  represents the marginal effect of  $X$  on  $v$ . The  $v$ -overall wage inequality gap between the countries  $A$  and  $B$  can be measured as follows:

$$(4) \quad \Delta_o^v = v_B(F_B) - v_A(F_A) = v_B - v_A$$

that can be decomposed into:

$$(5) \quad \Delta_o^v = (v_B - v_c) + (v_c - v_A) = \Delta_s^v + \Delta_x^v$$

The overall gap between Italy and each other country is decomposed into wage structure ( $\Delta_s^v$ ) and composition effect ( $\Delta_x^v$ ). The first term corresponds to the effect on  $v$  of a change from  $f_B(\cdot, \cdot)$  to  $f_A(\cdot, \cdot)$ , keeping the distribution of  $(X, \varepsilon)|G = B$  constant. The second term keeps constant the wage structure  $f_A(\cdot, \cdot)$  and measures the effect of changes from  $(X, \varepsilon)|G = B$  to  $(X, \varepsilon)|G = A$ . However, the key term for decomposing the total gap is the counterfactual distributional statistics  $v_c$  that represents the distributional statistic that would have prevailed if employees from the country  $A$  have the wage structure of those from the country  $B$ .

As regards the Gini coefficient, the distributional statistic  $v$  is defined as follows:

$$(6) \quad v^{GC}(F_Y) = 1 - 2\mu^{-1}R(F_Y)$$

where  $R(F_Y) = \int_0^1 GL(p; F_Y)dp$  with  $p(y) = F_Y(y)$  and the generalised Lorenz ordinate of  $F_Y$  is given by  $GL(p; F_Y) = \int_{-\infty}^{F^{-1}(p)} zdF_Y(z)$ . As demonstrated by Firpo et al. [4], the recentered influence function of Gini can be rewritten as:

$$(7) \quad RIF(y; v^{GC}) = 1 + 2\mu^{-2}R(F_Y) - 2\mu^{-1}[y[1 - p(y)] + GL(p(y); F_Y)]$$

The gaps in the Gini index between the two countries may be decomposed as in equation (5).

#### 4. Main results

This Section shows the main results of RIF-decomposition that allow evaluating the wage inequality gaps of Italy with each other country in both 2005 and 2013 and their evolution over time (Tables 1-2).

**Table 1:** RIF decomposition of Gini on log-wage. Gap between Italy and each other country – 2005

	2005					
	Spain-Italy	% share	Portugal-Italy	% share	Greece-Italy	% share
<b>Total Gap</b>	0.0080*** (0.0004)	-	0.0102*** (0.0006)	-	0.0044*** (0.0006)	-
<b>Composition Effect</b>	0.0030*** (0.0003)	38.24	-0.0040*** (0.0005)	-39.79	0.0004 (0.0006)	10.79
<b>Wage Structure</b>	0.0049*** (0.0004)	61.76	0.0142*** (0.0006)	139.79	0.0039*** (0.0008)	89.21

\*Significant at 10%; \*\*Significant at 5%; \*\*\*Significant at 1%. Standard errors in brackets.

**Table 2:** RIF decomposition of Gini on log-wage. Gap between Italy and each other country – 2013

	2013					
	Spain-Italy	% share	Portugal-Italy	% share	Greece-Italy	% share
<b>Total Gap</b>	0.0116*** (0.0006)	-	0.0037*** (0.0006)	-	0.0016** (0.0008)	-
<b>Composition Effect</b>	0.0047*** (0.0005)	40.12	-0.0034*** (0.0003)	-94.04	0.0021*** (0.0007)	131.61
<b>Wage Structure</b>	0.0069*** (0.0007)	59.87	0.0071*** (0.0007)	194.04	-0.0005 (0.0009)	-31.61

\*Significant at 10%; \*\*Significant at 5%; \*\*\*Significant at 1%. Standard errors in brackets.

Between 2005 and 2013, the overall Gini has increased in Spain, Italy and Greece, while it has slightly decreased for Portugal, changing the intensity of inequality gaps over time. In fact, while Italy shows the lowest level of wage inequality in both 2005 and 2013, the spatial gap has progressively reduced over time for Portugal (from 0.0102 to 0.0037) and Greece (from 0.0044 to 0.0016) and it has increased for Spain (from the 0.0088 in 2005 to 0.0116 in 2013). In detail, the decrease of wage inequality in Portugal makes smaller the spatial differential with Italy in 2013, and similarly, the spatial gap between Italy and Greece has narrowed because inequality has increased in Greece to a lesser extent than Italy. Instead, the harshest increase in Spanish wage inequality makes the spatial differential with Italy even wider in 2013.

For Spain and Portugal, a great deal of spatial inequality gap is attributable to the wage structure (2005: 81.01% and 139.22%, respectively; 2013: 62.61% and 202.6%, respectively), highlighting how their larger inequality depends on the lower capacity of their labour markets to reward the employees' characteristics. Some distinctions are appropriate regarding the composition effect of the two countries. In Spain, the composition effect plays a significant role (around 40%) in widening the inequality differential, showing that, beyond the changes in demanding skills, the divide is also potentially due to the different employees' endowments of the two countries. By contrast, in Portugal, the composition effect is even negative. It means that the disadvantage of the wage structure in widening the inequality gap between Italy and Portugal is partly compensated by the larger availability, in the Portuguese workforce, of high-skilled employees (as evidenced by the upgrading structure of its labour market) who usually earn higher salaries. The same composition effect plays a leading role to explain the gap between Italy and Greece in 2013 (188.2%). As regards the wage structure in the comparison with Greece, it is worthy of attention its evolution over time: while in 2005 it was the primary component of the total spatial gap, in 2013 it became no significant and the composition effect remains the only responsible of the higher wage inequality in Greece.

Briefly, the spatial comparison highlights how inequality gaps may also be explained by the differences in labour market structures and their ability to reward the investment in skills. The evolution of inequality divides over time shows how the well-defined structures of upgrading (Portugal) and job polarisation (Greece) seem to have an equalising effect on the countries' wage distribution, lessening both the overall inequality within countries and penalties respect to Italy. Instead, a more hybrid structure (Spain) exacerbates wage inequality where both the wage structure and composition effect play a key role to explain the spatial gap with Italy.

## References

1. Autor, D.: Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing. *Journal of Labor Economics*, 21 (1), 2003.
2. Castellano, R., Musella, G., Punzo, G.: Structure of the labour market and wage inequality: Evidence from European countries. *Quality&Quantity*, 2016.
3. Fernández-Macías, E.: Job Polarization in Europe? Changes in the Employment Structure and Job Quality, 1995-2007. *Work and Occupations*, 39 (2), 157-182, 2012.
4. Firpo, S. Fortin, N. Lemieux, T.: Unconditional Quantile Regressions. *Econometrica* 77(3), 953-973, 2009.
5. Goos, M., Manning, A., Salomons, A.: Job Polarization in Europe. *American Economic Review*, vol. 99, 58-63, 2009.
6. Hampel, F. R.: The Influence Curve and Its Role in Robust Estimation, *Journal of the American Statistical Association* 69, 383-393, 1974.
7. Tóth, I.G.: Revisiting Grand Narratives of Growing Inequalities: Lesson from 30 Country Studies. In: Nolan et al. (eds) *Changing Inequalities and Societal Impacts in Rich Countries: Thirty Countries' Experiences*, Oxford: Oxford University Press, 2014.

# Exploratory factor analysis of ordinal variables: a copula approach

## *Analisi fattoriale esplorativa di variabili ordinali: un approccio via copula*

Marta Nai Ruscone

**Abstract** Exploratory factor analysis attempts to identify the underlying factors that explain the pattern of correlations within a set of observed variables. The analysis is almost always performed with Pearson's correlations even when the data are ordinal, but this is not appropriate since they are not quantitative data. The use of Likert scales is increasingly common in the field of social research, so it is necessary to determine which methodology is the most suitable for analysing the data obtained as non quantitative measures. In this context, also by means of simulation studies, we aim to illustrate the advantages of using Spearman's grade correlation coefficient on a transformation operated by the copula function in order to perform exploratory factor analysis of ordinal variables. Moreover, by using the copula, we consider the general dependence structure, providing a more robust reproduction of the measurement model.

**Abstract** *L'analisi fattoriale esplorativa vuole identificare i fattori latenti che spiegano un insieme di variabili osservate. L'analisi quasi sempre utilizza la correlazione di Pearson, anche quando i dati sono di natura ordinale, ma questo non è appropriato in quanto questi dati non sono quantitativi. L'uso di scale Likert è sempre più comune nel campo della ricerca sociale, risulta quindi necessario determinare quale metodo risulta essere più idoneo per l'analisi di tali dati tenendo presente che spesso vengono analizzati utilizzando tecniche idonee solo per misure quantitative. In questo contesto, e mediante studi di simulazione, si illustrano i vantaggi nell'utilizzo dello Spearman grade correlation ottenuto mediante l'utilizzo dalla funzione copula anziché della correlazione di Pearson. Con l'utilizzo della copula, si considera così la struttura di dipendenza generale, fornendo così una misurazione più accurata*

**Key words:** Factor analysis, copula, ordinal variables, Likert scales, correlation

---

Marta Nai Ruscone

School of Economics and Management - LIUC - University Cattaneo, C.so Matteotti 22 - 21053 Castellanza (VA), Italy, e-mail: mnairuscone@liuc.it

## 1 Introduction

Exploratory factor analysis is a widely used statistical technique in the social sciences where the main interest lies in measuring the unobserved construct, such as emotions, attitudes, beliefs and behaviors. The main idea behind the analysis is that the latent variables (also named factors) account for the dependencies among the observed variables (also named items or indicators) in the sense that if the factors are held fixed, the observed variables would be independent. In exploratory factor analysis the goal is the following: for a given set of observed variables  $x_1, \dots, x_p$  one wants to find a set of latent factors  $\xi_1, \dots, \xi_k$ , fewer in number than the observed variables ( $k < p$ ), that contain essentially the same information. In its classical formulation [1], it concerns a set of continuous variables measured on a set of independent units. The data usually encountered in social sciences are of categorical nature (ordinal or nominal). The Likert Rating Scale [10], [11] is a simple procedure for generating measurement instruments which is widely used by social scientists to measure a variety of latent constructs, and meticulous statistical procedures have therefore been developed to design and validate these scales [3], [15]. However, most of these ignore the ordinal nature of observed responses and assume the presence of continuous observed variables measured at interval level. Evidence shows that, under relatively common circumstances, classical factor analysis (FA) yields inaccurate results characterizing the internal structure of the scale or selecting the most informative items within each factor [4], [7].

In the present work Spearman's grade correlation coefficient on a transformation operated by the copula function is employed, in order to take into account the ordinal nature of the data. The copula is a helpful tool for handling multivariate continuous distributions with given univariate marginals [14]. It describes the dependence structure existing across pairwise marginal random variables. In this way we can consider bivariate distributions with dependence structures, different from the linear one, that characterises the multivariate normal distribution.

So taking into account that the use of measurement instruments which require categorical responses from subjects is increasingly common in social research, and this implies the use of ordinal scales, the present work aims to point out a correct definition of dependence measure for ordinal variables rather than the Pearson correlation coefficient correctly applied to quantitative variables. Moreover, the use of several copulae with specific tail dependence allow us to obtain an index that weights the ordinal variables categories in several ways. In so doing we can address and recognize the ordinal nature of observed variables and estimate that weight directly from the data.

## 2 The copula function

The copula function is the key ingredient for handling multivariate continuous distributions with given univariate marginals. We will discuss this issue briefly below,



for further details and proofs, see for instance [14], [8] and [2]. It describes the dependence structure existing across pairwise marginal random variables. In this way we can consider bivariate distributions with dependence structures different from the linear one that characterises the multivariate normal distribution.

A bivariate copula  $C : I^2 \rightarrow I$ , with  $I^2 = [0, 1] \times [0, 1]$  and  $I = [0, 1]$ , is the cumulative bivariate distribution function of a random variable  $(U_1, U_2)$  with uniform marginal random variables in  $[0, 1]$

$$C(u_1, u_2; \theta) = P(U_1 \leq u_1, U_2 \leq u_2; \theta), \quad 0 \leq u_1 \leq 1 \quad 0 \leq u_2 \leq 1 \quad (1)$$

where  $\theta$  is a parameter measuring the dependence between  $U_1$  and  $U_2$ .

The following theorem by Sklar [14] explains the use of the copula in the characterization of a joint distribution. Let  $(X_1, X_2)$  be a bivariate random variable with marginal cdfs  $F_{X_1}(x_1)$  and  $F_{X_2}(x_2)$  and joint cdf  $F_{X_1, X_2}(x_1, x_2; \theta)$ , then there is always a **copula function**  $C(\cdot, \cdot; \theta)$  with  $C : I^2 \rightarrow I$  such that

$$F_{X_1, X_2}(x_1, x_2; \theta) = C(F_{X_1}(x_1), F_{X_2}(x_2); \theta), \quad x_1, x_2 \in \mathbb{R}. \quad (2)$$

Conversely, if  $C(\cdot, \cdot; \theta)$  is a copula function and  $F_{X_1}(x_1)$  and  $F_{X_2}(x_2)$  are marginal cdfs, then  $F_{X_1, X_2}(x_1, x_2; \theta)$  is a joint cdf.

If  $F_{X_1}(x_1)$  and  $F_{X_2}(x_2)$  are **continuous** functions then the copula  $C(\cdot, \cdot; \theta)$  is **unique**. Moreover, if  $F_{X_1}(x_1)$  and  $F_{X_2}(x_2)$  are continuous the copula can be found by the inverse of (2):

$$C(u_1, u_2) = F_{X_1, X_2}(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)), \quad (3)$$

with  $u_1 = F_{X_1}(x_1)$  and  $u_2 = F_{X_2}(x_2)$ . This theorem states that each joint distribution can be expressed in term of two separate but related issues, the marginal distributions and the dependence structures between them. The **dependence structure** is explained by the copula function  $C(\cdot, \cdot; \theta)$ . Moreover the (2) provides a general mechanism to construct new multivariate models in a straightforward manner. By changing the copula function we can construct new bivariate distributions with different dependence structures, with the association parameter indicating the strength of the dependence, also different from the linear one that characterizes the multivariate normal distribution.

Each copula is related to the most important measures of dependence: the Pearson correlation coefficient, the Spearman grade correlation coefficient and tail dependence parameters. The Spearman grade correlation coefficient (see [14] pp. 169-170 for the definition of the grade correlation coefficient for continuous random variables) measure the association between two variables and can be expressed as a function of the copula. More precisely, if two random variables are continuous and have copula  $C$  with parameter  $\theta$ , then the Spearman grade correlation is

$$\rho_s(C) = 12 \int_{I^2} C_\theta(u_1, u_2) du_1 du_2 - 3 = \frac{\text{Cov}(U_1, U_2)}{\sqrt{\text{Var}(U_1)} \sqrt{\text{Var}(U_2)}}. \quad (4)$$

For continuous random variables this is invariant with respect to the two marginal distributions, i.e. it can be expressed as a function of its copula. This property is also known as 'scale invariance'. Note that not all measures of association satisfy this property, e.g. Pearson's linear correlation coefficient [6]. Among all copulas  $C: I^2 \rightarrow I$  such that for every  $u, v \in I$ , three especially noteworthy ones are  $W(u, v) = \max(u + v - 1, 0)$ ,  $\Pi(u, v) = uv$ , and  $M(u, v) = \min(u, v)$ . These copulae correspond to perfect negative association ( $\rho_S(C) = -1$ ), independence ( $\rho_S(C) = 0$ ), and perfect positive association ( $\rho_S(C) = +1$ ) between the two random variables, respectively. For all  $(u, v) \in I^2$  it holds that  $W(u, v) \leq \Pi(u, v) \leq M(u, v)$ .

The tail dependence relationship can be measured by means of the upper and lower tail dependence parameters

$$\lambda_u = \lim_{u \rightarrow 1^-} P[X_2 > F_2^{-1}(u) | X_1 > F_1^{-1}(u)] = \lim_{u \rightarrow 1^-} \frac{C(u, u)}{u}, \quad (5)$$

$$\lambda_l = \lim_{u \rightarrow 0^+} P[X_2 \leq F_2^{-1}(u) | X_1 \leq F_1^{-1}(u)] = \lim_{u \rightarrow 0^+} \frac{1 - 2u + C(u, u)}{1 - u}. \quad (6)$$

If  $\lambda_u \in (0, 1]$  or  $\lambda_l \in (0, 1]$ , the random variables  $X_1$  and  $X_2$  present upper or lower tail dependence. If  $\lambda_u = 0$  or  $\lambda_l = 0$ , there is no upper or lower tail dependence. These parameters measure the dependence in the tails of the joint distribution, i.e. high/low values of one variable are associated with high/low values of the other one. They represent the probability that one variable is extreme given that the other is extreme. The Spearman grade correlation coefficient and both tail dependence parameters are directly associated with the parameters of some copula family [14].

### 3 Our proposal

Theory and methodology for exploratory factor analysis have been well developed for continuous variables, but in practice observed or measured variables are often ordinal.

Observations on an ordinal variable are assumed to have logical ordering categories. This logical ordering is typical when data are collected from questionnaires. A good example is the Likert Scale that is frequently used in survey research: 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Neutral*, 4 = *Agree*, and 5 = *Strongly agree*. Although a question is designed to measure a theoretical concept, the observed responses are only a discrete realization of a small number of categories and distances between categories are unknown. Following [13], [9] and others, it is assumed that there is a continuous variable  $x_{i*}$  underlying the ordinal variable  $x_i$ ,  $i = 1, \dots, p$ . This continuous variable  $x_{i*}$  represents the attitude underlying the order responses to  $x_i$  and it is assumed to have a range from  $-\infty$  to  $+\infty$ .

The underlying variable  $x_{i*}$  is unobservable. Only the ordinal variable  $x_i$  is observed. For an ordinal variable  $x_i$  with  $m_i$  categories, the connection between the ordinal variable  $x_i$  and the underlying variable  $x_{i*}$  is:

$$x_i \Leftrightarrow \tau_{i-1}^i < x_i^* < \tau_i^i, \quad i = 1, 2, \dots, m_i \tag{7}$$

where

$$-\infty = \tau_0^i < \tau_1^i < \tau_2^i < \dots < \tau_{m_i-1}^i < \tau_{m_i}^i = +\infty \tag{8}$$

are threshold parameters. For variable  $x_i$  with  $m_i$  categories, there are  $m_i - 1$  strictly increasing threshold parameters  $\tau_1^i < \tau_2^i < \dots < \tau_{m_i-1}^i$ .

Let  $x_i$  and  $x_j$  be the two ordinal variables with  $m_i$  and  $m_j$  categories respectively. We define now Spearman's grade correlation via copula. We consider a copula  $C_\theta$  associated with each pair  $(X_i^*, X_j^*)$  underlying the pair  $(X_i, X_j)$  in the set of ordinal items  $X_1, X_2, \dots, X_i$ , we thus assume that each pair  $(X_i, X_j)$  corresponds to a bivariate discrete random variable obtained by a discretisation of a bivariate continuous latent variable  $U_i = F(X_i^*), U_j = F(X_j^*)$  with support on the unit interval.

Let  $A_{ij} = [u_{i-1}, u_i] \times [v_{j-1}, v_j]$ ,  $i = 1, 2, \dots, m_i$   $j = 1, 2, \dots, m_j$ , be the rectangles defining the discretisation. Let  $p_{11}, \dots, p_{m_i m_j}$  be the joint probabilities of the ordinal variables corresponding to the rectangles  $A_{11}, \dots, A_{m_i m_j}$ . Let  $V_{C_\theta}(A_{11}, \dots, A_{m_i m_j})$  be the volumes of the rectangles under the copula  $C_\theta$ , then

$$V_{C_\theta}(A_{11}, \dots, A_{m_i m_j}) = p_{11}, \dots, p_{m_i m_j} \tag{9}$$

There exists a unique element in the family of copula for which (9) holds true. We apply this to each pair  $(X_i, X_j)$   $i \neq j$  in the set of the items.  $\theta$  can be estimated via maximum likelihood [5] [12]. The multivariate normality assumption pertaining to the underlying variables, assumed by polychoric correlation and Pearson correlation, is relaxed. To apply the index one needs only to specify the dependence structure of the variables by means of a copula family.

In this way the construct validity is analysed according to ordinal data obtained from Likert scales using the most suitable method. The factor results show a better fit to the theoretical model when the factorization is carried out using the Spearman's grade correlation via copula rather than Pearson correlation. Our focus here has been to identify the type of correlation that yields a factor solution more in keeping with the original measurement model, as we believe this to have great importances in terms of drawing correct substantive conclusions. When we conduct a FA our results can be summarized as follow:

- regardless of the number of dimensions and items with skewness, Pearson correlations are lower than Spearman's grade correlations. The results are more significant when all items are asymmetric.
- The model obtained is more consistent with the original measurement model when we factorize using the Spearman's grade correlation. This result does not depend on the number of dimensions and asymmetric items.

To summarize the factor results obtained when we use Spearman's grade correlation better reproduce the measurement model present in the data, regardless of the number of factors.

## References

1. Anderson, T.W.: An introduction to multivariate statistical analysis. Wiley, New York (2003)
2. Cherubini, U., Luciano, E., Vecchiato, W.: Copula methods in finance. John Wiley & Sons (2004)
3. DeVellis, R.F.: Scale development, theory and applications. Sage, Newbury Park (1991)
4. DiStefano, C.: The impact of categorization with confirmatory Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal* **9**, 327–346 (2002)
5. Ekstrom, J.: Contributions to the Theory of measures of association for ordinal variables. *Digital Comprehensive Summaries of Uppsala Dissertation from the Faculty of Social Sciences, Uppsala* (2003)
6. Embrechts, P., McNeil, A., Straumann, D.: Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 176–223 (2002)
7. Holgado-Tello, F.P., Chacón-Moscoso, S., Barbero-García, I., Vila-Abad E.: Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity* **44**, 153–166 (2010)
8. Joe, H.: Multivariate models and multivariate dependence concepts. CRC Press (1997)
9. Jöreskog, K. G.: New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity* **24(4)**, 387–404 (1990)
10. Likert, R.: A technique for the measurement of attitudes. *Archives of Psychology*, 44–45 (1932)
11. Likert, R., Sydney, R., Murphy, G.: A simple and reliable method of scoring Thurstone attitudes scales. *The journal of Social Psychology*, 228–238 (1934)
12. Martinson, E.O., Hamdan, M.A.: Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables. *J. Stat. Comput. Simul.* **1**, 45–54 (1971)
13. Muthén, B. O.: Full maximum likelihood analysis of structural equation models with polytomous variables. *Psychometrika* **9(1)**, 91–97 (1984)
14. Nelsen, R.B.: An introduction to copulas. Springer Science & Business Media (2013)
15. Spector, P.E.: Summating rating scale construction: an introduction. Sage, Newbury Park (1992)

# IPUMS Data for describing family and household structures in the world.

## *I dati IPUMS per la descrizione della struttura delle famiglie nel mondo.*

Fausta Ongaro, Silvana Salvini

**Abstract** Our research focuses on the change of the characteristics of families and households in developing (especially sub-Saharan countries) and developed countries (especially European countries) in the last twenty years. The choice of countries depends on the available variables in the different data sets.

Data used refer to IPUMS data base. IPUMS-International is dedicated to collecting and distributing census data from around the world. The database currently describes approximately 614 million persons recorded in 277 censuses taken from 1960 to the present. The database includes censuses from 82 countries. The data series includes information on a broad range of population characteristics, both at household and individual level, including fertility, mortality, occupational structure, education, ethnicity, and household composition. These last data are at the core of our contribution. The information available in each sample varies according to the questions asked in every census (<https://international.ipums.org/international/>).

**Abstract** *La nostra ricerca si focalizza sul cambiamento delle caratteristiche delle famiglie nei paesi in via di sviluppo (in particolare i paesi dell'Africa sub-Sahariana) e nei paesi sviluppati (in particolare quelli europei) negli ultimi 20 anni. La scelta dei paesi dipende sostanzialmente dalle variabili disponibili nei diversi data set. I dati utilizzati fanno parte del data base IPUMS dedicato alla raccolta e alla diffusione dei dati censuari dei diversi paesi del mondo. Il data base contiene approssimativamente 614 milioni di record riferiti a 277 censimenti dal 1960 ad oggi di 82 paesi. Le informazioni riguardano un grande numero di caratteristiche della popolazione, sia a livello familiare sia a livello individuale, sulla fecondità, la mortalità, la struttura occupazionale, l'istruzione e l'etnia, oltre alle informazioni sulla composizione familiare. E' su quest'ultima serie di dati che concentreremo la nostra attenzione per l'analisi.*

**Key words:** Census data, Family structure, Household characteristics, Developed and developing countries.

## 1 Introduction

Our research focuses on the change of the characteristics of families and households in developing (especially sub-Saharan countries; see Bongaarts, 2001 and Randall et al., 2015) and developed countries (especially European countries; see Keilman, 2006) in the last twenty years. The choice of countries depends on the available variables in the different data sets.

Data used refer to IPUMS data base. IPUMS-International is dedicated to collecting and distributing census data from around the world. The database includes censuses from 82 countries. The data series includes information on a broad range of population characteristics, both at household and individual level, including fertility, mortality, occupational structure, education, ethnicity, and household composition. These last data are at the core of our contribution. The information available in each sample varies according to the questions asked in every census (<https://international.ipums.org/international/>).

## 2 Data used

Micro-data will be used to analyze countries' regions using NUTS for Europe and other territorial subdivisions for developing countries. Methods used to detect similarities and dissimilarities of regions are represented by cluster analysis and other multivariate techniques apt to analyze large data sets.

The database currently describes approximately 614 million persons recorded in 277 censuses taken from 1960 to the present. The database we use includes censuses from both European and sub-Saharan countries (Austria, Belarus, Burkina Faso, Cameroon, Ethiopia, France, Germany, Ghana, Guinea, Greece, Hungary, Ireland, Italy, Kenya, Liberia, Malawi, Mali, Mozambique, Netherlands, Nigeria, Portugal, Romania, Rwanda, Senegal, Sierra Leone, Slovenia, South Africa, South Sudan, Spain, Sudan, Switzerland, Tanzania, Uganda, United Kingdom, and Zambia). For most of these countries, more than one census is present.

Most population data — especially census data — have traditionally been available only in aggregated tabular form. IPUMS-International is composed of microdata, which means that it provides information about individual persons and households. Since this data base includes most of the information originally recorded by census, users can construct a great variety of tabulations interrelating any desired set of variables and perform models directly using these variables.

## 3 Methods

After a preliminary analysis of the definition of household (and of the meaning of child and parent) which are used in the different censuses, individual (especially,

sex, age, family relationship) and household (especially, n. components, n. unrelated persons, n. families) variables are used to build macro variables able to describe the family and household structures in the countries of sub-Saharan region and in European Union.

We will use the macro-data information to conduct cluster analyses on family and household macro data at a sub-national geographic level of both Europe and sub-Saharan Africa to examine the regions that show similar characteristics and trends at family level (WenYang Yu et al., 2015).

## 4 Preliminary results

Descriptive analysis shows the following distributions of individuals and the mean number of persons in the households classified by country.

Large differences emerge between the two groups of countries (UN, 2004): European regions generally show a low number of person per household and, on the contrary, sub-Saharan countries present a mean number of person per household very high. In particular, we note the high value in Senegal, but also Guinea, Sierra Leone and Burkina Faso show mean number of persons per household higher than 8. On the contrary, in Europe the dimension of families is very lower, and Germany in particular shows a value lower than 3. Many countries, such as Italy and France, present values on a little higher.

**Table 1 - Number of persons, mean number of persons per household and standard deviation by country**

<b>Europe</b>			
Country	Mean	Nb.	St. Dev.
Austria	3.49	3929934	2.035
Belarus	3.26	990706	1.398
France	3.29	55880084	2.486
Germany	2.86	14623488	2.014
Greece	3.79	3749350	1.589
Hungary	3.54	2079868	1.792
Ireland	4.32	3377884	2.211
Italy	3.23	2990739	1.336
Portugal	3.72	2029940	1.797
Romania	3.89	6313566	1.809
Slovenia	3.47	179632	1.288
Spain	2.98	10162418	1.726
Switzerland	3.30	1337224	1.699
Ukraine	3.40	4889288	1.753

<b>sub-Saharan Africa</b>			
Cameroon	7.92	3406084	5.726
Ethiopia	5.86	15882990	2.597
Ghana	7.26	5669774	5.057
Guinea	9.74	1186908	6.869
Kenya	5.61	8016659	4.186
Liberia	5.25	498313	4.271
Malawi	5.97	3132039	4.627
Mali	8.58	3228570	5.417
Mozambique	5.79	3598565	3.013
Nigeria	6.39	426395	3.356
Rwanda	5.99	1586310	2.629
Senegal	13.32	1694761	8.313
Sierra Leone	8.56	494298	5.452
South Africa	5.34	12813070	3.132
South Sudan	7.59	542765	4.279
Sudan	6.90	5066530	3.336
Uganda	6.62	4045909	3.496
Tanzania	6.67	6043159	3.887
Burkina Faso	8.95	3383667	5.200
Zambia	7.26	3105551	3.857

## References

- Bongaarts J., (2001), Household Size and Composition in the Developing World, No. 144, <http://www.eldis.org/vfile /upload/1/document/0708/DOC9224.pdf>, consulted 5th march 2017.
- Keilman N. (2006), Households and Families: Developed Countries, [https://www.researchgate.net/publication/250928439\\_Households\\_and\\_Families\\_Developed\\_Countries](https://www.researchgate.net/publication/250928439_Households_and_Families_Developed_Countries), consulted 5th march 2017.
- IPUMS international data base, in <https://international.ipums.org/international/>, consulted 5th march 2017.
- Randall S, Coast E., Antoine P., Compaore N., Dial F., Fanghanel A., Gning S., Gnoumou Thiombiano B., Golaz V., and Ojiambo Wandera S. (2015), UN Census "Households" and Local Interpretations in Africa Since Independence, SAGE Open April-June 2015: 1–18.
- WenYang Yu, YuBing Yang, and XianWei Wu (2015) Technique of Cluster analysis in Data mining, International Conference on Applied Science and Engineering Innovation (ASEI 2015), Atlantis Press. [www.atlantis-press.com/php/download\\_paper.php?id=25836624](http://www.atlantis-press.com/php/download_paper.php?id=25836624).
- United Nations, (2004), Department of Economic and Social Affairs, Statistics Division, Demographic and Social Statistics Branch, United Nations Demographic Yearbook review, National reporting of household characteristics, living arrangements and homeless households - Implications for international recommendations, ESA/STAT/2004/6.



# Topological Summaries for Time-Varying Data

## *Sintesi Topologiche per Serie Storiche*

Tullia Padellini and Pierpaolo Brutti

**Abstract** Topology has proven to be a useful tool in the current quest for "insights on the data", since it characterises objects through their connectivity structure, in an easy and interpretable way. More specifically, the new, but growing, field of TDA (Topological Data Analysis) deals with Persistent Homology, a multiscale version of Homology Groups summarized by the Persistence Diagram and its functional representations (Persistence Landscapes, Silhouettes etc). All of these objects, however, are designed and work only for static point clouds. We define a new topological summary, the Landscape Surface, that takes into account the changes in the topology of a dynamical point cloud such as a (possibly very high dimensional) time series. We prove its continuity and its stability and, finally, we sketch a simple example.

**Abstract** *A causa della crescente complessità dei dati, diventa sempre più importante riuscire a sintetizzarli attraverso un numero ridotto di caratteristiche interpretabili. Lo studio delle invarianti topologiche si è dimostrato utile in questo senso, in quanto caratterizza un oggetto in termini della sua struttura di connettività. In particolare, lo studio della topologia dei dati viene condotto a partire da una versione multiscala dei gruppi omologici detti gruppi di omologia persistente, rappresentati da oggetti come il diagramma di persistenza, che rappresenta i generatori di tali gruppi, e le sue trasformazioni in spazi di funzioni. In questo lavoro introduciamo un nuovo strumento, costruito per studiare l'evoluzione delle caratteristiche topologiche di serie storiche multidimensionali, la "Landscape Surface". Dopo averne provato continuità e stabilità, accenneremo ad una sua applicazione in un semplice esempio.*

**Key words:** Persistent Homology, Time Series, Topological Inference

---

Tullia Padellini

Sapienza, Università di Roma, Piazzale Aldo Moro 5, e-mail: tullia.padellini@uniroma1.it

Pierpaolo Brutti

Sapienza, Università di Roma, Piazzale Aldo Moro, 5 e-mail: pierpaolo.brutti@uniroma1.it

## 1 Introduction to TDA

As we are dealing with increasingly complex data, our need for characterising them through a few, interpretable features has grown considerably. In recent years there has been quite some interest in the study of the “shape of data” [2]. Among the many ways a “shape” could be defined, topology is the most general one, as it describes an object in terms of its connectivity structure: connected components (topological features of dimension 0), cycles (features of dimension 1) and so on. There is a growing number of techniques (generally denoted as *Topological Data Analysis*) aimed at estimating the shape of a point-cloud through some topological invariant. In this work we extend those techniques to the case of multivariate time series, i.e. when, rather than considering only one point-cloud, we are dealing with a collection of point-clouds indexed by time, as for example in animal migration, player tracking in sports, EEG signals and most spatio-temporal data; our goal is to summarize in one object not only the shape of the data at each fixed time, but also how this shape changes with time.

Before introducing new objects, it is worth briefly reviewing what Topological Data Analysis (TDA) is, and how can we estimate the topology of data, or, to be more precise, the topology of the space  $\mathcal{M}$  data was sampled from. As a matter of fact, data itself, when in the form of a point cloud  $\mathbb{X} = \{X_1, \dots, X_n\}$ , has a trivial topological structure, consisting of as many connected components as there are observations and no higher dimensional features. The basic idea in the TDA is thus to use data to build “shape aware” estimates of  $\mathcal{M}$  and then compute topological invariants. One of the most common way of estimating  $\mathcal{M}$ , in TDA, is *Devroye-Wise support estimator*  $\widehat{\mathcal{M}}_\varepsilon$  built by centering a ball of fixed radius  $\varepsilon$  in each of the observations  $X_i$ , i.e.

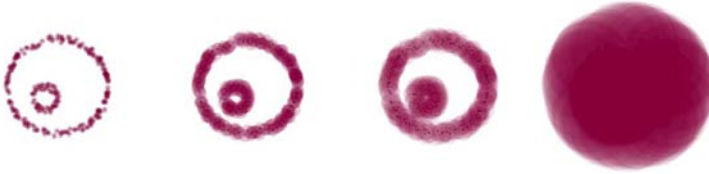
$$\widehat{\mathcal{M}}_\varepsilon = \bigcup_{i=1}^n B(X_i, \varepsilon)$$

where  $B(Y, \delta)$  denotes a ball of radius  $\delta$  and center  $Y$ . For each value  $\varepsilon$  we obtain a different estimate  $\widehat{\mathcal{M}}_\varepsilon$ , whose topology can be recovered by computing its Homology Groups. Persistent Homology, a multiscale version of Homology, then allows us to analyze how those Homology Groups change with  $\varepsilon$ .

Persistent Homology Groups can be summarized by the *Persistence Diagram*, a multiset  $D = \{(b_i, d_i), i = 1, \dots, m\}$  whose generic element  $(b_i, d_i)$  is the generator of the  $i$ -th Persistent Homology group. The space of persistence Diagrams  $\mathcal{D}$  is a metric space, when endowed with the *Bottleneck distance*, which, given two multisets  $A$  and  $B$ , is defined as

$$d_B(A, B) = \inf_{\gamma} \sup_{x \in A} \|x - \gamma(x)\|_\infty$$

where the infimum is taken over all bijections  $\gamma: A \rightarrow B$ .



**Fig. 1**  $\widehat{\mathcal{M}}_\varepsilon$  for different values  $\varepsilon$ . For small values of  $\varepsilon$  (left), the topology of  $\widehat{\mathcal{M}}_\varepsilon$  is close to the one of the point cloud itself. As  $\varepsilon$  grows more and more points start to be connected, until eventually (right) the corresponding  $\widehat{\mathcal{M}}_\varepsilon$  is homeomorphic to a point. Values  $\varepsilon_b, \varepsilon_d$  of  $\varepsilon$  corresponding to when two components are connected for the first time (*birth-step*) and when they are connected to some other larger component (*death-step*) are the generators of a Persistent Homology Group.

The Bottleneck distance allows us to compare Persistence Diagrams and to define their most important property: *stability* [4].

**Theorem 1.** *Let  $\mathbb{X}, \mathbb{Y}$  two point clouds, and  $D_{\mathbb{X}}, D_{\mathbb{Y}}$  their corresponding Persistence Diagrams, then*

$$d_B(D_{\mathbb{X}}, D_{\mathbb{Y}}) \leq 2d_H(\mathbb{X}, \mathbb{Y})$$

where  $d_H(A, B)$  is the Hausdorff distance between two topological spaces  $A$  and  $B$ .

Roughly speaking, this means that if two point clouds are similar, then their Persistence Diagrams will be as well, and is therefore instrumental for using them in statistical tasks such as classification or clustering.

Since Persistence Diagrams are general metric objects, it is usually advisable to transform them in order to work with more statistics-friendly spaces. The most famous transformations of the persistence diagram are the persistence landscape [1] and the persistence silhouette[3], which are functions built by mapping each point  $z = (b_i, d_i)$  of a Persistence Diagram  $D$  to a piecewise linear function called the “triangle” function  $T_z$ , defined as

$$T_z(y) = (y - b_i + d_i) \mathbb{1}_{[b_i - d_i, b_i]}(t) + (b_i + d_i - y) \mathbb{1}_{(b_i, b_i + d_i]}(y)$$

where  $\mathbb{1}_A(x) = 1$  if  $x \in A$  and  $\mathbb{1}_A(x) = 0$  otherwise. Informally a triangle function links each point of the diagram to the diagonal with segments parallel to the axes, which are then rotated of 45 degrees.

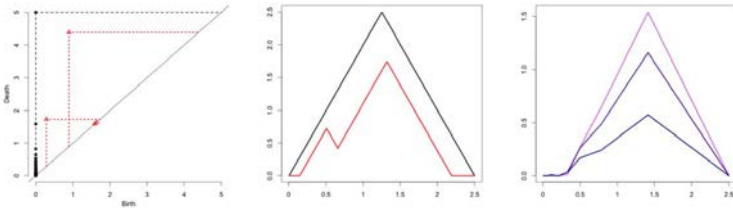
The blocks  $T_z$  can be combined in many different ways. If we take their  $k$ max, i.e. the  $k$ -th largest value in the set  $T_z(y)$ , we obtain the *Persistence Landscape*

$$\lambda_D(k, y) = k \max_{z \in D} T_z(y) \quad k \in \mathbb{Z}^+.$$

The persistence landscape is the collection of functions  $\lambda_D(k, y)$ . If we take the weighted average of the functions  $T_z(y)$ , we have the *Power Weighted Silhouette*

$$\phi_p(t) = \frac{\sum_{z \in D} w_z^p T_z(y)}{\sum_{z \in D} w_z^p}.$$

Although we are losing some information in going from Persistence Diagrams



**Fig. 2** Persistence Diagram (left), Persistence Landscape (center) and Persistence Silhouette for different values of  $p$  (right) of the data shown in Fig. 1

to Persistence Landscapes, the main result we had for Diagrams, i.e. stability, still holds [1].

**Theorem 2.** Let  $\mathbb{X}, \mathbb{Y}$  two point clouds,  $D_{\mathbb{X}}, D_{\mathbb{Y}}$  their corresponding Persistence Diagrams, and  $\lambda_{\mathbb{X}}, \lambda_{\mathbb{Y}}$  their corresponding Persistence Landscapes, then

$$d_{\Lambda}(\lambda_{\mathbb{X}}, \lambda_{\mathbb{Y}}) \leq d_B(D_{\mathbb{X}}, D_{\mathbb{Y}}) \leq 2d_H(\mathbb{X}, \mathbb{Y})$$

where  $d_{\Lambda}(\lambda_{\mathbb{X}}, \lambda_{\mathbb{Y}}) = \|\lambda_{\mathbb{X}} - \lambda_{\mathbb{Y}}\|_{\infty}$  is the  $L^{\infty}$  distance in the space of Persistence Landscapes.

## 2 The Landscape Surface

In order to study the evolution of the topological structure of time-varying data, we think of a multidimensional Time series  $\mathbb{X}(t)$  as a dynamic point cloud; for every fixed time  $t$  we can use the tools we have previously defined and build a Persistence Diagram  $D(t)$ , Landscape  $\lambda_{\mathbb{X}(t)}(k, y)$  and Silhouette. Intuitively we can consider this Persistence Landscape  $\lambda_{\mathbb{X}(t)}$  as a function of time  $t$  as well, which means that we can work with a surface, rather than just a curve. It is important to notice that although in the following we focus on Landscapes, the same results hold for Silhouettes as well.

**Definition 1.** Given a dynamic point cloud  $\mathbb{X}(t)$  we define the *Landscape Surface* as the function

$$\Lambda(t, k, y) = \lambda_{\mathbb{X}(t)}(k, y) \quad \forall t, k, y.$$

This surface is still a meaningful topological summary, as we can prove its stability.

**Theorem 3.** Let  $\{\mathbb{X}(t), \mathbb{Y}(t)\}$  with  $t \in (0, 1)$  two continuous dynamic point clouds,  $\Lambda_{\mathbb{X}}$  and  $\Lambda_{\mathbb{Y}}$  their corresponding Landscape Surfaces, then:

1.  $\Lambda_{\mathbb{X}}$  and  $\Lambda_{\mathbb{Y}}$  are continuous;
2.  $I_{\Lambda}(\Lambda_{\mathbb{X}}, \Lambda_{\mathbb{Y}}) \leq I_H(\mathbb{X}, \mathbb{Y})$

where  $I_{\Lambda} = \int_0^1 d_{\Lambda}(\lambda_{\mathbb{X}(t)}, \lambda_{\mathbb{Y}(t)})dt$  is the Integrated  $L^{\infty}$  distance on the space of Persistence Landscapes and  $I_H(\mathbb{X}, \mathbb{Y}) = \int_0^1 d_H(\mathbb{X}(t), \mathbb{Y}(t))dt$  is the Integrated Hausdorff distance for dynamic pointclouds.

The proof is a direct consequence of the Stability Theorem for Persistence Landscapes (Theorem 2), in fact:

1. For a fixed  $t$ , consider  $\mathbb{X}(t)$  and  $\mathbb{X}(t + \varepsilon)$  (same applies for  $\mathbb{Y}$ ). By Theorem 2 and the continuity of  $\mathbb{X}(t)$  we have

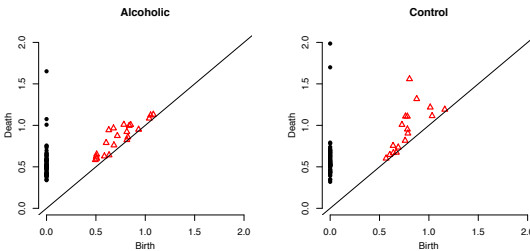
$$0 \leq \lim_{\varepsilon \rightarrow 0} d_{\Lambda}(\lambda_{\mathbb{X}(t)}, \lambda_{\mathbb{X}(t+\varepsilon)}) \leq \lim_{\varepsilon \rightarrow 0} 2d_H(\mathbb{X}(t), \mathbb{X}(t + \varepsilon)) = 0.$$

2. Since for a fixed  $t$  we have, by Theorem 2 we have

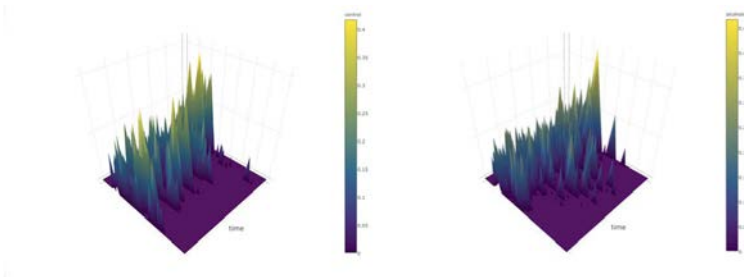
$$d_{\Lambda}(\lambda_{\mathbb{X}(t)}, \lambda_{\mathbb{Y}(t)}) \leq 2d_H(\mathbb{X}(t), \mathbb{Y}(t))$$

integrating both terms is enough to prove the result.

In order to show an example of this object with real data, we consider EEG data, which are signals recorded at a very high frequency through many different electrodes (64 in our case). We build the Persistence Surface using EEG signals from an alcoholic and a control patient, both under the same stimuli. As we can clearly see from Fig. 3 and 4 these two subjects show a very different behavior. While the signal from the control patient is strongly characterized by a few persistent features, in the alcoholic patient there is less structured, as there are many features but they all have a smaller persistence, and could therefore be interpreted as noise.



**Fig. 3** Persistence Diagram of the Alcoholic and Control subjects for a fixed time  $t$ .



**Fig. 4** Landscape Surface of dimension 1 for the EEG signal of a control patient (top) and an alcoholic (bottom)

## References

1. Bubenik, P.: Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, **16**(1), 77–102 (2015)
2. Carlsson, G.: Topology and data. *Bulletin of the American Mathematical Society*, **46**(2), 255–308 (2009)
3. Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L.: Stochastic convergence of persistence landscapes and silhouettes. In *Proceedings of the thirtieth annual symposium on Computational geometry*, ACM (2014)
4. Cohen-Steiner, D., Edelsbrunner, H., and Harer, J.: Stability of persistence diagrams. *Discrete and Computational Geometry* **37**(1), 103–120 (2007)
5. Edelsbrunner, H., Letscher, D., Zomorodian, A.: Topological persistence and simplification. *Discrete and Computational Geometry*, **28**(4), 511–533 (2002)
6. Munch, E.: Applications of persistent homology to time varying systems. *Diss. Duke University* (2013).

# Modeling of Complex Network Data for Targeted Marketing

## *Modellazione di Dati di Rete Complessi per il Marketing Mirato*

Sally Paganin

**Abstract** Developing strategies for targeted advertising of existing customers is a common goal in many business sectors, with usual practice focused on identifying shared acquisition patterns of products based on ownership data. We observe customers' behavior for multiple agencies within the same company, monitoring choices of specific products along with co-subscription networks representing multiple purchases. Our aim is to exploit co-subscription networks to efficiently inform targeted advertising of cross-sell strategies to currently mono-product customers. We address this goal by developing a Bayesian joint model for mixed domain data which adaptively clusters agencies characterized by a similar customer base, exploiting a cluster-dependent mixture of latent eigenmodels to describe multi-purchase networks. An application to data from the insurance market is presented.

**Abstract** *Lo sviluppo di strategie per la pubblicità mirata rivolta ai clienti esistenti è un obiettivo comune a diversi settori del mercato, in cui la prassi è quella di identificare modelli comuni di acquisto dei prodotti sulla base di dati di possesso. Si è osservato il comportamento dei clienti in diverse agenzie all'interno della stessa compagnia, monitorando le scelte riguardanti prodotti specifici e reti di sottoscrizione rappresentanti gli acquisti multipli. Lo scopo è quello di sfruttare le reti di sottoscrizione per individuare in maniera efficiente quali pubblicità mirate rivolgere ai clienti mono-prodotto correnti. Per raggiungere tale obiettivo, si propone un modello congiunto bayesiano per dati di natura mista in grado di raggruppare agenzie caratterizzate da una base clienti simile, utilizzando una mistura di modelli a distanze latenti dipendente dal gruppo per descrivere reti di acquisto multiplo. Si presenta un'applicazione a dati provenienti dal mercato assicurativo.*

**Key words:** Business Intelligence; Co-clustering; Cross-sell Marketing Strategies; Mixed domain data;

---

Sally Paganin

Department of Statistical Sciences, University of Padua, Via C. Battisti 241, Padova, Italy, e-mail: paganin@stat.unipd.it

## 1 Introduction

Business statistics is becoming more and more focused on developing new tools for the definition of cross-selling campaigns, targeting existing customers instead of attracting new ones. Adding value to the current customer base has been proved to be an efficient strategy for the growth of the company, enhancing customer retention by increasing the switching costs. For this reason, mono-product customers purchasing a single product from a company represent a key segment of the customer base, and companies are growing interest in expanding these customers purchases to additional products.

Current methods focus on identifying shared acquisition patterns of products, based on customer ownership data sometimes along with additional data such as demographics records or survey responses, aiming to provide some measure of product subscription propensity [3]. Even if such methods can lead to useful insight about the customers purchasing behavior, they are usually limited to provide analysis for a single portfolio while many companies possess dislocated agencies all over a territory. In such settings, efficient targeting of the customer and differentiation of the advertising may lead to better profits despite the higher costs. We propose a Bayesian joint model for mixed domain data which clusters agencies characterized by a similar composition of their mono-product customer choices as well as a comparable multi-purchase behavior. We built on the model presented in [2] providing a more general setting. In the next two sections we present the modeling framework, while in Section 4 an application to real data is discussed.

## 2 Definition of cross-sell strategies

Customers can be distinguished in mono and multi-product customers, having purchased one or multiple products among a number of  $V$  products. Let  $y_{is} \in \{1, \dots, V\}$  denote the product subscribed from a mono-product customer  $s$ ,  $s = 1, \dots, n_i$  within agency  $i$  for  $i = 1, \dots, n$ . We represent multi-purchase behavior in each agency  $i$  as a co-subscription network, described via a  $V \times V$  adjacency matrix  $A_i$ , with  $A_{i[uv]} = A_{i[vu]} = 1$  meaning that the product  $u$  and product  $v$  are subscribed together, and  $A_{i[uv]} = A_{i[vu]} = 0$  otherwise.

Definition of the edges between pairs of products may depend on the company requirements or, in absence of those, some threshold criteria to be fixed. In our application we defined an edge between two products  $v$  and  $u$  if the number of customers subscribing to both products exceeds the 10% of the total number of multi-product customers subscribed to at least one of the two. Hence a presence of an edge between two products suggests a preference of customers in agency  $i$  for that specific pair, controlling for the total number of multi-product customers subscribed to at least one of the two products.

We may exploit co-subscription networks in each agency to estimate the propensity of customer who subscribed to product  $v$  to additionally buy  $u \neq v$ , and pair each



product  $v = 1, \dots, V$  with the one, say  $u_{iv}$ , that maximizes such propensity as the best choice to offer to a customer who has already product  $v$ . However, in order to define an effective cross-selling strategy, it is important to take in consideration also the proportion of mono-product customer that would be targeted from that strategy. We denote such proportion as  $p_{iv} = pr(\mathcal{A}_i = v)$ , with  $\mathcal{A}_i$  being the random variable describing the choices of the mono-product customer in agency  $i$ . Depending on  $p_{iv}$ , strategy  $u_{iv}$  may target a small proportion of existing customers, resulting to be less efficient than a strategy characterized by a lower estimate of subscription propensity but targeting a wider portion of the customer base. To take in account the role of  $p_{iv}$ , we associate each strategy  $u_{iv}$  with a performance indicator  $e_{iv} = p_{iv}u_{iv}$  with  $u \neq v$  for  $i = 1, \dots, n$ . Strategies with a high  $e_{iv}$  will target a sizable proportion of the available mono-product customers in agency  $i$  with advertising for a new product likely to be appealing to them.

Since we observe data for agencies belonging to same company it is reasonable to expect them to share some pattern in the mono and multi-purchasing behavior. For groups of agencies having sufficiently similar customer bases, an identical strategy can be adopted to reduce administrative costs without decreasing effectiveness. On the basis of such assumptions, we introduce a clustering underlying mechanism in the modeling framework; efficient detection of clusters allows adaptive reduction of the total number of strategies from  $n$  to  $K < n$ .

We address this goal by proposing a Bayesian hierarchical joint model for the data  $\{(y_{i1}, \dots, y_{iV}), A_i\}$  for  $i = 1, \dots, n$  which characterizes the distribution across agencies of the mono-product customer subscriptions along with the co-subscription network for multi-product customers. The model is chosen to be flexible while automatically clustering agencies that have similar mono-product customer choices and co-subscription network profiles.

### 3 Model

Let  $G = (G_1, \dots, G_n)$  be the vector of cluster assignments, with  $G_i \in \{1, \dots, K\}$  indicating the cluster membership of agency  $i$ . Conditional to the cluster we provide a cluster-specific probabilistic representation of the mono-product customer choices, as well as a cluster-specific probabilistic generative mechanism underlying the co-subscription networks.

Let  $p_k = (p_{k1}, \dots, p_{kV})$  be a cluster-specific probability vector, with  $p_{kv}$  indicating the probability of subscription to the product  $k$  for a mono-product customer in an agency belonging to cluster  $k$ . Assuming independence of the mono-customer choices, the joint probability for the mono-product data given the cluster assignment is

$$p(\mathcal{A}_k = y_{i1})p(\mathcal{A}_k = y_{i2}) \cdots p(\mathcal{A}_k = y_{iV}) = \prod_{v=1}^V p_{kv}^{n_{iv}} \tag{1}$$

with  $n_{iv}$  the number of customers in agency  $i$  that subscribed to product  $v$ .

In defining a conditional model for the co-subscription networks within each cluster we leverage the probabilistic framework from [1] in which a flexible Bayesian model for a population of networks is provided via a mixture of latent eigenmodels. We refer to the original paper for details about the model and related prior specification. We complete this last by considering a conjugate Dirichlet prior distribution for mono-product choices probability and a nonparametric prior for the cluster assignment vector  $G$ .

In particular we consider a Chinese Restaurant representation of the Pitman-Yor process (CRP-PY) [5] with discount parameter  $d \in (0, 1)$  and concentration parameter  $\alpha > -d$ . Under such representation the conditional prior probability of allocating an observation  $i$  to one of the already existing clusters is  $\frac{n_{k,-i}-d}{\alpha+n-1}$  where  $n_{k,-i}$  is the number of observations in cluster  $k$  excluding the  $i$ th one. Instead the conditional prior probability of creating a new cluster is  $\frac{\alpha+dK^-}{\alpha+n-1}$  with  $K^-$  the total number of nonempty clusters after removing the  $i$ th observation. Setting  $d = 0$ , the Pitman-Yor process reduces to the Dirichlet Process with parameter  $\alpha$ , but as the discount parameter increases observations are less likely to be allocated to large clusters and more likely to be allocated to a new ones. The advantage of such more general specification, is that parameter  $d$  can be adapted to the company requirements, as for example penalizing the creation of new cluster in order to reduce administrative costs in advertising a minor number of cross-selling campaigns or, on the contrary, favoring it with the aim to provide more specific advertising.

Posterior computation is available via a simple Gibbs sampler which exploits results in [4] to allocate agencies to clusters under the CRP-PY prior and steps in [1] to update the quantities describing the co-subscription networks.

## 4 Application

We analyzed subscription data provided from  $n = 130$  agencies selling  $V = 15$  products belonging to a company operating in the insurance market. Initialization of network related quantities follows directions in [1], while we center the hyperparameters of the mono-product probabilities around the averaged preferences of mono-product customers in the entire company. We evaluate the clustering behavior under the CRP-PY prior by choosing different values for the hyperparameters  $d$  and  $\alpha$ . In particular we consider values for  $d \in \{0, 0.25, 0.5, 0.75\}$  and pick the corresponding values for  $\alpha$  such that the expected number of clusters a priori is  $t \in \{5, 15\}$ .

In our application the clustering behavior appears to be quite robust, producing a number of clusters varying between 20 and 25 despite of the prior expected number, changing according to the value of the discount parameter  $d$ . Results are characterized by the presence of 4 large groups comprising the 60% of the total number of agencies, with the parameter  $d$  affecting mostly small clusters of 2 or 3 observations.

In computing estimates of cluster-specific cross-selling strategies and associated performance indicators, different clusters result to share similar cross-sell strategies, with minor differences in mono- and multi-product customer profiles highlighted

across different clusters of agencies. This is a reasonable finding since data comes from agencies in the same company, and provides insights on which strategies could be advertised for all the agencies and which specific ones are potentially more profitable.

## References

1. Durante, D., Dunson, D. B., Vogelstein, J. T.: Nonparametric Bayes Modeling of Populations of Networks. *Journal of the American Statistical Association*. (2016) doi: 10.1080/01621459.2016.1219260
2. Durante, D., Paganin, S., Scarpa, B. and Dunson, D. B.: Bayesian modelling of networks in complex business intelligence problems. *J. R. Stat. Soc. C-App.* **66**, 555–580 (2017)
3. Kamakura, W. A., Wedel, M., de Rosa, F. and Mazzon, J. A.: Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *Int. J. Res. Mark.* **20**, 45–65 (2003)
4. Neal, R. M.: Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Stat.* **9**, 249–265 (2000)
5. Pitman, J., Yor, M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **4**, 855–900 (1997)



# Statistical categorization through archetypal analysis

## *Analisi archetipale per la definizione di categorie*

Francesco Palumbo and Giancarlo Ragozini

**Abstract** Human knowledge develops through complex relationships between categories. In the era of the *Big Data*, categorization implies data summarization in a limited number of well-separated groups that must be maximally internally homogeneous at the same time. This proposal exploits archetypal analysis capabilities in finding a set of extreme points that can summarize the entire data set in homogeneous groups. Archetypes are then used to identify the best prototypes according to the Rosch's definition. Finally, in the geometric approach to cognitive science, the Voronoi tessellation based on the prototypes is used to define a categorization. An example on the Forina's et al. well-known wine data set illustrates the procedure.

**Abstract** *La capacità di definire relazioni complesse fra categorie è la base della conoscenza umana. Nell'era dei Big Data la costruzione di categorie passa attraverso la capacità di riassumere i dati in un limitato numero di gruppi ben distinti fra loro e omogenei al loro interno. Questa proposta sfrutta l'analisi archetipale per individuare un insieme di punti estremi in grado di riassumere l'intero dataset in gruppi omogenei. Gli archetipi sono poi utilizzati per identificare i prototipi, secondo l'accezione proposta dalla Rosch. Infine, utilizzando una tassellazione di Voronoi si definisce la categorizzazione rispetto a ciascun prototipo. Un esempio sul data set wine messo a disposizione da Forina et. al. illustra la procedura.*

**Key words:** Categorization, Archetypal Analysis, Prototypes

## 1 Introduction

Knowledge consists basically of categorizations: humans learn new concepts very fast by building complex relationships between a set of complex items or categories.

---

Dept. of Political Sciences, Univerirsità degli Studi di Napoli Federico II  
e-mail: [[francesco.palumbo](mailto:francesco.palumbo)] [[giancarlo.ragozini@unina.it](mailto:giancarlo.ragozini@unina.it)]

Whilst the total number of objects that can be considered should remain limited to five/six, these objects can be described by several features defining a high grade of complexity. Categories are stored in our long-term memory, and it has been demonstrated that we recall the categories in the working memory developing connections among them that improve our knowledge [4]. In other words, few examples of a new concept are often sufficient for us to grasp its meaning. On the contrary, we are overwhelmed by a large amount of data and information. With the explosion of Big Data problems, statistical learning has become a very hot field in many scientific areas as well as marketing, finance, and in other environmental and behavioral disciplines. The huge amount of stored data represents an incredible source of knowledge, providing that they can be summarized in a (small) number of categories that are consistent with the human cognitive capabilities.

In the present paper we parallel the cognitive process of categorization through statistical learning techniques relying on the conceptual spaces framework [14], in which conceptual spaces are geometric structures, and the categorization mainly consists in a partitioning process of the conceptual spaces. The paper is structured in four sections besides this introduction: section 2 discusses that relationship between statistical learning and the construction of a categorization in the cognitive science. Section 3 presents the prototypes identification after the archetypal analysis; through a real data based example, section 4 presents the Voronoi tessellation [26] starting from the prototypes as tool to derive a categorization in the conceptual space; the last section presents some concluding remarks and future possible research directions.

## 2 Statistical learning and cognitive categorization

Statistical and machine learning can significantly speed up the human knowledge development helping to find the basic categories in a relatively short time. Exploratory Data Analysis (EDA) can be considered the forefather of statistical learning: it relies on the mind's ability to learn from the data and, in particular, it aims to summarize datasets through a limited number of interpretable latent features or clusters offering cognitive geometric models to define categorizations. It can also be understood as the implementation of the human cognitive process extended to large or huge amounts of data: the "*Big Data*" [16]. Factorial models belong to the former approach, they permit the representation of the original data into a reduced space by replacing the original variables with a reduced number of linear mixtures of independent components. These methods include principal component analysis (PCA), independent component analysis (ICA) and independent vector analysis, when dealing with multiple datasets. On the other hand, fuzzy and crisp clustering methods allow us to represent each statistical unit as a weighted sum of the means of the groups that minimize the overall model error.

However, EDA itself cannot answer to the questions: "*How many, and what are the categories to retain?*" and "*What are the observations that better than others can be understood and elaborated in the human cognitive processes?*". In cognitive

science, according to Rosch [24, 25], the best is related to the concept of typicality, in other words we must look for those elements that better than others can represent a category. We call these elements *prototypes* and measure their representativeness degree using a distance function to a salient entity of the category [11, 22]. These objects can be observed or unobserved (abstract), and they can be represented by a single value or by interval-valued variables. In many cases, in classification and clustering, and more generally in cognitive sciences, the concept of *prototype* has been unknowingly adopted to synthesize and represent categories [3, 2]. However, dealing with Big Data, the role of prototypes becomes more and more relevant, thus giving rise to a wide variety of studies in the literature on prototype-based clustering methods (see [17, Chapt 13]).

Identifying groups that can be connected to a related prototype does not fulfill the categorization process. Without any proper description, prototypes cannot be advantageous to learning. D’Esposito et al. (2012, 2013) [6, 7] and Ragozini et al. (2016) [22] considered the archetypal analysis, as proposed by Cutler and Breiman [5], to identify the prototypes in a geometric view. According to the idea of symbolic object [9], in [7] D’Esposito et al. (2013) proposed the prototype description in terms of symbolic objects. The present proposal grounds on the conceptual space framework and starting by the geometric properties of the proposed prototypes, exploits the Voronoi tessellation to obtain a data-driven categorization, i.e. a partition of the conceptual space in convex regions centered on the prototypes.

### 3 Prototype identification

In statistical literature, numerical techniques to find prototypes in a given multivariate dataset have been proposed and are based on several different criteria. The most widely used techniques are generally based on non-hierarchical clustering algorithms [8, 18]. However, in this proposal we present some recent results on the prototypes definition through the archetypal analysis. Archetypal analysis (AA), was firstly introduced by Cutler and Breiman [5], and it mainly is a matrix factorization method of a generic  $n \times p$ , random vector  $X$ , such that  $\min_{BG} \{\|X - \Gamma BX\|_F\}$ , where  $\Gamma$  and  $B$  represent the factorization matrices of order  $n \times k$  and  $k \times n$ , respectively, and  $\|\cdot\|_F$  states for the Frobenius norm. Matrices  $B$  and  $\Gamma$  have nonnegative entries and must satisfy the following constraints: *i*)  $B\mathbf{1}_n = \mathbf{1}_k$ ; *ii*)  $\Gamma\mathbf{1}_k = \mathbf{1}_n$ , where  $\mathbf{1}$  is a vector of ones. The  $k \times p$  matrix  $A = BX$  represents the  $k$  archetypes, where  $k$  is assumed as *a priori* defined. It is worth nothing that the matrix  $\Gamma$  defines a fuzzy allocation rule of each data point to the  $k$  archetypes: let us indicate with  $\gamma_j$  the general term of  $\Gamma$ , with  $i = 1, \dots, n$  and  $j = 1, \dots, k$ . As  $\sum_j \gamma_j = 1$ ,  $\gamma_j$  represents the membership degree of  $x_i$  to the archetype  $a_j$ .

Setting up structural constraints makes learning more efficient. In other words, one can constrain the learning process in a convex space. However, adding structural constraints often means that some form of information about the relevant domains or other dimension-generating structures is added. Consequently, this strategy pre-

sumes a conceptual level in the construction of the prototypes. Archetypal analysis exploits redundancies in input data, it finds the number (determined by the user) of archetypes in the input data that can be used to represent (approximate) data points. It is worth noting that archetypal analysis constraints ensure symmetrical relationship between archetypes and data points: archetypes are convex combinations of data points and data points are approximated in terms of convex combinations of archetypes.

In this view, we propose a geometric approach that allows prototypes identification as the most typical object within a group or a category. A prototype is the member within a group that best represents the other members (i.e., internal resemblance), and that at the same time differs from the members of the other groups or categories (i.e., external dissimilarity). This double semantics related to centrality and extremeness can be operationalized through a typicality index  $T(\cdot, \cdot)$  [23, 13, 19, 20].

Formally, given a set of  $n$  objects  $\Omega = \{x_i\}_{i=1, \dots, n}$ ,  $x_i \in \mathbb{R}^p$  and a partition  $\mathcal{C} = (C_1, \dots, C_K)$  of  $\Omega$  in  $K$  groups, an internal resemblance measure  $R(x_i, C_h)$  of  $x_i$  w.r.t.  $x_{i'} \in C_h$ , an external dissimilarity measure  $D(x_i, \overline{C}_h)$  of  $x_i$  w.r.t.  $x_{i'} \notin C_h$ , and a mixing function  $\Phi(\cdot)$  that combines both measures, a typicality index  $T(x_i, C_h)$  of  $x_i$  with respect to the class  $C_h$  is given by:

$$T(x_i, C_h) = \Phi(R(x_i, C_h); D(x_i, \overline{C}_h)). \quad (1)$$

The set of prototypes  $\mathcal{P} = (p_1, \dots, p_K)$  is then defined as:

$$\mathcal{P} = \{p_h \in \mathfrak{R}^p \mid p_h = \arg \max_{x_i} T(x_i, C_h), h = 1, \dots, K\}. \quad (2)$$

It is clear that, in this framework and setting, the prototype identification depends on the ways in which dissimilarity and resemblance are measured, and on the partition that is assumed to be known in advance. The main proposals in this direction for prototype identification assume that both resemblance and dissimilarity measures are based on the Euclidean distance. The semantic of prototypes is also strongly affected by the choice of the mixing or aggregating function  $\Phi(\cdot, \cdot)$ . If one considers only the internal resemblance, the prototypes will be the central elements of the groups; on the other hand, if one takes into account only the external dissimilarity, the prototypes will be the most extreme points. The mixing function  $\Phi(\cdot, \cdot)$  yields a compromise between these two instances.

## 4 Categorization by Voronoi tessellation: the wine data-set

In the conceptual space framework, the categorization problem can be solved by a partitioning of the space through the Voronoi tessellation starting by a given set of prototypes. In our approach, we provided a way to derive prototypes from data [22]. We note that the geometrical properties of our prototypes are congruent with the



conceptual space approach, and then we propose to use our data-driven prototypes for the Voronoi tessellation in order to obtain a categorization. In addition, in cognitive science it is often assumed that the number of prototypes and then typologies in the data is *a priori* known. However, in any real world cognitive study, things are completely different and the *true* number of typologies must be inferred by studying the groups in the data, albeit to decide the number of groups is one of the most widely addressed problems in cluster analysis, and most likely it has no satisfactory solution that can be generalized to any category of problem. Dealing with extreme data points, AA allows us to choose the number of archetypes according to the behavior of the loss function evaluated at different numbers of archetypes. The loss function is plotted on a Cartesian coordinate system, where the  $x$ -axis represents the number of archetypes and the  $y$ -axis the value of the loss function (decreasing by definition), the optimal number of archetypes should be revealed by an elbow of the function (graphically: the loss function begins to be parallel to the  $x$ -axis). However, the presence of multivariate outliers or highly correlated variables could mask the *true number* in favor of a redundant and not stable solution. Deeper investigations based on computationally intensive studies can reveal such a kind of situation.

In this section we consider the `wine` dataset. Firstly presented by Forina et al. [12], it contains data of 178 wines produced from three different Italian cultivars (*barbera*, *barolo* and *grignolino*) and described by 13<sup>1</sup> features that refer to organoleptic and chemical categories. As the three different varieties of wines are recognized as having their own specific properties, we assume that each of them represents a category and can be summarized by a prototype.

The first step consists in the archetypes identification. The package `archetypes` [10], available at CRAN repository, permits to identify the optimal number of archetypes, here we set the number of archetypes equal to three. We refer the interested reader to [22] for a more detailed description on the choice of the number of prototypes. Table 1 reports the three archetypes described by their thirteen original variables (expressed in their own original scales).

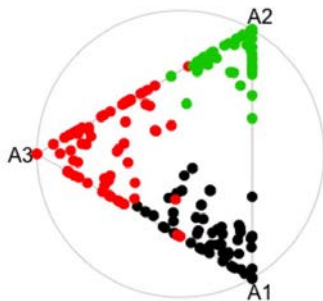
	Alc	Mal	Ash	Alk	Mag	Phe	Fla	NFla	Pro	Col	Hue	Dil	Prol
<b>a<sub>1</sub></b>	14.19	1.97	2.51	16.45	114.63	3.24	3.40	0.26	2.21	6.68	1.05	3.28	1316.07
<b>a<sub>2</sub></b>	13.22	3.78	2.48	22.12	97.47	1.56	0.65	0.49	1.05	7.69	0.63	1.51	621.94
<b>a<sub>3</sub></b>	11.79	1.41	2.07	20.04	86.50	2.26	1.97	0.34	1.61	2.15	1.20	3.08	406.40

**Table 1** Wine data: Archetypes as first solution.

The second step consists in grouping the points around the archetypes in the space defined by the matrix  $\Gamma$ . In such an example a crisp classification has been taken into account. A fuzzy allocation rule can also be taken into account, it can ensure a higher “*purity*” degree in the groups and (generally) produces an extra group with respect to the number of archetypes. The three groups, corresponding to the three

<sup>1</sup> 1) **Alcohol**, 2) **Malic acid**, 3) **Ash**, 4) **Alkalinity of ash**, 5) **Magnesium**, 6) **Total Phenols**, 7) **Flavanoids**, 8) **NonFlavanoid phenols**, 9) **Proanthocyanidins**, 10) **Color intensity**, 11) **Hue**, 12) **OD280/OD315 of Diluted wines**, 13) **Proline**.

archetypes, are visualized in the space spanned by the three columns of  $\Gamma$  in the figure 1.



**Fig. 1** Wine data set: groups around the archetypes obtained by the crisp allocation rule.

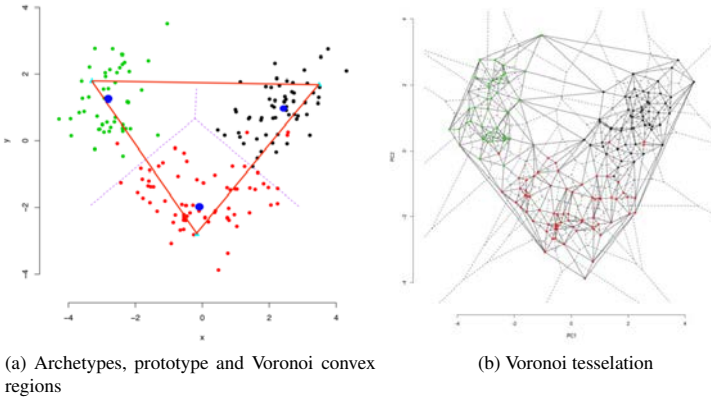
The group centroids are identified by the generalized compositional geometric mean of the group computed from the  $\gamma_{ij}$  membership scores. Exploiting the relationship between the geometric basis spanned by the archetypes and the original space [1], prototypes can be represented in the original variable space.

It can be shown that in a metric space the representation of properties is obtained as convex regions. Let us consider the set of prototypes  $\mathcal{P} = \{p_1, p_2, \dots, p_K\}$ , their representation in any conceptual space implies (according to the definition of prototype itself) that they are the central points in the categories they represent. The distance between any prototype point  $p$  and  $p'$  represents their *external dissimilarity*. If we assume that any generic point  $x_i$  belongs to the same category as the closest prototype, it can be shown that this rule will generate a partitioning of the space into convex regions [15, 21]. This partition/categorization is given by the Voronoi tessellation of the conceptual space based only on the prototypes. Note that this approach has also computational advantages. The tessellation is performed using only few points, i.e. the prototypes, and, given the geometric properties of the Voronoi tessellation, the allocation on new instances to a given category can be done in a very easy and efficient way.

The two plots in figure represent the Voronoi tessellation on the first two principal components (29% of the total variance). The plot (a) summarizes the entire categorization process: (i) the triangle vertices represent the three archetypes; (ii) the blue points (larger than the other points) refer to the prototypes; (iii) the dashed lines converging in the center define the convex regions associated to the three categories. It is worth noting that the prototypes appear more internal with respect to the corresponding archetypes.

The plot (b) on the right hand side shows the entire tessellation around the three pro-

totypes and developed with respect to the 178 observed points. It is easy to note that the categorization given by the tessellation reproduce well the three wine typologies.



**Fig. 2** Wine data set: Plots a) and b) represent the Voronoi tessellation and the convex geometric region on the first two principal components. In Figure (a) the red triangle vertices represent the archetypes, the blue points refer to the prototypes and the dashed lines represent the edges of the convex regions that correspond to three categories.

## 5 Conclusion

Several alternative cognitive approaches are grounded on the geometric representation between *properties* and *concepts* in convex conceptual spaces. Like in the Voronoi tessellation, our method allows a partitioning of the convex conceptual space into convex regions, which is based on the Euclidean metric. Thus, assuming that a Euclidean metric is defined on the subspace that is subject to categorization, a set of prototypes will generate a unique partitioning of the subspace into convex regions by this method. The upshot is that there is an intimate link between prototype theory and criterion. Furthermore, the metric is an obvious candidate for a measure of similarity between different objects. In this way, the Voronoi tessellation and archetypes categorization provide a constructive geometric answer to how a similarity measure and a set of prototypes determine a set of categories.

## References

1. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J., Pawłowsky-Glahn, V.: Logratio analysis and compositional distance. *Mathematical Geology* **32**(3), 271–275 (2000)
2. Bien, J., Tibshirani, R.: Prototype selection for interpretable classification. *The Annals of Applied Statistics* pp. 2403–2424 (2011)
3. Chang, F., Lin, C.C., Lu, C.J.: Adaptive prototype learning algorithms: Theoretical and experimental studies. *The Journal of Machine Learning Research* **7**, 2125–2148 (2006)
4. Cowan, N.: The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science* **19**(1), 51–57 (2010)
5. Cutler, A., Breiman, L.: Archetypal analysis. *Technometrics* **36**(4), 338–347 (1994).
6. D’Esposito, M., Palumbo, F., Ragozini, G.: Interval archetypes: a new tool for interval data analysis. *Statistical Analysis and Data Mining* **5**(4), 322–335 (2012)
7. D’Esposito, M.R., Palumbo, F., Ragozini, G.: *Archetypal Symbolic Objects*, pp. 41–49. Springer Berlin Heidelberg, Berlin, Heidelberg (2013).
8. Diday, E.: Optimization in non-hierarchical clustering. *Pattern Recognition* **6**(1), 17–33 (1974)
9. Diday, E.: Categorization in symbolic data analysis. In: H. Cohen, C. Lefebvre (eds.) *Handbook of Categorization in Cognitive Science*, pp. 845 – 867. Elsevier Science Ltd, Oxford (2005)
10. Eugster, M., Leisch, F., Seth, S.: archetypes: Archetypal analysis. R package version pp. 2–2 (2014)
11. Fordellone, M., Palumbo, F.: Prototypes definition through consensus analysis between fuzzy c-means and archetypal analysis. *Italian Journal of Applied Statistics* **26**(2), 141–162 (2014)
12. Forina, M., Armanino, C., Castino, M., Ubigli, M.: Multivariate data analysis as a discriminating method of the origin of wines. *Vitis* **25**, 189–201 (1986)
13. Frigui, H., Nasraoui, O.: Unsupervised learning of prototypes and attribute weights. *Pattern recognition* **37**(3), 567–581 (2004)
14. Gärdenfors, P.: *Conceptual spaces: the geometry of thought*. a bradford book. MIT Press **3**, 16 (2000)
15. Gärdenfors, P.: Concept learning and nonmonotonic reasoning, pp. 823–843. Elsevier (2005)
16. Grolemond, G., Wickham, H.: A cognitive interpretation of data analysis. *International Statistical Review* **82**(2), 184–204 (2014)
17. Hastie, T., Tibshirani, R., Friedman, J.: *The elements of statistical learning: data mining, inference, and prediction*. Springer (2011)
18. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. *ACM computing surveys (CSUR)* **31**(3), 264–323 (1999)
19. Lesot, M., Kruse, R.: Typicality degrees and fuzzy prototypes for clustering. In: *Advances in Data Analysis*, pp. 107–114. Springer (2007)
20. Lesot, M., Rifqi, M., Bouchon-Meunier, B.: Fuzzy prototypes: From a cognitive view to a machine learning principle. In: *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*, pp. 431–452. Springer (2008)
21. Okabe, A., Boots, B., Sugihara, K.: Nearest neighbourhood operations with generalized voronoi diagrams: a review. *International Journal of Geographical Information Systems* **8**(1), 43–71 (1994)
22. Ragozini, G., Palumbo, F., D’Esposito, M.R.: Archetypal analysis for data-driven prototype identification. *Statistical Analysis and Data Mining: The ASA Data Science Journal* (2016)
23. Rifqi, M.: Constructing prototypes from large databases. In: *International conference on Information Processing and Management of Uncertainty in knowledge-based systems, IPMU’96*. Granada, Spain (1996). URL <https://hal.archives-ouvertes.fr/hal-01075383>
24. Rosch, E.: Natural categories. *Cognitive psychology* **4**(3), 328–350 (1973)
25. Rosch, E.: Prototype classification and logical classification: The two systems. *New trends in conceptual representation: Challenges to Piaget’s theory* pp. 73–86 (1983)
26. Watson, D.F.: Computing the n-dimensional delaunay tessellation with application to voronoi polytopes. *The computer journal* **24**(2), 167–172 (1981)

# Inference with the Unscented Kalman Filter and optimization of sigma points

## *Inferenza con Unscented Kalman Filter e ottimizzazione dei punti sigma*

Michela Eugenia Pasetto, Umberto Noè, Alessandra Luati, Dirk Husmeier

**Abstract** We investigate the accuracy of inference in a chaotic dynamical system (Duffing oscillator) with the Unscented Kalman Filter and quantify the dependence on the sample size and the signal to noise ratio. In order to improve convergence to the true parameters in the case of a bad initialization of the algorithm, we optimize the location of sigma points with Bayesian optimisation.

**Abstract** *Si studia l'accuratezza d'inferenza in un sistema dinamico caotico (oscillatore di Duffing) con l'Unscented Kalman Filter e si quantifica la dipendenza dalla numerosità campionaria e dal rapporto segnale-rumore. Per migliorare la convergenza ai veri parametri nel caso di una cattiva inizializzazione dell'algoritmo, si ottimizza la posizione dei punti sigma in modo Bayesiano.*

**Key words:** Bayesian filtering, Unscented Kalman Filter, Chaotic dynamical system, Bayesian optimisation, Gaussian Process

## 1 Introduction

We analyse the deterministic Duffing process, defined as

$$dx_{1t}/dt = x_{2t}, \quad dx_{2t}/dt = -(cx_{2t} + \alpha x_{1t} + \beta x_{1t}^3), \quad (1)$$

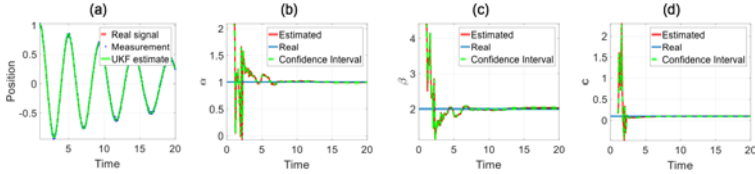
where  $x_{1t}$  and  $x_{2t}$  are the position and the velocity, respectively, of the oscillation at time  $t$ ,  $g(x) = \alpha x_{1t} + \beta x_{1t}^3$  is a restoring force,  $\alpha$  is the natural frequency of the vibration,  $\beta$  the mode of the restoring force (hard or soft spring), and  $c$  is the damping term. The Duffing system (1) describes a periodically forced oscillator

---

Michela Eugenia Pasetto and Alessandra Luati  
University of Bologna, Bologna, Italy. e-mail: michela.pasetto2@unibo.it

Umberto Noè and Dirk Husmeier  
University of Glasgow, Glasgow, United Kingdom

with a nonlinear elasticity, and has been widely used in physics, economics and engineering (Kovacic and Brennan, 2011). A characteristic feature is its chaotic behaviour, which makes statistical inference challenging. In the present paper we present an approach based on the Unscented Kalman Filter (UKF).



**Fig. 1** UKF estimates for the deterministic Duffing system with SNR=31 and  $n = 1000$ . (a) Signal estimate. (b) Estimate of parameter  $\alpha$ . (c) Estimate of parameter  $\beta$ . (d) Estimate of parameter  $c$ .

## 2 Methodology

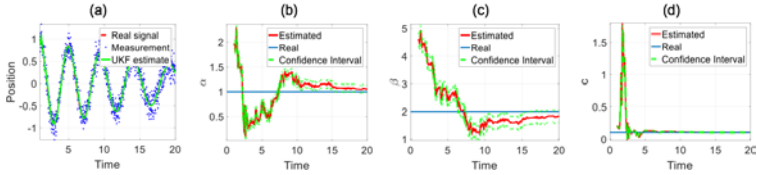
The UKF algorithm is a non-linear generalization of Kalman filter which relies on the unscented transform (Julier and Uhlmann (2004)) in order to construct a Gaussian approximation to the filtering distribution. The UKF performs a Bayesian estimation of a state-space model:

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}) + \boldsymbol{\varepsilon}, \quad \mathbf{y}_t = h(\mathbf{x}_t) + \boldsymbol{\eta}, \quad (2)$$

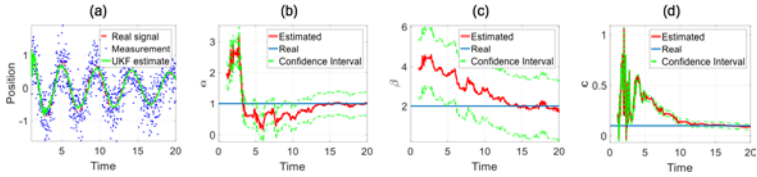
where  $\mathbf{x}_t \in \mathbb{R}^M$  is the (hidden) state at time  $t$ ,  $\mathbf{y}_t \in \mathbb{R}^D$  is the measurement,  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$  is the Gaussian system noise and  $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$  is the Gaussian observation noise. The non-linear differentiable functions  $f$  and  $h$  are, respectively, the transition and observation models. UKF passes a deterministically chosen set of points (sigma points) through  $f$  to obtain the predictive distribution  $p(\mathbf{x}_t | \mathbf{y}_{1:t-1})$ . Then, the sigma points are transformed using model  $h$  to compute the filtering distribution  $p(\mathbf{x}_t | \mathbf{y}_{1:t})$ . As suggested in Sitz et al. (2002), we merge the signal with the parameter vector  $\boldsymbol{\lambda} = [\alpha \ \beta \ c]^T$  in a joint state vector  $\mathbf{j}_t = [\mathbf{x}_t, \boldsymbol{\lambda}_t]^T = [(f(\mathbf{x}_{t-1}, \boldsymbol{\lambda}_{t-1}) + \boldsymbol{\varepsilon}), \boldsymbol{\lambda}_{t-1}]^T$ , and  $\mathbf{y}_t = h(\mathbf{j}_t) + \boldsymbol{\eta}$ . In our case, the function  $f$  of model (2) is given by the numerical solution of system (1),  $h$  is the identity function, and  $\boldsymbol{\varepsilon} = \mathbf{0}$ .

## 3 Simulations

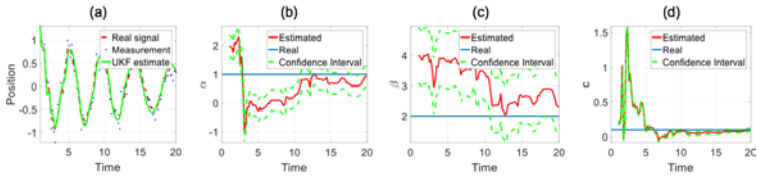
We simulate system (1) through the `ode23` MATLAB function with a stepsize of integration  $\delta t = 0.01$  and starting values for the numerical integration  $[1, 0]$ . Measurements are obtained from the first component,  $x_{1t}$ , by adding observational noise  $\eta_t \sim N(0, \sigma_{\eta}^2)$  with known variance. The time interval is  $t = 1, \dots, 20$ , and the presented results are averaged over 10 simulations. The UKF algorithm is performed



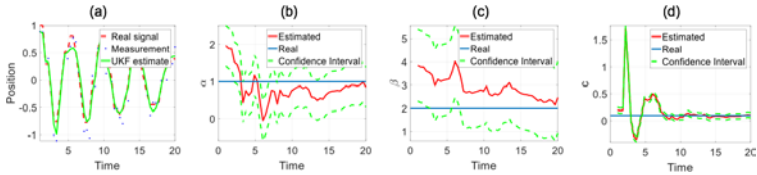
**Fig. 2** UKF estimates for the deterministic Duffing system with SNR=10 and  $n = 1000$ . (a) Signal estimate. (b) Estimate of parameter  $\alpha$ . (c) Estimate of parameter  $\beta$ . (d) Estimate of parameter  $c$ .



**Fig. 3** UKF estimates for the deterministic Duffing system with SNR=1 and  $n = 1000$ . (a) Signal estimate. (b) Estimate of parameter  $\alpha$ . (c) Estimate of parameter  $\beta$ . (d) Estimate of parameter  $c$ .



**Fig. 4** UKF estimates for the deterministic Duffing system with SNR=10 and  $n = 100$ . (a) Signal estimate. (b) Estimate of parameter  $\alpha$ . (c) Estimate of parameter  $\beta$ . (d) Estimate of parameter  $c$ .



**Fig. 5** UKF estimates for the deterministic Duffing system with SNR=10 and  $n = 50$ . (a) Signal estimate. (b) Estimate of parameter  $\alpha$ . (c) Estimate of parameter  $\beta$ . (d) Estimate of parameter  $c$ .

with the EKF/UKF toolbox of Hartikainen et al. (2011). To investigate the behaviour of the Duffing process and the UKF performance, we have simulated several scenarios, varying the Signal to Noise Ratio,  $SNR \in \{30, 10, 1\}$ , and the sample size,  $n \in \{1000, 100, 50\}$  (Figures 1–5). To evaluate the impact of initialization, we considered different offsets (low, medium and high) as starting values for the parameters. The offsets are sampled randomly from a Gaussian distribution in which the

**Table 1** Impact of the initialization for the deterministic Duffing system for different offsets (as percentage of the true parameter values) in term of Euclidean norm prior inference and post inference. Default sigma points.

	$\alpha$		$\beta$		$c$	
	Prior	Post	Prior	Post	Prior	Post
100%	1.05	0.49	2.37	1.69	0.08	0.02
250%	2.71	0.22	5.16	9.24	0.22	0.02
400%	3.83	1.94	8.27	9.10	0.54	1.79

mean is defined by a percentage deviation from the true parameter values and the variance is 10% of the mean (Table 1).

#### 4 Optimization of sigma points

The sigma points location in the UKF algorithm is parametrised by three scalar values  $\theta = (\alpha_{\text{ukf}}, \beta_{\text{ukf}}, k_{\text{ukf}})$ . These parameters are heuristically set by the algorithm, and the default values for model (1) are  $\alpha_{\text{ukf}} = 1$ ,  $\beta_{\text{ukf}} = 0$ ,  $k_{\text{ukf}} = -3$ . However, the positioning of the sigma points affects the overall inference performance of the UKF method and its convergence. We optimize the sigma points location by minimising the loss function  $L(\theta)$  using Bayesian optimisation, in order to improve the convergence of UKF to the true differential equation parameters even in the case of a bad initialization. The Bayesian optimisation algorithm iteratively maintains a statistical emulator of the objective function  $L$  in Figure (6) and chooses the next “best” point  $\theta = (\alpha_{\text{ukf}}, \beta_{\text{ukf}}, k_{\text{ukf}})$  by maximising an auxiliary acquisition function derived from the current emulator. The emulator of the loss function  $L$  is given by a Gaussian Process (GP) with constant mean function and Matérn  $\nu = 5/2$  covariance function, which leads to twice differentiable sample paths. The GP parameters are estimated by maximum log marginal likelihood. Given the GP at the current iteration,  $\hat{L} \sim \text{GP}(m, s)$ , the acquisition function used in this study is given by a weighted version of the Expected Improvement (EI):

$$\text{EI}(\theta) = (L_{\min} - m(\theta))\Phi\left(\frac{L_{\min} - m(\theta)}{s(\theta)}\right) + s(\theta)\phi\left(\frac{L_{\min} - m(\theta)}{s(\theta)}\right), \quad (3)$$

where  $\Phi$  and  $\phi$  denote the cdf and pdf of a  $N(0, 1)$  random variable. We follow the approach discussed in Noè et al. (2017) which weights the EI acquisition (3) with the probability of a successful objective function evaluation. This allows us to account for failure in the evaluation of  $L$  due to matrix singularities, and still optimise it when standard optimization algorithms would fail to. The weighted EI acquisition function balances *exploitation*, where the GP mean  $m(\theta)$  predicts a low function value, and *exploration* where the GP predicts high uncertainty  $s^2(\theta)$ . The acquisition function is optimized using the Nelder-Mead algorithm on the 10 start points having lowest acquisition function value between  $10^4$  random starting points.

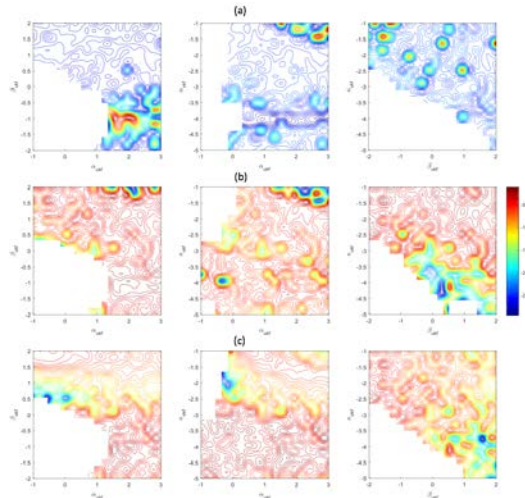


**Table 2** Euclidean norm prior inference and post inference with Bayesian optimisation.

	$\alpha$		$\beta$		$c$	
	Prior	Post	Prior	Post	Prior	Post
100%	1.05	0.37	2.37	1.25	0.08	0.02
250%	2.71	1.82	5.16	6.05	0.22	0.06
400%	3.83	0.19	8.27	5.46	0.54	0.09

**Table 3** Standardized Euclidean norm in the parameter space: comparison between the default algorithm parameters and Bayesian optimisation (BO).

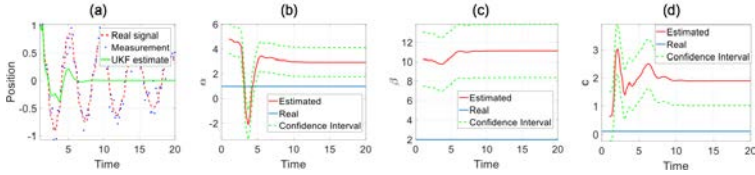
	Default		BO	
	Prior	Post	Prior	Post
100%	1.76	1.01	1.76	0.76
250%	4.33	4.63	4.33	3.59
400%	7.83	18.62	7.83	2.90



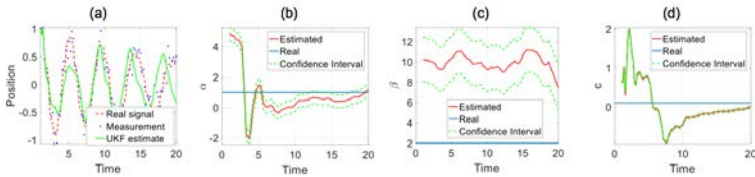
**Fig. 6** Loglikelihood of measurements for different offset. (a) High offset. (b) Medium offset. (c) Low offset. The white spaces are due to numerical instability when inverting the Kalman gain matrix.

## 5 Results

Figures 1-5 show that the UKF successfully learns the parameters from the noisy data, and that at the end of the filtering phase the true parameters always lie within the predicted standard error around the estimate. This suggests that Bayesian filtering offers a successful paradigm for inference in chaotic dynamical systems. The



**Fig. 7** UKF estimates for the deterministic Duffing system with default sigma points. (a) Signal estimate. (b) Estimate of parameter  $\alpha$ . (c) Estimate of parameter  $\beta$ . (d) Estimate of parameter  $c$ .



**Fig. 8** UKF estimates for the deterministic Duffing system with optimized sigma points in the case of high offset. (a) Signal estimate. (b) Estimate of parameter  $\alpha$ . (c) Estimate of parameter  $\beta$ . (d) Estimate of parameter  $c$ .

prediction uncertainty depends on the sample size  $n$ , and the level of noise, quantified by the SNR. As one would expect, the uncertainty increases with decreasing  $n$  and decreasing SNR, i.e. as information in the data is lost, and our study allows a quantification of this trend. The increase in uncertainty particularly affects the parameter  $\beta$ , which is associated with the nonlinear term and the source of the chaotic behaviour. Tables 1-3 show improvement in the convergence to the true values, measured in terms of the Euclidean distance in parameter space, due to the optimization of sigma points through Bayesian optimisation. This distance is consistently reduced with the optimized sigma points, suggesting that, even in the case of a bad initialization of the UKF algorithm, optimize the sigma points will improve the inference results.

## References

- Hartikainen, J., Solin, A., Särkkä, S. (2011). EKF/UKF Toolbox for MATLAB. <http://becs.aalto.fi/en/research/bayes/ekfukf/>.
- Julier, S. J., and Uhlmann, J. K. (2004). Unscented Filtering and Nonlinear Estimation. *Proceedings of the IEEE*, 92(3), 401-422.
- Kovacic I., Brennan, M. J., eds. (2011). *The Duffing Equation: Nonlinear Oscillators and their Behaviour*. New York: John Wiley & Sons.
- Noè, U., Chen, W., Filippone, M., Hill, N.A. and Husmeier, D. (2017). Inference in a Partial Differential Equations Model of Pulmonary Arterial and Venous Blood Circulation using Statistical Emulation *Submitted to Lecture Notes in Bioinformatics, Springer*.
- Sitz, A., Schwarz, U., Kurths, J. and Voss, H. U. (2002). Estimation of parameters and unobserved components for nonlinear systems from noisy time series. *Phys. Rev.*, E 66(1), 016-210.

# Pairwise Likelihood Inference for Parameter-Driven Models

## *Inferenza basata sulla verosimiglianza a coppie per modelli ‘parameter-driven’*

Xanthi Pedeli and Cristiano Varin

**Abstract** This paper discusses likelihood-type inference in parameter-driven models for regression analysis of non-normal data in presence of serial correlation. Since the ordinary likelihood function involves an intractable high-dimensional integral, we consider a pairwise likelihood approach that requires to approximate a limited set of two-dimensional integrals. Maximization of the pairwise likelihood is carried out with a pairwise version of the expectation-maximization algorithm. The methodology is illustrated with surveillance data to evaluate the relationship between influenza and meningococcal infections. Results are in close agreement with Bayesian inference based on the integrated nested Laplace approximation.

**Abstract** *Questo articolo discute l’inferenza basata sulla verosimiglianza a coppie nei modelli ‘parameter-driven’ usati per analisi di regressione con dati non normali in presenza di dipendenza temporale. Siccome la verosimiglianza ordinaria è data da un integrale di alta dimensionalità che non ha soluzione in forma chiusa, abbiamo considerato un approccio basato sulla verosimiglianza a coppie che richiede di approssimare un limitato insieme di integrali bivariati. La massimizzazione della verosimiglianza a coppie è effettuata tramite una versione a coppie dell’algoritmo ‘expectation-maximization’. La metodologia è illustrata con l’analisi di dati di sorveglianza epidemiologica per valutare l’associazione fra l’influenza e le infezioni da meningococco. I risultati in questa applicazione sono molto simili a quelli ottenuti in ambito Bayesiano usando il metodo ‘integrated nested Laplace approximation’.*

**Key words:** Expectation-maximization algorithm; Pairwise likelihood; Parameter-driven models; Surveillance; Time series of counts.

---

Xanthi Pedeli  
Ca’ Foscari University of Venice, Via Torino 155, 30170 Venezia Mestre,  
e-mail: xanthi.pedeli@unive.it

Cristiano Varin  
Ca’ Foscari University of Venice, Via Torino 155, 30170 Venezia Mestre,  
e-mail: cristiano.varin@unive.it

## 1 Introduction

Parameter-driven models [1] are frequently used for regression analysis of non-normal data in presence of serial correlation. This class of models assumes that time series observations  $Y_t$  are independent random variables conditionally to a latent process  $U_t$  designed to describe the serial correlation. Let  $p(y_t|u_t; \theta)$  be the conditional density or probability function of  $Y_t$  given  $\{U_t = u_t\}$  and  $p(u_1, \dots, u_n; \theta)$  the joint density function of the latent variables, commonly assumed to be multivariate normal. The distributions depend on a  $p$ -dimensional parameter  $\theta$ . The likelihood function for  $\theta$  is the  $n$ -dimensional integral obtained by integrating out the latent variables

$$L(\theta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{t=1}^n p(y_t|u_t; \theta) p(u_1, \dots, u_n; \theta) du_1 \cdots du_n. \quad (1)$$

A variety of simulation and non-simulation methods have been proposed for approximate inference in parameter-driven models, see [2] for a review.

In this paper, we study inference in parameter-driven models through the pairwise likelihood of order  $d$  [3], constructed by pooling together bivariate marginal distributions

$$L_2^{(d)}(\theta) = \prod_{t=d+1}^n \prod_{i=1}^d \log p(y_t, y_{t-i}; \theta), \quad (2)$$

where each component is a two-dimensional integral

$$p(y_t, y_{t-i}; \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(y_t|u_t; \theta) p(y_{t-i}|u_{t-i}; \theta) p(u_t, u_{t-i}; \theta) du_t du_{t-i}.$$

The merit of the pairwise likelihood is to replace the intractable  $n$ -dimensional integral of the full likelihood (1) with a set of bivariate integrals. Under model conditions, the maximum pairwise likelihood estimator of order  $d$  is consistent with asymptotic normal distribution

$$\mathbf{G}(\theta)^{1/2}(\hat{\theta}^{(d)} - \theta) \rightarrow_d \text{MVN}(\mathbf{0}, \mathbf{I}_p),$$

where  $\text{MVN}(\mathbf{0}, \mathbf{I}_p)$  is a  $p$ -dimensional multivariate standard normal distribution and  $\mathbf{G}(\theta)$  is the Godambe information [3],

$$\mathbf{G}(\theta) = \mathbf{H}(\theta)\mathbf{J}(\theta)^{-1}\mathbf{H}(\theta),$$

with  $\mathbf{H}(\theta) = -\mathbb{E}\{\nabla^2 \ell_2^{(d)}(\theta)\}$  and  $\mathbf{J}(\theta) = \text{var}\{\nabla \ell_2^{(d)}(\theta)\}$  where  $\ell_2(\theta) = \log L_2(\theta)$  is the log-pairwise likelihood.

The maximum pairwise likelihood estimator can be computed using a pairwise version of the expectation-maximization algorithm. The algorithm constructs a sequence of estimates  $\hat{\theta}^{(k)}$  through maximization of the expected pairwise complete log-likelihood,

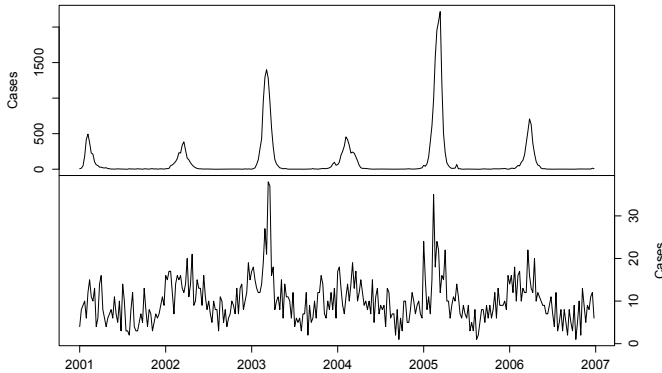
$$Q(\theta|\hat{\theta}^{(k-1)}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \log p(y_t, y_{t-i}, u_t, u_{t-1}; \theta) p(u_t, u_{t-1} | y_t, y_{t-i}; \hat{\theta}^{(k-1)}) du_t du_{t-1}.$$

The bivariate integrals involved in the expected pairwise complete log-likelihood are approximated with a double Gauss-Hermite quadrature. The pairwise expectation-maximization algorithm is particularly convenient for inference in parameter-driven models because  $\hat{\theta}^{(k)}$  is partially available in closed-form.

## 2 Application

The analysis of time series of infectious disease counts is of particular interest owing to the special features that they present, which include long-term trends, seasonality and occasional outbreaks. Apart from these features, special links between several infectious diseases form an additional source of information for biosurveillance purposes. A characteristic example is the association between influenza infection and meningococcal disease with the former being a well-known risk factor for the latter, see for example [4].

Figure 1 displays the weekly counts of meningococcal disease and influenza infection cases in Germany for the years 2001 - 2006. The data come from the German national surveillance system for notifiable diseases, administered by the Robert Koch Institute, and consist of  $n = 312$  observations. It is clear in Figure 1 that both



**Fig. 1** Weekly counts of influenza infection (top panel) meningococcal disease (bottom panel) cases in Germany for the period 2001-2006.

diseases display a seasonal pattern with evident outbreaks during the winters of 2003 and 2005, while there is no indication of trend.

The standard analysis of this type of surveillance data assumes that the meningococcal disease counts  $Y_t$  are marginally distributed as independent Poisson random variables with mean  $\exp(\eta_t)$  specified in way to account for annual seasonality and the potential association with influenza infection,

$$\eta_t = \beta_0 + \beta_1 \cos\left(2\pi \frac{t}{52}\right) + \beta_2 \sin\left(2\pi \frac{t}{52}\right) + \beta_3 \log(\text{Flu}_t + 1). \quad (3)$$

The transformation  $\log(x+1)$  is used for reducing the right skewness of influenza infection data after its transformation to strictly positive values.

In order to handle for the presence of serial correlation, we use the pairwise expectation-maximization algorithm for fitting a parameter-driven model that assumes that  $Y_t$  follows a Poisson distribution with mean  $\exp(\eta_t + U_t)$ , where the linear predictor  $\eta_t$  is specified as in (3) and  $U_t$  is the first-order autoregressive model

$$U_t = \phi U_{t-1} + \sigma \varepsilon_t, \quad \varepsilon_t \sim N(0, 1).$$

The order  $d$  of the pairwise likelihood to be maximized is chosen among a reasonable set of candidate orders based on the criterion of efficiency.

The parameter estimates and the corresponding standard errors obtained with the standard analysis and the parameter-driven model are displayed in Table 1. Inference for the parameter-driven model is based on a pairwise likelihood of or-

**Table 1** Parameter estimates (standard errors) for models fitted to the weekly counts of meningococcal infections in Germany for the period 2001-2006. PL ( $d = 7$ ) stands for pairwise likelihood of order  $d$ .

	Poisson GLM	Parameter-driven	
		PL ( $d = 7$ )	INLA
$\hat{\beta}_0$	2.10 (0.04)	2.12 (0.04)	2.11 (0.07)
$\hat{\beta}_1$	0.14 (0.03)	0.16 (0.02)	0.15 (0.05)
$\hat{\beta}_2$	0.24 (0.04)	0.27 (0.02)	0.27 (0.07)
$\hat{\beta}_3$	0.06 (0.02)	0.05 (0.02)	0.05 (0.02)
$\hat{\sigma}^2$	-	0.03 (0.01)	0.02 (0.01)
$\hat{\phi}$	-	0.70 (0.09)	0.68 (0.12)

der  $d = 7$  that gives the higher efficiency among all orders from 1 to 10. For comparison purposes, we consider also Bayesian inference using the integrated nested Laplace approximation (INLA) [5], as implemented in the R [6] package R-INLA ([www.r-inla.org](http://www.r-inla.org)).

Results of the standard analysis indicate significant seasonality and risk effect of influenza infection for meningococcal disease. The fitted parameter-driven model

confirms these findings and provide further evidence of significant autocorrelation. Pairwise likelihood and INLA provide similar parameter estimates, but standard errors of the maximum pairwise likelihood estimates are sensibly smaller than those based on INLA thus suggesting a higher precision for the proposed method in this particular application.

**Acknowledgements** This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 699980.

## References

1. Cox, D.R.: Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics* **8**, 93–115 (1981).
2. Davis, R.A., Dunsmuir, W.T.M.: State space models for count time series. In: *Handbook of Discrete-Valued Time Series*, Chapman & Hall/ CRC (2016).
3. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–41 (2011).
4. Cartwright, K.A.V., Jones, D.M., Kaczmarski, E., Smith, A.J., Stuart, J.M., Palmer, S.R.: Influenza A and meningococcal disease. *The Lancet* **338**, 554 - 557 (1991).
5. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B* **71**, 319 -392 (2009).
6. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2016).





# Social emotional data analysis. The map of Europe

## *Analisi emozionale dei Social Network. La mappa dell'Europa*

Felicia Pelagalli, Francesca Greco and Enrico De Santis

**Abstract** In this paper we present an investigation of the emotional content conveyed by words in online conversations captured on Twitter. A multivariate technique applied to co-occurrence of words together with Correspondence Analysis is adopted in order to find clusters of meaningful words detecting emotional categories that provide meaning to everyday events. Specifically, given the current historical period, where the European Union has to gain trust in its citizens, a corpus of 155000 tweets selected through the Italian keywords “Europa” and “EU” is analyzed. Results show clearly how the textual content is structured according to the different emotional expressions.

**Abstract** *In questo articolo è presentata un'analisi testuale che esplora il contenuto emozionale delle parole nelle conversazioni su Twitter. È stata adottata una tecnica di analisi multivariata applicata alla co-occorrenza delle parole assieme all'analisi delle corrispondenze al fine di raggruppare le parole in cluster di significato e individuare le categorie e le emozioni che danno senso agli eventi – ossia, i significati attribuiti agli eventi dagli attori partecipanti a un determinato contesto. Dato il particolare periodo storico in cui versa l'Unione Europea, che si trova a dover guadagnare la fiducia dei propri cittadini, è stato preparato ed analizzato un corpus di 155000 tweet selezionati attraverso le keyword “Europa” ed “EU”. I risultati mostrano chiaramente come il contenuto testuale è strutturato secondo le differenti espressioni emozionali del fenomeno.*

---

Felicia Pelagalli

Culture s.r.l., Piazza Capranica, 95 00186 ROMA, Italia,  
Scuola di Specializzazione in Psicologia della Salute, Sapienza Università degli Studi di Roma, Via degli Apuli, 1 - 00185 Roma, Italia, e-mail: feliciapelagalli@yahoo.it

Francesca Greco

Dipartimento di Psicologia Dinamica e Clinica, Sapienza Università degli Studi di Roma, Piazzale Aldo Moro, 5, 00185 Roma, Italia, e-mail: francesca.greco@uniroma1.it

Enrico De Santis

Department of Computer Science, Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada, e-mail: enrico.desantis@ryerson.ca

**Key words:** Text Mining, Social Data Mining, Multivariate Analysis, Correspondence Analysis, Clustering.

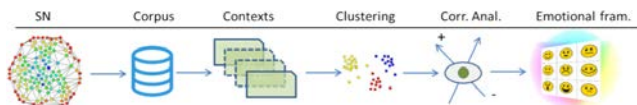
## 1 Introduction

With the spread of social networks and micro-blogging platforms, statistical methodologies boosted with machine learning techniques find their natural habitat in the sea of available online data. In fact, related techniques enable us to perceive the feeling that runs through the network. An overwhelming quantity of conversations are exchanged, mostly through words in a written form. If from one side it can be possible grasping the opinions underlying the online social exchanges, from the other it is clearly interesting to have a measure of the emotional significance that gives meaning to social phenomena. Now more than ever, this knowledge can help institutions and community managers to realize people needs and problems. It is the emotion that drives us in making relation with the objects of a given context on the basis of affective symbolizations and social representations. Hence, in conveying emotions, words show the functioning structure of the mind-brain, according to a dual logic [1]: i) the asymmetrical conscious thought which allows entering in a relationship with a context or event; ii) the symmetrical emotional thinking that the context or the events immediately arouses within us. Thus, the content analysis of conversations has to catch and externalize the emotional “density” conveyed by words or chains of words, through suitable knowledge models substantiated by statistical techniques, such as the multivariate analysis. In fact, the latter, as an unsupervised technique, can find recurrences, relations between nodes of a network or can help grouping words in meaningful clusters, detecting emotional categories that provide meaning to everyday events. According to this framework, the linguistic communication can be interpreted not only on the basis of its semantic elements but also through the emotional framework that yields value to a given text. This context fits with the co-occurrence analysis of words, used as the first step of our investigation, to find associative links among words. In this study we analyze online conversations trying to discover how they are organized within the current social context and upon a given object represented by a set of keywords. Specifically, the corpus consists in 155000 tweets gathered, in the time period ranging from January 11, 2017, to February 11, 2017, through the Twitter API, filtering the stream by the Italian keywords “*Europa*” and “*UE*”. The corpus is analyzed through a pipeline of statistical and learning techniques briefly described in next section. Specifically, in order to obtain a thematic analysis based on the co-occurrence of lexical units upon the corpus at hand, a mapping of the latter in the Vector Space Model (VSM) [2] is performed. The  $k$ -means algorithm is then adopted obtaining a suitable partition through the cosine dissimilarity measure between word vectors. Finally, the Boolean contingency matrix, describing documents membership to the retrieved clusters, is analyzed with the well-known Correspondence Analysis (CA) technique. The current paper is organized as follows. In Sec. 2 we provide a brief summary

of the adopted methodology, while in Sec. 3 main results are discussed. Finally, conclusion are drawn in Sec. 4.

## 2 Material and methods

To finalize the herein proposed investigation, data is cleaned and pre-processed. In particular, instead of raw words, lemmas as main categories are used. Subsequently, the the most common words and the very rare words are filtered out. Lemmatization and filtering allows to obtain a more compact VSM, reducing even the sparsity of the model. We note that in the current section the formal terms “document” and “context” are interchangeable, such as “term”, “word” or “lexical units”. Following



**Fig. 1** Schematic diagram of the adopted methodology for measuring the emotional structures underlying the online conversations.

the diagram of Fig. 1 the analysis presented is centered on the VSM [2], a particular vector or distributional model of meaning. VSM is based on a co-occurrence matrix, i.e. the word-document matrix, that is a way of representing how often words co-occur. From a methodological point of view the VSM embeds information retained within a corpus in a vector space representation, substantiating the distributional hypothesis according to which words that occur in similar contexts tend to have similar meanings. Lets define the term-document matrix  $\mathbf{X} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_D]$  where the content of each document vector  $\mathbf{d}_j = [w_1, w_2, \dots, w_V]$  is represented as a vector in the term space of dimension  $V$  that is usually the dimension of the vocabulary. A standard weighting scheme, used in the current work for  $w_i$ , is the the tf-idf (term frequency-inverse document frequency) [3], that provides higher weights to terms or words that are frequent in the current  $j$ -th document but rare overall in the collection.

In order to measure the similarity between two documents  $\mathbf{d}_p$  and  $\mathbf{d}_q$  enabling the cluster analysis, a well-suited similarity measure is used. It is the cosine similarity, that is  $sim(\mathbf{d}_p, \mathbf{d}_q) = \cos(\mathbf{d}_p, \mathbf{d}_q) = \frac{\mathbf{d}_p \cdot \mathbf{d}_q}{\|\mathbf{d}_p\| \|\mathbf{d}_q\|}$ .

The  $k$ -means algorithm is a *partitional* clustering algorithm [4, 5] based on squared error optimization approach. Specifically, given a set of objects (word vectors)  $\mathbf{X} = \{\mathbf{d}_j\}_{j=1}^D \in \mathbb{R}^V$ , where  $V$  is the dimension of data vectors, it finds a suitable partition  $P = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k\}$  so that the sum of the squared distances between objects in each cluster and the respective representative element is minimized:

$$\arg \min_P \sum_{i=1}^k \sum_{\mathbf{x}_j \in \mathcal{C}_i} \|\mathbf{x}_j - \mathbf{c}_i\|^2, \quad (1)$$

where  $\mathbf{c}_i$  is the representative of the  $i$ -th cluster  $\mathcal{C}_i$ . Belonging to the family of the NP-Hard problems, a complete analytical solution is not known and  $k$ -means as greedy algorithm, can only converge to a local minimum.

CA is a statistical method useful for data visualization that is applicable to cross-tabular data such as counts, compositions or any ratio-scale data. In this work, it is performed on the Boolean contingency matrix describing the partition  $P$  [6]. Let  $\mathbf{P}$  denote a  $q_r \times q_c$  data matrix with non-negative elements that sum up to 1, i.e.  $\mathbf{1}_{q_r}^T \mathbf{P} \mathbf{1}_{q_c} = 1$ , where in general  $\mathbf{1}_q$  is a  $q$ -dimensional vector of ones and  $T$  is the transpose operator. The CA is formulated as the following least-squares problem:

$$\min_{\mathbf{A}, \mathbf{B}} \left\| \tilde{\mathbf{P}} - \mathbf{D}_r^{1/2} \mathbf{A} \mathbf{B}^T \mathbf{D}_c^{1/2} \right\|^2, \quad (2)$$

where  $\tilde{\mathbf{P}} = \mathbf{D}_r^{-1/2} (\mathbf{P} - \mathbf{r} \mathbf{c}^T) \mathbf{D}_c^{-1/2}$ ,  $\mathbf{r} = \mathbf{P} \mathbf{1}_{q_c}$ ,  $\mathbf{c} = \mathbf{P}^T \mathbf{1}_{q_r}$ ,  $\mathbf{D}_r$  and  $\mathbf{D}_c$  are corresponding diagonal matrices. The column coordinate matrices  $\mathbf{A}$  and  $\mathbf{B}$  are of rank  $k$  that is the dimensionality of the approximation. By imposing  $\mathbf{B}^T \mathbf{D}_c \mathbf{B} = \mathbf{I}_k$ , it is possible obtaining a solution through the well-known Singular Value Decomposition:  $\tilde{\mathbf{P}} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T$ , where  $\mathbf{\Lambda}$  is a diagonal matrix with in descending order the singular values on the leading diagonal and  $\mathbf{U}$  and  $\mathbf{V}$  are orthonormal matrices. A least-squares approximation of  $\tilde{\mathbf{P}}$  is obtained by selecting the first  $k$  columns of  $\mathbf{U}$  and  $\mathbf{V}$  and the corresponding singular values in  $\mathbf{\Lambda}$ . Finally, the coordinate matrices are  $\mathbf{A} = \mathbf{D}_r^{-1/2} \mathbf{U} \mathbf{\Lambda}$  and  $\mathbf{B} = \mathbf{D}_c^{-1/2} \mathbf{V}$ , so that  $\mathbf{A}^T \mathbf{D}_r \mathbf{A} = \mathbf{\Lambda}^2$ . Given the coordinate matrices the row coordinates are referred to as principal coordinates whereas the column coordinates are standard coordinates. The two sets of coordinates are also known as biplot and the inner-product  $\mathbf{D}_r^{1/2} \mathbf{A} \mathbf{B}^T \mathbf{D}_c^{1/2}$  in (2) approximates the data. If the matrix  $\mathbf{P}$  constitutes a contingency table,  $\tilde{\mathbf{P}}$  is the matrix of standardized residuals, i.e. the matrix of standardized deviations from the independence model. Hence, a low-dimensional approximation of these standardized residuals is given by the biplot coordinates in  $\mathbf{A}$  and  $\mathbf{B}$ . In other words, it can be shown that this biplot will approximate, by euclidean distances on the plot, chi-square distances in  $\mathbf{P}$ . Chi-square distance is mathematically the euclidean distance inversely weighted by the marginal totals.

### 3 Results

As concerns the cluster analysis the cardinality  $k$  of the partition  $P$  is set to 5. In Tab. 1 are reported the explained variances for each principal components that hereinafter are named “factors”. In Fig. 2 we can appreciate the emotional *map* of the Europe coming out from Italian tweets. It shows how discovered clusters are placed in the factorial space, whereas in Tab. 2 is reported the factors–clusters matrix that

summarizes our main findings. The emerging *map* shows on the horizontal plane a sharp contrast between the “political power” and the “populist protest”. The cluster of words  $\mathcal{C}_1$  sees the chill and sooty European institutional places that are perceived as a remote center of power in which citizens do not definitely recognize themselves. The theme is the election of Antonio Tajani as president of the European parliament and Pittella defeat. Congratulations words, but even disappointment and irritation for who does not feel represented (*dividere, urtare, sensibilità, impera*). On the opposite side, a strong sense of helplessness regarding the big problems, such as immigrants and the economic crisis.  $\mathcal{C}_3$  is characterized by the UE plan proposed in order to stop the sea blockade in front of Libyan territories. We have also tweets where the Italian Economy ministry is perceived as “unable”, while the former prime minister Matteo Renzi together with Angelino Alfano (current Italian foreign minister) are considered “hypocrites”. Another emerging contrast on the vertical plane is the “success of the economic power” and “people problems”. From  $\mathcal{C}_2$  it emerges a two-speed Europe and the “economic power” represented by Germany with the chancellor Anghela Merkel and the president of the European Central Bank Mario Draghi. It is a strong power (*velocità, vincere*) that cohabits/forgets the human tragedies (*permettere, vergognarsi*). On the opposite side,  $\mathcal{C}_5$  refers to the necessity of funds for places hit by the earthquake. Furthermore, it shows clearly the arising of new political movements, such as the one referred to Marine Le Pen in France, evidencing tension, betrayal, isolation and risks for Europe. Finally, in  $\mathcal{C}_4$  (in a middle position on the map) we find the ambivalence fear/anguish related to the dichotomy opening–closing, where closing seems to prevail together with the fantasy of closing themselves off in the localism to avoid chaos. This is a cluster full of fears that undermine the Altiero Spinelli’s project for a united Europe.  $\mathcal{C}_4$  is close the origin of axes on the factorial map, in fact it contains basic emotions that seem to span all the facets of the underlying discourse.

**Table 1** Explained variance for each factor.

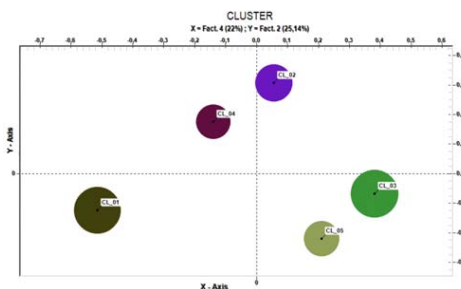
Ind	Eigenvalues	%	Cumul. %
1	0.1538	29.54	29.5438
2	0.1308	25.14	54.683
3	0.1214	23.32	77.9995
4	0.1145	22.00	100

## 4 Conclusion

The current paper presents an analysis of a huge corpus of tweets in Italian language based on a set of statistical techniques, specifically a Cluster analysis and a Correspondence Analysis. Unlike the current sentiment analysis techniques, the proposed

**Table 2** The factor-clusters matrix.

	Factor 1	Factor 2	Factor 3	Factor 4	
$\mathcal{C}_1$	–	–	changing	–	Problems related to the political power unable to drive the changing.
	–	-0,2485	0,2177	-0,5145	
$\mathcal{C}_2$	hopes	power	changing	–	The success is related to the economic power represented by A. Merkel and M. Draghi.
	-0,5119	0,6132	0,2352	0,0558	
$\mathcal{C}_3$	alert	–	changing	injustice	Alert generated by the changing related to balance of powers.
	0,3088	-0,1367	0,2492	0,3808	
$\mathcal{C}_5$	alert	–	changing	injustice	The idea about the union with the social power of foreign countries because of the loss of identity.
	0,4653	0,3534	-0,5362	-0,1396	
$\mathcal{C}_5$	hopes	problems	product	–	The European genesis has a cost that causes problems: economic request for <i>help</i> and the rejection to <i>give</i> .
	-0,5718	-0,4382	-0,4911	0,2101	

**Fig. 2** The map of the Europe.

methodology takes into account the conversations on social networks like structured corpora, in which the relationships between words can be described beyond the evaluative bias (positive/negative or agree/disagree), giving rise to a dense structure of meaning. Results show clearly how the textual content is structured according to the different emotional expressions.

## References

- [1] Matte Blanco. *L'inconscio come insiemi infiniti*. Biblioteca Einaudi, 2000.
- [2] Gerard Salton. The smart retrieval system – experiments in automatic document processing, 1971.
- [3] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *J. documentation*, 28(1):11–21, 1972.
- [4] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *J. ACM (JACM)*, 51(3):497–515, 2004.
- [5] Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.*, 31(8):651–666, June 2010.
- [6] M. van de Velden, A. Iodice D’Enza, and F. Palumbo. Cluster correspondence analysis. *Psychom.*, 82(1):158–185, 2017.

# Differential Interval-Wise Testing for the Inferential Analysis of Tongue Profiles

## *Test Intervallare Differenziale per l'analisi Inferenziale di Profili Linguali*

Alessia Pini, Lorenzo Spreafico, Simone Vantini and Alessandro Vietti

**Abstract** Motivated by the functional data analysis of a data set of tongue profiles, we describe in this talk the differential interval-wise testing (D-IWT), i.e., a local non-parametric inferential technique for testing the distributional equality of two samples of functional data. The described method can impute significant differences between the two samples to specific intervals of the domain and to specific orders of differentiation. D-IWT based inference provides a highly informative and detailed representation of the regions of the tongue where a significant difference between manners of articulation is located.

**Abstract** *Motivati dall'analisi di un data set funzionale di profili linguali, descriviamo un test intervallare differenziale (differential interval-wise testing o D-IWT), una tecnica inferenziale non parametrica locale per testare l'uguaglianza in distribuzione di due campioni di dati funzionali. Il metodo descritto permette di imputare le eventuali differenze significative tra i due campioni a specifici intervalli del dominio e a specifici ordini di differenziazione. L'inferenza ottenuta tramite il D-IWT fornisce una rappresentazione altamente informativa e dettagliata delle regioni della lingua che presentano una differenza significativa tra diversi modi di articolazione.*

**Key words:** Functional data analysis, Derivatives, Non-parametric inference, Local inference, Articulatory phonetics

---

Alessia Pini and Simone Vantini

MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy e-mail: [alessia.pini@polimi.it](mailto:alessia.pini@polimi.it), [simone.vantini@polimi.it](mailto:simone.vantini@polimi.it)

Lorenzo Spreafico and Alessandro Vietti

ALPs - Alpine Laboratory of Phonetics and Phonology Free University of Bozen-Bolzano, piazza Università, 1, 39100, Bolzano, Italy e-mail: [lorenzo.spreafico@unibz.it](mailto:lorenzo.spreafico@unibz.it), [alessandro.vietti@unibz.it](mailto:alessandro.vietti@unibz.it)

## 1 Introduction

Speech sounds are produced with a mechanism involving three main steps. The lungs pump an airflow to vibrate the vocal folds; then the vocal folds generate audible pulses; finally, pulses are “fine tuned” by the speech organs (e.g., tongue, lips, palate). The tongue plays a central role in this final step, and it is involved in the production of most of speech sounds in the world languages. In fact, it is very flexible, precise, and fast. This work aims at describing a statistical comprehensive approach to infer if and how the tongue position and shape change while different sounds are pronounced by the same speaker. The analysis focuses on the statistical comparison of tongue profiles corresponding to different manners of articulation the /R/ sound in the Tyrolean dialect, i.e., a German dialect spoken in South Tirol (Italy).

The comparison between the groups of curves can be naturally embedded within the framework of functional data analysis [12, 4, 6]. The literature dealing with inference of functional data has pursued different approaches. Most of them are global, i.e., they provide the analyst with a “simple” rejection or non-rejection of the null hypothesis (e.g., [5, 2, 3, 6]). Recently, some local methods have been proposed, providing the analyst with portions of the domain where the null hypothesis is rejected or not rejected (e.g., [1, 9, 14, 10]).

In this work we present an overview of a non-parametric local method, that is the differential interval-wise testing (D-IWT), described in detail in [11]. The D-IWT is a technique that tests differences between groups of functional data jointly taking into account the curves and their derivatives. Its output is an adjusted  $p$ -value function for each explored derivative order that can be used to select intervals of the domain imputable for the rejection of a null hypothesis.

## 2 Methodology

Assume to observe two independent samples of functional data  $\xi_{ji}$ ,  $j = 1, 2$ ,  $i = 1, \dots, n_j$  embedded in the Sobolev space  $H^d(T)$  of all real-valued squared-integrable functions on the domain  $T$  with squared-integrable derivatives up to order  $d \geq 1$  (where  $T$  is an open interval of  $\mathbb{R}$ ). Assume  $\{\xi_{1i}\}_{i=1, \dots, n_1} \sim \text{iid } \xi_1$  and  $\{\xi_{2i}\}_{i=1, \dots, n_2} \sim \text{iid } \xi_2$ , where  $\xi_1$  and  $\xi_2$  are two independent random elements of  $H^d(T)$ . We aim at performing the following family of tests, each focusing on a specific order of differentiation  $k = 0, \dots, d$ :

$$H_0 : \xi_1 \stackrel{d}{=} \xi_2 \text{ against } H_1^k : \mathbb{E}[D^k \xi_1] \neq \mathbb{E}[D^k \xi_2]. \quad (1)$$

The outputs of the D-IWT are:

- $d + 1$  **partial adjusted  $p$ -value functions**  $\tilde{p}_{D^k} : T \rightarrow [0, 1]$ , one for each order of differentiation, for testing separately the partial hypotheses (1), computed by applying the interval-wise testing [10] to every test of the family (1). For every  $k$ , the  $p$ -value  $p_{D^k}^\mathcal{I}$  of the restriction of test (1) on every interval of the domain



$\mathcal{J} \subseteq T$  is computed by means of a non-parametric permutation test [8]. The adjusted  $p$ -value function  $\tilde{p}_{D^k}(t)$  of order  $k = 0, \dots, d$  is defined as:

$$\tilde{p}_{D^k}(t) = \sup_{\mathcal{J} \ni t} p_{D^k}^{\mathcal{J}}. \quad (2)$$

- **$d + 1$  multi-derivative adjusted  $p$ -value functions**  $\tilde{p}_{D^k} : T \rightarrow [0, 1]$ , one for each order of differentiation, for testing jointly the partial hypotheses (1), computed by adjusting the  $d + 1$  partial  $p$ -value functions  $\tilde{p}_{D^k}(t)$  by means of a closed testing procedure [7]. In detail, for all possible combinations of differentiation orders indexed by  $\mathbf{k} = \{k_1, k_2, \dots, k_Q\}$  with  $\forall q : k_q \in \{0, \dots, d\}$  and  $Q \in \{2, \dots, d + 1\}$ , a  $Q$ -variate IWT is performed by means of permutation tests based on Sobolev norms (or semi-norms) on the corresponding orders of differentiation. The adjusted  $p$ -value functions  $\tilde{p}_{D^{\mathbf{k}}}(t)$  are computed according to formula (2) based on the obtained  $p$ -values of multi-derivative tests. Finally, the  $d + 1$  adjusted multi-aspect  $p$ -value functions  $\tilde{\tilde{p}}_{D^{\mathbf{k}}}(t)$  are calculated by taking for each order of differentiation the point-wise maximum of all adjusted  $p$ -value functions  $\tilde{p}_{D^{\mathbf{k}}}(t)$  involving that order:

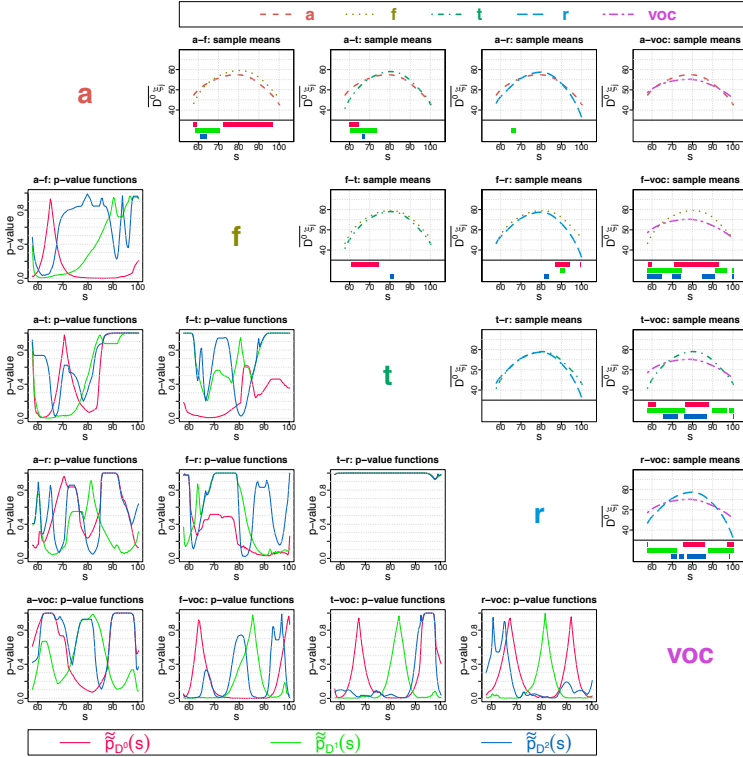
$$\tilde{\tilde{p}}_{D^{\mathbf{k}}}(t) = \sup_{\mathbf{k} \ni k} \tilde{p}_{D^{\mathbf{k}}}(t) \quad (3)$$

The  $p$ -value functions  $\tilde{\tilde{p}}_{D^{\mathbf{k}}}(t)$  can be thresholded at level  $\alpha$  to select the intervals of the domain presenting significant differences between the two populations on the  $k$ th order of differentiation. The properties of the D-IWT in terms of control of the family-wise error rate and consistency are proven in [11].

### 3 Data Analysis

To better understand the potential of the D-IWT in the practice, we summarize the results of the analysis of tongue profiles. The aim of the analysis is to test the differences between five different manners of articulating the uvular /R/:

- f FRICATIVE: produced by constricting airflow through a narrow channel at the place of articulation. There is contact between tongue and palate.
- a APPROXIMANT: shares with  $f$  the way of transmission of the sound, although there is no contact between tongue and palate.
- r TRILL: produced by directing air over the tongue so that it vibrates. There is contact between tongue and palate.
- t TAP: produced with a single contraction of the muscles so that the tongue, is thrown against the palate. There is contact between tongue and palate.
- voc VOCALIZATION: the airstream proceeds along the sides of the tongue but is blocked by the tongue from going through the middle of the mouth. There is no contact between tongue and palate.



**Fig. 1** Scatter matrix of pairwise differences between the five groups. Groups are identified in the diagonal. For each couple of groups: the upper-diagonal box indicates the two sample means (upper part) and the significant intervals at 5% level (lower part); the lower diagonal box indicates the three multi-derivative adjusted  $p$ -value functions  $\tilde{p}_{D^0}(t)$ ,  $\tilde{p}_{D^1}(t)$ , and  $\tilde{p}_{D^2}(t)$ .

Data were collected by ultrasound imaging techniques at the Alpine Laboratory of Phonetic Sciences and Phonology of the Free University of Bozen - Bolzano, Italy. For a detailed description of the data set, see [13]. The functional data have been obtained by a penalized B-spline smoothing of order six. The penalization parameter was computed via generalized cross-validation criterion [12].

We perform a D-IWT-based analysis of tongue profiles, in order to identify the possible pairwise differences between the five variants in the curves and the first two orders of differentiation ( $d = 2$ ). Figure 1 displays the results. For each comparison, the upper diagonal plots show the two sample mean curves and the lower diagonal panel shows the three multi-derivative adjusted  $p$ -value functions. The three bars in the lower part of each upper diagonal plot indicate the intervals with associated

adjusted  $p$ -value lower than 5%. The color of the bars is consistent with the one of the three adjusted  $p$ -value functions.

Inference in terms of D-IWT provides a highly informative and detailed representation of the regions of the tongue where a significant difference is located. As expected, we observe more pronounced differences when comparing  $a$  or  $voc$  (produced without touching the palate) with  $f$ ,  $t$ , or  $r$  (produced by touching the palate), while there are less pronounced differences when comparing  $a$  with  $voc$  and when comparing two variants of the group  $f$ ,  $t$ , and  $r$ . For instance, there are no significant difference between trill /R/ ( $r$ ) and tap /R/ ( $t$ ) in all orders of differentiation. Conversely, at  $\alpha = 1\%$ , approximant /R/ ( $a$ ) and fricative /R/ ( $f$ ) (second panel of the first row) are pointed out as not identically distributed. Fricative /R/ is produced by touching the palate, while approximant /R/ is produced without touching the palate. Coherently, we observed significant differences in vertical position between the two variants, with  $f$  reaching higher vertical positions than  $a$ . In addition, having a lower degree of constriction, approximant /R/ has a lower slope in the back part of the tongue. A more detailed analysis of the results, as well as a simulation study assessing the performances of the D-IWT can be found in [11].

## References

1. F. Abramovich and R. Heller. Local functional hypothesis testing. *Mathematical Methods of Statistics*, **14** (3), 253 (2005)
2. H. Cardot, A. Goia, and P. Sarda. Testing for no effect in functional linear regression models, some computational approaches. *Communications in Statistics - Simulation and Computation*, **33** (1), 179–199 (2004)
3. A. Cuevas, M. Febrero, and R. Fraiman. An ANOVA test for functional data. *Comput. Statist. Data Anal.*, **47** (1): 111–122 (2004)
4. F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer (2006).
5. P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, **89** (2), 359–374 (2002).
6. L. Horváth and P. Kokoszka. *Inference for functional data with applications*. Springer (2012)
7. R. Marcus, E. Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63** (3), 655–660 (1976)
8. F. Pesarin and L. Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons Inc (2010)
9. A. Pini and S. Vantini. The interval testing procedure: a general framework for inference in functional data analysis. *Biometrics*, **72**, 835–845 (2016)
10. A. Pini and S. Vantini. Interval-Wise Testing for Functional Data. *Journal of Nonparametric Statistics*. To appear, (2017)
11. A. Pini, L. Spreafico, S. Vantini and A. Vietti. Multi-aspect local inference for functional data: analysis of ultrasound tongue profiles. Tech. Rep. 2017 *MOX, Politecnico di Milano* (2017)
12. J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer, New York (2005)
13. A. Vietti, L. Spreafico, and V. Galatà. An ultrasound study of the phonetic allophony of Tyrolean /r/. *ICPhS 2015 Proceedings* (2015)
14. O. Vsevolozhskaya, M. Greenwood, and D. Holodov. Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. *Ann. Appl. Stat.*, **8** (2), 905–925 (2014)



# Hotelling meets Hilbert: inference on the mean in functional Hilbert spaces

## *Un incontro con Hotelling e Hilbert: inferenza per la media in spazi di Hilbert funzionali*

Alessia Pini, Aymeric Stamm, Simone Vantini

**Abstract** The talk will focus on the problem of finite-sample null hypothesis significance testing on the mean element of a random variable that takes value in a generic separable Hilbert space. For this purpose, we will present a definition of Hotelling's  $T^2$  statistic that naturally expands to any separable Hilbert space. In detail, after having recalled the notion of Gelfand-Pettis integral in separable Hilbert spaces and introduced the definition of random variables in Hilbert spaces, and the derived concepts of mean and covariance in such spaces, we will present a unified framework for making inference on the mean element of Hilbert populations based on Hotelling's  $T^2$  statistic, using a permutation-based testing procedure. We will then present the theoretical properties of the procedure (i.e., finite-sample exactness and consistency) and show the explicit form of Hotelling's  $T^2$  statistic in the case of some famous spaces used in functional data analysis like Sobolev and Bayes spaces. We will finally demonstrate the importance of the space into which one decides to embed the data by means of simulations and a case study.

**Abstract** *La relazione verterà sul problema della verifica delle ipotesi per campioni finiti relativamente all'elemento medio di una variabile aleatoria a valori in un generico spazio di Hilbert separabile. A questo scopo, presenteremo una definizione del  $T^2$  di Hotelling che lo generalizza naturalmente a qualsiasi spazio di Hilbert separabile. In particolare, dopo aver ricordato la nozione di integrale di Gelfand-Pettis in un generico spazio di Hilbert separabile e introdotto la definizione di vari-*

---

Alessia Pini

MOX - Dept. of Mathematics, Politecnico di Milano, Milan, Italy

e-mail: alessia.pini@polimi.it

Aymeric Stamm

MOX - Dept. of Mathematics, Politecnico di Milano, Milan, Italy

CRL, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

e-mail: aymeric.stamm@polimi.it

Simone Vantini

MOX - Dept. of Mathematics, Politecnico di Milano, Milan, Italy

e-mail: simone.vantini@polimi.it

*abili aleatoria a valori in un generico spazio di Hilbert e contestualmente anche i concetti derivati di media e covarianza in tali spazi, presenteremo un approccio per fare inferenza permutazionale sull'elemento medio di popolazioni Hilbertiane basato sul  $T^2$  di Hotelling. Approfondiremo in seguito le proprietà teoriche della procedura (ovvero l'esattezza per campioni finiti e la consistenza) e mostreremo la forma esplicita assunta dal  $T^2$  di Hotelling nel caso di alcuni spazi frequentemente utilizzati nell'ambito dell'analisi di dati funzionali: gli spazi di Sobolev e di Bayes. Concluderemo mostrando (per mezzo di simulazioni e di un caso di studio) l'importanza dello spazio in cui vengono rappresentati i dati funzionali.*

**Key words:** Functional Data Analysis, Object-oriented Data Analysis, Null hypothesis testing

## 1 Motivation

Statisticians are more and more confronted with the analysis of *complex* data, where *complexity* often take the form of a data analysis which pertains to analyzing data that are represented with abstract mathematical constructs, often belonging to some space on which a Hilbert structure is assumed (i.e., Object-oriented Data Analysis, OODA, [12, 14]). For example, the advent and development of technologies able to capture high-frequency measurements has provided the statistician with data that can be viewed as functions which are the foundations of functional data analysis (FDA [17, 7]). While FDA and OODA are expanding rapidly, the theoretical study of statistical tools for making inference in such spaces is still a lively area of methodological investigation ([6, 20, 5, 19, 1, 9, 3, 4, 10, 18, 2, 8, 13, 15]).

## 2 Talk Outline

The talk will focus on the inferential problem of constructing a statistical test for the means of random variables belonging to Hilbert spaces of possibly infinite dimension. After having recalled the notion of Gelfand-Pettis integral in separable Hilbert spaces ([11]) and introduced the definition of random variables in Hilbert spaces, and the derived concepts of mean and covariance in such spaces, starting with Hotelling's  $T^2$  statistic widely used in multivariate data analysis for testing the mean, we will show that Hotelling's  $T^2$  statistic can be coherently defined in any Hilbert space independently from its dimensionality and the sample size. We will then discuss all the theoretical properties pertaining to Hotelling's  $T^2$  statistic as hereby defined, its explicit form in the case of some famous spaces used in functional data analysis like Sobolev spaces ([17, 7]) and Bayes spaces ([?]), and the development of new null hypothesis significance testing procedures for making inference on the mean element in Hilbert spaces based on this statistic. We will con-

clude presenting a theoretical and empirical comparison on simulated and real data with other state-of-the-art procedures and discussing the importance of the space into which one decides to embed the data by means of simulations and a case study. The presented work is fully detailed in [16].

### 3 Acknowledgement

This work was partly supported by the 2014 Polimi International Fellowship program.

### References

1. Cardot, H., Prchal, L., Sarda, P.: No effect and lack-of-fit permutation tests for functional regression. *Computational Statistics*. **22**, 371–390 (2007).
2. Corain, L., Melas, V. B., Pepelyshev, A., Salmaso, L.: New insights on permutation approach for hypothesis testing on functional data. *Advances in Data Analysis and Classification*. **8**, 339–356 (2014).
3. Cox, D. D., Lee, J. S.: Pointwise testing with functional data using the Westfall–Young randomization method. *Biometrika*. **95**, 621–634 (2008).
4. Cuesta-Albertos, J. A., Febrero-Bande, M.: A simple multiway ANOVA for functional data. *Test*. **19**, 537–557 (2010).
5. Cuevas, A., Febrero, M., Fraiman, R.: An ANOVA test for functional data. *Computational statistics & data analysis*. **47**, 111–122 (2004).
6. Fan, J., Lin, S. K.: Test of significance when data are curves. *Journal of the American Statistical Association*. **93**, 1007–1021 (1998).
7. Ferraty, F., Vieu, P.: *Nonparametric functional data analysis: theory and practice* (2006). Springer, New York.
8. Galeano, P., Esdras, J., Lillo, R.E.: The Mahalanobis distance for functional data with applications to classification. *Technometrics*. **57**(2), 281–291 (2015).
9. Hall, P., Van Keilegom, I.: Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*. **17**, 1511 (2007).
10. Horváth, L., Kokoszka, P.: *Inference for functional data with applications*. Springer, New York (2012).
11. Hsing, T., Eubank, R.: *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons (2015).
12. Marron, J. S., Alonso, A. M.: Overview of object oriented data analysis. *Biometrical Journal* (2014).
13. Menafoglio, A., Petris, G.: Kriging for Hilbert-space valued random fields: The operatorial point of view. *Journal of Multivariate Analysis*. **146**, 84–94 (2016).
14. Menafoglio, A., Secchi, P.: Statistical analysis of complex and spatially dependent data: A review of object oriented spatial statistics. *European Journal of Operational Research* (2016).
15. Pini, A., Vantini, S.: The interval testing procedure: A general framework for inference in functional data analysis. *Biometrics* (2016).
16. Pini, A., Stamm, A., Vantini, S.: Hotellings  $T^2$  statistic and test in separable Hilbert spaces. MOX Technical report 10/2017, Dept. of Mathematics, Politecnico di Milano, available at <https://mox.polimi.it/publications/> (2017).
17. Ramsay, J. O., Silverman, B. W.: *Functional data analysis*. Springer, New York (2005).

18. Secchi, P., Stamm, A., Vantini, S.: Inference for the mean of large  $p$  small  $n$  data: a finite-sample high-dimensional generalization of Hotellings theorem. *Electronic Journal of Statistics*. **7**, 2005–2031 (2013).
19. Shen, Q., Faraway, J.: An F test for linear models with functional responses. *Statistica Sinica*. 1239–1257 (2004).
20. Spitzner, D. J., Marron, J. S., Essick, G. K.: Mixed-model functional ANOVA for studying human tactile perception. *Journal of the American Statistical Association*. **98**, 263–272 (2003).



# Accounting for measurement error in small area models: a study on generosity.

## *Modelli per piccola area con errore di misurazione: uno studio sulla generosità*

Silvia Poletti and Serena Arima

**Abstract** In this paper we focus on a recently documented effect of economic inequality, namely that higher income individuals tend to be less generous than poorer individuals, *but only in contexts where macro-level economic inequality is high*, or is perceived as high. We consider data from the Measuring Morality study, a nationally representative survey of United States residents, that contains a validated behavioural measure of generosity (the dictator game) along with the household income of respondents. We fit a small area model to this data with the aim of investigating the role of economic inequality on generosity in the US. We observe that model covariates (reported income and Gini index) are subject to measurement error and investigate the effect of introducing the measurement error in this model.

**Abstract** *Il lavoro considera il ruolo della disuguaglianza economica sulla generosità, a partire da uno studio recente secondo cui gli individui con redditi più elevati tendono ad essere meno generosi degli individui meno abbienti, ma solo in contesti di grande disuguaglianza economica. I dati analizzati provengono dal Measuring Morality study, un'indagine effettuata negli USA in cui viene rilevato il reddito e una misura validata di generosità (dictator game). Per ogni area di residenza è stato anche ricavato l'indice di Gini, come misura di disuguaglianza economica. In questo lavoro si stima la generosità mediante un modello per piccole aree con reddito e disuguaglianza come variabili ausiliarie. Il modello viene esteso al fine di considerare l'errore di misurazione nelle variabili ausiliarie, sia continue che discrete.*

**Key words:** small area estimation, measurement error, misclassification, Bayesian inference.

---

Silvia Poletti

Dip. di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza Università di Roma, via del Castro Laurenziano, 9, e-mail: [silvia.poletti@uniroma1.it](mailto:silvia.poletti@uniroma1.it)

Serena Arima

Dip. di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza Università di Roma, via del Castro Laurenziano, 9, e-mail: [serena.arima@uniroma1.it](mailto:serena.arima@uniroma1.it)

## 1 Introduction

There is an increasing interest in understanding the implications of income for behaviour, in particular generosity toward others. Well grounded literature on this topic has portrayed a picture of higher-income individuals as consistently more selfish than poorer individuals [13]. A different perspective is reported in a recent paper [6], where the relationship between economic inequality, income, and generosity is tested. Analysing data from the Measuring Morality study (a nationally representative survey of United States residents), as well as a follow-up experiment, the authors identify a previously undocumented effect of economic inequality, namely that higher income individuals in the US tend to be less generous than poorer individuals, *but only in contexts where macro-level economic inequality is high*, or is perceived as high. The Authors comment that the results obtained challenge the prevailing view in the literature that higher income individuals are necessarily less generous and conclude that “inequitable resource distributions undermine collective welfare” and that redistributive policies may “attenuate, or even reverse, the negative relationship between income and generosity, in turn increasing the generosity of those individuals who have the most to give”.

The Measuring Morality study data contain a validated behavioural measure of generosity (the dictator game) along with the household income of respondents; moreover, Gini indices were available from the American Community Survey. The authors fit a mixed effects model to these data, where significant, negative, interaction between income and inequality is found. Using a Bayesian approach, we consider the same model, in a small area context and speculate on the fact that both income and the Gini index are subject to measurement error for different reasons: indeed income is self reported and the Gini index is estimated from another survey. As stressed in the literature, ignoring the measurement error in the covariates may lead to inconsistent estimates and can severely invalidate inferences.

The paper is organized as follows: in Section 2 we introduce the problem of measurement error in small area estimation and propose a small area model accounting for measurement error in covariates and present. In Section 3 we present and discuss the results obtained when the model is applied to the generosity data.

## 2 A measurement error small area model for generosity data

In this paper, we focus on unit level small area models, within a Bayesian framework. Unit level small area models relate the unit values of the study variable to unit-specific auxiliary variables with known area means. See [11] for an up-to-date review.

Suppose there are  $m$  areas and let  $N_i$  be the known population size of area  $i$ . We denote by  $Y_{ij}$  the response of the  $j$ -th unit in the  $i$ -th area ( $i = 1, \dots, m$ ;  $j = 1, \dots, N_i$ ). A random sample of size  $n_i$  is drawn from the  $i$ -th area. The goal is to predict the small area means  $\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}$ ,  $i = 1, \dots, m$ , based on the available

data. To develop reliable estimates, auxiliary information is introduced as covariates and usually a mixed effects model is specified as

$$Y_{ij} = \alpha + \beta w_{ij} + u_i + \varepsilon_{ij} \quad i = 1, \dots, m; \quad j = 1, \dots, N_i \quad (1)$$

with  $\varepsilon_{ij}$  and  $u_i$  independent,  $\varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$  and  $u_i \stackrel{iid}{\sim} N(0, \sigma_u^2)$ . [8] and [9] were the first to consider the problem of measurement error in small area models for unit-level data. They assume that the true, area-level, covariate,  $w_i$ , is measured with error as

$$S_{ij} = w_i + \eta_{ij}, \quad \eta_{ij} \stackrel{iid}{\sim} N(0, \sigma_\eta^2) \quad i = 1, \dots, m; \quad j = 1, \dots, n_i \quad (2)$$

where  $\varepsilon_{ij}$ ,  $u_i$  and  $\eta_{ij}$  are taken mutually independent. [8] also assumed that  $w_i \stackrel{iid}{\sim} N(\mu_w, \sigma_w^2)$ , defining the structural measurement error model. They considered both an empirical Bayes and a hierarchical Bayes approach to derive predictors of small area means  $\theta_i$ . [12] extended the approach in [8] including sample information on the covariate values. [8] also proposed a fully Bayesian approach, by specifying a hierarchical model, with vague prior distributions for all the model parameters, whose posterior distributions are estimated via Gibbs sampling. [1, 3] extended the above approach, proposing to use the Jeffreys' prior on the model parameters. The aforementioned literature considers the case in which the measurement error only affects continuous variables, according to the measurement error model of equation (1). For discrete covariates, measurement error means misclassification. To allow for auxiliary discrete covariates measured with error, [4] propose to model the misclassification mechanism through an unknown transition matrix  $P$  and estimate all the unknown parameters in a fully Bayesian framework. Following [4], for each unit in each area, we consider the following covariates:  $t_{ij}$  – the vector of  $p$  continuous or discrete covariates measured without error,  $w_i$  and  $x_{ij}$  – respectively, a vector of  $q$  continuous covariates and  $h$  discrete variables (with a total of  $K$  categories), both measured with error. Denote by  $s_{ij}$  and  $z_{ij}$  the observed values of the latent  $w_i$  and  $x_{ij}$ , respectively. Without loss of generality, in what follows we assume  $h = 1$ .

Following the notation in [8], the proposed measurement error model can be written in the usual multi-stage way: for  $j = 1, \dots, n_i$ ,  $i = 1, \dots, m$  and for  $k, k' = 1, \dots, K$

- Stage 1.  $y_{ij} = \theta_{ij} + e_{ij}$   $e_{ij} \stackrel{iid}{\sim} N(0, \sigma_e^2)$
- Stage 2.  $\theta_{ij} = t'_{ij}\delta + w'_i\gamma + \sum_{k=1}^K I(x_{ij} = k)\beta_k + u_i$   $u_i \stackrel{iid}{\sim} N(0, \sigma_u^2)$
- Stage 3.  $S_{ij}|w_i \stackrel{iid}{\sim} N(w_i, \Sigma_s = \text{diag}(\sigma_{s_1}^2, \dots, \sigma_{s_q}^2))$   $W_i \stackrel{iid}{\sim} N(0, \Sigma_w = \text{diag}(\sigma_{w_1}^2, \dots, \sigma_{w_q}^2))$   
 $\Pr(Z_{ij} = k|X_{ij} = k') = p_{k'k}$ ,  $p_{k'k} \sim \text{Dir}(\alpha_{k',1}, \dots, \alpha_{k',K})$   $\Pr(X_{ij} = k') = \frac{1}{K}$
- Stage 4.  $\beta$ ,  $\delta$ ,  $\gamma$ ,  $\sigma_e^2$ ,  $\sigma_u^2$ ,  $\sigma_{s_1}^2, \dots, \sigma_{s_p}^2$  are, loosely speaking, a-priori mutually independent.

Stage 3 defines the measurement error model for both continuous and discrete covariates. For the discrete covariates, the misclassification mechanism is specified according to the  $K \times K$  matrix  $P$ , whose  $(k', k)$  element,  $p_{k'k}$ , denotes the probability

that the observable variable  $Z_{ij}$  takes the  $k$ -th category when the true unobservable variable  $X_{ij}$  takes the  $k'$ -th category. We also assume that the misclassification probabilities are the same across subjects and that all the categories have the same prior probability  $\frac{1}{K}$  to occur. Over each row of  $P$ , we place a Dirichlet  $Dir(\alpha_{k',1}, \dots, \alpha_{k',K})$  prior distribution, with known  $\alpha_{k',1}, \dots, \alpha_{k',K}$ . In Stage 4 we assume Normal priors for  $\beta$ ,  $\delta$ , and  $\gamma$  and inverse gamma distributions for  $\sigma_e^2$  and  $\sigma_u^2$  and  $\sigma_s^2$ . Hyperparameters have been chosen to have flat priors. Finally, we fix  $\Sigma_w$  and  $(\alpha_{k',1}, \dots, \alpha_{k',K})$ . According to the above assumptions, we can estimate the transition matrix  $P$  and the measurement error variance  $\sigma_s^2$  jointly with all the other model parameters. As the posterior distribution cannot be derived analytically in closed form, we obtain samples from the posterior distribution using Gibbs sampling.

### 3 Results and conclusions

We fit a unit level small area model with measurement error in covariates, which also allows us to evaluate the relationship between economic inequality, income and generosity. We use data from the Measuring Morality study, a nationally representative survey of United States residents consisting of a sample of 1498 respondents in the US. For each respondent, income and some personal and demographic variables (such as age, gender, education, ...) have been collected. Respondents completed a validated behavioural measure of generosity: the dictator game. Respondents learned that they had been randomly assigned the role of *decider* and had received 10 tickets, each worth one entry in a raffle to win a monetary prize of either 10 or 500. They could transfer any number of tickets to the next participant, a *receiver* who did not have any tickets. By giving tickets, respondents could benefit another person at a cost to themselves in a zero-sum opportunity to win money. This measure of generosity was administered to individuals with different incomes residing in areas (US states plus the District of Columbia) that vary in levels of inequality, measured according to the Gini's coefficient. The number of respondents in each area ( $m = 9$  divisions) ranges from 72 to 286. In the proposed model we take generosity as the response variable and income, standardized Gini coefficients and their interaction as auxiliary variables. According to the survey design, household income was collected as a 19-classes variable; for ease of interpretation in the application we recoded it into five classes ( $C_1$ : less than 12500;  $C_2$ : [12500, 30000),  $C_3$ : (30000, 60000],  $C_4$ : (60000, 125000],  $C_5$ : over 125000). Since income is self reported and the Gini index is estimated using data from the 2012 American Community survey, we can suspect that both auxiliary variables are subject to measurement error. In order to evaluate the impact of accounting for this source of error, we fit both the standard model that ignores the measurement error and the model proposed in Section 2. Figure 1 shows the posterior distribution of the model parameters. The left panel reports the posterior distribution of the regression parameters under the proposed measurement error model: income is the only factor that significantly impacts on the response variable, since for all the other pa-

rameters the 95% credible intervals contain the zero value ( $CI_{Gini} : [-0.207, 0.349]$ ,  $CI_{C1*Gini} : [-0.632, 0.241]$ ,  $CI_{C2*Gini} : [-0.542, 0.217]$ ,  $CI_{C3*Gini} : [-0.533, 0.189]$ ,  $CI_{C4*Gini} : [-0.827, -0.028]$ ). With respect to the income, it is apparent that generosity increases with income, with the exception of the last class, in which the effect on generosity is comparable to that of the second one. This actually means that the richest are less generous with respect to the others, which is line with findings in the mainstream literature on the subject. On the other hand, when one ignores the measurement error, all the covariates and their interactions seem to be significant (Figure 1, right panel). In particular, income exhibits a positive effect on generosity, with no distinctions between income classes, which contradicts the economic theories; moreover, an unexpectedly positive effect of inequality is found. With respect to the measurement error for income, the posterior distribution of  $P_{1,1}$  is concentrated around 0.5 and almost uniformly distributed over the other categories. This is an empirical evidence that income is often underreported by the respondents. The distributions of the other diagonal elements of  $P$  are concentrated around 0.9 and credible intervals do not contain 1. We conclude that measurement error has a significant impact on income. The small area estimates produced under the model with and without measurement error are reported in Table 1. As can be seen, allowing for measurement error in both continuous and categorical covariates also impacts on estimation of the small area means in both point estimates (in particular for the first division, which is one of the smallest ones) and measures of uncertainty. Also, although the posterior means are not very different for the large areas, the ranking of the divisions varies. As can be seen, allowing for measurement error in both continuous and categorical covariates also impacts on estimation of the small area means. Although the posterior means are not very different, the ranking of the divisions varies. In conclusion, our application reveals that ignoring the measurement error in covariates may drive inferences and yield misleading conclusions.

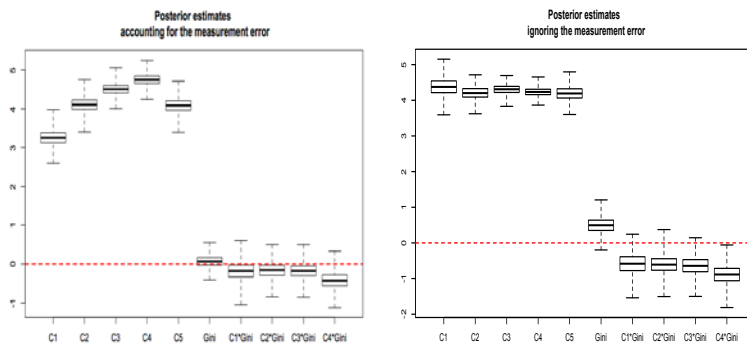
**Table 1** Small area estimates: posterior means of the small area means obtained with the model that does not account for the measurement error (first row) and the model that accounts for it (second row). Standard deviations in brackets.

Division	1	2	3	4	5	6	7	8	9
$\theta_{NoErr}$	4.17 (0.27)	4.11 (0.33)	4.25 (0.18)	4.44 (0.20)	4.19 (0.24)	4.28 (0.10)	4.25 (0.14)	4.37 (0.16)	4.22 (0.23)
$\theta_{Err}$	4.27 (0.36)	4.09 (0.41)	4.26 (0.38)	4.43 (0.37)	4.17 (0.40)	4.30 (0.33)	4.25 (0.34)	4.38 (0.32)	4.23 (0.40)

## References

1. Arima, S., Datta, G.S., Liseo, B.: Objective Bayesian analysis of a measurement error small area model. *Bayesian Analysis*, **72** (2),363–384 , (2012)

**Fig. 1** Posterior distribution of the model parameters. Left panel: posterior distributions obtained from the proposed model. Right panel: posterior distributions from the model that ignores the measurement error.



2. Arima, S., Datta, G.S., Liseo, B.: Bayesian Estimators for Small Area Models when Auxiliary Information is Measured with Error. *Scandinavian Journal of Statistics*, **42** (2),518–529, (2014)
3. Arima, S., Datta, G.S., Liseo, B.: Models in Small Area Estimation when Covariates are Measured with Error, in *Analysis of Poverty Data by Small Area Estimation*, 151–170, (2015)
4. Arima, S., Poletini, S.: A unit-level small area model with misclassified covariates, arXiv:1611.02845 [stat.ME],(2016)
5. Carroll, R.J., Ruppert, D., Stefanski, L., Crainiceanu, C.: *Measurement error in nonlinear models: a modern perspective*. 2nd edn. Chapman & Hall, CRC, (2006)
6. Côté, S., House, J., Willer, R.: High economic inequality leads higher-income individuals to be less generous . *Ann. PNAS*, **112**, 52, 15838–15843 (2015)
7. Engel, C.: Dictator Games: A Meta Study. *Experimental Economics* **14**(4), 583?-610, (2011)
8. Ghosh, M., Sinha, K. and Kim, D.: Empirical and Hierarchical Bayesian estimation in finite population sampling under structural measurement error model. *Scandinavian Journal of Statistics*, **33**(3), (2006)
9. Ghosh, M., Sinha, K.: Empirical Bayes estimation in finite population sampling under functional measurement error models. *Journal of Statistical Planning Inference*,**137**, 2759–2773,(2007)
10. Poletini, S., Arima, S.: Small area estimation with covariates perturbed for disclosure limitation. *Statistica*, **25** (1), 57–72, (2015)
11. Rao, J.N.K. and Molina, I.: *Small Area Estimation*, 2nd Edition, Wiley, Hoboken, New Jersey, (2015).
12. Torabi, M., Datta, G.S. and Rao, J.N.K. Empirical Bayes estimation of small area means under nested error linear regression model with measurement error in the covariates, *Scandinavian Journal of Statistics*, **36**, 355–368, (2009).
13. Trautmann, S.T., van de Kuilen, G. and Zeckhauser, R.J.: Social class and (un)ethical behavior: A framework, with evidence from a large population sample *Perspectives on Psychological Science* **8**(5):487–497, (2013).
14. Ybarra, L.M.R., Lohr, S.L.: Small area estimation when auxiliary information is measured with error. *Biometrika*, **95**(4), 919–931, (2008).

# **Structural changes in the employment composition and wage inequality: A comparison across European countries**

## ***Cambiamenti della struttura occupazionale e disuguaglianza salariale: Un confronto a livello europeo***

Gennaro Punzo and Mariateresa Ciommi

**Abstract** For several years many countries have been experienced a progressive impoverishment of middle-skill jobs that has led to structural changes in their labour markets (job polarisation, upgrading or downgrading of occupations). This paper investigates how the shifts in the workforce affect wage inequality comparatively for a selection of European countries. The RIF regression, tested on the EU-SILC data (2005-2013), enables us to assess how much of inequality differentials over time is accounted for by workers' endowments rather than the capability of country's labour market to capitalise skills. An outright deterioration of all jobs, irrespective of skill levels required, and the lack of a well-defined structure of the labour market may jeopardise wage distribution and the return effect plays a leading role in this process.

**Abstract** *Molti paesi stanno assistendo ad un'alterazione nella composizione della propria forza lavoro in seguito ad una progressiva diminuzione delle occupazioni con livelli intermedi di competenze. Alla luce di questi cambiamenti strutturali, si propone un'analisi comparativa della disuguaglianza salariale in Europa. La metodologia RIF, applicata a dati EU-SILC (2005-2013), permette di incorporare, dai differenziali di disuguaglianza, la quota imputabile alle dotazioni dei lavoratori da quella attribuibile alla capacità dei mercati di valorizzare tali risorse. L'assenza di una struttura di mercato ben definita, cui spesso si associa un deterioramento di tutte le occupazioni, può incidere seriamente sulla distribuzione salariale e, in tale processo, la componente "ritorno" riveste un ruolo fondamentale.*

**Key words:** Labour market, wage inequality, European countries, RIF regression

---

<sup>1</sup> Gennaro Punzo, University of Naples Parthenope, Department of Economics and Law Studies; e-mail: gennaro.punzo@uniparthenope.it

Mariateresa Ciommi, Università Politecnica delle Marche, Department of Economic and Social Science; e-mail: m.ciommi@univpm.it

## 1. Introduction

A basic prerequisite of the Kuznets theory holds that inequality tends to decline with the economic progress [13]. Hence, substantial changes of global macroeconomic environment create a general inequality climate for both developed and developing countries (Galbraith and Kum, 2005). For instance, the US income distribution suffered a hard shock during the Great Depression of the 1930s and the Second World War (1939-1945) with permanent fallouts in the years ahead. The US income inequality was still comparatively high in the 1970s and continued to grow until the US has reached the top of the rich country inequality pyramid [12].

The ongoing global crisis – the worst since 1930 – has produced painful effects for most Europe, especially for countries with weaker economies. As detailed by Eurostat (*on-line database*), the Eurozone unemployment increased from 7.5% to 11.3% between 2007 and 2013, while Mediterranean and Central/Eastern European countries were affected by unemployment more severely [14]. It is for these emergencies that at least three of the goals of Europe 2020 strategy for smart, sustainable and inclusive growth relate directly to employment, productivity and inequality. With the purpose of reaching the employment rate of 75% for 20-64-year-olds, increasing at least 40% of 30-34-year-olds completing tertiary education and lifting 20 million people out of poverty by 2020, the strategy focuses on the target of “new skills for new jobs” taking the headline idea of “more and better jobs” from the earlier Lisbon agenda.

However, within the same country, workers with varying levels of skills suffered at different extent and intensity. In particular, as discussed by Eurofound [5], the relatively recent trends identified major declines of the demand for jobs in the middle of skills hierarchy. This has resulted in structural shifts in the composition of labour force that give rise to varying labour market outcomes and income inequality trajectories [2]. In other words, changes in income inequality may be contextualised in the structure of the country’s labour market in terms of job polarisation, upgrading or downgrading of occupations. Specifically, job polarisation consists of a relative expansion in the demand of jobs occupying the top and bottom of the skills hierarchy and shrinking of the jobs in the middle, while the upgrading favours high-qualified activities with respect to low- and middle-skill jobs [1,10]. More rarely, low-skilled jobs grow faster than the rest, leading to downgrading of occupations [11].

In this field, the paper aims at identifying regularities in the structural shifts in the labour market comparatively for ten European countries and their potential relationships with the changes in wage distribution. Borrowing the geographical classification by Nolan et al. [14], which approximately corresponds to the standard welfare regimes typology [4], the following countries were selected:

- 1) The “Big Three” of Europe: France, Germany, and the United Kingdom
- 2) The four Mediterranean countries: Italy, Greece, Portugal, and Spain
- 3) Three Central/Eastern countries: Czech Republic, Hungary, and Poland.

The Recentered Influence Function (RIF) regression [7,8] allows: *i*) exploring the primary driving forces of wage inequality, *ii*) decomposing inequality gaps into the



effect attributable to individuals' characteristics (endowment) and the effect due to the capability of country labour market to valorise workers' characteristics (return), *iii*) splitting these components to evaluate the contribution that each factor gives to inequality change.

## 2. Changes in the countries' labour markets

This Section describes how the structure of employment changed between 2005 and 2013 in the selected European countries and how these shifts produced varying patterns in their labour markets. The choice of 2005 and 2013 as the reference years allows us to obtain clues about the socio-economic scenarios that foreshadowed the global crisis and their role in affecting the structure of the country's labour markets and patterns of wage inequality.

The data are from the EU-SILC (European Union-Survey on Income and Living Conditions), which is currently the main European reference source for comparable socio-economic statistics at both the household and individual levels. Moving from the assumption that inequality starts in the labour market, changes in the wage distribution become the key factors behind inequality trends. Therefore, our analysis focuses on employees, aged 16-64, irrespective of their activity sector, excluding those employed in military occupations. They are classified in the three distinct groups of high-, middle- and low-skilled employees based on the level of expertise required to perform their specific job. Given the strong correlation between the current average education level and skills required to perform that job (Eurostat, 2010), the average level of education is selected as a measure of the skills needed.

Figure 1 shows the percentage changes in the composition of employment shares between 2005 and 2013 for each of the three broad categories of employees by skills level in the selected countries. The results allow the countries to be classified according to the patterns of the labour market sketched over time in terms of job polarisation, the upgrading of occupations or neither of the two.



**Figure 1:** Percent changes 2005–2013 in employment shares by skill levels by country

The French, British and Portuguese labour markets are mostly characterised by the *upgrading of occupations*. They share a growth in professions that demand high skills and the simultaneous contraction of low- and middle-skill activities. In France, the above-mentioned jobs have both decreased by 32%, whereas the share of high-skill workers has increased by 27%. The United Kingdom follows a similar trend albeit with less intensity. To forehad of a reduction of low- and middle-skill jobs of about one-quarter, Portugal shows the largest proliferation of high-skill activities.

In Poland, instead, the drastic reduction in the demand for low- (-71%) and middle-skill (-48%) jobs is opposed only a slow-growing of highly specialised jobs (+2%). One specific point deserves the Czech Republic where there has been the simultaneous growth of all jobs regardless of the level of skills required, and surprisingly, the demand for high-skill jobs has practically doubled (+98%). In brief, the structural changes in the Polish and Czech labour markets provide evidence of two patterns that can potentially evolve in the future but, at present, are *relatively upgraded*.

In Germany, middle-skill jobs have declined as a share of employment by about 40 percent with slightly increasing levels of high-skill occupations. Instead, in Hungary, the small decrease of middle-skill jobs goes together with an important expansion of jobs for employees at the high (+47%) and low (+23%) end of the skill spectrum. Similarly, Greece has seen a large increase in the share of its low- and high-skilled employees (+40% and 57%, respectively) and the shrinkage of middle-skill jobs by 6%. Accordingly, if the patterns of the Hungarian and Greek labour markets may be classified as *purely polarised*, the German is however *relatively polarised*.

Finally, as regards Italy and Spain, changes in 2005-2013 do not enable us to determine whether one phenomenon prevails over the other. More precisely, it is not possible to define which structure succeeds because the share of employees has decreased for each of the three groups, in contrast to both job polarisation, where only the middle-skill jobs fall, and the upgrading of occupations, which comprises a decrease in the share of low- and middle-skill jobs with a simultaneously growth in high-skill activities. In both countries, the strong deterioration in the employment structures, even more severe for Italy, sketches *hybrid* patterns of their labour markets. However, both the Italian and Spanish high-skilled employees suffer relatively smaller declines than their low- and middle-skill counterparts.

### 3. RIF decomposition

The Recentered Influence Function regression [7,8] of Gini on (log of) gross individual wage replaces the log-wage as the dependent variable with the recentered influence function of the Gini coefficient  $v(F)$  and directly estimates the impact of the explanatory variables on Gini. First, the RIF methodology allows the exploration of the primary driving forces of the inequality-generating process by country. Second, the overall Gini change between 2005 and 2013 is decomposed by country into the endowment and return effects. Third, the latter two components are

computed for each covariate to identify the factors that are quantitatively more significant to make inequality differentials over time.

The RIF approach overcomes the two main limitations of the Oaxaca-Blinder method: *i*) the estimations of the return and endowment effects can be misleading if the linear model is unspecified; *ii*) the contribution of each covariate to the return effect is highly sensitive to the choice of the base group. Moreover, the Oaxaca-Blinder method enables the decomposition to be applied only to the mean, while the RIF approach also allows the decomposition of Gini (or median, quantiles, and variance). The Juhn, Murphy and Pierce method and the quantile-based decomposition by Machado and Mata overcome these drawbacks, but they are unable to trace the contribution provided by each covariate to the endowment effect, whenever they are used to compute the decomposition for various distributional statistics [7].

The observed wage ( $Y_i$ ) can be written without imposing a specific functional form considering the wage determination function of observed components  $X_i$  and some unobserved components  $\varepsilon_i$ :

$$Y_{gi} = f_g(X_i, \varepsilon_i), \quad \text{for } g = 0, 1 \quad (1)$$

$g = 1$  for workers observed in group 1 and  $g = 0$  for those in group 0. In this work, the two groups are composed of employees at time 2005 and 2013.

Let  $v(F_Y)$  be the generic distributional statistic to study (in this work, Gini), the first-order directional derivative is known as its influence function  $IF(y, v)$  so that it measures the relative effect of a small perturbation in the underlying outcome distribution on the statistic of interest. The recentered influence function (RIF) is:

$$RIF(Y; v) = IF(Y; v) + v \quad (2)$$

The unconditional expectation of the  $RIF(y, v)$  can be modelled as a linear function of the covariates:

$$E[RIF(Y; v)|X] = X\gamma + \varepsilon \quad (3)$$

the parameters  $\gamma$ , which are the marginal effect of  $X$  on  $v$ , can be estimated by OLS.

As regards the Gini coefficient, the distributional statistic  $v$  is defined as:

$$v^{GC}(F_Y) = 1 - 2\mu^{-1}R(F_Y) \quad (4)$$

where  $R(F_Y) = \int_0^1 GL(p(y); F_Y) dp$  with  $p(y) = F_Y(y)$  and the Generalised Lorenz ordinate of  $F_Y$  is given by  $GL(p(y); F_Y) = \int_{-\infty}^{F^{-1}(p)} z dF_Y(z)$ . As demonstrated by Firpo et al. (2007), the recentered influence function of Gini can be rewritten as:

$$RIF(y; v^{GC}) = 1 + 2\mu^{-2}R(F_Y) - 2\mu^{-1}[y[1 - p(y)] + GL(p(y); F_Y)] \quad (5)$$

The Gini ( $v^{GC}$ ) gap between the periods 0 and 1 is decomposed as:

$$\widehat{\Delta}_0^{v^{GC}} = \bar{X}_1(\hat{\gamma}_{1,v^{GC}} - \hat{\gamma}_{0,v^{GC}}) + (\bar{X}_1 - \bar{X}_0)\hat{\gamma}_{0,v^{GC}} = \widehat{\Delta}_S^{v^{GC}} + \widehat{\Delta}_X^{v^{GC}} \quad (6)$$

The overall inequality gap ( $\widehat{\Delta}_0^{v^{GC}}$ ) is disentangled into the return effect ( $\widehat{\Delta}_S^{v^{GC}}$ ) and the endowment effect ( $\widehat{\Delta}_X^{v^{GC}}$ ). The first term corresponds to the effect on  $v^{GC}$  of a change from  $f_1(\cdot, \cdot)$  to  $f_0(\cdot, \cdot)$  while keeping the distribution of  $(X, \varepsilon)|G = 1$  constant. Conversely, the endowment effect keeps the return effect  $f_0(\cdot, \cdot)$  constant and measures the effect of changes from  $(X, \varepsilon)|G = 1$  to  $(X, \varepsilon)|G = 0$ . Further methodological details on the contribution of a single covariate in the decomposition can be found in Firpo et al. [7] and Fortin et al. [8]. However, the key term for decomposing  $v^{GC}$  is the counterfactual distributional statistic  $v_c^{GC}$ , which is the distributional statistic that would have prevailed if the workers observed in group 1 had the return effect of period 0. Using the counterfactual distribution, the above mentioned components can be rewritten as:

$$\widehat{\Delta}_S^{v^{GC}} = \bar{X}_1(\hat{\gamma}_{1,v^{GC}} - \hat{\gamma}_{0,v^{GC}}^C) \quad \text{and} \quad \widehat{\Delta}_X^{v^{GC}} = (\bar{X}_0^C - \bar{X}_0)\hat{\gamma}_{0,v^{GC}} \quad (7)$$

The estimation of the coefficients,  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_c$ , requires first estimating the weighting functions  $\omega_1(G)$ ,  $\omega_0(G)$  and  $\omega_c(G, X)$ . In order to have weights summing up to one, the normalisation procedure is used (details in DiNardo et al. [3] and Firpo et al. [7]).

#### 4. Main results

Once the RIF regressions of Gini on log-wage are estimated to explore primary factors (individual, human capital, and job-related) that drive the observed wage inequality by country, the overall Gini differences in 2005-2013 are disentangled into the endowment (composition effect) and return effects (wage structure) (Tables 1-3). The composition effect assesses the portion of Gini change attributable to the employees' endowments. The wage structure explores the capability of the country's labour market to transform individual skills into job opportunities and earnings and explains why employees with the same individual characteristics are rewarded differently. Standard errors of components are computed according to the method detailed in Fortin et al. [8].

The overall Gini has declined over time for France and Germany as part of the most developed European economies even though the magnitude of the fall has been more pronounced for the transition countries of Central/Eastern Europe. Conversely, the United Kingdom and Italy (that together with Germany and France form the "Big Four" of Europe) show a rise in the overall inequality, which is however far larger for Italy. In line with the literature [14], Italy is an equal country if compared to others with similar patterns of growth, but relatively less unequal than the other Mediterranean countries. In fact, Greece still keeps harsher levels of inequality

Structural changes in the employment composition and wage inequality: A comparison 807  
 despite the Gini index has increased in 2005-2013 less than in Italy. Wage inequality has largely increased also in Spain; it has remained constant for Portugal, coherently with the literature [14] that shows a reversal of the previous increase in income inequality since 2005, although the decrease has not been large enough to compensate for the strong growth of inequality during 1989-1994.

**Table 1** RIF decomposition of Gini on log-wage. Gap 2005–2013. Western countries

	<i>France</i>		<i>Germany</i>		<i>The UK</i>	
Total gap	-0.0041***	–	-0.0043***	–	0.0011*	–
Composition	-0.0021***	51.2%	-0.0039***	90.7%	-0.0004	-36.4%
Wage structure	-0.0020***	48.8%	-0.0004	9.3%	0.0015**	136.4%

\*Significant at 10%; \*\*Significant at 5%; \*\*\*Significant at 1%.

**Table 2** RIF decomposition of Gini on log-wage. Gap 2005–2013. Mediterranean countries

	<i>Italy</i>		<i>Greece</i>		<i>Portugal</i>		<i>Spain</i>
Tot. gap	0.0064***	–	0.0037***	–	-0.0001	0.0101***	–
Comp.	0.0004	6.3%	0.0019***	52.6%	0.0041	-0.0021***	-20.8%
Wage str	0.0060***	93.7%	0.0017***	47.4%	-0.0042	0.0122***	120.8%

\*Significant at 10%; \*\*Significant at 5%; \*\*\*Significant at 1%.

**Table 3** RIF decomposition of Gini on log-wage. Gap 2005–2013. Central/Eastern countries

	<i>Czech Republic</i>		<i>Hungary</i>		<i>Poland</i>	
Total Gap	-0.0060***	–	-0.0119***	–	-0.0176***	–
Composition	-0.0043***	71.47	-0.0077***	64.71	-0.0040***	22.73
Wage structure	-0.0017***	28.73	-0.0042***	35.29	-0.0136***	77.27

\*Significant at 10%; \*\*Significant at 5%; \*\*\*Significant at 1%.

In countries that have experienced a decline of wage inequality, a great deal of the total changes is due to the composition effect. Therefore, up to more than 90% for Germany (where the wage structure is even not significant), three-quarter for the Czech Republic and two-third for Hungary of the reduction of wage inequality depends on the changes in workers' characteristics happened over time. In other words, in these countries, the endowments in employees' characteristics and potentialities have contributed more effectively to decrease, or at least not to increase, wage inequality. Instead, the wage structure plays a leading role (Spain) – if not exclusive (the United Kingdom, Italy) – in increasing wage inequality, stressing the low capacity of the countries' labour markets to transform inputs into better job-related careers and higher earnings. Not only the skill endowments but the ways in which they are rewarded in the labour market are crucial in explaining differentials in wage inequality over time. A more detailed analysis (whose results are not reported for brevity) has identified the human capital endowments and job-related characteristics as the individual resources that mostly contribute in shaping, in one direction or another, wage inequality differentials within the two components of composition effect and wage structure.

In sum, those countries that experienced a decrease (or at least a not increase) in wage inequality – France, Portugal, Poland, the Czech Republic, Hungary and

Germany – share shifts in the employment composition between 2005 and 2013 that have led to more explicit and clearly defined structures of their labour markets (upgrading or relatively upgrading, polarisation or relatively polarisation). Probably, the employment changes, which have led the labour markets towards more upgraded or polarised structures, usually less unequal, discontinued the inequality growth within the country with an equalising effect on the wage distribution. In Greece, the employment changes towards a more polarised pattern have only slowed the growth of inequality within the country, mainly due to the recent crisis that has hit Greece so even harder. Conversely, in Italy and Spain, where the distribution of occupations by skill levels appears to be more ambiguous, the increasing differentials in wage inequality are mostly attributable to the lower efficiency of their labour markets to offer better job opportunities and careers, and thus, better salaries for employees. In other words, the outright deterioration of all jobs, irrespective of skill levels required, and the lack of a clear structure of the Italian and Spanish labour markets have exacerbated disparities among the three sub-groups of employees, increasing the overall inequality within countries.

## References

1. Autor, D.: Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing. *Journal of Labor Economics*, 21 (1) (2003).
2. Castellano, R., Musella, G., Punzo, G.: Structure of the labour market and wage inequality: evidence from European countries. *Quality & Quantity*, 1-28 (2016).
3. DiNardo, J., Fortin, N., Lemieux T.: Labor Market Institutions and the Distribution of Wages, 1973-1993: A Semi-Parametric Approach, *Econometrica* 64, 1001–1045 (1996).
4. Esping-Andersen, G.: *Social Foundation of Post-Industrial Economies*, Oxford University Press, Oxford (1999).
5. Eurofound: Drivers of recent job polarisation and upgrading in Europe: *European Jobs Monitor 2014*, Publications Office of the European Union, Luxembourg (2014).
6. Eurostat: Educational Intensity of Employment and Polarization in Europe and the US, Eurostat Methodologies and Working Paper (2010).
7. Firpo, S., Fortin, N., Lemieux, T.: Decomposing wage distributions using recentered influence function regressions, University of British Columbia (2007).
8. Fortin, N., Lemieux, T., Firpo, S.: Decomposition Methods in Economics, *Handbook of Labor Economics* (2011).
9. Galbraith, J. K., Kum, H.: Estimating the inequality of household incomes: a statistical approach to the creation of a dense and consistent global data set. *Review of Income and Wealth* 51(1), 115-143 (2005).
10. Goos, M., Manning, A.: Lousy and Lovely Jobs: The Rising Polarization of Work in Britain, *The Review of Economics and Statistics* 89, 118-133 (2007).
11. Hurley, J., Storrie, D., Jungblut, J.M.: Shifts in the job structure in Europe during the Great Recession, Luxembourg: Publications Office of the European Union (2015).
12. Kenworthy, L., Smeeding, T.: The United States: high and rapidly-rising inequality. *Inequality and its impacts* 2, 695-717 (2013).
13. Kuznets, S.: Economic growth and income inequality. *The American economic review*, 1-28 (1955).
14. Nolan, B., Salverda, W., Checchi, D., Marx, I., McKnight, A., Tóth, I. G., van de Werfhorst, H. G. (Eds.): *Changing inequalities and societal impacts in rich countries: thirty countries' experiences*. OUP Oxford (2014).

# Official Statistics 4.0 – learning from history for the challenges of the future

## *Statistica Ufficiale 4.0 - apprendere dalla storia per le sfide del futuro*

Walter J. Radermacher, La Sapienza University Rome, wjr@outlook.de

### **Abstract**

The quantity of digital data created, stored and processed in the world has grown exponentially. The demand for statistics and the power of facts has never been so apparent. In the process of adapting to the new reality, Official Statistics will continue to focus on relevant products, efficient production processes and quality of statistical information. Quality, trust and authority are central to Official Statistics in a modern democratic society, holding a neutral and impartial position between the political decision-makers and citizens. The article underlines the importance of statistical governance and quality management.

*Abstract La quantità di dati digitali creati, memorizzati e elaborati nel mondo è cresciuta in modo esponenziale. La domanda di statistica e la potenza dei fatti non è mai stata così notevole. Nel processo di adattamento alla nuova realtà, le statistiche ufficiali continueranno a concentrarsi sui prodotti rilevanti, sui processi di produzione efficienti e sulla qualità delle informazioni statistiche. La qualità, la fiducia e l'autorità sono al centro della statistica ufficiale di una società democratica moderna, con una posizione neutra e imparziale tra i decisori politici ed i cittadini. L'articolo sottolinea l'importanza della governance statistica e della gestione della qualità.*

**Key words:** Official Statistics, Statistical Governance, Quality management, Data Revolution

### **A (simplified) model of reality: statistics**

There is no alternative to facts!” or “Science belongs to everyone” were slogans of the March for Science on ‘Earthday’ 22. April 2017, a “*call for science that upholds the common good and for political leaders and policy makers to enact evidence based*

*policies in the public interest.*”(MarchforScience 2017). Hopefully, this initiative will be inspired by scientific reflexions concerning the character, role and limits of science (Benessia et al. 2016), thus representing a ‘new enlightenment’. (EuropeanAlpbachForum 2016)

*"No substantial part of the universe is so simple that it can be grasped and controlled without abstraction. Abstraction consists in replacing the part of the universe under consideration by a model of similar but simpler structure. Models, formal or intellectual on the one hand, or material on the other, are thus a central necessity of scientific procedure. ... That is, in a specific example, the best material model for a cat is another, or preferably the same cat. In other words, should a material model thoroughly realize its purpose, the original situation could be grasped in its entirety and a model would be unnecessary."*(Rosenblueth and Wiener 1945)

In social sciences this relation between reality and a model of it has been introduced by Max Weber in form of what he called "Idealtypen". "According to Weber's definition, "an ideal type is formed by the one-sided accentuation of one or more points of view" according to which "concrete individual phenomena ... are arranged into a unified analytical construct" (Gedankenbild); in its purely fictional nature, it is a methodological "utopia [that] cannot be found empirically anywhere in reality" .... Keenly aware of its fictional nature, the ideal type never seeks to claim its validity in terms of a reproduction of or a correspondence with reality. Its validity can be ascertained only in terms of adequacy, which is too conveniently ignored by the proponents of positivism."(Kim 2012)

It is of fundamental importance for Official Statistics, that conceptual models are designed in such a way, that they are 'adequate' abstractions of reality. This leads to the question, what 'adequate' concretely means or which criteria and which processes are offered by statistical methodology to answer this question. Compared to other quality components of statistical information (e.g. sampling errors), this area is however less covered by statistical theory.

In German statistical terminology, 'Adäquation' (Grohmann 1985) represents the design-phase within the process of statistical knowledge building, which contains basically the choice of model parameters according to the purpose of the research, available resources, time constraints etc. "*Data quality is depending on ... developing operational methods corresponding as much as possible to theory and - intensively controlling and monitoring the survey procedure in process.*"(Radermacher 1992)

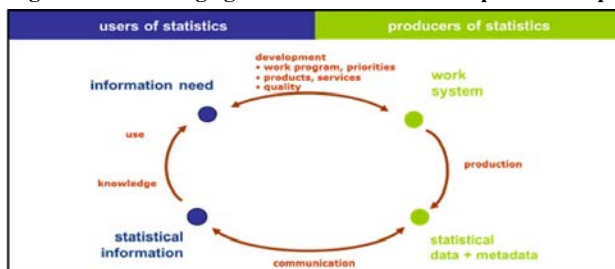
This will say, statistical information is produced with two main ingredients: methodology and conventions. On the one hand, "*the notion of statistics as a primarily mathematical discipline really developed during the 20th century, perhaps up to around 1970, during which period the foundations of modern statistical inference were laid*"(Hand 2009). On the other, the final products of statistical processes depend essentially on their conceptual design, which, like for other (manufactured) products, depends essentially on the fact whether the questions raised by stakeholders can be answered by statistics and whether they are answered in a satisfactory manner.



## The production of Official Statistics

A simplified circular process chart describing the interaction between users and producers of information should help to understand the main features:

**Figure 1: Knowledge generation and statistical production process**



The key-processes within the production sphere of this chart are

- D: development and design,
- P: production,
- C: communication/dissemination,

which corresponds to widely accepted standards, such as the Generic Statistical Business Process Model (GSBPM) (UNECE 2013) or the Generic Statistical Information Model (GSIM) (UNECE 2017). In addition, it is essential to include explicitly the interaction with stakeholders and the following process on the user side

- U: creation of knowledge and application.

The ultimate goal of statistical evidence is to contribute to better informed decisions of all kind and for all types of users, which can only be achieved when all four processes are taken into account and integrated in a comprehensive conceptual approach. Each of them should contribute to excellent information quality. Each of them can of course also fail and contribute to errors, misunderstandings and underperformance:

- The process D has an external part (dialogue with users) and an internal part (development and testing of methods). Intensive cooperation with users is crucial for the adequacy of the entire process chain that follows.
- During the production process P the methods agreed in the preceding development phase are implemented. It is relatively straightforward to measure the quality of this process and its sub-processes against these predefined norms.
- Communication processes C represent the other end of the user interaction. They can also be grouped in an internal part (preparing the results from the production process for different channels, access points etc.) and an external part (interaction with users in all formats and through all channels). The internal part does also

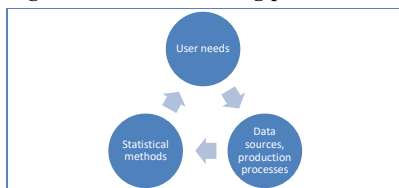
belong to the set of predefined methods and is in that way similar and closely linked to production.

- The processes of application and use  $U$  are not under any kind of control or influence by statisticians. It is however obvious that users might not be sufficiently prepared or trained to interpret and use statistics in the best possible manner. Statistical literacy is therefore an area of interest also for statistical producers. Furthermore, statisticians should carefully observe cases of wrong interpretation and they must protect their information against misuse.

### ***Evolution and continuous adaptation***

This process chart helps us to understand the planning of Official Statistics as an evolutionary process, as a sequence of learning cycles and feed-back loops:

**Fig. 2: Statistical learning process**



Over time, changes might be started from all three angles. New demands and political issues trigger new statistical developments, as new data sources or new methodologies do. Historically, it can be observed that these driving forces are also mutually influencing each other's, thus stimulating new episodes in official statistics (Desroisières 1998).

It is therefore essential to link the communication process of today with the development of statistics for tomorrow. Partly, this loop could be a short one, if user feedbacks can lead to quick fixes and improvements in services. Partly, it might however take time, since changes in a programme need profound preparations and even more profound developments.

This evolutionary development of statistics is confronted with several limiting factors, which could be practical limitations, such as:

- Clandestine, non-observable phenomena
- Statistical items in the future and elsewhere, relevant for decisions now and here (e.g. capital goods, depreciation, trade chains, sustainable development)
- Values and prices for non-market-goods (Can we simulate non-existing markets?)
- Limitations by resource or time constraints; in cases where only a limited amount of information and data are available or where limited time is given for the decision-making process.

Limitations could also relate to the understanding and use of data and information:

- Innumeracy, statistical and data illiteracy

- High and too high expectations
- (No) Appetite for high quality information (Davies 2017)

Since the beginning of Official Statistics in the nineteenth century, the boundaries have been substantially stretched. Continuous improvement has opened new opportunities so that today many subjects (e.g. quality of life), which were impossible to observe only a few years ago, are fully integrated elements of the standard statistical program. Nevertheless, it is crucial to understand that basic principles must be respected, if the fundament, on which trust in Official Statistics is built, shall not be damaged. This is for example the reason to refrain from monetising natural resources and their services, if they are not valued by market transactions.

### ***Consultation of users***

Official Statistics is a special application and form of statistics that belongs to the public infrastructure of (modern) states. Working methods in Official Statistics reflect both, their political and administrative position as well as the status and development of societies, (i.e. the specific relationship between state and citizens).

In a modern and democratic definition, Official Statistics is no longer a knowledge tool in the hands of the powerful and mighty. Rather, it must follow principles of neutrality and impartiality, whereby this information infrastructure becomes an important democratic pillar, equally available and accessible for everyone. (Radermacher and Bischoff 2017 forthcoming)

#### ***Article 338 TFEU (European Union 2012)***

*1. Without prejudice to Article 5 of the Protocol on the Statute of the European System of Central Banks and of the European Central Bank, the European Parliament and the Council, acting in accordance with the ordinary legislative procedure, shall adopt measures for the production of statistics where necessary for the performance of the activities of the Union.*

*2. The production of Union statistics shall conform to impartiality, reliability, objectivity, scientific independence, cost-effectiveness and statistical confidentiality; it shall not entail excessive burdens on economic operators.*

The interaction with stakeholders must be governed by principles of transparency, democratic control/supervision and public/legislative form of all kind of conventions. In particular, the programme of work must emerge from a democratic decision making process, at the end of which a choice is made in favour of the 'Pareto-optimal' composition of statistical tasks. Priority setting in this context has an important role to play, as it must facilitate the annual adaptation of the program following changes in user needs.

The way, in which this consultation and decision making was organised so far, relies mainly on the functioning of 'official' procedures concerning the preparation of

legislation and political decisions. Modern societies ask however for more; more in terms of wider consultation (more room for all active contributions from civil societies), new forms (collection of user needs through social media) and faster (quicker adaptation of the program).

## **Official Statistics 3.0 – the last 25 years**

Since the beginning of the 1990<sup>th</sup>, the environment around statistics has dramatically changed due to several factors (Radermacher 2014b), such as:

- Pressure on the public sector; major cuts in budgets and human resources
- Reduced willingness to respond to statistical surveys; response burden as a political target
- Exponentially growing importance of ICT and new data sources (e.g. administrative data, GIS)
- New political demands (e.g. environment, globalization, migration) and crises (e.g. financial)

These changes are expressed in the Regulation of European Statistics: “*Whereas 14: The operation of the ESS (the author: European Statistical System) also needs to be reviewed as more flexible development, production and dissemination methods of European statistics and clear priority-setting are required in order to reduce the burden on respondents and members of the ESS and improve the availability and timeliness of European statistics.*” (Eurostat 2015b)

### ***Changing the business model***

A widely-supported starting point concerning the strategic orientation for Official Statistics is: “*Our output has traditionally been determined by the demands of our respective governments and other customers. The process is one of reasoning back from the output desired to survey design because often few or no pre-existing data were available. This paradigm has shaped the way official statistics are designed and produced. ... In the future it will become increasingly unrealistic to expect meaningful statistics from this approach, even when results are collected and transmitted electronically.*” (Vale 2017)

### **Towards multiple source – mixed mode design**

Since the end of the 1990<sup>th</sup>, a re-engineering of the business model is ongoing, according to which the single statistical production lines are bundled and integrated, common technical tools are developed and terminology is standardised, thus minimising redundancies, inefficiencies and sources of incoherence. Information is generated by (re-)using available data as far as possible, aiming at minimising response burden and costly surveys.

In terms of the above-mentioned learning process this means that the current ‘development loop’ is driven by changes on the production side, which will lead to substantial improvements.

Nevertheless, one must take the implications for the other actors of the learning cycle into account. For example, it was not difficult in the past and with the traditional business model to organise functioning user-producer dialogues, since participants of these dialogues shared the interest and knowledge of same subject matter: Agricultural statistics was discussed between the specialists for agricultural policies and the technicians in a specialised branch of the statistical office; the same applied for labour market, population, health statistics and so forth; a balanced agreement sufficient for static and narrow user needs. As long as statistical offices did not have to cope with substantial resource scarcities (and rapidly changing user needs), it was therefore not necessary to establish an overall program-planning, to decide on priorities etc. The program was just the sum of a great number of partial solutions in each separated area; both users and producers were generally satisfied; users with their tailor-made products and producers with their control of the entire production process. This inefficient ‘spaghetti-bowl-business-model’ of the past is replaced the new ‘industrialised-process-model’: multiple-source inputs, standardised production, multiple purposes output. The new business model of production cannot be ‘administrated’ in a traditional manner. It needs to be ‘managed’, including the development of planning tools, a catalogue of products / services, marketing and cost accounting, which means not less than a complete overhaul of the traditional culture in Official Statistics.

### **New components in the statistical programme**

Firstly, new products and services will be generated on this way. For example, if populations censuses are moving towards a new design, it is possible to produce results not only every five or ten years but annually, which would better fit to information needs in times of high population dynamics (e.g. migration, ageing). (Kyi and Knauth 2012)

Secondly, the more integrative approach has stressed the fact that the different statistical bits and pieces should form a coherent information system, in which different types of products have their place. Within this system, basic statistics, macro-economic accounts and indicator sets have different functions and fulfil different roles.

Finally, genuine new ‘metrics’ are requested for new political debates, consequences of crises or other pressing demands from different stakeholders, which could be demonstrated with two examples:

- Sustainable development: In 2009 the Commission on the Measurement of Economic and Social Progress issued a report, which highlights the need to go ‘beyond GDP’ (Stiglitz, Sen, and Fitoussi 2009), in 2015 the General Assembly of the United Nations adopted the ‘Sustainable Development Goals’ (United Nations 2016), both having already caused significant changes in statistical programs.
- Globalisation: Recent reports from Sturgeon (Sturgeon 2013) Bean (Bean 2016) and the ‘Economic Statistics Review Group’(ESRG 2016) highlight the

shortcomings of economic statistics concerning the monitoring of globalised production processes. New indicators have been proposed, which will have consequences also for routines in basic statistics and ‘national’ accounts.

### **Strengthening the statistical governance**

Complexity of the production process is significantly increased when the new business model is introduced. At the same time, statistical information has grown into a powerful tool in the political arena. New opportunities come with new risks (Radermacher 2017 forthcoming), which are taken into account by an adaptation of firstly the legal frames of Official Statistics, secondly the Codes of Conduct and thirdly the broadening and deepening of stakeholder consultation.

### ***European Statistics***

European political developments have asked for customised statistical solutions. The introduction of an European single market has created the need for another form of external trade statistics (i.e. Intrastat), the Maastricht treaty asked for special statistical monitoring (i.e. EDP statistics), the European Central Bank requested solid and comparable price statistics (i.e. HICP) (Radermacher 2012).

European statisticians were at the forefront of the international modernisation activities. With ‘Vision 404’ (Eurostat 2009a) a strategy for the next years was outlined, which contained all three dimensions: process, product and governance. With the communication ‘GDP and Beyond’ (Eurostat 2009b) the European Commission has set up a work program aiming at substantial improvements of statistical information.

Meanwhile, these plans have been implemented:

- The governance of the European Statistical System was substantially revised (Eurostat 2015b; Radermacher 2014a),
- the multi-annual programme adapted to the new business model (Eurostat 2013b);
- Sustainable Development indicators have been developed (Eurostat 2016b);
- the accounting layer has been modernised and broadened (Eurostat 2013c, 2016a);
- basic statistics have been re-engineered (e.g. demographic statistics, Census HUB, HICP, integrated social and agricultural statistics, FRIBS).

Close cooperation amongst the partners in the European Statistical System was the enabling factor to forcefully implement the strategy that was outlined in the ESS Vision 2020 (Eurostat 2013a).

In terms of user orientation important new initiatives were taken, such as:

- Relaunch of the website, new visualisation tools, active use of social media, DIGICOM (Eurostat 2017b);
- Public consultation in the preparation of new legislation (e.g. (Eurostat 2015a));
- Indicators as user interface (Eurostat 2017a) and Conferences of European Statistical Stakeholders (Smedt 2016).

## **Official Statistics 4.0 – data revolution**

The quantity of digital data created, stored and processed in the world has grown exponentially. Broad consensus reigns regarding the wonderful opportunities, which 'Big Data' can bring in relation to the statistics acquired from traditional sources such as surveys and administrative records. Much faster and more frequent dissemination of data; responses of greater relevance to the specific requests of users since the gaps left by traditional statistical production are filled; refinement of existing measures, development of new indicators and the opening of new avenues for research; a substantial reduction in the burden on persons or businesses approached and a decrease in the non-response rate are all possibilities potentially offered. Access to Big Data could considerably reduce the costs of statistical production.

However, the Big Data phenomenon also poses a certain number of challenges: These data are not the result of a statistical production process designed in accordance with standard practice. They do not fit the methodologies, classifications and definitions, and are therefore difficult to harmonise and convey in statistical structures. Complex aggregates, such as the GDP or the Consumer Price Index aim at measuring macro-economic indicators for the nation as a whole (Lehtonen 2015); their (immediate) substitution by Big Data sources seems to be out of reach. In addition to this, major legal issues are raised: security and confidentiality of data, respect for private life, data ownership, etc. All the above means that, at least for now, Big Data can be used only to a limited degree to supplement rather than replace sources of traditional data in certain statistical fields. Their integration is – besides all technical aspects, a challenge for an “informational governance” (Soma et al. 2016).

### **Conclusions and guiding principles (Radermacher and Baldacci 2016)**

#### **Statistics is a key for people empowerment**

High-quality statistics strengthen democracy by allowing citizen access to key information that enhances accountability. Access to solid statistics is a fundamental "right" that permits choices and decision based on information. Without statistics there cannot be a well grounded and participated democracy.

Statisticians should be aware of the power of data which lies in their transformation of information services for knowledge.

#### **Open data are fundamental for open societies**

Statistics are the cornerstone of public open data. They are the basis of open government. In the EU Open Data Data Portal, Eurostat statistical database accounts for the bulk of data offered. Enhancing access to statistics in open formats enables the free use of data, its interoperability and consumption in integrated modalities. Open statistics as a result allow to make sense of complex phenomena and help in their interpretation without borders and limits. As such open statistics are a key sources of free dialogue in our societies.

Statisticians should ensure open and transparent access to data and metadata and measure their actual use for information and knowledge.

#### **Datacy is a key enabler for citizens**

Statistical literacy is critical to ensure that individuals can benefit from the power of data and can make use of open access to statistical information and its associated services. Data literacy is not limited to knowledge of basic statistical information, it entails knowing the limit of statistics and their use/misuse. Capabilities to understand statistics and how they are produced are a fundamental skill for a whole individual and an aware citizen.

Statisticians should proactively invest in data capabilities in society at large and measure the results of statistical literacy.

#### **The future is smart statistics**

The value of data is in the statistical methods which ensure quality services. In the digital ecosystem where data are abundant and a commodity, the value of information is increasingly based on algorithms that generate tailored insights for users.

Statisticians should continue to invest in methods and algorithms that enhance the quality of data for statistical services tailored to users' needs.

#### **More influence means more responsibilities**

As statistical information is increasingly used for policy decisions, statisticians need to investigate how their services are used, the ethical implications and the impact of evidence use on the policy cycle.

It is a duty of statisticians to explore the link between statistics, science and society and lead intellectual reflections on the possible risk of reliance on data-centrism.

### **References**

- Bean, Charles. 2016. *Independent Review of UK Economic Statistics* (UK Government: London).
- Benessia, A., S. Funtowicz, M. Giampietro, A. Guimaraes Pereira, J. Ravetz, A. Strand Saltelli, R., and J.P. van der Sluijs. 2016. *The rightful place of science: science on the verge* (Consortium for Science, Policy and Outcomes: Tempe, AZ).
- Davies, William. 2017. 'How statistics lost their power – and why we should fear what comes next', *The Guardian*.
- Desroisières, Alain. 1998. *The Politics of Large Numbers - A History of Statistical Reasoning* (Harvard University Press: Cambridge Massachusetts).
- ESRG. 2016. *REPORT OF THE ECONOMIC STATISTICS REVIEW GROUP (ESRG)* (CSO Ireland: Dublin).
- EuropeanAlpbachForum. 2016. 'An Introduction to “New Enlightenment”'. <https://www.alpbach.org/en/forum2016/programme-2016/new-enlightenment-an-introduction-by-the-presidents-of-the-european-forum-alpbach/>.
- EuropeanUnion. 2012. "THE TREATY ON THE FUNCTIONING OF THE EUROPEAN UNION (TFEU)." In, edited by European Commission. Brussels: Official Journal of the European Union C326/193.
- Eurostat. 2009a. "COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL on the production method of EU statistics: a vision for the next decade." In, edited by EuropeanCommission. Brussels.



- . 2009b. "GDP and beyond - Measuring progress in a changing world - COMMUNICATION FROM THE COMMISSION TO THE COUNCIL AND THE EUROPEAN PARLIAMENT." In. Brussels: European Commission.
- . 2013a. 'ESS Vision 2020 - Building the Future of European Statistics', Eurostat. <http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020>.
- . 2013b. "REGULATION (EU) No 99/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 15 January 2013 on the European statistical programme 2013-17." In, edited by Uropean Commission. Brussels 9.2.2013: Official Journal of the European Union.
- . 2013c. "REGULATION (EU) No 549/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 May 2013 on the European system of national and regional accounts in the European Union." In, edited by European Commission. Brussels: Official Journal of the European Union.
- . 2015a. 'IMPLEMENTATION OF THE FRAMEWORK REGULATION INTEGRATING BUSINESS STATISTICS (FRIBS)', Eurostat. <http://ec.europa.eu/eurostat/about/opportunities/consultations/fribs>.
- . 2015b. "REGULATION (EC) No 223/2009 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL." In *2009R0223 — EN — 08.06.2015 — 001.001 — 1*. Luxembourg: Eurostat.
- . 2016a. "REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL on the implementation of Regulation (EU) No 691/2011 on European environmental economic accounts." In. Brussels: European Commission.
- . 2016b. *Sustainable development in the European Union - A STATISTICAL GLANCE FROM THE VIEWPOINT OF THE UN SUSTAINABLE DEVELOPMENT GOALS 2016 Edition*.
- . 2017a. *Communicating through indicators* (Eurostat).
- . 2017b. 'DIGICOM – USERS AT THE FOREFRONT', Eurostat. <http://ec.europa.eu/eurostat/web/ess/digicom>.
- Grohmann, Heinz. 1985. 'Vom theoretischen Konstrukt zum statistischen Begriff - Das Adäquationsproblem', *Allgemeines Statistisches Archiv*, Allg. Statist. Archiv 69: 1 - 15.
- Hand, David J. 2009. 'Modern statistics: the myth and the magic', *Journal of the Royal Statistical Society*, J. R. Statist. Soc. A (2009): 287–306.
- Kim, Sung Ho. 2012. 'Max Weber.' in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Metaphysics Research Lab, Stanford Universit: Stanford).
- Kyi, Gregor, and Bettina Knauth. 2012. "A census is a census is a census ?" In *UNECE-Eurostat Expert Group Meeting on Censuses Using Registers*. Geneva: UNECE Conference of European Statisticians.
- Lehtonen, Markku. 2015. 'Indicators: tools for informing, monitoring or controlling?' in Andrew J. Jordan and John R. Turnpenny (eds.), *The Tools of Policy Formulation - Actors, Capacities, Venues and Effects* (Edward Elgar Publishing.).
- MarchforScience. 2017. 'March for Science'. <https://www.marchforscience.com/>.

- Radermacher, Walter. 1992. 'Methoden und Möglichkeiten der Qualitätsbeurteilung von statistischen Informationen aus der Fernerkundung / Methods and Possibilities of Assessing the Quality of Statistical Data of Remote Sensing', *Jahrbücher für Nationalökonomie und Statistik*: 169-79.
- . 2014a. 'The European Statistics Code of Practice as a pillar to strengthen public trust and enhance quality in official statistics', *Journal of the Statistical and Social Inquiry Society of Ireland*, 43: 27-33.
- Radermacher, Walter, and Pierre Bischoff (ed.) (eds.). 2017 forthcoming. *Article 338* (Springer).
- Radermacher, Walter J. 2012. 'Zahlen zählen - Gedanken zur Zukunft der amtlichen Statistik in Europa', *ASStA Wirtsch Sozialstat Arch*, 2012: 285–98.
- . 2014b. 'New challenges facing official statistics', *Statistical Journal of the IAOS*, 30(2014): 3-16.
- . 2017 forthcoming. '3S: Science, Statistics and Society', *International Journal of Data Science and Analytics*.
- Radermacher, Walter J., and Emanuele Baldacci. 2016. "Official Statistics for Democratic Societies - Dinner speech at the CESS 2016, Budapest." In *Conference of European Statistical Stakeholders*. Budapest.
- Rosenblueth, Arturo, and Norbert Wiener. 1945. 'The Role of Models in Science', *Philosophy of Science*, 12: 316-21.
- Smedt, Marleen De. 2016. 'Addressing the Needs of Official Statistics Users: The Case of Eurostat', *Journal of Official Statistics*, 32: 913-16.
- Soma, Katrine, Bertrum H. MacDonald, Catrien JAM Termeer, and Paul Opdam. 2016. 'Introduction article: informational governance and environmental sustainability', *Current opinion in Environmental Sustainability*, 2016: 131–39.
- Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi. 2009. *Report by the Commission on the Measurement of Economic and Social Progress*.
- Sturgeon, Timothy J. . 2013. *Global Value Chains and Economic Globalization - Towards a New Measurement Framework* (Eurostat: Luxembourg).
- UNECE. 2013. 'Generic Statistical Business Process Model (GSBPM)'. <http://www1.unece.org/stat/platform/display/GSBPM/Generic+Statistical+Business+Process+Model>.
- . 2017. 'The Generic Statistical Information Model (GSIM)', UNECE. <http://www1.unece.org/stat/platform/display/gsim/Generic+Statistical+Information+Model>.
- UnitedNations. 2016. 'Sustainable Development Goals'. <http://www.un.org/sustainabledevelopment/sustainable-development-goals/>.
- Vale, Steven. 2017. 'Strategic vision of the HLG-MOS', UNECE High-Level Group for the Modernisation of Official Statistics. <http://www1.unece.org/stat/platform/display/hlgbas/Strategic+vision+of+the+HLG-MOS - StrategicvisionoftheHLG-MOS-I.Introduction>.

# Comparison of contingency tables under quasi-symmetry

## *Confronto di tabelle di contingenza sotto l'ipotesi di quasi-simmetria*

Fabio Rapallo

**Abstract** In this work we define a test to compare several square contingency tables under the quasi-symmetry model. Working within the class of log-linear models, we present a suitable model and an exact test to verify if two or more tables fit a common quasi-symmetry model. The exact test is then defined through classical tools of Algebraic Statistics, namely the computation of a Markov basis and the application of a MCMC algorithm.

**Abstract** *In questo lavoro viene definito un test per confrontare tabelle di contingenza quadrate sotto l'ipotesi di quasi indipendenza. Rimanendo nella classe dei modelli log-lineari, si definisce un appropriato modello e un test esatto per verificare se due o più tabelle possono soddisfare un comune modello di quasi-simmetria. Il test esatto è definito tramite i classici strumenti della Statistica Algebrica, ossia il calcolo di una base di Markov e l'applicazione di un algoritmo di tipo MCMC.*

**Key words:** Algebraic Statistics, exact tests, Markov bases, MCMC algorithms

## 1 Introduction

Complex models for contingency tables have received an increasing interest in the last decades from researchers and practitioners in different fields, from Biology to Medicine, from Economics to Social Science. As general references for the statistical models for contingency tables see [1] and [7]. Quasi-symmetry and quasi-independence models are well known log-linear models for square contingency tables. Under these models, it is possible to fix the diagonal counts, or even to analyze incomplete tables where the diagonal counts are undefined or unavailable. In the

---

Fabio Rapallo

Department DISIT, University of Piemonte Orientale, viale Teresa Michel 11, 15121 Alessandria, Italy, e-mail: fabio.rapallo@uniupo.it

next section we will recall the basic facts on the quasi-symmetry model, while for a full presentation and an historical overview the reader can refer to [3] and [6].

Since quasi-symmetry model is a log-linear model, one can test the goodness of fit through the classical chi-squared approximation of the Pearson or the likelihood ratio test statistics. Alternatively, exact tests have been introduced for quasi-symmetry and for other classes of log-linear models. Apart from the independence model, exact tests are usually difficult to implement, and the new techniques introduced with Algebraic Statistics has allowed a noticeable progress through the notion of Markov basis and the definition of the Diaconis-Sturmfels (D-S) algorithm. Notice that the exact approach is particularly important for quasi-symmetry models, because the asymptotic approximation fail also with moderately large sample sizes, as noted in [9].

Algebraic Statistics has been a very growing research area, with major applications to the analysis of contingency tables. In addition to a general algorithm for exact inference, Algebraic Statistics provides an easy description of complex log-linear models for multi-way tables and it represents the natural environment to define statistical models for contingency tables with structural zeros, through the notion of toric models. Toric models are generalization of log-linear models allowing also zero-probability cells. As general references on the use of Algebraic Statistics for contingency tables, see [5] and [2]. Some specific statistical models related to quasi-symmetry in the framework of Algebraic Statistics can be found in [10].

In this paper, we use classical techniques from Algebraic Statistics in order to compare several contingency tables under the quasi-symmetry model. In particular, we present an exact test to verify if two or more tables fit a common quasi-symmetry model, versus the alternative hypothesis that each table follows a specific quasi-symmetry model with its own parameters. This is accomplished by the construction of a suitable three-way table and the definition of new log-linear models for this new table. The exact test is then derived by applying the D-S algorithm.

The material is organized as follows. In Sect. 2 we recall some definitions and basic results about log-linear models and toric models, with special attention to quasi-symmetry. In Sect. 3 we show how to study define suitable log-linear models to compare two or more square tables under quasi-symmetry, together with the description of the D-S algorithm for this application. Finally, Sect. 4 is devoted to the illustration of a real-data example and some pointers to future works.

## 2 Log-linear models and quasi symmetry

A probability distribution on a finite sample space  $\mathcal{X}$  with  $K$  elements is a normalized vector of  $K$  non-negative real numbers. Thus, the most general probability model is the simplex

$$\Delta = \left\{ (p_1, \dots, p_K) : p_k \geq 0, \sum_{k=1}^K p_k = 1 \right\}.$$

A statistical model  $\mathcal{M}$  is therefore a subset of  $\Delta$ .

A classical example of finite sample space is the case of a multi-way contingency table where the cells are the joint counts of two or more random variables with a finite number of levels each. In the case of square two-way contingency tables, where the sample space is usually written as a cartesian product of the form  $\mathcal{X} = \{1, \dots, I\} \times \{1, \dots, I\}$  we will use the notation  $p_{i,j}$  to ease the readability.

Following the classical theory of log-linear models, under the Poisson sampling scheme the cell counts are independent and identically distributed Poisson random variables with means  $Np_1, \dots, Np_K$ , where  $N$  is the sample size, and the statistical model specifies constraints on the raw parameters  $p_1, \dots, p_K$ . A model is log-linear if the log-probabilities lie in an affine subspace of the vector space  $\mathbf{R}^K$ . Given  $d$  real parameters  $\alpha_1, \dots, \alpha_d$ , a log-linear model is described, apart from normalization, through the equations:

$$\log(p_k) = \sum_{r=1}^d A_{k,r} \alpha_r \tag{1}$$

for  $k = 1, \dots, K$ , where  $A$  is the model matrix (or design matrix), see in [8]. Exponentiating Eq. (1), we obtain the expression of the corresponding toric model

$$p_k = \prod_{r=1}^d \zeta_r^{A_{k,r}} \tag{2}$$

for  $k = 1, \dots, K$ , where  $\zeta_r = \exp(\alpha_r)$ ,  $r = 1, \dots, d$ , are the new non-negative parameters. It follows that the model matrix  $A$  is also the matrix representation of the minimal sufficient statistic of the model. The matrix representation of the toric models as in Eq. (2) is widely discussed in, e.g., [11] and [5]. Note that from Eq. (1) it follows that different model matrices with the same image as vector space generate the same log-linear model.

The log-linear form of the quasi-symmetry model is

$$\log(p_{i,j}) = \mu + \alpha_i^{(X)} + \beta_j^{(Y)} + \gamma_{(i,j)} \tag{3}$$

with the constraints

$$\sum_{i=1}^I \alpha_i^{(X)} = 0, \quad \sum_{j=1}^I \beta_j^{(Y)} = 0, \quad \gamma_{i,j} = \gamma_{j,i}, \quad i, j = 1, \dots, I.$$

In Eq. (3), the  $\alpha_i^{(X)}$  are the parameters of the row effect, the  $\beta_j^{(Y)}$  are the parameters of the column effect, while the parameters  $\gamma_{i,j}$  force the quasi-symmetry. Comparing Equations (1) and (3) it is easy to explicitly write the model matrix  $A_{qs}$  for the quasi-symmetry model.

### 3 Comparison of several tables under quasi-symmetry

As outlined in the introduction, in this section we define two log-linear models to compare two or more square tables under quasi-symmetry. Let us consider  $H$  tables ( $H \geq 2$ ) and define a three-way contingency table  $T$  by stacking the  $H$  tables. Conversely, each original table is a layer of the  $T$ . Let  $K' = HI^2$  be the number of cells of  $T$ . The two models are defined as follows. The first model  $M_0$  is defined by

$$(M_0) \quad \log(p_{h,i,j}) = \mu + \mu_h + \alpha_{i,h}^{(X)} + \beta_{j,h}^{(Y)} + \gamma_{h,i,j} \quad (4)$$

with the constraint  $\sum_{h=1}^H \mu_h = 0$  in addition to the constraints on  $\alpha_{i,h}^{(X)}$ ,  $\beta_{j,h}^{(Y)}$  and  $\gamma_{h,i,j}$  naturally derived from the basic quasi-symmetry model in Eq. (3). The second model  $M_1$  is defined by

$$(M_1) \quad \log(p_{h,i,j}) = \mu + \mu_h + \alpha_i^{(X)} + \beta_j^{(Y)} + \gamma_{i,j} \quad (5)$$

with the constraint  $\sum_{h=1}^H \mu_h = 0$  in addition to the constraints on  $\alpha_i^{(X)}$ ,  $\beta_j^{(Y)}$  and  $\gamma_{i,j}$  naturally derived from the basic quasi-symmetry model in Eq. (3). It easy to see that  $M_1 \subset M_0$ . Under the model  $M_0$  we assume that each layer of the table follows a quasi-symmetry model with its own parameters, while under the model  $M_1$  we assume that all the layers follow a common quasi-symmetry model. In terms of the model matrix, the models  $M_0$  and  $M_1$  have a simple block structure. In fact:

$$A_{M_0}^t = \left( \begin{array}{c|c|c} A_{qs}^t & & \\ \hline & \ddots & \\ \hline & & A_{qs}^t \end{array} \right) \quad \text{and} \quad A_{M_1}^t = \left( \begin{array}{c|c|c} A_{qs}^t & \cdots & A_{qs}^t \\ \hline \mathbf{1}_K & & \\ \hline & \ddots & \\ \hline & & \mathbf{1}_K \end{array} \right)$$

where  $\mathbf{1}_K$  is a row vector of 1's with length  $K = I^2$ , and each empty block means a block filled with 0's.

Let  $f$  be the observed table of counts, and write  $f$  as a vector of length  $K'$  according to the row labels of  $A_{M_0}$ . The test for nested models with null hypothesis  $H_0 : p \in M_1 \subset M_0$  versus  $H_1 : p \in M_0$  can be done using the log-likelihood ratio statistic

$$G^2 = 2 \sum_{k=1}^{K'} f_k \log \left( \frac{\hat{f}_{1k}}{\hat{f}_{0k}} \right),$$

where  $\hat{f}_{0k}$  and  $\hat{f}_{1k}$  are the maximum likelihood estimates of the expected cell counts under the models  $M_0$  and  $M_1$  respectively. In the asymptotic theory, the value of  $G^2$  must be compared with the quantiles of the chi-square distribution with the appropriate number of degrees of freedom, depending on the dimensions of the table.

We introduce here a procedure for exact inference via Markov bases and the D-S algorithm, see [5] for details. Given the observed table  $f$ , the key idea of the D-S algorithm is to make the reference set of a given table

$$\mathcal{F}(f) = \left\{ f' \in \mathbf{N}^{K'} : A_{M_1}^t f' = A_{M_1}^t f \right\}$$

connected. This is done through a set of moves  $\mathcal{M}_1$ , i.e., integer-valued tables with null value of the sufficient statistics  $A_{M_1}$ . If the set  $\mathcal{M}_1$  contains enough moves such that each table of  $\mathcal{F}(f)$  can be reached from  $f$  in a finite number of steps by adding/subtracting moves, we say that  $\mathcal{M}_1$  is a Markov basis. Once a Markov basis for the model is available, the D-S algorithm is a MCMC algorithm and proceeds as follows. At each step:

1. let  $f$  be the current table;
2. choose with uniform probability a move  $m \in \mathcal{M}_1$  and a sign  $\varepsilon = \pm 1$  with probability  $1/2$  each;
3. define the candidate table as  $f_+ = f + \varepsilon m$ ;
4. generate a random number  $u$  with uniform distribution over  $[0, 1]$ . If  $f_+ \geq 0$  and

$$\min \left\{ 1, \frac{\mathcal{H}(f_+)}{\mathcal{H}(f)} \right\} > u$$

then move the chain in  $f_+$ ; otherwise stay at  $f$ . Here  $\mathcal{H}$  denotes the hypergeometric distribution on  $\mathcal{F}(t)$

After an appropriate burn-in-period and taking only tables at fixed times to reduce correlation between the sampled tables, the proportion of sampled tables with test statistics greater than or equal to the test statistic of the observed one is the Monte Carlo approximation of  $p$ -value of the log-likelihood ratio test. The results in the next sections are based on Monte Carlo samples of size 10,000.

### 4 Example

As a simple numerical example we consider the data reported in Tab. 1 (adapted from [4] and originally collected during the “Indagine Longitudinale sulle Famiglie Italiane (Italian Household Longitudinal Survey), where the inter-generational social mobility has been recorded on a sample of 4,343 Italian workers in 1997. The data take into account the gender, and thus we have separate tables for men and women. There are 4 categories of workers.  $A$ : “High level professionals”;  $B$ : “Employees and commerce”;  $C$ : “Skilled working class and artisans”;  $D$ : “Unskilled working class”. In [4] these data are analyzed extensively with a thorough presentation of a lot of models to analyze special patterns of mobility. Here we merely use the simplified version displayed in Tab. 1 to show the practical applicability of the methodology introduced in Sect. 3.

The relevant Markov basis has been computed with the software `4ti2` [12] and it consists of 151 moves. If we consider the two tables separately, we obtain exact  $p$ -values computed through the D-S algorithm are equal to 0.051 and 0.088 respectively ( $G^2 = 6.703$  and  $G^2 = 8.279$  respectively, with 3 df). Running the test

**Table 1** Table of social mobility in Italy (1997). Columns represent the father's occupation, rows represent the son's (or daughter's) occupation. Male respondents in the left panel, female respondents in the right panel.

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	172	31	31	28	<i>A</i>	137	52	29	15
<i>B</i>	108	49	24	46	<i>B</i>	78	46	14	23
<i>C</i>	174	84	301	272	<i>C</i>	142	100	124	145
<i>D</i>	225	148	236	664	<i>D</i>	164	181	141	35

described above to test a unique quasi-symmetry model, the D-S test produces a  $p$ -value equal to 0 ( $G^2 = 112.687$  with 13 df), meaning that there is a strong departure from the null hypothesis. Combining these results, one can conclude that the two tables have strong differences in terms of patterns of mobility.

Among the future directions of this research we mention: (a) the theoretical characterization of the relevant Markov bases in order to apply our technique also to large tables; (b) the use of this technique to make inference on other measures of mobility based on log-linear models, see [13] and [4] for an introductory overview of these measures with several examples from surveys in European nations.

## References

1. Agresti, A.: *Categorical Data Analysis*, 3 edn. Wiley, New York (2013)
2. Aoki, S., Hara, H., Takemura, A.: *Markov Bases in Algebraic Statistics*. Springer, New York (2012)
3. Bishop, Y.M., Fienberg, S., Holland, P.W.: *Discrete multivariate analysis: Theory and practice*. MIT Press, Cambridge (1975)
4. Breen, R.: *Social Mobility in Europe*. Oxford University Press, Oxford (2007)
5. Drton, M., Sturmfels, B., Sullivant, S.: *Lectures on Algebraic Statistics*. Birkhauser, Basel (2009)
6. Goodman, L.A.: Contributions to the statistical analysis of contingency tables: Notes on quasi-symmetry, quasi-independence, log-linear models, log-bilinear models, and correspondence analysis models. *Ann. Fac. Sci. Toulouse* **11**(4), 525–540 (2002)
7. Kateri, M.: *Contingency Table Analysis. Methods and Implementation Using R*. Springer, New York (2014)
8. Pistone, G., Riccomagno, E., Wynn, H.P.: *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman&Hall/CRC, Boca Raton (2001)
9. Rapallo, F.: Algebraic Markov bases and MCMC for two-way contingency tables. *Scand. J. Statist.* **30**(2), 385–397 (2003)
10. Rapallo, F.: Algebraic exact inference for rater agreement models. *Stat. Methods Appl.* **14**(1), 45–66 (2005)
11. Rapallo, F.: Toric statistical models: Parametric and binomial representations. *Ann. Inst. Statist. Math.* **59**(4), 727–740 (2007)
12. 4ti2 team: 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces. Available at [www.4ti2.de](http://www.4ti2.de) (2008)
13. Xie, Y.: The log-multiplicative layer effect model for comparing mobility tables. *American Sociological Review* **57**, 380–395 (1992)



# Testing Beta-Pricing Models Using Large Cross-Sections

## *Test per modelli di pricing multifattoriali utilizzando cross-sections di grandi dimensioni*

Valentina Raponi, Cesare Robotti and Paolo Zaffaroni

**Abstract** Building on the Shanken (1992) estimator, we develop a new methodology for estimating and testing beta-pricing models when a large number of assets  $N$  is available but the number of time-series observations is small. We show empirically that our large  $N$  framework poses a serious challenge to common empirical findings regarding estimated risk premia and validity of beta-pricing models. We generalize our theoretical results to the more realistic case of unbalanced panels. The practical relevance of our findings is confirmed via Monte Carlo simulations.

**Abstract** *Partendo dal timatore di Shanken (1992), il paper introduce una nuova metodologia per stimare e testare modelli di mercato quando il numero di assets  $N$  disponibili per l'analisi è molto elevato, ma la dimensione temporale  $T$  è piccola. Dal punto di vista empirico, viene mostrato come questa nuova metodologia sia in grado di fornire risultati molto diversi da quelli che solitamente si otterrebbero applicando la metodologia standard. I risultati teorici, sia in termini di stima che di test, vengono generalizzati al caso più realistico di panel non bilanciati. L'importanza dei risultati teorici viene confermata anche da un esercizio di simulazione Monte Carlo.*

**Key words:** Beta-pricing models; Ex-post risk premia; Two-pass cross-sectional regression; Large  $N$  asymptotics; Specication test; Unbalanced panel.

## 1 Introduction

Tens of thousands of stocks are traded every day in financial markets, providing an extremely rich information set to validate and estimate asset pricing models. At the same time, both academics and practitioners could be reluctant to use time-series spanning long time periods to avoid the risk of including structural breaks and to

---

Valentina Raponi  
Imperial College London, e-mail: v.raponi13@imperial.ac.uk

avoid the additional difficulty of parameterizing time variation of betas and risk premia.

Therefore, it is important to have a methodology that allows for carrying out statistically correct inference on risk premia and testing the validity of beta-pricing models, by exploring the large available cross-sectional variation of returns across  $N$  individual securities and, at the same time, relying on a limited number of time-series observations,  $T$ . Our main contribution in this paper is to develop such a formal methodology, built on the large- $N$  estimator proposed by Shanken (1992).

Theoretically, we provide a rigorous statistical analysis of the Shanken (1992) estimator of the ex-post risk premia and, more in general, provide a formal methodology for estimation and testing of beta-pricing models in a large- $N$  environment. To provide further motivation to our analysis, we show that the Shanken estimator is an element, with a special property, of a class of OLS bias-adjusted estimators of ex-post risk premia. In particular we demonstrate mathematically that it is the only element of this class not requiring a preliminary estimation of the bias-adjustment. This is a particularly convenient feature because it avoids, for example, any pre-testing biases and, at the same time, it does not require sacrificing data for preliminary estimation. We then focus on the asymptotic properties of the Shanken estimator for large  $N$  and fixed  $T$ : under mild, easily verifiable assumptions, in particular permitting a degree of cross-correlation among returns, we establish  $\sqrt{N}$ -consistency and asymptotic normality. Moreover, we derive an explicit, and easy to interpret, expression for its asymptotic covariance matrix, showing how it can be consistently estimated and used to conduct inference on the risk premia estimates.

In addition to estimation, we provide a new test for the validity of the asset-pricing restrictions and characterize its distribution for large  $N$  and fixed  $T$ , under the null hypothesis that the model is correctly specified. Noticeably, our test has power, that is, it is able to discriminate whether the beta-pricing model is correctly specified, despite being built on the ex-post pricing errors, which are, necessarily, contaminated by the unexpected factor's outcomes.

To further motivate the importance of our methodology, we also explore the finite- $N$  properties of the Shanken estimator and of our test statistic via Monte Carlo experiments, and compare them with the properties of traditional methodologies. Our simulations highlight how, when  $N$  is much larger than  $T$ , the inference, based on the traditional t-statistics, can be severely misleading, even when accounting for the correct large- $T$  standard errors. In contrast, the t-statistics of the Shanken estimator, based on our large- $N$  standard errors, are correctly sized.

We demonstrate the usefulness of our methodology by means of an empirical analysis that employs individual monthly stock returns from the CRSP database over overlapping three-, six- and 10-year periods from 1966 until 2013. The three prominent beta-pricing specifications that we consider are the Capital Asset Pricing Model (CAPM) of Sharpe (1964) and Lintner (1965), the three-factor Fama and French (1993) model, and the recently proposed five-factor Fama and French (2015) model. We find significant pricing ability for all the factors, for most periods, for each of the three models, even when using a relatively short time window of three years. In contrast, the same risk premia appear insignificantly different from

zero when using the traditional approach of estimating risk premia based on large- $T$  asymptotics. In terms of asset pricing test, our methodology tends to reject the CAPM even when using a short time window, in contrast to the traditional large- $T$  approach of Gibbons, Ross and Shanken (1989).

Although computationally appealing, it remains to verify whether the Shanken estimator  $\hat{\Gamma}^*$  exhibits desirable (asymptotic) statistical properties. This is studied in the next Section, where we provide a formal asymptotic analysis of  $\hat{\Gamma}^*$ .

## 2 Asymptotic Analysis

The analysis in this section assumes that  $N \rightarrow \infty$  and  $T$  is fixed. We first establish the limiting distribution of the Shanken bias-adjusted estimator  $\hat{\Gamma}^*$  and explain how its asymptotic covariance matrix can be consistently estimated. We then characterize the limiting behavior of our test  $\mathcal{S}^*$  of the asset-pricing restriction.

### 2.1 Asymptotic Distribution of the Shanken Estimator

In this subsection, we study the asymptotic distribution of  $\hat{\Gamma}^*$ , under the assumption that the model is correctly specified, namely that exact no-arbitrage holds (Assumption 4 in Appendix A).

Let  $\Sigma_X = \begin{bmatrix} 1 & \mu'_\beta \\ \mu_\beta & \Sigma_\beta \end{bmatrix}$ ,  $\sigma^2 = \lim \frac{1}{N} \sum_{i=1}^N \sigma_i^2$ ,  $U_\varepsilon = \lim \frac{1}{N} \sum_{i,j=1}^N E \left[ \text{vec}(\varepsilon_i \varepsilon'_i - \sigma_i^2 I_T) \text{vec}(\varepsilon_j \varepsilon'_j - \sigma_j^2 I_T)' \right]$ ,  $M = I_T - D(D'D)^{-1}D'$ , where  $I_T$  is a  $T \times T$  identity matrix,  $D = [1_T, F]$ ,  $Q = \frac{1}{T} - \mathcal{P} \gamma_1^p$ , and  $Z = (Q \otimes \mathcal{P}) + \frac{\text{vec}(M)}{T-k-1} \gamma_1^p \mathcal{P}' \mathcal{P}$ , where all the limits are finite by our assumptions, as  $N \rightarrow \infty$ . In the following theorem, we provide the rate of convergence and the limiting distribution of  $\hat{\Gamma}^*$ .

**Theorem 1.**

(i) Under Assumptions 1–5 (listed in Appendix A),

$$\hat{\Gamma}^* - \Gamma^P = O_p \left( \frac{1}{\sqrt{N}} \right). \tag{1}$$

(ii) Under Assumptions 1–6 (listed in Appendix A),

$$\sqrt{N} (\hat{\Gamma}^* - \Gamma^P) \xrightarrow{d} \mathcal{N} (0_{K+1}, V + \Sigma_X^{-1} W \Sigma_X^{-1}), \tag{2}$$

where

$$V = \frac{\sigma^2}{T} \left[ 1 + \gamma_1^{p'} (\tilde{F}' \tilde{F} / T)^{-1} \gamma_1^p \right] \Sigma_X^{-1} \tag{3}$$

and

$$W = \begin{bmatrix} 0 & 0'_K \\ 0_K & Z'U_\varepsilon Z \end{bmatrix}. \tag{4}$$

**Proof:** See Appendix C and Lemmas 1 to 5 in Appendix B.

To conduct statistical inference, we need a consistent estimator of the asymptotic covariance matrix  $V + \Sigma_X^{-1}W\Sigma_X^{-1}$ . Let  $M^{(2)} = M \odot M$ , where  $\odot$  denotes the Hadamard product operator. In addition, define

$$\hat{\sigma}_4 = \frac{\frac{1}{N} \sum_{t=1}^T \sum_{i=1}^N \hat{\varepsilon}_{it}^4}{3\text{tr}(M^{(2)})}, \tag{5}$$

and let

$$\hat{Z} = (\hat{Q} \otimes \mathcal{P}) + \frac{\text{vec}(M)}{T - K - 1} \hat{\gamma}_1^* \mathcal{P}' \mathcal{P}, \tag{6}$$

with

$$\hat{Q} = \frac{1}{T} - \mathcal{P} \hat{\gamma}_1^*. \tag{7}$$

The following theorem provides a consistent estimator of the asymptotic covariance matrix of the  $\hat{\Gamma}^*$  estimator.

**Theorem 2.** *Under Assumptions 1–5 (listed in Appendix A), we have*

$$\hat{V} + (\hat{\Sigma}_X - \hat{\Lambda})^{-1} \hat{W} (\hat{\Sigma}_X - \hat{\Lambda})^{-1} \xrightarrow{p} V + \Sigma_X^{-1}W\Sigma_X^{-1}, \tag{8}$$

where

$$\hat{V} = \frac{\hat{\sigma}^2}{T} \left[ 1 + \hat{\gamma}_1^* (\bar{F}'\bar{F}/T)^{-1} \hat{\gamma}_1^* \right] (\hat{\Sigma}_X - \hat{\Lambda})^{-1}, \tag{9}$$

$$\hat{W} = \begin{bmatrix} 0 & 0'_K \\ 0_K & \hat{Z}'\hat{U}_\varepsilon\hat{Z} \end{bmatrix}, \tag{10}$$

and  $\hat{U}_\varepsilon$  is a consistent plug-in estimator of  $U_\varepsilon$  described in Appendix D.

**Proof:** See Appendix C and Lemmas 1 to 6 in Appendix B.

## 2.2 Limiting Distribution of the Specification Test

The null hypothesis underlying the asset-pricing restriction can be formulated as

$$H_0: e_i = 0 \quad \text{for every } i = 1, 2, \dots, \tag{11}$$

where  $e_i = E[R_{it}] - \gamma_0 - \beta'_i \gamma_1$  is the pricing error associated with asset  $i$ . The null hypothesis  $H_0$  easily follows by simply rewriting Assumption 4. Let  $X_i = [1, \beta'_i]$ ,  $\hat{X}_i = [1, \hat{\beta}'_i]$ , and denote by  $\hat{e}_i^p$  the ex-post sample pricing error for asset  $i$ . Then, we have

$$\hat{e}_i^P = \bar{R}_i - \hat{X}_i \hat{\Gamma}^* \quad (12)$$

Since we estimate  $\Gamma^P$  via OLS cross-sectional regressions, we propose a test based on the sum of the squared ex-post sample pricing errors, that is,

$$\hat{\mathcal{Q}} = \frac{1}{N} \sum_{i=1}^N (\hat{e}_i^P)^2. \quad (13)$$

Consider the centered statistic

$$\mathcal{S} = \sqrt{N} \left( \hat{\mathcal{Q}} - \frac{\hat{\sigma}^2}{T} (1 + \hat{\gamma}_1' (\bar{F}' \bar{F} / T)^{-1} \hat{\gamma}_1^*) \right). \quad (14)$$

The following theorem provides the limiting distribution of  $\mathcal{S}$  under  $H_0 : e_i = 0$  for all  $i$ .

**Theorem 3.** *Under Assumptions 1–6 (listed in Appendix A), implying that  $H_0 : e_i = 0$  holds for all  $i$ , we have*

$$\mathcal{S} \xrightarrow{d} \mathcal{N}(0, \mathcal{V}), \quad (15)$$

where  $\mathcal{V} = Z_Q' U_\varepsilon Z_Q$  and  $Z_Q = (Q \otimes Q) - \frac{\text{vec}(M)}{T-k-1} Q' Q$ .

**Proof:** See Appendix C and Lemmas 1 to 5 in Appendix B.

### 3 Empirical Analysis

In this section, we empirically estimate the risk premia associated with some prominent beta-pricing models, using individual stock return data, and investigate their performance. This demonstrates how the empirical results obtained using our large  $N$  methodology, illustrated in the previous section, can differ, even dramatically, from the results obtained with the more traditional large  $T$  methodologies. We consider three linear beta-pricing models: (i) the single-factor CAPM, (ii) the three-factor model of Fama and French (1993, FF3), and (iii) the five-factor model recently proposed by Fama and French (2015, FF5). The data on the above factors is available from Kenneth French's website. We use monthly data on individual stocks from the CRSP database, available from January 1966 to December 2013. We carry out the empirical analysis using balanced panel with three different time windows of, respectively, three-, six- and ten-year (i.e.,  $T=36, 72$ , and  $120$ , respectively). For each of these time windows, we estimate each of the above beta-pricing models by rolling the window one month at the time. In this way, we obtain time-series of estimated risk premia and of the test statistic based on overlapping time windows of fixed length  $T$ .

We document a sizeable difference between the results of our large  $N$  approach and the results of the conventional large  $T$  approaches. This outcome is a combination of two elements, the extremely small standard errors associated with a very

large  $N$  and the bias-correction of the Shanken estimator that leads to an increment, on average of about 20% but sometimes up to 50%, of the risk premia estimates over the OLS estimator.

We finally consider the performance of our specification test  $\mathcal{S}^*$ . The result is, again, rather striking: in contrast to our large- $N$  test, the GRS test is almost always unable to reject the CAPM at 5% when considering the shortest time window of three years of data ( $T = 36$ ). The CAPM will be rejected about half of the time for the six years window and will almost always be rejected for the long time window of 10 years.

## 4 Conclusion

This paper is concerned with estimation of risk premia and testing of beta-pricing models when data is available for a large cross-section of securities  $N$  but only for a limited number of time periods. Because in this context the CSR OLS estimator of the risk premia is asymptotically biased and inconsistent, the focus of the paper is on the bias-adjusted estimator of the ex-post risk premia proposed by Shanken (1992). In terms of estimation, we demonstrate that the Shanken estimator exhibits desirable properties, such as  $\sqrt{N}$ -consistency and asymptotic normality, as  $N$  diverges. In terms of testing, we propose a new test of the no-arbitrage asset pricing restriction and establish its asymptotic distribution (assuming that the restriction holds) as  $N$  diverges. Finally, we show how our results can be extended to deal with the more realistic case of unbalanced panels, allowing us to take advantage of the large cross-sections of stocks existing only for certain time periods. Monte Carlo simulations corroborate our theoretical finding, both in terms of estimation and in terms of testing for the asset pricing restriction.

The usefulness of our methodology is demonstrated by means of an empirical analysis that employs individual monthly stock returns from the CRSP database. We find some convincing pricing ability for all the factors, to different degrees, for each of the three models, even when using a relatively short time window of three years, for most periods.

## References

1. Fama, E. F., and K. R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.
2. Gagliardini, P., E. Ossola, and O. Scaillet, 2016, Time-varying risk premium in large cross-sectional equity datasets, *Econometrica*, 84, 985–1046.
3. Kan, R., C. Robotti, and J. Shanken, 2013, Pricing model performance and the two-pass cross-sectional regression methodology, *Journal of Finance* 68, 2617–2649.
4. Shanken, J., 1992, On the estimation of beta-pricing models, *Review of Financial Studies* 5, 1–33.

# On the use of predictive methods for ship fuel consumption analysis from massive on-board operational data

## *Metodi di analisi predittiva dei consumi di una nave mediante dati massivi di navigazione*

Marco Seabra dos Reis<sup>1</sup>, Biagio Palumbo, Antonio Lepore, Ricardo Rendall, Christian Capezza

**Abstract** Measuring, reporting and verification of ship fuel consumption are the main requirements imposed by upcoming European regulations. However, the massive amount of navigation data resulting from ship computerization is not easily handled by shipping operators because of the lack of standardized solutions. In this context, modern statistical and machine learning techniques provide effective methods to exploit the massive operational data available on modern ships and, in particular, can be used for building predictive models to estimate fuel consumption. With resort to real operational data collected from a Ro-Pax cruise ship owned by the Italian shipping company Grimaldi Group, this paper presents an extensive comparison study of modern predictive analytical methods (e.g. variable selection, penalized regression, latent variable methods and tree-based ensembles) in order to explore new directions in the analysis of ship fuel consumption.

**Abstract** *Le nuove regolamentazioni europee impongono, tra i principali requisiti, il monitoraggio, la documentazione e la verifica del consumo di carburante delle navi. L'enorme quantità di dati di navigazione disponibile mediante i moderni sistemi di acquisizione installati a bordo delle navi non è del tutto utilizzata dalle compagnie armatoriali per la mancanza di opportune tecniche di analisi. In tale scenario, è sempre più evidente la necessità di esplorare tecniche statistiche e di apprendimento automatico al fine di poter adeguatamente interpretare tali dati. Il presente lavoro propone un confronto critico dei metodi di analisi predittiva (e.g., metodi di selezione delle variabili, regressione penalizzata, analisi delle variabili latenti e metodi ensemble) mediante dati di navigazione reali acquisiti a bordo di una nave da carico e passeggeri, di proprietà della società armatoriale italiana Grimaldi Group S.p.a..*

**Key words:** fuel consumption prediction, variable selection method, penalized regression, latent variable methods, tree-based ensembles.

---

<sup>1</sup>Marco Seabra dos Reis  
Department of Chemical Engineering, University of Coimbra, email: marco@eq.uc.pt

## 1 Introduction

In the last decades, increasing emissions of greenhouse gases by maritime transport has urged the European Commission to implement a new regulatory regime, requiring shipping companies to adopt procedures for the measuring, reporting, and verification of CO<sub>2</sub> emissions. This is achieved by analyzing fuel consumption data, which is directly related to CO<sub>2</sub> emissions. In this regard, the marine engineering literature mainly relies on empirical curves that quantify fuel consumption as a function only of the vessel's speed over ground. These curves are based on dedicated experiments where external/noise factors can be controlled and set to standard conditions [8], but they are not applicable in real environments where a large number of other variables also influence fuel consumption. On the other hand, modern ships can record a great amount of multi-sensor operational data through on-board automatic acquisition systems. These data can be analyzed using statistical and machine learning techniques in order to develop new solutions to the fuel prediction challenge. Thus, in this paper, modern predictive analytics based on four classes of methods (variable selection, penalized regression, latent variable methods and tree-based ensembles) are compared and discussed in order to explore new directions in the analysis of ship fuel consumption based on operational data (i.e. under non-standard conditions).

## 2 Data Description and Comparison Framework

Operational data were collected from a Ro-Pax ship owned by the Italian shipping company Grimaldi Group. The data cover one year's worth of relevant observations and for confidentiality reasons, the ship's name, route and voyage dates are intentionally omitted. The response variable is the fuel consumption per hour (FCPH) for each voyage. Table 1 shows the variables used to describe the ship's operating conditions, which also serve as predictor variables in order to estimate FCPH. Further details can be found in [2].

In order to explore the operational dataset, the statistical and machine learning literature offers a diverse set of predictive methods that can be applied to predict fuel consumption based on data collected during the ship's voyage. In this paper, the following classes of methods are investigated: variable selection, penalized regression, latent variable and tree-based ensemble methods. Note that multiple linear regression (MLR), one of the most tested and studied methods, is not considered because it does not cope with some characteristics of the dataset, namely the high collinearity among some predictors, which leads to unstable estimations of the regression coefficients and poor prediction intervals [4]. In this scenario, alternative methods are preferred to overcome the limitations of MLR.

**Variable selection methods** stand on the assumption that only some variables have relevant predictive power, while the others can be discarded [1]. Forward stepwise



regression (FSR) is selected as the representative method of this class and is implemented by a sequential algorithm that evaluates, at each step, the predictive power of incrementally larger and smaller models through appropriate statistics. The statistic used to choose or discard variables is the p-value of a partial F-test and variables with p-values smaller than a specified threshold ( $p_{in}$ ) are included in the model while variables already in the model but with p-values bigger than a tolerance ( $p_{out}$ ) are removed.

**Penalized regression methods** introduce a penalty on the magnitude of the regression coefficients to make them stable and smaller. This increases model bias, but stabilizes the estimator variance [7]. In this class, four methods were considered: support vector regression (SVR), elastic net (EN), ridge regression (RR) and least absolute shrinkage and selection operator (LASSO). SVR minimizes the sum of squared regression coefficients to reduce model variance while constraining the prediction error to be below a threshold  $\epsilon$ , although slack variables are used to allow some errors to be above  $\epsilon$  [11]. The EN model is obtained by solving the following optimization problem:

$$\hat{\mathbf{b}}_{EN} = \arg \min_{\mathbf{b}=[b_0, \dots, b_j]^T} \left\{ \sum_{i=1}^n (y(i) - \hat{y}(i))^2 + \gamma \left( \alpha \sum_{j=1}^p |b_j| + \frac{1-\alpha}{2} \sum_{j=1}^p b_j^2 \right) \right\}.$$

The two hyper-parameters are  $\gamma$  that controls the bias-variance tradeoff and  $\alpha$  that weights the squared ( $b_j^2$ ) and the norm  $|b_j|$  penalties. RR is obtained by setting  $\alpha = 0$ , while LASSO is obtained by setting  $\alpha = 1$ .

**Latent variable methods** are based on the assumption that the variability shared by predictors and response variables can be explained by a set of unmeasured quantities, called latent variables, estimated as linear combinations of the measured variables and used to predict the response. In this class, principal component regression (PCR), principal component regression with the scores added in a forward stepwise fashion (PCR\_FS), and PLS regression are considered. PCR [9] decomposes the predictor space using principal component analysis (PCA), since most of its variability can often be explained by a number of principal components ( $\alpha_{PCR}$ ) smaller than the number of predictors. Then, the principal components are used as predictors of the response variable and MLR is used to estimate their regression coefficients. In PCR\_FS, the principal components are selected using the forward stepwise algorithm based on the p-value of the partial F-test, following an iterative process similar to FSR. PLS regression [10] chooses  $\alpha_{PLS}$  latent variables that maximize covariance between predictors and response variable.

**Tree-based ensemble methods** [3] iteratively split the predictors' space into smaller regions, reducing the response variability. In this work, trees are built until a minimum of five samples for each region are obtained and three methods were considered: bagging of regression trees (BaRTs), random forests (RFs) and boosting of regression trees (BoRTs). BaRTs and RFs are based on bootstrap to generate many datasets, which are used to build regression trees. The predicted response is the prediction average from all trees in the ensemble. The number of trees,  $T_{BRT}$  and

$T_{RF}$ , are the hyper-parameters that control the bias-variance tradeoff for BaRT and RF, respectively. Moreover, the BoRT method [5] exploits the residuals from previous trees to fit new regression trees. Its parameters are the learning rate  $u$  (fixed as  $u=0.02$ ) and the number of trees  $T_{BT}$ .

The prediction performances of each regression technique is evaluated through the root mean squared error of double cross-validation [6]  $RMSE_{dev}$ . The dataset is randomly partitioned into a training and a test set: the former (80% of the data) is used to select the model hyper-parameters (Table 2), whereas the latter (20% of the data) is used for assessing prediction performance and to compute the  $RMSE_{dev}$ . Since this measure is affected by the initial split of the dataset, the procedure is iterated 40 times in order to obtain a measure of variability. Moreover, 10-fold cross-validation is adopted for selecting hyper-parameters during model training and variables are transformed to zero mean and unit variance.

**Table 1:** Operational variables measured for each voyage.

	<i>Variable</i>	<i>Description</i>
1	$SG_p$	Shaft generator power (port) [kW]
2	$SG_s$	Shaft generator power (starboard) [kW]
3	$\Delta P$	Power difference between port and starboard propeller shafts [kW]
4	$\Delta SG$	Power difference between two shaft generators [kW]
5	$V$	Speed Over Ground (SOG) [kn]
6	$W_F$	Following wind [kn]
7	$W_H$	Head wind [kn]
8	$W_S$	Side wind [kn]
9	$T_{FD}$	Departure draught (fore perpendicular) [m]
10	$T_{AD}$	Departure draught (aft perpendicular) [m]
11	$T_{PD}$	Departure draught (midship section - port) [m]
12	$T_{SD}$	Departure draught (midship section - starboard) [m]
13	$T_{FA}$	Arrival draught (fore perpendicular) [m]
14	$T_{AA}$	Arrival draught (aft perpendicular) [m]
15	$T_{PA}$	Arrival draught (midship section - port) [m]
16	$T_{SA}$	Arrival draught (midship section - starboard) [m]
17	$\sigma_V^2$	SOG Variance [kn <sup>2</sup> ]
18	$Trim_D$	Departure trim [m]
19	$Trim_A$	Arrival trim [m]
20	$\Delta$	Displacement [Mt]

### 3 Results and discussion

Prior to the application of the comparison procedure described above and to the analysis of the best predictive methods, a pre-analysis of the operational data is conducted to identify potentially predictive variables and to summarize their main

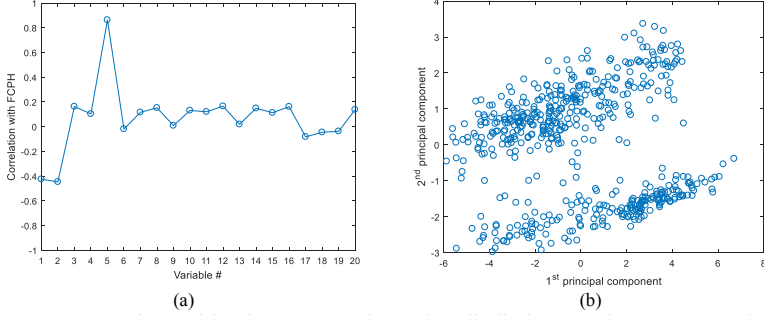
**Table 2:** Summary table of the comparison framework with hyper-parameter value(s) considered for model training. For each of the 40 iterations, suitable values are selected by 10-fold cross-validation.

<i>Class</i>	<i>Method</i>	<i>Hyper-parameters</i>	<i>Hyper-parameter value(s)</i>
Variable selection	FSR	$p_{in}$	0.05
		$p_{out}$	0.1
Penalized regression	RR	$\gamma$	0.002; 0.02; 0.2; 2; 20
	LASSO	$\gamma$	0.001; 0.01; 0.1; 1; 10
	EN	$\alpha$	0.001; 0.01; 0.1; 1
	EN	$\gamma$	0.002; 0.02; 0.2; 2; 20
	SVR	$\epsilon$	0.001; 0.005; 0.01; 0.05; 0.1
Latent variable	PCR	$\alpha_{PCR}$	[1:min(20, n, p)]
	PCR_FS	$p_{in}$	0.05
		$p_{out}$	0.1
PLS	$\alpha_{PLS}$	[1:min(20, n, p)]	
Tree based ensemble	BRT	$T_{BRT}$	50; 100; 500; 1000; 5000
	RF	$T_{RF}$	50; 100; 500; 1000; 5000
	BT	$T_{BT}$	50; 100; 500; 1000; 5000

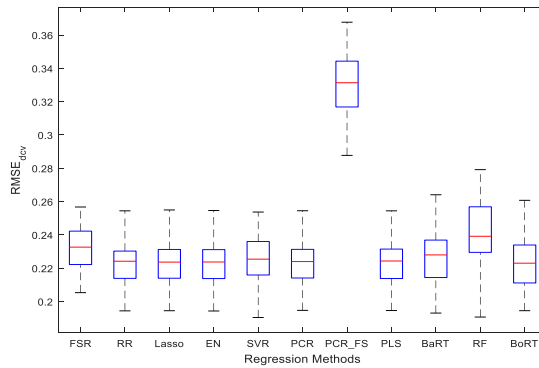
characteristics. The first goal is achieved by computing the Pearson correlation between each predictor variable and FCPH as reported in Figure 1.a, where one can observe that speed over ground (variable #5) has the highest correlation coefficient followed by the starboard shaft generator power (variable #2) and the port generator power (variable #1). Thus, quantifying FCPH based only on speed over ground, as used for building empirical curves, is most likely the best univariate approach. However, by adopting multivariable methods, the information content of other predictors can also be used for obtaining more reliable predictions of FCPH.

In order to perform an exploratory analysis, PCA was applied to the set of predictor variables. The first and second principal components are presented in Figure 1.b. Analyzing Figure 1.b, one can notice the existence of two clusters, which correspond to different levels of the port and starboard shaft generator (variable #1 and #2, respectively). These clusters occur because the generator is turned off in some voyages and suggest that the regression methods might be applied separately to each cluster. However, as will be presented shortly, variable #1 and #2 were not important for the regression models and no clusters were observed for the response variable. To assess the prediction performance of the various regression methods included in the comparison study, Figure 2 presents the distribution of  $RMSE_{dev}$  obtained over 40 iterations of double cross-validation.

Analyzing Figure 2, one can note that most regression methods present similar prediction errors (the median  $RMSE_{dev}$  is close to 0.22) and only PCR\_FS and RF are poor choices since their prediction errors are higher than the other methods. In other words, the general conclusion is that all methods can predict FCPH equally well, except PCR\_FS and RF, and the choice should fall on the simpler method, that is the method that has a smaller number of model parameters. Thus, the recommended method is LASSO as it tends to discard irrelevant variables and produce a sparse structure in regression coefficients. In order to identify important variables,



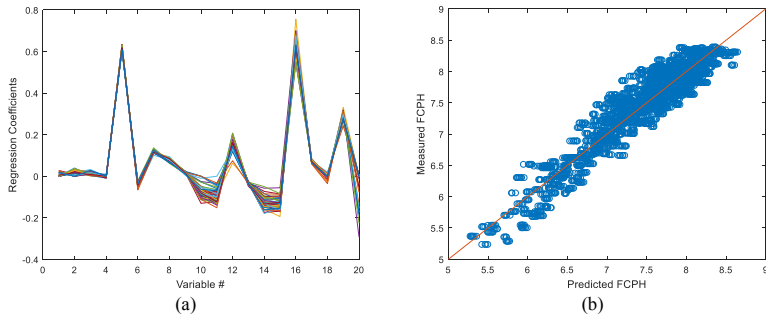
**Figure 1.** Assessing variable importance and samples' distribution: (a) the Pearson correlation coefficient between each predictor and FCPH and (b) the first two principal components from PCA.



**Figure 2.** Distribution of  $RMSE_{dev}$  in 40 iterations of double cross-validation.

Figure 3.a presents the regression coefficients obtained for LASSO in the 40 iterations of double cross-validation. One can observe that speed over ground (variable #5) is one of the most important variables, as expected from its high correlation with FCPH. Furthermore, arrival draught (variable #16) has, in most iterations, the same weight as speed over ground and arrival trim (variable #19) consistently contributes to the model.

In terms of irrelevant predictors, variables #1-4 have regression coefficients very close to zero and can be discarded from the model. Lastly, Figure 3.b further corroborates the validity of the LASSO models and presents the predicted and measured FCPH over all 40 iterations of double cross-validation. Figure 3.b shows a good agreement between predicted and measured values, the values fall close to the identity line and no clusters are observed. Furthermore, the median coefficient of determination over all iterations is 0.88, corroborating the ability of the developed models to predict FCPH.



**Figure 3.** Results obtained for LASSO during model training: (a) the regression coefficients and (b) the predicted and measured FCPH.

## 4 Conclusions

In this work, we assessed the potential for using operational data collected from a Ro-Pax ship to predict its fuel consumption per hour (FCPH). The collected dataset covers a period of one year and contained 20 predictor variables. In order to build regression models, four classes of regression methods were considered: variable selection, penalized regression, latent variable and tree-based ensembles. Within each class, representative methods were selected in order to account for a wide range of a priori assumptions regarding the distribution of predictors, response variable and the relation between them.

The regression methods were compared using 40 iterations of double cross-validation. In each iteration, the root mean squared error ( $RMSE_{dev}$ ) was computed. The distribution of  $RMSE_{dev}$  allows the estimation of the variability in the methods' performance, resulting in a more robust comparison.

The application of the regression methods to the collected dataset revealed that the choice of the regression method is not particularly important since most methods, except PCR\_FS and RF, presented a similar distribution of  $RMSE_{dev}$ . Nevertheless, the recommended method was LASSO as it often eliminates irrelevant variables by the application of a suitable penalty to the magnitude of regression coefficients. Furthermore, the good agreement observed between predicted and measured FCPH was confirmed also by the median coefficient of determination over the 40 iterations of double cross-validation (0.88).

## References

1. Andersen, C.M., Bro R.: Variable selection in regression—a tutorial. *J. Chemom.* 24, 728-737 (2010) doi: 10.1002/cem.1360
2. Bocchetti, D., Lepore A., Palumbo B., Vitiello L.: A Statistical Approach to Ship Fuel Consumption Monitoring. *J. Ship Res.* 59, 162-171 (2015) doi: 10.5957/JOSR.59.3.150012
3. Breiman, L., Friedman J., Stone C.J., Olshen R.A.: Classification and regression trees. CRC press (1984)
4. Draper, N.R., Smith H.: Applied regression analysis. John Wiley & Sons (2014)
5. Elith, J., Leathwick J.R., Hastie T.: A working guide to boosted regression trees. *J. Anim. Ecol.* 77, 802-813 (2008) doi: 10.1111/j.1365-2656.2008.01390.x
6. Filzmoser, P., Liebmann B., Varmuza K.: Repeated double cross validation. *J. Chemom.* 23, 160-171 (2009) doi: 10.1002/cem.1225
7. Hesterberg, T., Choi N.H., Meier L., Fraley C.: Least angle and  $\ell_1$  penalized regression: A review. *Stat. Surv.* 2, 61-93 (2008) doi: 10.1214/08-SS035
8. IMO: 15016:2015 - Ships and marine technology - Guidelines for the assessment of speed and power performance by analysis of speed trial data, ISO/TC 8/SC 6. 2015.
9. Jolliffe, I.: Principal component analysis. Wiley Online Library (2002)
10. Martens, H., Naes T.: Multivariate calibration. John Wiley & Sons (1989)
11. Smola, A.J., Schölkopf B.: A tutorial on support vector regression. *Stat. Comput.* 14, 199-222 (2004) doi: 10.1023/B:STCO.0000035301.49549.88

# Twitter as a Statistical Data Source: an Attempt of Profiling Italian Users' Background Characteristics

## *Twitter come fonte di dati: un tentativo di individuare le caratteristiche di background degli utilizzatori italiani*

Righi Alessandra and Gentile Mauro Mario

**Abstract** Social media (SM) are becoming an important data source about the opinions and the sentiment of their users because they allow to capture in real-time and in a not solicited way what the users think about a certain topic. In Italy Twitter appears to be one of the most used media and it has a greater accessibility and allows a more readily text analysis. This paper presents an attempt of profiling the Italian twitterers (for research purposes only) carried out at national level using a REST API downloading method. This knowledge would allow to better know the representativeness of users and, consequently, to correct the strong selectivity of the SM users. The technological/statistical approach used and the main results are presented.

**Abstract** *I social media (SM) stanno diventando una fonte di dati importante circa le opinioni e il sentiment dei loro utenti perché consentono di acquisire in tempo reale e in modo non sollecitato ciò che gli utenti pensano su un determinato argomento. In Italia Twitter sembra essere uno dei più utilizzati, ha una maggiore accessibilità e permette agevolmente l'analisi dei testi. Questo articolo presenta un tentativo di profilazione dei twitterers italiani (solo per scopi di ricerca) svolto a livello nazionale utilizzando un metodo di scaricamento REST API. La conoscenza di queste informazioni permetterebbe di identificare meglio la rappresentatività degli utilizzatori e, di conseguenza, di correggere la forte selettività degli utenti di Twitter. L'approccio tecnologico / statistico utilizzato e i principali risultati vengono presentati.*

**Key words:** Big Data, Social media.

---

<sup>1</sup>

Righi Alessandra, Istat; email: righi@istat.it

Gentile Mauro Mario, Istat external collaborator; email:mauro.gentile@iesse.net

## 1 Introduction

In recent years Social media (SM) are becoming an important data source about the opinions and the sentiment of their users because they allow to capture in real-time and in a not solicited way what the users think about a certain topic. In Italy Facebook and Twitter appear to be the most used media and the latter has a greater accessibility and allows a more readily text analysis (Della Ratta et al, 2016).

Twitter is a microblogger service which let users post 140 characters length tweets about anything. Created in 2006, the worldwide service today, according Statista Website, averages out at 319 million monthly active users and, according Alexa Website, Twitter is the ninth most visited site in the world.

At national level some estimates, such as the Total Digital Audience provided by Audiweb (a private impartial entity monitoring the Internet audience data in Italy) and Nielsen, quantify in more than 6.4 million Italians “active” Twitter users.

Anyway, before using Social media as a real statistical source some challenges should be faced. They concern the representativeness and the guarantee of time stability of the source and the need to know who the users are in terms of socio-economic characteristics. This would allow to correct the strong selectivity of the Social media users (Daas et al, 2016).

Unfortunately, official information about who are the users at national level is not available but, as the tweets are associated to metadata, some background characteristics information on the of the users and are publicly available.

This paper presents an attempt of profiling the Italian twitterers (for research purposes only) carried out at national level. Twitter data description and the technological and statistical approach used are in Section 2; the main results regarding the background characteristics (particularly on gender, location and active /non active status) of the users making use of the Italian language in their posts are in Section 3.

## 2 Data and Methods

Obtaining auxiliary information from units in Twitter is challenging especially because data are becoming widely available to researchers to predict financial tangibles as well as intangible assets, such as reputation (Bollen et al, 2011).

A method called ‘profiling’ is an interesting option to do this (Daas et al, 2016). We consider only data available on public Twitter and we use a REST API<sup>1</sup> downloading method to get information.

---

<sup>1</sup> Twitter exposes its data via an Application Programming Interface (API), the REST APIs provide programmatic access to read and write Twitter data and give responses in JSON format. The REST APIs allow to perform historical search queries on recently posted tweets and to retrieve lists of users, followers etc. Access to the API is associated to the personal Twitter account of the developer. Users



We used information derived from Socialbakers.com, a worldwide portal providing real-time statistics on twitterers with the largest audience in each country (in general terms and by specific topic, e.g. sport, entertainment, etc.). We started the downloading of the profiles of the followers of the Top ten Italian most popular twitter accounts, then continuing with those of the top accounts of each category in a decreasing order until when the finding of new profiles became increasingly limited and difficult. Besides that, for each downloaded user, we have downloaded the most recent tweets to evaluate the active status.

Downloaded data is thus divided in two different logic groups: on the one hand, we have some attributes supplied by the user, as user's name, nickname, a biographic description and the number of users that respectively follow and are followed by the user (giving indication of the popularity of the user). On the other hand, the full text of the tweets and some other information, as time and date of creation of the tweet, the place from where the tweet has been posted).

The user's name does not necessary display a person's name, on the contrary, both the user's name and the nickname can be filled with fancy names or brands. There are no specific variables indicating gender, age, or status of the user.

As for the search strategy, scripts have been developed in Python extensively exploiting Tweepy library.

### 3 Main results

Using the developed software, we obtained a 11 million sample of unique twitter usernames of Italian speaking users (the work is still in progress).

Our first goal with this data was to set the active/ non active status of the users defining as "active" someone having posted at least a tweet in the last four months. In this processing we used a subset of the first 3.7 million users' profiles downloaded and we verified that a very high proportion of users has never posted a tweet (around 1/3 of the sample) and only 30% of people posting tweets has sent at least one tweet in the last 4 months. Nevertheless, it was impossible to distinguish the proportion of users who just passively follows the exchange of tweets of other users, a mode of use which seems to be widespread enough.

The second goal was to try to define the user's gender as no one direct information is in the user's profile. Thus, we tried to set it from other information supplied by the user, mainly from the name, the nickname and the bio description (unfortunately, available for one out of four users only).

In the first attempt we used a wide list of the most frequent male/female Italian names to compare it with the users names to set the gender. Due to the presence in the sample of many foreigners living in Italy and writing tweets in Italian, we added a subset of foreign names to our list (around 1,500 names). In this way, we could

---

have to request a personal key for accessing the API. Moreover, the REST API poses limits in the number of requests that can be issued from a same user and in the number of the tweets that are returned.

assign a gender to around 68% of the users in the sample, calculating a share of 56% men and 44% women among the twitterers. We found also 8% of ambiguous cases and 23% of unclassifiable users.

Willing to complete the picture, a machine learning (ML) algorithm aimed at assigning a gender to the unclassified/ambiguous cases was developed, using the contents of two profile's fields "name" and "bio description". After having normalized texts (with text mining techniques as tokenization, elimination of stop words, punctuations, etc.) we weighted each term occurrence through an information retrieval technique called term frequency-inverse document frequency (Tf-idf) to give more relevance to terms containing highest discriminatory information and, then, we applied a ML algorithm for Logistic Regression. The algorithm has been tuned through Python GridSearch object to find the optimal set of parameters for the logistic regression model using a 5-fold stratified cross-validation. we used the n-gram range (from unigram to trigram), the applied regularization (none, L1, L2) and the regularization strength C as tuning parameters.

We divided the sample in two sub-set: the training set (with information relating to 70% of the users for whom we determined the gender) and the test set (with the remaining 30% of users). The best model found through GridSearchCV was applied to the test set and it led to a 75% accuracy score for our general imputation. Using this ML process we were able to predict the gender for a part of the subset of unclassified/ ambiguous cases, but some unclassifiable cases still remains (21% of the sample). It should be further investigated on these cases for understanding if this sub-set is composed of brands, associations or what else.

Anyway, this second process of imputation has slightly diminished the share of males in the sample to 55%.

An evaluation of the use of the self-declared professional status in the profile's bio description with information on to define the user's occupation was performed also. Unfortunately, the bio description item (containing information on the user's activity) is filled in 25% of cases only, even though some information could be erroneously contained in other profile's items (e.g., location, screen name). Even in a lesser percentage of cases the self-declared professional status is reported in the text of the tweets in a meaningful way. The most recurring professional conditions are students, journalists, architects, photographers, and even managers. It seems, however, difficult to get statistical information on a large scale in this way.

Even the "location" of the user in the Twitter profile is filled in a limited number of cases (24%) and the text expresses mostly the concept of domicile / place of activity in a playful way (e.g., "a place in the world" or " around the world") or in an unspecified way. In order to properly use these data (expressed as names of geographical places or even as geographic coordinates) a ML approach is needed trying to search for the geographical terms in various profile's items. Moreover, user's localization might be inferred from the tweets. However, only 15% of downloaded users allows the geolocation of tweets.

## 4 Conclusions

As the Social media access may cover a selective part of the target population, auxiliary information explaining the missingness of some sub-populations should be used to quantify and correct for selectivity. This work showed that among the commonly used auxiliary variables using the user's name and the short biography only the gender of the user can be determined for the entire sample with an acceptable degree of reliability. The result is that males are overrepresented among the Twitter's users. Even though possible distortions due to the download method can occur: starting from the most popular twitter accounts followers, in fact, the profiles of those who are more "isolated" in the social network could be achieved with greater difficulty or even not achieved.

This work is still in progress in order to complete the sample of Twitter's users but it seems to be promising. Among the open issues for future work there is the extension of the search for the professional status and the location to other profile's fields where this information could be unexpectedly found (due to the errors in the entry), or using the texts of the tweets. Moreover, a clear detection of the number of firm's accounts is also needed.

## References

1. Bollen J., Mao H, Zeng X.: Twitter Mood Predicts the Stock Market. *J.Comput.Sci.* 2(1), 1-8 (2011).
2. Daas P., Burger J.: Profiling Big Data Sources to Assess their Selectivity. Presented at *NTTS 2015* [https://ec.europa.eu/eurostat/cros/system/files/Daasetal\\_NTTS%202015%20abstract%20unblinded-v3z\\_903.pdf](https://ec.europa.eu/eurostat/cros/system/files/Daasetal_NTTS%202015%20abstract%20unblinded-v3z_903.pdf)
3. Daas P., Burger J., Le Q., ten Bosch O., Puts M.J.H.: Profiling of Twitter Users: a Big Data Selectivity Study. CBS Discussion paper 6 (2016) <https://www.cbs.nl/-/.../2016-profiling-of-twitter-users-a-big-data-selectivity-study-1.pdf>
4. Della Ratta F., Pontecorvo M.E., Vaccari C., Virgillito A.: Big Data and Textual Analysis: a Corpus Selection from Twitter. Rome Between the Fear of Terrorism and the Jubilee (2016) [https://www.researchgate.net/publication/303843023\\_Big\\_data\\_and\\_textual\\_analysis\\_a\\_corpus\\_selection\\_from\\_Twitter\\_Rome\\_between\\_the\\_fear\\_of\\_terrorism\\_and\\_the\\_Jubilee](https://www.researchgate.net/publication/303843023_Big_data_and_textual_analysis_a_corpus_selection_from_Twitter_Rome_between_the_fear_of_terrorism_and_the_Jubilee)



# Quality issues when using Big Data in Official Statistics

## *Aspetti di qualità statistica quando si usano i Big Data nella Statistica Ufficiale*

Paolo Righi, Giulio Barcaroli, Natalia Golini

**Abstract** The use of Big Data (BD) for improving the statistics and reducing the costs is a great opportunity and challenge for the National Statistical Offices (NSOs). Often the debate on BD is focused on the IT issues to deal with their volume, velocity, variety. Nevertheless, the NSOs have to be assured that the estimates have a good level of accuracy as well. This paper evaluates when estimators using Internet web scraped variables from a list of enterprise websites, suffering from selectivity concerns, are competitive with respect to a survey sampling estimators. A Monte Carlo simulation using a synthetic population based on real data is implemented to compare predictive estimators based on BD, survey estimators and blended estimators combining predictive and survey estimators.

**Key words:** Big Data, sampling estimation, selectivity, Big Data quality framework

## 1. Introduction

The opportunities of producing enhanced statistics and the declining budgets, make using Big Data (BD) in National Statistical Offices (NSOs) appealing. Often the debate on these sources is focused on volume, velocity, variety and on IT capability to capture, store, process and analyze BD for statistical production. Nevertheless,

---

<sup>1</sup> Paolo Righi, Istat; [parighi@istat.it](mailto:parighi@istat.it)  
Giulio Barcaroli, Istat; [barcarol@istat.it](mailto:barcarol@istat.it)  
Natalia Golini, Istat

other features have to be taken into account, especially in the NSOs, such as veracity (data quality as selectivity and trustworthiness of the information) and validity (data correct and accurate for the intended use). Veracity and validity affect the accuracy (bias and variance) of the estimates and, therefore, question if high amount of data produces necessarily high quality statistics. This paper evaluates when the estimators using Internet as BD source and suffering from selectivity concerns, are competitive with a survey sampling estimator. Design based estimators [2,3] and supervised model based estimators [5] using scraped data are compared (Section 2). A simulation study based on real 2016 Istat “Survey on ICT usage and e-Commerce in Enterprises” data (ICT survey) has been carried out. A synthetic enterprise population with websites has been built up (Section 3.1). Target and scraped from the website variables have been generated according to the distributions observed in ICT survey. Section 3.2 describes the set-up of the simulation. The performances of the estimators are shown in terms of bias, variance and mean square error (Section 3.3). Section 4 is devoted to short conclusions.

## 2. Notation and sampling strategy

Let  $U$  be the reference population of  $N$  elements and let  $U_d$  ( $d = 1, \dots, D$ ) be an estimation domain, where the  $U_d$ 's partition  $U$ .  $U_d$  is a sub-population of  $U$  with  $N_d$  elements, for which separate estimates are calculated. Let  $y_k$  denote the value of the interest variable attached to the  $k$ -th population unit ( $k=1, \dots, N$ ). The parameters to be estimated are  $Y_d = \sum_{k \in U_d} y_k$  and  $Y = \sum_{k \in U} y_k$ .

For defining the estimation procedure let us introduce a further partition of  $U$ . Let  $U^v$  ( $v=1, \dots, V$ ) be a sub-population of size  $N^v$  that distinguish itself for the set of auxiliary information, for instance a sub-population in which auxiliary variables from BD source are available. Let  $\mathbf{x}_k^v$  be the auxiliary variable vector from BD source and  $\mathbf{z}_k^v$  be the auxiliary variable vector known from the frame list for unit  $k$ . For simplicity  $\mathbf{z}_k^v = \mathbf{z}_k \forall v$ ,  $V=2$  and if  $k \in U^1$  the vector  $(\mathbf{x}_k^1, \mathbf{z}_k^1)$  is known, while for  $v=2$  only  $\mathbf{z}_k^2$  is known. Then the totals  $\mathbf{Z}_d = \sum_{k \in U_d} \mathbf{z}_k$  are known. The  $U^v$ 's cross cut the  $U_d$ 's, then  $U_d^v = U_d \cap U^v$ . We assume known the totals  $\mathbf{Z}_d^v = \sum_{k \in U_d^v} \mathbf{z}_k$ .

In the sampling strategy,  $y_k$  is observed with a random sample  $s$  of size  $n$ . The sample could be affected by non-response. Let  $r$  be the number of respondents in  $s$  and let  $r_d$  and  $r^v$  be respectively the number of respondents belong to  $U_d$  and  $U^v$ . In the observed sample, we can estimate a model  $\tilde{y}_k = f(\mathbf{x}_k^v, \mathbf{z}_k^v)$  for predicting the  $y$  variable. Table 1 introduces the estimators  $\hat{Y}_d$  of  $Y$  that are compared in the simulation. The derivations of the  $\hat{Y}_d$  of  $Y_d$ , are straightforward.

The list of estimators is not exhaustive but broadly maps possible estimators.

**Table 1:** General description of the estimators used in the simulation.

<i>Estimator</i>	<i>Expression</i>	<i>Description</i>	<i>Note</i>
Mod1	$\hat{Y} = \sum_{(U^1-r^1)} \tilde{y}_k b_k + \sum_r y_k b_k$	$b_k = N/(N^1 + r - r^1)$	Model based est.
Mod2	$\hat{Y} = \sum_{(U^1-r^1)} \tilde{y}_k w_k + \sum_r y_k w_k$	$w_k$ calibration [3] of $b_k$ 's defined in Mod1 being $\sum_{r,d} z_k w_k = \mathbf{Z}_d \forall d$	Pseudo-calibration model based est.
Des1	$\hat{Y} = (n/r) \sum_r y_k b_k$	$b_k$ is the sampling basic weight	Horvitz-Thompson est. corrected by no-response
Des2	$\hat{Y} = \sum_r y_k w_k$	$w_k$ calibration [3] of $b_k$ 's defined in Des1 being $\sum_{r,d} z_k w_k = \mathbf{Z}_d \forall d$	Calibration est.
Comb 1	$\hat{Y} = \sum_{(U^1-r^1)} \tilde{y}_k + \sum_{r^1} y_k + (n/r) \sum_{(r-r^1)} y_k b_k$	$b_k$ is the sampling basic weight	Combined est. Mod1 and Des1
Comb 2	$\hat{Y} = \sum_{(U^1-r^1)} \tilde{y}_k + \sum_{r^1} y_k + \sum_{(r-r^1)} y_k w_k$	$w_k$ calibration [3] of $b_k$ 's defined in Des1 being $\sum_{(r,d-r^1)} z_k w_k = \mathbf{Z}_d^v \forall d$	Combined est. Mod1 and Des2

### 3. Simulation study

Accuracy of statistical estimates is traditionally decomposed into bias (systematic error) and variance (random error) components. While variance can be estimated, bias is not observable if the parameter of interest is unknown.

We studied the accuracy of a set of estimators via Monte Carlo simulation. A synthetic population based on the 2016 ICT survey data has been created. The estimators have been taken into account can be distinguished with respect to:

- a. the origin of the exploited auxiliary information, coming from the frame list, from a BD source or both;
- b. the inferential approach (design based, model based and a combination of both).

#### 3.1 Target population

We consider the set of the Italian enterprises with 10 to 249 employed persons in activities of manufacturing, electricity, gas and steam, water supply, sewerage and waste management, construction and non-financial services (near 180,000 units). The population and a  $\mathbf{z}$  vector of auxiliary variables (location, unit size, and economic activity) are identified by the Italian Business Register (BR).

Currently, Istat uses this register as frame list for drawing the yearly ICT survey. The frame list (BR) is updated with information relating to two years before the survey time reference. Among the target estimates of the ICT survey there are a number of

characteristics related to the functionalities of the websites: for instance the presence of online ordering (e-commerce) or job application facilities. The simulation focuses on a single binary variable i.e. e-commerce, denoted as  $y$  variable, being  $y_k = 1$  if unit  $k$  does e-commerce and  $y_k = 0$  otherwise. The target parameters are the count of  $y_k = 1$  at domain of level (type of economic activity by size class of employed persons),  $Y_d$  ( $d = 1, \dots, 16$ ) and total level,  $Y$ . In particular, the type of economic activities are denoted as M1, M2 M3 and M4 and the size class of employees are denoted as c11 (small), c12 c13 and c14 (large). Since the survey estimates show that about 30% of BR units have not website we exclude these units from the analysis and remaining units define the target population  $U$ . The discarded units follow the distribution observed in the 2016 ICT survey in the 16 domains. We note that in practice the size of  $U$  should be treated as random. The  $y$  variable is unknown in  $U$ , so we create the probability  $p(y_k = 1)$  for each unit by means of logistic model,  $\text{logit}(y_k) = \alpha + \mathbf{z}_k' \boldsymbol{\beta}$  (hereinafter denoted as true model) where  $\alpha$  and  $\boldsymbol{\beta}' = (\beta_1, \dots, \beta_d, \dots, \beta_{16})$  are known regression coefficient and  $\mathbf{z}_k' = (z_k, \dots, z_{dk}, \dots, z_{16k})$ , being  $z_{dk} = 1$  if  $k \in U_d$  and  $z_{dk} = 0$  otherwise. We fix  $\alpha$  and  $\boldsymbol{\beta}$  such that, the sum over the  $U_d$ 's of  $p(y_k = 1)$  reflects observed distribution in the last 2016 Istat ICT survey (Table 2, column  $p$ ).

The population  $U$  is partitioned in 3 sub-populations,  $W^1, W^2$  and  $W^3$  :

- $W^1$ , the enterprises with website address (URL) available;
- $W^2$ , the enterprises with wrong URL or website not allowing automatic scraping;
- $W^3$ , the enterprises having website but the URL is not available;

We generated the distribution in the 3 sub-populations following the evidences:

- Istat has got a second list of business units where the website address (URL) is available. The inclusion in the URL-list is on volunteer basis and it does not cover all the business register (101,000 enterprises,  $W^1 \cup W^2$ );
- in a concrete application of automatic web scraping procedure 68,676 websites have been investigate ( $W^1$ ) and 32,320 have been not ( $W^2$ ).

We assume the URL-list suffers from selectivity problems, that is the distribution of target variable within the URL-list ( $W^1 \cup W^2$ ) differs from the distribution of the unit out this list,  $W^3$ . This reflects the hypothesis that if an enterprise uses actively its website for business (for instance doing e-commerce) then it has interest to increase its reachability, and therefore the probability to be in the Url-list. Table 2 shows the sizes and the expected  $p(y_k = 1)$  for the 3 sub-populations.

The simulation works with  $U^1 = W^1$  and  $U^2 = W^2 \cup W^3$ .

For completing the synthetic population we generate the output of the web scraping so that Internet is the BD source of the simulation.

The automatic scraping is not able to observe the variable  $y$ , but instead it collects all texts from websites and, in a second step, based on the use of text mining and natural processing techniques, relevant terms are detected to play the role of predictors (for instance: “add to cart”, “credit card”, “order”, etc.) [1]. We assume to observe, at the end of the process, 12 binary variables (presence/absence), denoted by the  $\mathbf{x}$  vector.



**Table 2:** Population size by domains and  $W^1$ ,  $W^2$  and  $W^3$  and the related probability of doing e-commerce

Domain	Population Size			$U$	Expected probability of e-commerce			
	$W^1$	$W^2$	$W^3$		$p^1$	$p^2$	$p^3$	$p$
M1 c11	23,519	10,995	11,435	45,949	0.170	0.170	0.048	0.140
M1 c12	3,146	1,499	1,595	6,240	0.154	0.154	0.023	0.120
M1 c13	1,873	887	853	3,613	0.218	0.218	0.014	0.170
M1 c14	922	440	370	1,732	0.333	0.333	0.000	0.261
M2 c11	1,122	565	578	2,265	0.138	0.138	0.037	0.110
M2 c12	237	97	82	416	0.124	0.124	0.027	0.110
M2 c13	146	71	84	301	0.151	0.151	0.009	0.110
M2 c14	120	53	44	217	0.222	0.222	0.000	0.181
M3 c11	5,408	2,486	2,992	10,886	0.050	0.050	0.013	0.040
M3 c12	382	176	206	764	0.026	0.026	0.004	0.020
M3 c13	168	78	81	327	0.039	0.039	0.002	0.030
M3 c14	65	27	27	119	0.025	0.025	0.000	0.020
M4 c11	26,525	12,574	11,289	50,388	0.319	0.319	0.103	0.270
M4 c12	2,430	1,144	890	4,464	0.379	0.379	0.081	0.320
M4 c13	1,527	712	507	2,746	0.396	0.396	0.036	0.330
M4 c14	1,086	516	371	1,973	0.396	0.396	0.000	0.321
<b>Total</b>	<b>68,676</b>	<b>32,320</b>	<b>31,404</b>	<b>132,400</b>	<b>0.235</b>	<b>0.235</b>	<b>0.061</b>	<b>0.194</b>

We underline that in practical application this number can be much larger. Nevertheless, a larger set of variables would only complicate the simulation without adding information. "Good" estimates are achieved when the target variable and the set of auxiliary variables (large or small) have a strong relationship: this result in high levels of performance indicators of models.

We generate the 12 auxiliary variables according to two scenarios:

- 1- weak dependence with the target variable (harmonic mean of precision and recall indicators equal to 63%);
- 2- strong dependence with the target variable ((harmonic mean of precision and recall indicators equal to 96%).

In particular, the first scenario seems closest to the evidences observed on the real 2016 ICT data. Scenario 2 remains a benchmark in evaluation analysis.

### 3.2 The simulation process

The simulation implements a feasible and reasonable estimation process. We consider a supervised approach, such that the target variable is observed in a sample, for instance in the ICT sample. We assume a stratified simple random sampling design with four strata defined by the size classes, c11, ..., c14. The sample of size  $n=23,229$ , is allocated with 16,307 units for c11, 1,820 units for c12, 1,061 units for c13 and 4,041 units for c14. Largest inclusion probabilities are assigned to the large enterprises in terms of employees reflecting the real sampling allocation. We generate unit non respondents, assuming homogeneous response probability in each stratum (c11 response probability= 0.45, c12 response probability= 0.88, c13 response probability=

0.95, c14 response probability= 0.97). The sample of respondents,  $r$ , has expected size of about 13,800 units (as in the 2016 ICT survey).

At domain level the sample size is not planned. We had three domain types: Large (L), Small (S) Very Small (VS) (see Table 3).

**Table 3:** Expected size and e-commerce frequency in the observed sample

Domain	Size	e-commerce	Type
M1 c1	3.074,09	430,45	L
M1 c2	845,37	101,37	L
M1 c3	520,21	88,42	L
M1 c4	1.681,42	438,14	L
M2 c1	151,53	16,63	S
M2 c2	56,36	6,19	VS
M2 c3	43,34	4,78	VS
M2 c4	210,66	38,07	L
M3 c1	728,30	29,16	S
M3 c2	103,50	2,06	VS
M3 c3	47,08	1,43	VS
M3 c4	115,53	2,34	VS
M4 c1	3.371,07	910,32	L
M4 c2	604,77	193,47	L
M4 c3	395,37	130,51	L
M4 c4	1.914,43	613,66	L
<b>Total</b>	<b>13.863,04</b>	<b>3.007,00</b>	Total

The estimation process follows these steps:

1. Collect the  $y$  variable for respondent units with website;
2. Make the web scraping for the units in  $U^1$  and collect the  $\mathbf{x}$  variables;
3. Model  $y$  on  $\mathbf{x}$  in  $r^1$ ;
4. Produce the estimate according to a given estimator.

For estimators Des1 and Des2 (Table 1), steps 2. and 3. are skipping.

The simulation compares 6 different estimators of Table 1. We note that:

- Mod1, Mod2, Comb1 and Comb2:  $\hat{y}_k = \hat{p}(y_k = 1)$  is predicted with a working logistic model using the  $\mathbf{x}$  variable;
- Des1: uses an incorrect MCAR [4] model for the non-response weight adjustment;
- Des2: calibration performs a correct weight adjustment for non-response;
- Comb1, Comb2: produce estimates for  $U^1$  (using Mod1) and  $U^2$  (using Des1 or Des2);
- Comb2: calibration performs a correct weight adjustment for non-response in  $U^2$ .

### 3.3 Results

The simulation takes into account the methodological frameworks of the respective estimators. For the model based estimators the  $y$  variable is treated as random, and then selected the sample, the  $y$  values change over the iteration. In the design based estimator the  $y$  values are fixed, and then in each iteration a new random sample is selected. The simulation implements 1,000 iterations and computes for each iteration

the estimates  $\hat{Y}_{j,d,i}$  for the  $j$ -th estimators, the  $d$ -th domain in the  $i$ -th iteration. The following statistics are considered for Mod1, Mod2, Des1 and Des2:

- the relative bias,  $RB(\hat{Y}_{j,d}) = \frac{\frac{1}{1,000} \sum_{i=1}^{1,000} (\hat{Y}_{j,d,i} - Y_d)}{Y_d}$ ;
- the coefficient of variation,  $CV(\hat{Y}_{j,d}) = \frac{\sqrt{\frac{1}{1,000} \sum_{i=1}^{1,000} (\hat{Y}_{j,d,i} - \bar{\hat{Y}}_{j,d})^2}}{Y_d}$ , being  $\bar{\hat{Y}}_{j,d} = 1/1,000[\sum_{i=1}^{1,000} \hat{Y}_{j,d,i}]$ ;
- the relative root mean square error,  $RRMSE(\hat{Y}_{j,d}) = \frac{\sqrt{[RB(\hat{Y}_{j,d}) Y_d]^2 + [CV(\hat{Y}_{j,d}) Y_d]^2}}{Y_d}$ .

For the estimators Comb1 and Comb2 the numerator of the  $RB$  becomes  $1/1,000[\sum_{i=1}^{1,000} \sum_v (\hat{Y}_{j,d,i}^v - Y_d^v)]$ , the numerator of the  $CV$  becomes  $\{1/1,000[\sum_{i=1}^{1,000} \sum_v (\hat{Y}_{j,d,i}^v - \bar{\hat{Y}}_{j,d}^v)^2]\}^{1/2}$  where  $\bar{\hat{Y}}_{j,d}^v = 1/1,000[\sum_{i=1}^{1,000} \sum_v \hat{Y}_{j,d,i}^v]$  in which  $\hat{Y}_{j,d,i}^v$  is the  $j$ -th estimator in the  $i$ -th iteration of  $Y_d^v = \sum_{k \in U_d^v} y_k$ . Table 4a shows the model based estimators produce biased estimates for all the domain types. These results convey that if we use a predictive model estimated on a sample representing a specific population ( $W^1$ ) such model does not fit for the other populations (such as  $W^3$ ). Calibration in Mod2 estimator, partially correct the bias. Discrepancies between Scenario 1 and 2 confirm the importance of using a good working model for improving the accuracy (bias). Table 4b shows the two design based estimators. Focusing on the calibration estimator (Des2), the correct weight adjustments produce nearly unbiased estimates but high  $CV$  and  $RRMSE$  especially for VS and S domains.

**Table 4a:** Maximum values of accuracy indicators observed in the simulation for model based estimators

Estimator	Statistic	Domain Type			
		VS	S	L	Total
Mod1 Scenario1	CV	112.90	10.54	25.42	1.82
	RBIAS	629.80	313.08	74.46	28.47
	RRMSE	632.17	313.26	77.97	28.53
Mod1 Scenario2	CV	111.24	8.63	25.54	0.65
	RBIAS	85.35	44.75	74.72	19.11
	RRMSE	135.34	45.36	77.43	19.12
Mod2 Scenario1	CV	65.47	10.51	14.75	1.83
	RBIAS	628.42	342.75	70.70	27.72
	RRMSE	630.26	342.91	70.82	27.78
Mod2 Scenario2	CV	64.65	8.79	14.86	0.67
	RBIAS	90.87	55.11	25.63	17.54
	RRMSE	99.44	55.66	26.03	15.56

Table 4c show the accuracy of blended estimates, combining the model and design based estimates. We note that Comb1 - Scenario 2 is highly competitive with respect to Des1 estimators. We underline that both estimators do not adjust correctly the weights of the  $r - r^1$  sampled units. Comparing Comb2-Scenario 2 with Des2 the first estimator seems better for S domain, competitive for VS, L and Total domains.

**Table 4b:** Maximum values of accuracy indicators observed in the simulation for design based estimators

<i>Estimator</i>	<i>Statistic</i>	<i>Domain Type</i>			
		<i>VS</i>	<i>S</i>	<i>L</i>	<i>Total</i>
Des1	CV	142.30	18.08	14.75	1.62
	RBIAS	61.61	-25.88	62.56	-9.36
	RRMSE	153.85	31.57	62.73	9.50
Des2	CV	89.33	23.97	9.25	1.92
	RBIAS	-1.59	-1.68	0.39	-0.02
	RRMSE	89.33	24.03	9.25	1.92

**Table 4c:** Maximum values of accuracy indicators observed in the simulation for combined estimators

<i>Estimator</i>	<i>Statistic</i>	<i>Domain Type</i>			
		<i>VS</i>	<i>S</i>	<i>L</i>	<i>Total</i>
Comb1 Scenario1	CV	83.27	9.95	12.79	1.48
	RBIAS	391.99	156.88	41.97	1.46
	RRMSE	399.29	157.19	42.78	2.08
Comb1 Scenario2	CV	81.99	10.16	12.961	1.13
	RBIAS	101.40	-18.48	32.20	-3.82
	RRMSE	130.36	21.09	34.70	3.99
Comb2 Scenario1	CV	81.94	12.74	12.59	1.58
	RBIAS	368.97	165.38	25.61	5.39
	RRMSE	373.27	165.80	26.26	5.62
Comb2 Scenario2	CV	80.64	12.91	12.71	1.26
	RBIAS	63.43	13.79	7.90	0.11
	RRMSE	102.59	17.72	14.97	1.26

## 4. Conclusion

Big Data represent a concrete opportunity for improving the official statistics. Nevertheless, their use has to carefully evaluate. In this paper, we show in a simulation that also the use of auxiliary variables coming from the Internet BD source highly correlated with the target variable (Scenario 2) does not guarantee enhancement of the quality of the estimates if selectivity issue affect the source. Analyse the BD variables and study the relationship between populations covered or not by the BD source is a fundamental step to know how to use and which framework implement to assure high quality output.

## References

1. Barcaroli G. et al.: Machine learning and statistical inference: the case of Istat survey on ICT. *Proceedings 48th Scientific Meeting Italian Statistical Society* (2016).
2. Cochran. W.G.: *Sampling Techniques*. Wiley. New York (1977).
3. Deville. J.-C., Särndal C.-E.: Calibration estimators in survey sampling. *Journal of the American Statistical Association*. 87. 376–382 (1992).
4. Little. R. J. A. and Rubin. D. B.: *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley (2002).
5. Valliant R., Dorfman A. H., Royall R. M.: *Finite Population Sampling and Inference: A Prediction Approach*. Wiley. New York (2000).

# Indicators for the representativeness of survey response as well as convenience samples

## *Indicatori di rappresentatività dei dati di indagine in presenza di non-risposta o di campioni non-probabilistici*

Emilia Rocco

**Abstract** Non-response bias has long been a concern for surveys, even more so over the past decades with the increasing decline of the response rates. A similar problem concerns the surveys based on non-representative samples, the convenience and cost-effectiveness of which has increased with the recent technological innovations that allow for collecting large numbers of highly non-representative samples via on-line surveys. These two cases may be considered jointly since in both it must be assumed that the bias is the result of a self-selection process and, for both, quality indicators are needed to measure the impact of this process. In this study we analyze, in different scenarios, the combined use of two indicators that have been suggested in the non-response context, but which may work as well for convenience samples .

**Abstract** *La distorsione per non risposta é da sempre una delle principali fonti di errore non campionario e ancora di piú negli ultimi anni per la crescente riduzione dei tassi di risposta. Un'analogo problema di distorsione riguarda i campioni non-probabilistici la cui convenienza é aumentata con le recenti innovazioni tecnologiche che consentono, tramite sondaggi on-line, di raccogliere facilmente dati su campioni non rappresentativi. In entrambi i casi si può assumere che la distorsione sia il risultato di un processo di autoselezione e sono necessari degli indicatori di qualità per valutare l'impatto di tale processo sulle stime. Qui analizziamo, sotto diversi scenari, l'uso congiunto di due diversi indicatori che sono stati proposti per la non-risposta ma possono essere utilizzati piú in generale per problemi di autoselezione.*

**Key words:** response probability, self selection bias, survey participation, weighting-adjustment methods

---

Emilia Rocco

Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti", Università degli Studi di Firenze, Viale Morgagni, 59 - 50134 Firenze, e-mail: emilia.rocco@unifi.it

## 1 Indicators of non-response bias

As response rates have declined over the past decades, the statistical benefits of probabilistic sampling have diminished. Assuming that a representative sample is initially selected, low response rates mean that those who ultimately supply the target data might not be representative. Moreover, with recent technological innovations, it is increasingly convenient and cost-effective to collect large numbers of highly non-representative samples via online surveys.

The main problem caused by non-representative survey data is that estimators of population characteristics must be assumed to be biased unless convincing evidence to the contrary is provided. This problem influences the data coming from a probability sample affected by non-response and the data obtained with a convenience sample in the same way. Hence, in both the cases, the same quality indicators may be used in order to evaluate the impact of non-representativeness and the same post-survey adjustment methods may be used to deal with it.

In recent literature, various indicators have been proposed as indirect measures of non-response bias in surveys. Wagner (2012) provides a taxonomy of such measures based on the types of data used to estimate each one. More in detail he describes three types of alternative indicators: (1) indicators involving the response indicator; (2) indicators involving the response indicator and auxiliary data that are known for all sample units and may stem from sampling frame data, administrative data and data about the data collection process; (3) indicators involving the response indicator, auxiliary data and survey data (i.e. the data for respondents). It is well-known finding in survey methodology that the only indicator of the first type, the response rate, by itself is a poor indicator of non-response bias. Indicators of the second type use auxiliary data for predicting the response indicator and provide a single measure of the risk of non-response bias for the whole survey, relying on the implicit assumption that the auxiliary variables used to create them are correlated with all the survey estimates. The fact of providing a single measure for the whole survey is a strength of such indicators since allows them to be used as tools for comparing different surveys and surveys over time, and for a comparison of different data collection strategies and modes. However, it is also a weakness, because, a single measure of the risk of non-response bias for the whole survey could lead to incorrect conclusions for the survey statistics for which the implicit assumption of correlation with the auxiliary data used to create such risk measure is not likely to be true. Indicators of the third type, which, in addition to the response indicator and the auxiliary variables, use the observed survey data are defined at a statistic level. Since non-response bias occurs at the level of the statistics, if the models assumption on which the indicator relies is good, it allows for directly adding information about the bias. However the definition of such indicators at statistic level is also a weakness of them. Given that most surveys have multiple objectives, there would be more indicators that makes the computation process more complex than for the other two types of indicators and could lead to potentially different conclusion about data-collection strategy.

In this study, in order to measure the risk of non-response bias, we examine the ef-

fectiveness in different scenarios of a prominent indicator of the second type, the "R-Indicator" suggested by Schouten et al. (2009), and suggest the combined use of this indicator with another one of the third type which relies on the variation of respondent means across the percentiles of the response probabilities predicted for estimating the R-indicator.

## 2 Theoretical framework and notation

Let  $U$  be a population of  $N$  units ( $i = 1, \dots, N$ ),  $s$  a probability sample drawn by employing the sampling design  $p(s)$  and  $r$  the set of responding units. Denote with

- $\pi_i$  the first order inclusion probability for unit  $i$ ;
- $\delta_i$  the response indicator so that  $\delta_i = 1$  if unit  $i$  responds and  $\delta_i = 0$  otherwise.

We shall suppose that the target of inference is a population mean of a survey variable taking value  $y_i$  for unit  $i$  and that the data available for estimation purposes consist of the values  $\{y_i; i \in r\}$  of the survey variable and the values  $\{\mathbf{x}_i = (x_{1,i}, \dots, x_{K,i}); i \in s\}$  of a vector of auxiliary variables that may influence the non-response mechanism and/or the survey variable. Moreover, we assume that the response mechanism is MAR and that, given the sample, the response indicators are independent random variables with:

$$pr(\delta_i = 1 | i \in s, y, \mathbf{x}) = \rho(\mathbf{x}_i) \equiv \rho_i \tag{1}$$

The basic idea of the R-indicator is that a response subset is representative with respect to  $\mathbf{x}$  when response propensities are constant for  $\mathbf{x}$ . Relying on this idea, it measures the extent to which the response probabilities  $\rho(\mathbf{x}_i)$  vary as follows:

$$R_\rho = 1 - 2S_\rho \tag{2}$$

where  $S_\rho$  is the standard deviation of the individual response propensities. Therefore, it will be higher when the variability among the response probabilities is lower. In practice, the response propensities are unknown. However, when auxiliary data are available at a sample level, it is possible to estimate them for all sampled units and to replace  $R_\rho$  with the estimator:

$$\hat{R}_\rho = 1 - 2\sqrt{\frac{1}{N-1} \sum_{i \in s} \frac{(\hat{\rho}_i - \hat{\rho})^2}{\pi_i}} \quad \text{where} \quad \hat{\rho} = \frac{1}{N} \sum_{i \in s} \frac{\hat{\rho}_i}{\pi_i} \tag{3}$$

The response propensities,  $\hat{\rho}_i$ , are commonly estimated with explicit or implicit models linking the response occurrences to the auxiliary variables, for instance, by using a logistic or a probit regression model, or the weighting within cell method. It is evident from ( 3) that  $\hat{R}_\rho$ , as already stated for all indicators of the second type, provides a single measure on the risk of bias for the whole survey and does not give any direct information about the real bias of a single survey statistic. Therefore, in

a multi-purpose survey  $\hat{R}_\rho$  could be a better indicator for some survey statistics and a less effective one for others. In fact, in a survey with several survey variables it would be unlikely to identify a set of auxiliary variables correlated together with the response probability and with any survey variable. In such a situation,  $\hat{R}_\rho$  could, in any case, be a useful quality measure of the survey data collection process as a whole. Moreover, when some auxiliary variables, relevant for describing the survey population, are available, it is reasonable to ask whether the subset of respondents is representative, at least, with respect to these, and  $\hat{R}_\rho$  can provide the answer. Finally, if the model used to estimate the response propensities is correct and  $\hat{R}_\rho$  has been used for adapting the data-collection process in order to achieve a highly representative response set, i.e. a value of  $\hat{R}_\rho$  close to 1, it is likely that the risk of non-response bias is negligible even for those statistics that are not correlated with the auxiliary variables used to estimate the response propensities. When a low value of  $\hat{R}_\rho$  is obtained, more investigations are needed. In fact, as empirically shown in Section 3, if a statistic is correlated with auxiliary variables used to estimate  $\hat{R}_\rho$ , then a low value of  $\hat{R}_\rho$  corresponds to a high bias of the statistic. Conversely, for a statistic that is not correlated with auxiliary variables used to estimate  $\hat{R}_\rho$ , the bias may be negligible even in correspondence with a low value of  $\hat{R}_\rho$ . Hence for moderate or low values of  $\hat{R}_\rho$  we recommend to create, in addition to it, an indicator estimated at the statistic level for each statistic. A simple indicator of this type is the variation of means across the percentiles of estimated response propensities (Olson, 2006). You could simply plot these means, or build a synthetic index as the ratio (denoted as  $\hat{R}_y$ , below) of their deviance and the total deviance of the respondents values.

### 3 Simulation Study

Our aim is to empirically explore the conditions in which  $\hat{R}_\rho$  is able to predict the bias of the unweighted mean estimator of a survey variable and how the variation of means across the percentiles of estimated response propensities may be useful for identifying its effectiveness. To this end we perform a simulation study by reproducing the simulation setting used by Little and Vartivarian (2005), to provide, in the set of weighting for an estimate of a survey mean based on adjustment cells, empirical proof of the fact that the non-response weighting adjustments are effective in reducing bias if the auxiliary information used for their estimation is related to both the non-response mechanism and the outcome of interest.

Simulation setting:

- $x$  is a categorical variable with 10 categories that identify 10 cells of adjustment;
- conditional on the sample size, the sampled cases have a multinomial distribution over the  $(10 \times 2)$  contingency table based on the classification of the response indicator,  $\delta$ , and  $x$ , with cell probabilities

$$\begin{aligned} pr(\delta = 1, x = c) &= pr(\delta = 1)pr(x = c | \delta = 1) \\ pr(\delta = 0, x = c) &= (1 - pr(\delta = 1))pr(x = c | \delta = 0) \quad c = 1, \dots, 10 \end{aligned} \quad (4)$$



given in Table 1 for two marginal response rates, 70% and 52%, and three conditional distributions of  $\delta$  given  $x$  corresponding to high, medium and low association between the two variables.

- The simulated distribution of  $y$  given  $\delta = h$ , ( $h = 0, 1$ ), and  $x = c$  have the form:

$$[y|\delta = h, x = c] \sim N(\beta_0 + \beta_1x, \beta_0), \tag{5}$$

and three sets of values of  $(\beta_1, \sigma^2)$  corresponding to high, medium and low association between  $y$  and  $x$  are considered and shown in Table 2. The intercept  $\beta_0$  is chosen so that the overall mean of  $y$  is 26.3625 for each scenario.

- 10,000 replicate samples of size 400 were simulated for each combination of parameters in Tables 1 and 2.
- for each replica the following estimates have been produced: (1) the unweighed mean of the respondents; (2) the response probability, for each unit in the sample, using the weighting within cell method means with the cells corresponding to the 10 categories of  $x$ ; (3)  $\hat{R}_p$  and (4)  $\hat{R}_y$ , considering the variation of unweighted means of the respondents across 5 percentiles of the estimated response probabilities.

**Table 1** Percent of samples in cell  $x \times \delta$

Response Rate = 52%											
association $x$ and $\delta$	$x$	1	2	3	4	5	6	7	8	9	10
High	$\delta = 1$	0.55	1.00	4.01	4.52	5.04	5.55	6.06	6.58	9.14	9.96
	$\delta = 0$	8.69	9.00	6.01	5.53	5.04	4.54	4.04	3.54	1.02	0.20
Medium	$\delta = 1$	2.77	3.50	4.01	4.52	5.04	5.55	6.06	6.58	7.11	7.62
	$\delta = 0$	6.47	6.50	6.01	5.53	5.04	4.54	4.04	3.54	3.05	2.54
Low	$\delta = 1$	4.62	5.15	5.21	5.28	5.34	5.40	5.45	5.52	5.58	5.64
	$\delta = 0$	4.62	4.85	4.81	4.77	4.73	4.69	4.65	4.60	4.57	4.52

Response Rate = 70%											
association $x$ and $\delta$	$x$	1	2	3	4	5	6	7	8	9	10
High	$\delta = 1$	0.55	3.00	6.51	7.04	7.55	8.07	8.59	9.11	9.64	9.96
	$\delta = 0$	8.69	7.00	3.51	3.02	2.52	2.02	1.52	1.01	0.51	0.20
Medium	$\delta = 1$	4.44	5.30	5.81	6.33	6.85	7.37	7.88	8.40	8.93	9.45
	$\delta = 0$	4.80	4.70	4.21	3.72	3.22	2.72	2.22	1.72	1.22	0.71
Low	$\delta = 1$	6.19	6.85	6.91	6.98	7.05	7.11	7.17	7.24	7.31	7.37
	$\delta = 0$	3.05	3.15	3.11	3.07	3.02	2.98	2.93	2.88	2.84	2.79

The empirical relative bias of the unweighed mean, the median across the replications of  $\hat{R}_p$  and the median across the replications of  $\hat{R}_y$  are reported in Table 3 from which we note that: (1) When the association between  $\delta$  and  $x$  is low, the  $\hat{R}_p$  value is high and the bias of the unweighed mean, even though it decreases with the decreasing of association between  $y$  and  $x$ , is always very low. (2) On the con-

**Table 2** Parameters  $\beta_1$  and  $\sigma^2$  for outcome model ( 5)

association between $x$ and $y$	$\beta_1$	$\sigma^2$
High	4.75	46
Medium	3.70	122
Low	0.00	234

trary, a low value of  $\hat{R}_\rho$  does not necessarily mean a high bias of the unweighted mean since when the association between  $y$  and  $x$  is low, the bias of the unweighted mean is negligible irrespective of the value of  $\hat{R}_\rho$ . (3) A value of  $\hat{R}_y$  close to zero allows for identifying the situations in which, given a low association between  $y$  and  $x$ , the bias of the unweighted mean is negligible. (4) If the model used to estimate the response propensities is correct, the two indicators, considered jointly, allow for discriminating between the statistics for which the risk of non-response bias is higher ( $\hat{R}_\rho$  is closer to 0 or  $\hat{R}_y$  is closer to 1) from those for which it is lower ( $\hat{R}_\rho$  is closer to 1 and  $\hat{R}_y$  is closer to 0).

**Table 3** Summaries of results based on 10,000 replicate samples for each of 18 scenarios

Response Rate = 52%					Response Rate = 70%					
association $x$ and $\delta$	association $x$ and $y$	emp. bias	$\hat{R}_\rho$	$\hat{R}_y$	association $x$ and $R$	association $x$ and $y$	emp. bias	$\hat{R}_\rho$	$\hat{R}_y$	
High	High	27.24%	0.43	0.63	High	High	19.10%	0.44	0.64	
	Medium	21.23%	0.43	0.35		Medium	Medium	19.91%	0.44	0.35
	low	0.02%	0.43	0.02		low	low	0.08%	0.44	0.01
Medium	High	14.76%	0.68	0.64	Medium	High	11.32%	0.70	0.67	
	Medium	11.48%	0.68	0.38		Medium	Medium	8.86%	0.70	0.40
	Low	0.05%	0.68	0.02		Low	Low	0.04%	0.70	0.01
Low	High	2.16%	0.85	0.31	Low	High	2.16%	0.86	0.31	
	Medium	1.68%	0.85	0.19		Medium	Medium	1.68%	0.86	0.19
	Low	0.00%	0.85	0.01		Low	Low	0.00%	0.86	0.01

## References

1. Little, R.J.A., Vartivarian, S.: Does weighting for nonresponse increase the variance of survey mean. *Surv. Meth.* **31**, 161-168 (2005)
2. Olson, K.: Survey participation, nonresponse bias, measurement Error bias and total bias. *Public Opinion Quarterly* **70**, 737-758 (2006)
3. Schouten, B., Cobben, F., Betleehem, J.: Indicators for the representativeness of survey response. *Surv. Meth.* **35**, 101-113 (2009)
4. Wagner, J.: A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly* **76**, 555-575 (2012)

# A sampling design for the evaluation of earthquakes vulnerability of the residential buildings in Florence

## *Un disegno campionario per la stesura di una mappa della vulnerabilità sismica dell'edificato residenziale della città di Firenze*

Emilia Rocco, Bruno Bertaccini, Giulia Biagi and Andrea Giommi

**Abstract** The assessment of earthquakes vulnerability of buildings is a key step in the analysis of seismic risk of a territory. The aim of this study is the identification of an appropriate sampling design for the analysis of the earthquake vulnerability of the residential buildings in the city of Florence. In order to identify such a design we have considered that the buildings are statistical units selected from a territory and therefore could be spatially correlated. Since in these cases it is advantageous to select units well spread over the territory, we propose a spatial balanced sampling design. In addition to the information on the geographical location of each building, the suggested sampling design takes into account other auxiliary information on characteristics of the buildings that may affect their vulnerability.

**Abstract** *La valutazione della vulnerabilità sismica delle costruzioni è un passo fondamentale nelle analisi del rischio sismico di un territorio e nella definizione di scenari di danno per terremoti di diverse intensità. Lo scopo di questo studio è l'individuazione di un appropriato disegno di campionamento per l'analisi della vulnerabilità sismica degli edifici residenziali della città di Firenze. Per individuare un tale disegno è importante tener conto che gli edifici come tutte le unità statistiche selezionate da un territorio sono generalmente correlati spazialmente e questa loro caratteristica rende vantaggiosa la selezione di unità ben diffuse sul territorio. Pertanto, suggeriamo di utilizzare un disegno di campionamento spazialmente bilanciato. Oltre alle informazioni sulla posizione geografica di ciascun edificio, il disegno campionario proposto sfrutta anche altre informazioni ausiliarie sulle caratteristiche degli edifici che possono influenzare la loro vulnerabilità sismica.*

**Key words:** auxiliary information, balanced sampling, spatial correlation, units well spread over the territory

---

Emilia Rocco, Bruno Bertaccini, Giulia Biagi and Andrea Giommi  
Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti", Università degli Studi di Firenze, Viale Morgagni, 59 - 50134 Firenze, e-mail: emilia.rocco@unifi.it - bruno.bertaccini@unifi.it - biagi@disia.unifi.it - andrea.giommi@unifi.it

## 1 Introduction

The seismic hazard of a territory is defined as the frequency and strength of earthquakes that could affect such territory. The consequences of an earthquake, however, also depend on the capacity of resistance to the actions of a seismic shock of the buildings in the territory. The predisposition of a building to be damaged is called vulnerability. More is the vulnerability of buildings to earthquakes, greater will be the consequences. Therefore, the assessment of buildings vulnerability to earthquakes is a key step in the analysis of seismic risk and in the definition of hazard scenarios for earthquakes of different level of intensity. In this framework, several researchers of the University of Florence, belonging to various disciplinary areas including Statistics, Earth Sciences, Architecture and Engineering, have defined a research project, denominated SISMED, whose goal is to draft a map of the seismic vulnerability of the residential buildings located in the city of Florence. The realization of this map should be based on the georeferencing of a seismic vulnerability index calculated for each residential building of the municipality. But this census evaluation is unfeasible since the collection for each building of all the data needed for calculating its index of seismic vulnerability is complex and reliable estimates predict that a complete analysis could take about 30,000 man-days. Therefore it is necessary to select a sample of residential buildings and the aim of this study, which represents a preliminary phase of SISMED, is the definition of an appropriate sampling strategy.

In the following Section we describe the characteristics of the target population and of the frame available for the sample selection, whereas in Section 3 we present our sampling design proposals.

## 2 The spatial population under study

During last years, the Department of Earth Sciences of University of Florence analyzed in details the geological and geophysical settings of the Florence underground. These studies revealed which parts of the city area are interested by different levels of amplification of the seismic energy due to site effects. The seismic amplification affects the infrastructures. Consequently, the earthquake vulnerability of the buildings depends on the substrate on which the building lies, the ground seismic response and the building dynamic response. This last is complex to evaluate and requires complex site-inspection by technical specialized staff, but it is also highly correlated with some characteristics of the buildings, that may be deduced from statistical or administrative fonts, such as the year of construction, the type of construction (masonry/concrete) and their height.

The number of buildings in the municipality of Florence, observed by the ISTAT 2011 Census, was 47,509 and about 65% of them has a “residential” destination. Residential buildings are mostly in the suburbs of the city, whereas buildings having a “commercial” or “service” use are predominant in the historical center. About

58% of the residential buildings was already present at the end of the 19th century and therefore passed the test of the last big earthquake in Florence (1985). However, during the years, the renovations of them may have modified their predisposition to be damaged by a new earthquake. On the other hand, almost 40% of the buildings, built from 1895 to 1981, has never suffered the “testing” of an earthquake and most of them were built without any anti-seismic regulations.

For economic and time constrains the study will be initially limited to a sub-area of the city. Our target population is therefore limited to the approximately 4,300 residential buildings lying within the zone just outside of the Poggi<sup>1</sup> boulevards perimeter and inside to the railway and Arno river boundaries, extended towards the highway junction of Florence South (see Fig. 1).



**Fig. 1** Geographical area under study

### 3 Sampling spatial populations

Statistical units selected from a territory are generally spatially correlated, which means that nearby units are more similar than units further apart. This is likely to happen also for the residential buildings lying in the city of Florence: nearby buildings not only share the substrate on which they lie but they often have been built in the same period and/or have other common characteristics; thus they could be more similar also in their level of seismic vulnerability. It is a well-known finding in the literature on sampling from spatial populations that in a situation of this type it

<sup>1</sup> The main Florence boulevards are identified by the name of the architect Giuseppe Poggi who designed them in 1865.

is advantageous to select units well spread over the territory (e.g. Stevens and Olsen, 2004; Grafström, 2012; Grafström et al., 2012). A well-spread sample is usually said to be spatially balanced. Different types of spatially balanced sampling designs have been suggested in literature for sampling populations with spatial trends in the variables of interest, for example different types of systematic designs. For many of these designs however, it is problematic to select the units with unequal probability but there are situations in which the use of equal selection probabilities does not appear reasonable. For this reason, we consider a spatial design recently introduced by Grafström et al. (2012), the Spatially Balanced Sampling through the Pivotal Method (SBStPM), that allows to select a spatially balanced sample with equal or unequal inclusion probabilities and can be used for any number of dimensions. For a detailed description of it we refer to (Grafström et al. 2012). The different inclusion probabilities may depend on either (i) the type of sampling procedure (for example stratification) or (ii) the probabilities may be imposed by the researcher to obtain better estimates by including more important units with higher probability. Both cases require the availability of auxiliary information. From the 2011 census dataset, for each residential buildings lying in the city of Florence we know the following auxiliary variables :

- the year of construction (grouped into the following classes: until 1918, [1919 – 1945], [1946 – 1960], [1961 – 1970], [1971 – 1980], [1981 – 1990], [1991 – 2000], [2001 – 2005]);
- the number of storeys above ground
- the type of construction (masonry / concrete).

All these variables may affect the buildings' vulnerability, therefore it is opportune to select a sample representative with respect to them. Given the categorical nature of these variables, they cannot be used directly in order to assign to each unit an unequal inclusion probabilities proportional to them. However we can stratify the buildings, according to these three variables and successively define inclusion probabilities equal within each stratum and different between strata following a non-proportional rule of allocation. Since the strata are used only to define the inclusion probabilities, their number can be relatively high with respect to the total sample size which, due to the modest available resources, cannot overcome 150 units. In order to define the strata we grouped the categories of the variable year of construction in 4 classes and those of the variable number of storeys above ground in 3 classes. Therefore we obtained  $4 \times 3 \times 2 = 24$  strata. The resulting sizes of the strata are very different, the most of the buildings is concentrated in few strata and there are many strata with few units: three strata with less than ten buildings. For this reason we adopted the optimal allocation criterium proposed by Kish (1988), that allows to assign higher inclusion probabilities to strata with a small number of buildings.

Our proposal is to select our sample by means of a "variant" of the SBStPM design that ensure the selection of at least one unit in the strata where the sum of the inclusion probabilities is at least one. This is achieved updating, at each step, the inclusion probabilities through the local pivotal method of Grafström et al. (2012) for two nearby units belonging to the same stratum. When, in the stratum remains

only a units with a probability mass greater than 0 and lesser than 1, the strata are collapsed in a hierarchical fashion until a nearby units is found. Other conditions for collapsing the strata can be evaluated in order to select samples which meet different objectives/needs.

## References

1. Grafström, A.: Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference* **142**, 139-147 (2012)
2. Grafström, A., Lunnström, N.L.P., Schelin, L.: Spatially balanced sampling through the pivotal method. *Biometrics* **68**, 514-520 (2012)
3. Kish, L.: Multipurpose sample designs. *Survey Methodology* **14**, 19-32 (1988)
4. Stevens, D.L., Jr. and Olsen, A.R.: Spatially balanced sampling of natural resource. *Journal of the American Statistical Association* **99**, 262-278 (2004)





# A local regression technique for spatially dependent functional data: an heteroskedastic GWR model

## *Un modello di regressione geografico eteroschedastico per dati funzionali spazialmente dipendenti*

Elvira Romano and Jorge Mateu

**Abstract** In this paper we propose a localized regression technique to account for spatial non-stationarity in functional data relationships by generalising a geographical weighted regression model. We present an heteroskedastic version of the geographically weighted regression model for functional data which allows the residual variance to vary across the space. In particular we propose to calibrate the variance of the model by replacing it by a continuous mean smoothing over the space. In addition, in order to deal with the calibration problem and to define and measure the so-called closeness in the spatial functional dimension, this paper proposes an alternative back-fitting approach. Several simulation studies and an application on real data show the performances of the proposed method.

**Abstract** *In questo lavoro viene proposto un modello di Regressione Geografica Pesata (Geographically Weighted Regression (GWR)) per dati funzionali spazialmente dipendenti. Il modello proposto rappresenta un'estensione del modello di regressione pesata al caso in cui si assuma la presenza di dipendenza spaziale nella variabilità degli errori. L'idea di base quella di definire una stima smoothing della varianza spazio-funzionale degli errori collocandosi in un contesto puramente funzionale. Inoltre, dal momento che la procedura di stima locale del modello prevede la scelta di una metrica, viene introdotto e generalizzato un algoritmo iterativo di back-fitting per ottimizzare tale scelta. Le caratteristiche e le performance della metodologia proposta sono state illustrate mediante l'applicazione della stessa a numerosi data set simulati ed a dati reali.*

**Key words:** spatially dependent functional data, Geographically Weighted Regression, heteroskedastic

---

Elvira Romano

Department of Mathematics and Physics, Università della Campania Luigi Vanvitelli, Caserta, Italy,  
e-mail: elvira.romano@unicampania.it

Department of Mathematics, Campus Riu Sec, University Jaume I, Castellon, Spain e-mail:  
mateu@mat.uji.es

## 1 Introduction

In this paper we focus on the problem of non-stationarity in the parameter estimation and propose a generalisation of a geographically weighted regression model (GWR) for spatially correlated sample curves [4] obtained as realisations of a spatio-temporal stochastic process.

GWR [1] can be defined as an approach that makes the spatial sub-samples of data by means of a kernel function.

The basis of this model concerns that it looks for local variation in space by moving a weighted window over the data, estimating one set of coefficient values at every chosen fit point. Thus it follows local representations by modeling the process showing directional variation in the spatial distance decay.

In the functional framework [7], the first and unique attempt to generalise the approach of [1] was done by [6].

They considered the functional regression model [2] and defined a geographical weight in a similar way of GWR. In particular by adapting the basic regression model of [1], they incorporate the estimation of the weight matrix into the procedure of the estimation of the functional coefficients. A kernel function was used to define geographical weight in terms of spatial correlation, and a Montecarlo simulation was used to establish the spatial parameter for controlling the spatial variability.

As in classical GWR it is assumed that the variance of the error term is fixed, and spatial weighting function, defined by using the classical Euclidean distance, is applied equally at each calibration point. However, in many real cases the assumption of constant error variance and the use of the Euclidean distance only for determining the weights not be realistic and reasonable.

To face these potential problems we address its stationary residual variance by generalising an heteroskedastic version (H-GWR) [3] to the functional framework which allows the residual variance to vary across space. We evaluate the choice of an appropriate distance metric by generalising a back-fitting approach in functional framework to calibrate a GWR model with parameter-specific distance metrics.

## 2 An heteroskedastic GWR model for spatially dependent functional data

Let us assume that we have a functional response variable  $Y_s = \{Y_s(t), t \in T\}$  observed at a location  $s \in D \in \mathbb{R}$ , whose realisation as a function of  $t \in T$  is a functional data, where  $T$  is a compact subset of  $\mathbb{R}$  [2].

Let  $\{\boldsymbol{\chi}_s(t), t \in T, s \in D \in \mathbb{R}\}$  be a multivariate functional random field.

Given  $s_i \in D, i = 1, \dots, n$ , and  $K$  functional covariates (with  $k = 1, \dots, K$ ) we have the realization

$$\boldsymbol{\chi}_s(t) = [(\boldsymbol{\chi}_{s_1}, \dots, \boldsymbol{\chi}_{s_{n_1}}), \dots, (\boldsymbol{\chi}_{s_1K}, \dots, \boldsymbol{\chi}_{s_{n_K}})]^T = [\boldsymbol{\chi}_{s_1}(t), \dots, \boldsymbol{\chi}_{s_K}(t)]^T$$

The aim of GWR is to predict the functional response variable starting from the set of functional covariates by allowing local variations in rates of change [1]. The model is defined by

$$Y_{s_i}(t) = \beta_{0s_i} + \sum_{k=1}^K \int_T \chi_{s_{ik}}(t) \beta_{s_{ik}}(v, t, s_i) dv + \varepsilon_{s_i}(t) \quad i = 1, \dots, n \quad (1)$$

where the function  $\beta_{0s_i}(t)$  is the mean function at location  $s_i$ ,  $\beta_{s_{ik}}(v, t, s_i)$  is the regression function for the  $k$ -th covariates at location  $s_i$ , and  $\varepsilon_{s_i}(t)$  is a random error function at point  $s_i$

The calibration process defined as a trade-off between bias and standard error and obtained by minimising the sum of integrated square residuals defined by

$$LMISE = \sum_{i=1}^n \int_T [Y_{s_i}(t) - \beta_{0s_i} - \sum_{k=1}^K \int_T \chi_{s_{ik}}(t) \beta_{s_{ik}}(v, t, s_i) dv]^2 dt \quad (2)$$

Suppose that the functional data can be approximated by a set of basis functions:  $\phi_k(v) = (\phi_{k1}(v), \dots, \phi_{kG}(v))^T$  and  $\varphi_k(t) = (\varphi_{k1}(t), \dots, \varphi_{kG}(t))^T$  and assume that they are centered. These can be expanded as  $\chi_s(v) = \mathbf{C}_k^T(v) \phi_k(v)$ ,  $y_s(t) = \mathbf{D}^T(v) \varphi(t)$ ,  $\beta_{s_{ik}}(v, t, s_i) = \phi_k(v)^T \mathbf{B}_{s_{ik}} \varphi(t)$

where  $\mathbf{C}_k$ ,  $\mathbf{D}$ ,  $\mathbf{B}_{s_{ik}}$  are matrices of dimension  $n \times H_\phi$ ,  $n \times H_\varphi$ ,  $H_\phi \times H_\varphi$ , with a number of basis functions respectively equal to  $H_\phi \cdot H_\varphi$ .

Then the 2 becomes

$$LMISE = trace\left\{ \left( D - \sum_{k=1}^K C_k J_{\phi_k} B_{s_{ik}} \right) J_\varphi \left( D - \sum_{k=1}^K C_k J_{\phi_k} B_{s_{ik}} \right)^T \right\} \quad (3)$$

where  $J_{\phi_k} = \int \phi_k(v) \phi_k(v)^T dv$ ,  $J_\varphi = \int \varphi(t) \varphi(t)^T dt$  can be solved by choosing  $B_{s_{ik}}$  which minimizes the expression

$$(C_k J_{\phi_k}) W_{s_i} \left( \sum_{k=1}^K C_k J_{\phi_k} B_{s_{ik}} \right) J_{\phi_k} = (C_k J_{\phi_k}) W_{s_i} D J_\varphi \quad (4)$$

Where  $\mathbf{W}_{s_i(n \times n)}$  is a diagonal weight matrix with a generic element defined as:

$$w_{s_i} = w_{s_i, s_k} = \exp\left(\frac{-d_{s_i, s_k}}{h}\right) \quad (5)$$

where  $d_{s_i, s_k}$  is the Euclidean distance between location  $s_i$  and location  $s_k$ , and  $h$  is a non-negative parameter known as bandwidth, selected by a cross-validation criterion. The method suffers from the problem that the accuracy of prediction of the model for functional data does not improve, although the goodness of fit improves by adding the weights in the functional regression model. In addition, the Euclidean

distance only for determining the weights may not be realistic because the attribute effects of the focal point and its neighbours are totally ignored.

To face these problems, with the aim of providing a spatial prediction technique to deal with the spatial non-stationarity of the functional coefficients, we propose a local model that can be seen as local model of variance since the local non-stationarity is consequence of spatial variance heterogeneity.

Borrowing the idea from [3], we introduce a model calibration by means of local estimation of the squared residuals.

Especially we suppose that the variance of the residual model depends on the spatial location.

The GWR prediction variance at a generic location  $s_i$ , without any assumption of spatial dependence, is defined as

$$\sigma_{GWR_{s_i}}^2(t) = var\{\hat{Y}_{s_i}(t) - Y_{s_i}(t)\} = \hat{\sigma}^2(t)[1 + S_{s_i}(t)] \tag{6}$$

where:

- $\hat{\sigma}^2(t) = RSS(t)/(n - ENP)$ , where  $RSS(t)$  is the residual sum of squares and  $ENP$  is the effective number of parameters of the GWR fit. This is a function independent from the spatial location.
- $S_i(t)$  are the element of the matrix  $\mathbf{S} = (C_k J_{\phi_k}) W_{s_i} (\sum_{k=1}^K C_k J_{\phi_k} B_k) J_{\phi_k}$

We propose to calibrate the variance of the model  $\sigma_{GWR_{s_i}}^2(t)$  by replacing  $\sigma(t)$  with  $\sigma_{s_i}(t)$ .

If we assume that  $\sigma_{s_i}(t)$  is a continuous function over the space, we estimate it by a mean smoother. The final variance  $\hat{\sigma}_{s_i}^2(t)$  replaces  $\hat{\sigma}^2(t)$  to give

$$\sigma_{GWR_{s_i}}^2(t) = var\{\hat{Y}_{s_i}(t) - Y_{s_i}(t)\} = \hat{\sigma}_{s_i}(t)^2 [1 + S_{s_i}(t)] \tag{7}$$

For the local variance estimation, we need to model the relationship with the local means. Thus we define by a local smoother the local mean

$$m_{s_i}(t) = \sum_{i=1}^n w_{s_i} y_{s_i}(t) / \sum_{i=1}^n w_{s_i} = \sum_{i=1}^n \sum_{l=0}^L w_{s_i} a_l(t) f_l(s_i) / \sum_{i=1}^n w_{s_i} \quad \mathbf{s} \in D, t \in T \tag{8}$$

where  $f_l(\cdot)$ ,  $l = 1, \dots, L$  are known functions of the variable  $\mathbf{s}$  and  $a_l(\cdot)$ ,  $l = 1, \dots, L$  are functional coefficients independent from the spatial location.

Thus the dependence of the mean from the space is related to the function  $\{f_l(\cdot)\}_{l=1, \dots, L}$  and to the weights  $w_{s_i}$ .

These functions are obtained by the same kernel function specified with GWR in order to allow the rate of the spatial variation according to the same criteria.

Consequently the local variance smoother becomes

$$L_{\sigma_{s_i}^2}(t) = \sum_{i=1}^n w_{s_i} (y_{s_i}(t) - m_{s_i}(t))^2 / \sum_{i=1}^n w_{s_i} \tag{9}$$

It is a mean smoothing over the observed square residuals able to provide the following local variance estimation

$$\hat{\sigma}_{s_i}^2(t) = \sum_{i=1}^n w_{s_i} (y_{s_i}(t) - \sum_{i=1}^n \sum_{l=0}^L w_{s_i} a_l(t) f_l(s_i)) / \sum_{i=1}^n w_{s_i}^2 / \sum_{i=1}^n w_{s_i}$$

The logic is the same to the use of weighted least squares (WLS) in multiple linear regression (MLR) to stabilise a non-constant residual variance. The algorithm is applied with updated estimates of  $\hat{\beta}_{s_{ik}}(v, t, s_i)$  and until an acceptable level of convergence is reached. As the parameter estimates, the H-GWR prediction at  $s_i$  is also updated. Assuming that each independent/dependent functional variable pair in the H-GWR model may correspond to different optimal distance metrics we propose to calibrate H-GWR with parameter-specific distance metrics by a generalization of a back-fitting procedure [5].

The procedure consists in evaluating different specific distance metrics for estimating their corresponding parameters and in choosing the one that has the best fitting value by an iterative procedure. Practically it is performed in three main steps: initialize the response variable for several distance metrics; compute the distance among the response variable and the model calculated using a specific distance; compute the residuals and chose the best model until the residual sum of squares converges. It enables to understand the relationship among the variables over space as well as over time with major locally-accurate measures of prediction uncertainty. Based on minimal assumptions, and as demonstrated in many simulations and real data study, the proposed HGWR shows significant improvement over the GWR in terms of AIC measures and parameter estimation [8].

## References

1. Brunson, C., Fotheringham, S., Charlton, M.: Geographically weighted regression: modelling spatial non-stationarity. *Journal of the Royal Statistical Society. Series D (The Statistician)* 47(3): 431-443, (1998).
2. Ferraty, F. and P. Vieu. *Non Parametric Functional Data Analysis. Theory and Practice*. New York: Springer, (2006.)
3. Fortigam, A., S. Brunson, C., Charlton, M.E.: *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, New York, (2002).
4. Delicado, P., Giraldo, R., Comas, C. and Mateu, J.: Statistics for spatial functional data: some recent contributions. *Environmetric*, 21: pp.224-239, (2010)
5. Lu, B., Harris, P., Charlton, M., Brunson, C., *Calibrating a Geographically Weighted Regression Model with Parameter-specific Distance Metrics*. *Procedia Environmental Sciences*, 26:109-114, (2015).
6. Yamanishi, Y., Tanaka, Y.: Geographically weighted functional multiple regression analysis: A numerical investigation. *Journal of Japanese Society of Computational Statistics* 15, 307-317, (2003).
7. Ramsay, J.E., Silverman, B.W.: *Functional Data Analysis*, (Second ed.) Springer (2005)
8. Romano E., Mateu J., Butzbach, O.: *An heteroskedastic geographical weighted regression model for functional data*. (2017) (submitted)



# Models for jumps in trading volume

## *Modelli per i salti nel trading volume*

Eduardo Rossi and Paolo Santucci de Magistris

**Abstract** In finance theory the log-price is often supposed to follow an Ito semimartingale while no explicit assumptions are made on the dynamic evolution of trading volumes. Trading volume is a measure of the quantity of shares that change owners for a given security. The amount of daily volume on a security can fluctuate on any given day depending on the amount of new information available about the company. We assume that the dynamic evolution of trading volume is represented as a semimartingale. Analogously to stock prices, the stochastic process for trading volume might be characterized by jump components. We distinguish between two classes of widely used processes: Brownian semimartingales plus jumps and pure-jump models. The relative contribution of each of two components is estimated by means of alternative nonparametric methods. We also analyze if the jump component is a stochastic process of finite or infinite variation. Finally, alternative parametric models are estimated and compared.

**Abstract** *Nella teoria della finanza si assume che il processo stocastico del log-prezzo segua una semimartingala di Ito mentre non sono esplicitate le ipotesi sulla dinamica dei volumi scambiati (trading volume). Il trading volume di un titolo azionario è il numero di azioni scambiate. L'ammontare di volume giornaliero relativo al singolo titolo pu fluttuare ogni giorno in funzione delle nuove informazioni disponibili. Si assume che l'evoluzione dinamica del trading volume possa essere rappresentata da una semimartingala. Analogamente a quanto si suppone per i log-prezzi, il processo stocastico per il trading volume caratterizzato dalla presenza di una componente di salto. Nel lavoro si distingue tra due classi di processi: semimartingale browniane con salti e modelli di salto. Il contributo relativo di ognuna*

---

Eduardo Rossi

European Commission, Joint Research Centre, Directorate Innovation and Growth, B.01, Italy.  
Dipartimento di Scienze Economiche ed Aziendali, University of Pavia, 27100 Pavia, Italy. e-mail: eduardo.rossi@unipv.it

Paolo Santucci de Magistris

Department of Economics and Business Economics and CREATES, Aarhus University, Denmark,  
email: e-mail: psantucci@econ.au.dk

*delle componenti stimato con tecniche non parametriche. Si indaga anche se la componente di salto un processo stocastico di variazione finita o infinita. Infine, sono stimati e comparati modelli parametrici alternativi.*

**Key words:** Trading Volume, Jumps, Activity level, Infinite variation.

## 1 Introduction

For each market equilibrium we have an equilibrium price and quantity. In finance theory the price is often supposed to follow an Itô semimartingale while no explicit assumptions are made on the dynamic evolution of trading volumes. Trading volume is a measure of the quantity of shares that change owners for a given security. The amount of daily volume on a security can fluctuate on any given day depending on the amount of new information available about the company, whether options contracts are set to expire soon, whether the trading day is a full or half day, and many other possible factors. Of the many different elements affecting trading volume, the one which correlates the most to the fundamental valuation of the security is the new information provided. This information can be a press release or a regular earnings announcement provided by the company, or it can be a third party communication, such as a court ruling or a release by a regulatory agency pertaining to the company. The news release can generate large variations in the trading volume. The trading volume can be measured instantaneously for each trade or cumulated for a given time interval. This implies that for longer time intervals the trading volume is an increasing process. This is not the case for the price process.

As in the case of prices, we assume that the dynamic evolution of trading volume is represented as an Itô semimartingale ( $SM$ ) defined on a filtered probability space  $(\Omega; \mathcal{F}; (\mathcal{F})_{t \in [0, T]}; \mathcal{P})$  satisfying usual conditions, evolving as

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s ds + \sum_{s \leq t} \Delta X_s$$

where

$$\Delta X_s = X_s - X_{s-}$$

is the size of the jump at time  $s$ . Even when the whole path of  $X$  is observed over  $[0, T]$  one can infer neither the drift nor the Lévy measure. With a finite  $T$  we can only infer the behavior of the Lévy measure near 0. For a semimartingale the activity index takes values in the interval  $[0, 2]$ . For a Lévy process the jump activity index coincides with the Blumenthal-Gettoor index of the process [1, 2]. The index takes its values in  $[0; 2]$  and allows to distinguish different classes of stochastic processes. The Blumenthal-Gettoor index is zero for finite activity jump processes (which have finite number of jumps in any finite interval) and it is equal to two for continuous (local) martingales. Stochastic processes with Blumenthal-Gettoor indices in  $(0; 2)$



are infinitely active pure-jump processes, with paths of infinite variation if and only if the index is larger than unity. When the process is the result of the sum of a jump component and a continuous process driven by Brownian motion, its activity index will take a value of 2 independently from the activity of the jumps. In general, the jump activity of a superposition of different Ito semimartingales is equal to the Blumenthal-Gettoor index of the most active component. If  $X$  is a stable process  $\beta$  is also the stable index of the process.  $\beta$  captures the level of the activity: when  $\beta$  increases the (small) jumps tend to become more and more frequent.

The main research question of this paper is: which process best approximates the trading volume dynamics? In other words, we want to distinguish between two classes of widely used processes in modeling the dynamics of financial prices: Brownian semimartingales plus jumps (with Blumenthal-Gettoor index equal to 2) and pure-jump models (with Blumenthal-Gettoor index less than 2). The study of the trading volume (TV) dynamics allows to better understand the role played by small and large jumps in equilibrium and on the microstructure of financial markets.

## 2 Which jumps in trading volume?

We assume that the observations are collected at a discrete sampling interval  $\Delta_n$ , which means that there are  $[T/\Delta_n]$  observed increments of  $X$  on  $[0, T]$ , i.e.

$$\Delta_n^i X = X_{i\Delta_n} - X_{(i-1)\Delta_n}.$$

Let  $\mu$  the jump measure of  $X$  and  $\nu$  its predictable compensator, Lévy measure. Both positive measure on  $\mathbb{R}_+ \times \mathbb{R}$ .

$$\text{Small jumps} = \int_0^t \int_{|x| \leq \varepsilon} x(\mu - \nu)(ds, dx)$$

$$\text{Big jumps} = \int_0^t \int_{|x| > \varepsilon} x\mu(ds, dx)$$

where the cutoff level  $\varepsilon > 0$  is arbitrary, but fixed. A  $SM$  will always generate a finite number of big jumps on  $[0, T]$  but it may give rise to either a finite or infinite number of small jumps, i.e.

$$\nu([0, t] \times (-\infty, -\varepsilon) \cup (\varepsilon, +\infty)) < \infty$$

whereas

$$\nu([0, t] \times [-\varepsilon, \varepsilon])$$

may be finite or infinite.

Using the methodology of power variation:

$$V(p) = \int_0^T |\sigma_s|^p ds$$

$$J(p) = \sum_{s \leq T} |\delta X_s|^p \quad p > 0.$$

1.  $V(p)$  is finite  $\forall p > 0$ , and  $V(p) > 0$  on  $\Omega_T^W$ .
2.  $J(p)$  is finite if  $p \geq 2$  but often not when  $p < 2$ .

The realized power variations proposed by Ait-Sahalia & Jacod [2],

$$B(p, u_n, \Delta_n) = \sum_{i=1}^{[T/\Delta_n]} |\Delta_i^n X|^p \mathbf{1}_{\{|\Delta_i^n X| \leq u_n\}}$$

where  $u_n$  is a sequence of truncation levels. With  $T$  fixed, the asymptotics are all with respect to  $\Delta_n \rightarrow 0$ . Since  $u_n$  has to converge to 0,  $u_n = \alpha \Delta_n^\varpi$ ,  $\varpi \in (0, 1/2)$ , and  $\alpha > 0$ . With  $\varpi < 1/2$  we keep all the increments that mainly contain a Brownian contribution. The in-fill asymptotics:

$$p > 2, \forall X \implies B(p, \infty, \Delta_n) J(p)$$

$$\forall p, \text{ on } \Omega_T^c \implies \frac{\Delta_n^{1-p/2}}{m_p} B(p, \infty, \Delta_n) V(p)$$

$m_p$  is the  $p$ th absolute moment of  $z \sim N(0, 1)$ . When  $p > 2$

$$B(p, \infty, \Delta_n) \xrightarrow{P} J(p)$$

the jump component dominates. If there are jumps the limit  $J(p)_t > 0$  is finite. If there are no jumps,  $X$  is continuous, then

$$J(p) = 0 \quad B(p, \infty, \Delta_n) \xrightarrow{P} 0$$

at rate  $\Delta_n^{p/2-1}$ . We can exploit the different asymptotic behavior of  $B(p, u_n, \Delta_n)$  by varying the tuning parameters:

1. the power  $p$ : to isolate either the continuous or jump components or to keep both.
  - $p < 2$  emphasizes the continuous component
  - $p > 2$  accentuates the jump component
  - $p = 2$  equal treatment
2. the truncation level  $u_n$ . The assumption is that there exists a finite number of large jumps with fixed size. As  $\Delta_n \rightarrow 0$ ,  $u_n$  becomes smaller than the large jumps which are thus no longer part of  $B(p, u_n, \Delta_n)$ . Alternatively, we can truncate to eliminate the Brownian component using the upward power variation

$$U(p, u_n, \Delta_n) = \sum_{i=1}^{[T/\Delta_n]} |\Delta_i^n X|^p \mathbf{1}_{\{|\Delta_i^n X| > u_n\}}$$

- the sampling frequency  $\Delta_n$ . Sampling at different frequencies we can distinguish three cases based on the asymptotic behavior of the ratio

$$\frac{B(p, u_n, k\Delta_n)}{B(p, u_n, \Delta_n)} \quad k \geq 2$$

As  $\Delta_n \rightarrow 0$ , the limiting behavior can be

- $\frac{B(p, u_n, k\Delta_n)}{B(p, u_n, \Delta_n)} = 1$ ,  $B(p, u_n, k\Delta_n)$  converges to a finite limit
- $\frac{B(p, u_n, k\Delta_n)}{B(p, u_n, \Delta_n)} < 1$ ,  $B(p, u_n, k\Delta_n)$  diverges to infinity
- $\frac{B(p, u_n, k\Delta_n)}{B(p, u_n, \Delta_n)} > 1$ ,  $B(p, u_n, k\Delta_n)$  converges to 0

The model includes three components: a continuous part, a small jumps part and a big jumps part. Accordingly we can describe the possible behavior by means of sets defined pathwise on  $[0, T]$

- $\Omega_T^c = \{X \text{ is continuous in } [0, T]\}$
- $\Omega_T^j = \{X \text{ has jumps in } [0, T]\}$
- $\Omega_T^f = \{X \text{ has finitely many jumps in } [0, T]\}$
- $\Omega_T^i = \{X \text{ has infinitely many jumps in } [0, T]\}$
- $\Omega_W^i = \{X \text{ has a Wiener component in } [0, T]\}$
- $\Omega_{noW}^i = \{X \text{ has no Wiener component in } [0, T]\}$

We should also note that we observe a time series originating in a given unobserved path in  $\Omega_T$  and wish to determine in which sets the path is likely to be. Any such time series can be obtained by discretization of a continuous path and also of a discontinuous one.

The jump activity index at time  $t$  is the random number (see [1])

$$\beta_t^i = \inf \left\{ r > 0 : \int_{\mathbb{R}} (|x|^r \wedge 1) F_s(dx) < \infty \right\}$$

Following [3]  $u_n = \alpha \Delta_n^\omega$  and  $u'_n = \alpha' \Delta_n^\omega$

$$\hat{\beta} = \frac{\log(U(0, u_n, \Delta_n)/U(0, \gamma u_n, \Delta_n))}{\log(\gamma)}$$

$$\gamma = \alpha' / \alpha$$

By using the statistic  $U$ , which simply counts the number of large increments, defined as those greater than  $\alpha \Delta_n^\omega$ , we are retaining only those increments of  $X$  that are not predominantly made of contributions from its continuous semimartingale part, which are  $O_p(\Delta_n^{1/2})$ , and instead are predominantly made of contributions due to a jump. When  $X$  has only finitely many jumps, the index is  $\beta = 0$  and  $U(p, u_n, \Delta_n)$  converges to the number of jumps between 0 and  $t$ , irrespective of the value of  $\alpha$ , so  $\hat{\beta} = 0$  for all  $n = \lceil T/\Delta_n \rceil$  large enough.

The paper presents and discuss the results of the techniques shown above to high-frequency data of SPY and individual stocks traded on the NYSE.

## References

1. Aït-Sahalia, Y., Jacod, J.: Estimating the degree of activity of jumps in high frequency data. *Annals of Statistics*, **37**(5A), 2202–2244 (2009)
2. Aït-Sahalia, Y., Jacod, J.: Analyzing the Spectrum of Asset returns: Jump and Volatility components in High Frequency Data. *Journal of Economic Literature*, **50**(4), 1007–1050 (2012)
3. Aït-Sahalia, Y., Jacod, J.: *High-frequency financial econometrics*. Princeton University Press, Princeton and Oxford (2014)

# On a failure process driven by a self-correcting model in seismic hazard assessment

## *Un processo di rotture guidato da un modello self-correcting nella valutazione della pericolosità sismica*

Rotondi Renata and Varini Elisa

**Abstract** Two widely noted features of earthquake generation process are the following: a) earthquakes tend to occur in clusters, and b) fault ruptures that generate earthquakes decrease the amount of strain present along the fault and hence the probability that another shock occurs in the near future. These diametrically opposed features have been widely studied separately in the literature by two classes of models: self-exciting and self-correcting models. To reconcile these contrasting trends we propose a new stochastic model which distinguishes strong events - *leaders* - from those of lower magnitude. The former follow a stress release model; conditioned on their occurrence, the remaining events constitute a set of ordered times of minor ruptures occurring in the time interval between two consecutive leader-events.

**Abstract** *Due ben note caratteristiche del processo di generazione di terremoti sono le seguenti: a) i terremoti tendono a verificarsi in clusters, e b) le rotture di faglia che generano terremoti riducono lo sforzo presente sulla faglia e quindi la probabilità che si abbia un altro evento nell'immediato futuro. Queste caratteristiche diametralmente opposte sono state ampiamente studiate separatamente in letteratura attraverso due classi di modelli: self-exciting e self-correcting. Per conciliare queste tendenze contrastanti proponiamo un nuovo modello stocastico che distingue eventi forti - leaders - da quelli di magnitudo inferiore. I primi seguono un modello di rilascio di sforzo; condizionato al loro accadimento, gli eventi rimanenti costituiscono un insieme di tempi di rotture ordinati che si verificano nell'intervallo di tempo tra due eventi leader consecutivi.*

**Key words:** point processes, bathtub shaped hazard function, generalized Weibull distributions, Bayesian inference

---

Rotondi Renata  
CNR-IMATI, Via Bassini 15, 20133 Milano (I) e-mail: reni@mi.imati.cnr.it

Varini Elisa  
CNR-IMATI, Via Bassini 15, 20133 Milano (I) e-mail: elisa@mi.imati.cnr.it

## 1 Stochastic point processes in seismology

Earthquakes are natural disasters by far the most powerful on the Earth. Each year thousands of earthquakes are recorded; among these about 60 are classified as able to cause fatalities and remarkable damages and about 20 are of major intensity with magnitude larger than 7. The potential cost of earthquakes is growing because of increasing urban development in seismically active areas and the vulnerability of older buildings, which may not have been built or upgraded to current building codes. Generally only point-information - origin time, epicentre, magnitude - is available for each earthquake; this idealization enables us to study the earthquake process through point process models. Explorative analysis of historical catalogues collecting the seismic activity recorded in countries like China, Greece, Italy, over centuries shows evidence of clustering in space, time, and size domains. Clustering in time is primarily, but not only, associated with the increase of seismic activity immediately after large earthquakes leading to aftershock sequences. This phenomenological aspect implies that the occurrence probability should increase immediately after an earthquake, to then decrease; this is a specific property of the class of self-exciting models, like the widely applied Epidemic-Type Aftershock Sequence (ETAS) model. As for physical models on earthquake generation processes, the elastic rebound theory by Reid was the first theory to satisfactorily explain earthquakes: the far field plate motions cause the rocks in the region of the locked fault to accrue gradually elastic deformation. When the accumulated strain is great enough to overcome the strength of the rocks, an earthquake occurs. According to this theory the occurrence probability should lower immediately after a strong earthquake to then increase, since the next event will happen only when enough stress will be built up along the fault. Vere-Jones and others introduced in a series of papers [11, 1] a stochastic translation of Reid's theory - the so-called stress release model - which belongs to the class of self-correcting models.

Each of aforesaid models gathers a feature of the phenomenon but it is not able to explain how it evolves in its entirety. Some attempts have been done to marry the two contrasting trends in a unique model: the simplest solution was proposed by Schoenberg and Bolt [7] who combine the conditional intensities  $\lambda_{tr}(t)$  and  $\lambda_{sr}(t)$  which characterize trigger and strain-release point process models respectively; in this way, since it is unknown which model each event follows, one assumes that every event is generated by both the models and the ratio of the cumulative intensities over the time interval  $(0, T)$ ,  $\Lambda_i(T)/[\Lambda_{tr}(T) + \Lambda_{sr}(T)]$ , represents the percentage of events due to the triggering effect, if  $i = tr$ , or to the strain-release component, if  $i = sr$ . The large difference between the scales at which the triggering and strain-release mechanisms appear to operate may be a misleading element. Another approach consists in assuming that the different trends correspond to different phases of the seismic activity and the dynamics of their activation times is driven by an unobserved pure jump Markov process; in this perspective a seismic sequence can be considered as a realization of a series of three marked point processes: Poisson, stress release and trigger models [9]. The comparison on simulated datasets shows

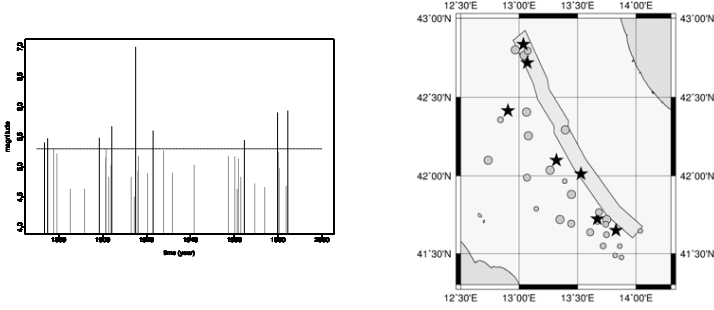
that about 70% of the events are correctly classified but the model is hardly able to fit the abrupt changes of state.

This leads to think that it is more reasonable to assume that the different behavioural trends (models) are superimposed rather than consecutive. In this perspective we suppose that the first level (model) concerns the most amount of released energy and guides the remaining seismic activity distributed within the intervals between each pair of consecutive I type quakes. The timing of the secondary (or II type) events could suggest their classification as aftershocks, isolated events (background) and foreshocks, that matches well with a bathtub-shaped hazard function. In Section 2 we propose a criterion for partitioning the data into two categories; other criteria can be adopted according to the available information.

## 2 Superimposed point processes: failure process and self-correcting model

In this study we consider two databases: the Database of Individual Seismogenic Sources (DISS, version 3.0.2; [5]) and the Parametric Catalog of Italian Earthquakes CPTI04 [4] which reflect the same level of knowledge at the end of 2002. DISS is a large repository of geological, tectonic and active fault data for Italy and the surrounding areas; in particular it contains 74 composite seismogenic sources (CSS) located in Italy. A CSS is essentially an active structure where an entire fault system is identified on the basis of geological data. One of the most active sources is the CSS-025 located in central Apennines region (Fig. 1 *right*); among the earthquakes of moment magnitude  $M_w \geq 4.5$  recorded in CPTI04, 50 may be associated with this fault system. To guarantee a satisfactory level of completeness of the data set we just consider the events occurred since 1870 and, in particular the set of 35 earthquakes covering the time interval from 1873 to 1985. Then we partition this data set into the set  $(t_i, M_w^{(i)})_{i=1}^n$  of  $n = 9$  *leaders* events of magnitude exceeding the threshold  $M_{tr} = 5.3$ , and into the sets  $(s_{ij}, M_w^{(ij)})_{j=1}^{J_i}$ ,  $\forall i = 1, \dots, n-1$ , of 26 *subordinates* events with  $t_i < s_{ij} < t_{i+1}$ , being  $t_i$  and  $s_{ij}$  the occurrence times and  $M_w$  the respective magnitudes (Fig. 1 *left*).

We assume that the *leader* events follow the stress release (SR) model. This model assumes that the stress  $x$  increases linearly with time at a constant loading rate  $\rho$  imposed by tectonic movements; hence the stress level at time  $t$  is given by:  $x(t) = x_0 + \rho t - s(t)$ , where  $x_0$  is the level of stress at the beginning of the analysed period and  $s(t)$  is the accumulated stress released by the earthquakes in the area at times  $t_i$ , which is  $s(t) = \sum_{i: t_i < t} x_i$ . By the word ‘stress’ we indicate any quantity that governs the state of the system; in this case we choose as proxy measure of the earthquake size the scaled energy  $E/M_0$ , ratio of the energy  $E$  and the seismic moment  $M_0$ , that turns out to be the best measure to use in SR models [10]. Therefore we have:



**Fig. 1** (Left) Time vs magnitude plot of the data: in black the *leaders* events and in grey the *subordinates* ones - (Right) Map of the composite seismogenic source CSS-025 with the epicentres of the *leaders* (black stars) and *subordinates* (grey circles) earthquakes.

$$x = \frac{E}{M_0} \propto \frac{M_0^{1/5}/\sqrt{A}}{M_0}$$

where  $A$  is the rupture area of the earthquake and the seismic moment is linked to the moment magnitude through the relation  $\log_{10} M_0 = 1.5 M_w + 9.1$ . As every point process, the SR model is characterized by its conditional intensity function:

$$\lambda_s(t | \mathcal{H}_t) = \exp \left\{ \alpha + \beta \left[ \rho t - \sum_{i: t_i < t} x_i \right] \right\}, \quad (1)$$

a monotonically increasing function of the stress level with parameters  $\alpha$ ,  $\beta$ ,  $\rho$ , where  $\mathcal{H}_t$  is the previous history of the process consisting in the set  $(t_i, M_w^{(i)})$ ,  $t_i < t$ ,  $i = 1, \dots, n$ .

For the magnitude we adopt the exponential distribution, that in seismology is inspired by the Gutenberg-Richter law,  $\log_{10} N = a - bM$ , which expresses the relationship between a magnitude value  $M$  and the number  $N$  of events of at least that magnitude. In our case we assign the exponential distribution  $g(M_w) = b \exp\{-b(M_w - m_0)\}$  on the interval  $[m_0, +\infty)$  with  $m_0 = 4.5$ , so that the density function of the magnitude of the *leaders* events is:

$$g_l(M_w | M_w \geq M_{lr}) = b e^{-b(M_w - M_{lr})} = \frac{g(M_w)}{[1 - G(M_{lr})]}, \quad M_w \in [M_{lr}, +\infty). \quad (2)$$

Given the interval  $(t_i, t_{i+1})$ ,  $i = 1, \dots, n-1$ , let us consider the number  $J_i$  of *subordinates* events  $(s_{ij}, M_w^{(ij)})$ ,  $j = 1, \dots, J_i$ , such that  $t_i < s_{ij} < t_{i+1}$  and  $M_w^{(ij)} < M_{lr}$ . If we indicate the probability of exceeding the magnitude threshold by  $p = 1 - G(M_{lr})$ , then  $J_i$  can be meant as the number of failures (events of  $M_w < M_{lr}$ ) before the next success (event of  $M_w \geq M_{lr}$ ); hence  $J_i$  follows a geometric distribution with parameter  $p = \exp(-bM_{lr})$ :



$$Pr\{J_i = j_i\} = p(1-p)^{j_i}, \quad j_i = 0, 1, \dots \tag{3}$$

The occurrence times  $s_{ij}$  of the *subordinates* events constitute a sample of  $J_i$  minor rupture times  $\tau_{ij} = s_{ij} - t_i$  with  $\tau_{ij} \in (0, t_{i+1} - t_i)$ ; being  $\tau_{ij} < \tau_{i(j+1)}$ , we assume that  $\tau_{ij}, j = 1, \dots, J_i$ , are the order statistics of a random sample drawn from the density function  $f(\cdot)$ . The length of the time interval between two consecutive strong earthquakes is managed by the stress release model, but we think that the process of secondary ruptures is the same once that their times are unit-based normalized in  $(0, 1)$ , that is  $\tau_{ij} = \frac{s_{ij} - t_i}{t_{i+1} - t_i}$  are identically distributed  $\forall i = 1, \dots, n - 1$ , and  $j = 1, \dots, J_i$ .

To be able to fit any trend, the probability distribution of  $\tau_{ij}$  should have so flexible hazard function as to exhibit not only monotonic shapes, but also unimodal, bathtub and modified bathtub (or N-shape) shapes. The Weibull distribution is one of the most cited lifetime distributions in reliability engineering and other disciplines; it describes failure times observed in many phenomena, with increasing, constant, or decreasing hazard rate. The class of generalized Weibull distributions [6] includes many Weibull related distributions with differently shaped hazard function; among them we have considered in particular two: the additive and flexible Weibull distributions. The additive Weibull distribution:

$$F(\tau) = 1 - e^{-(\tau/\beta_1)^{\alpha_1}} e^{-(\tau/\beta_2)^{\alpha_2}}, \quad \alpha_1, \alpha_2, \beta_1, \beta_2 > 0, \tag{4}$$

is a twofold competing risks model that can therefore represents jointly the decreasing seismic activity after the main shock and the increase of activity before the next strong earthquake. Involving two Weibull distributions its hazard function:

$$h(\tau) = (\alpha_1/\beta_1)(\tau/\beta_1)^{\alpha_1-1} + (\alpha_2/\beta_2)(\tau/\beta_2)^{\alpha_2-1}. \tag{5}$$

can be increasing, if both shape parameters are greater than 1 ( $\alpha_1 > 1$  and  $\alpha_2 > 1$ ), decreasing if  $\alpha_1 < 1$  and  $\alpha_2 < 1$ , or  $h(\tau)$  has a bathtub shape if  $\alpha_1 < 1$  and  $\alpha_2 > 1$ .

The flexible Weibull distribution has the following survival function:

$$\bar{F}(\tau) = \exp(-e^{\gamma\tau - \delta/\tau}) \tag{6}$$

and hazard function:

$$h(\tau) = (\gamma + \delta/\tau^2) \exp(\gamma\tau - \delta/\tau). \tag{7}$$

Unlike other generalized Weibull distribution, this distribution has rather simple hazard rate and Bebbington *et al.* [2] showed that its shape is a modified bathtub (i.e.,  $h$  is first increasing followed by a bathtub shape) if  $\gamma\delta < 27/64$ .

Finally the probability distribution of the magnitude of the *subordinate* events is defined on  $[m_0, M_{tr})$ ; hence we have:

$$g_s(M_w | m_0 \leq M_w < M_{tr}) = \frac{b e^{-b(M_w - m_0)}}{1 - e^{-b(M_{tr} - m_0)}}. \tag{8}$$

If we indicate by  $\theta = (\alpha, \beta, \rho, b, \alpha_1, \beta_1, \alpha_2, \beta_2)$  (or  $\theta = (\alpha, \beta, \rho, b, \gamma, \delta)$ ) the parameter vector, the likelihood is given by:

$$\begin{aligned} \mathcal{L}(\theta, data) = & \prod_{i=1}^n \lambda_s(t_i) \exp \left\{ - \int_{t_1}^{t_n} \lambda_s(u) du \right\} \times \prod_{i=1}^n \frac{g(M_w^{(i)})}{1 - G(M_{th})} \times \\ & \prod_{i=1}^{n-1} [1 - G(M_{th})] [G(M_{th})]^{J_i} \times \\ & \prod_{i=1}^{n-1} \left[ J_i! \prod_{j=1}^{J_i} \frac{f(\tau_{i(j)})}{(t_{i+1} - t_i) F(1)} \right] \times \prod_{i=1}^{n-1} \prod_{j=1}^{J_i} \frac{g(M_w^{(ij)})}{G(M_{th})} \end{aligned} \quad (9)$$

where the first line gives the probability of  $(t_i, M_w^{(i)})$ , the second the probability of the number of *subordinates* events, and the third the probability of  $(s_{ij}, M_w^{(ij)})$ . We note that the factor  $J_i!$  is due to the fact that there are  $J_i$  unordered samples from  $F(\tau)$  corresponding to the ordered sequence of observed times  $s_{ij}$  in  $(t_i, t_{i+1})$ ,  $\forall i = 1, \dots, n-1$ .

### 3 Bayesian estimation

We estimate the model parameters following the Bayesian approach. According to this paradigm, the parameters  $\theta$  are considered as random variables and our beliefs about their variability are formalized through prior distributions. Prior knowledge on the model parameters arises generally from the literature and previous experience. Our model is formulated for the first time and some parameters are not strictly related to easily measurable physical quantities; therefore we assign the prior distributions according to an objective Bayesian perspective, by combining the empirical Bayes method and the use of vague-proper prior distributions [3]. We choose the prior distribution of each parameter in agreement with its support, and we express the parameters of this prior distribution (called *hyperparameters*) as functions of the prior mean  $\mu_0$  and variance  $\sigma_0^2$  of the corresponding model parameter so that, assigned  $\mu_0$  and  $\sigma_0^2$ , we also have the hyperparameters. According to the empirical Bayes method, preliminary values of the prior means are obtained by maximizing the marginal likelihood and by setting the standard deviations to 90% of the corresponding means to avoid that the estimates provided for the variances through the maximization are too close to zero. This implies a double use of the data in assigning the hyperparameters and in evaluating the posterior distributions. To avoid this drawback we choose priors that ‘span the range of the likelihood function’ [3], that is, by varying the hyperparameters around their preliminary estimates and choosing those values that include most of the mass of the likelihood function, but that do not extend too far.

In the Bayesian framework, the estimate of a parameter is typically given by its posterior mean, obtained, together with measures of its uncertainty, by its posterior distribution. If, through Bayes’ theorem, an explicit formulation for the posterior

distribution is not available, we resort to methods of stochastic simulation based on constructing a Markov chain that has the desired distribution as its equilibrium distribution (*Markov chain Monte Carlo* (MCMC) methods). In this study we have applied the Metropolis-Hasting MCMC algorithm to generate Markov chains converging to the posterior distributions of the model parameters.

## 4 Results

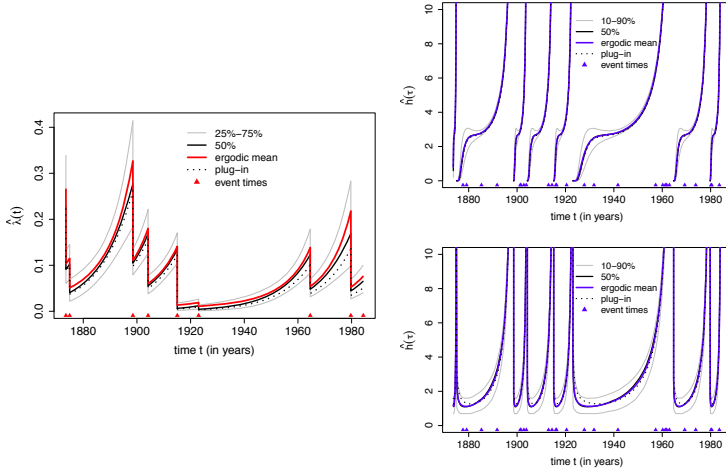
Fig. 2 provides a graphical summary of the results; the picture on left shows the conditional intensity function of the stress release model (1), whereas the pictures on right represent the hazard functions of the flexible (*top*) and additive (*bottom*) Weibull models respectively. The estimates of these functions have been obtained (i) by replacing the parameter estimates in their expressions (plug-in estimate) and (ii) through the ergodic mean of the values obtained by replacing each parameter with the elements of the respective Markov chain generated by the MCMC algorithm. This second way of estimation provides a sequence of values at each instant  $t$ ; e.g., for the conditional intensity function (1), one has  $\{\lambda_s(t | \theta^{(l)}, \mathcal{H}_n)\}_{l=1}^R$ , where  $R$  is the number of elements of the Markov chain  $\{\theta^{(l)}\}_{l=1}^R$ . Through this sequence we can also obtain the median and the quartiles of the pointwise estimate of  $\lambda_s(t)$ .

Different approaches can be adopted to evaluate the goodness of fit of a model to the data and to compare pairs of models; according to the Bayesian approach we quantify the evidence in favour of a model through the Bayes factor. Given two models  $\mathcal{M}_1, \mathcal{M}_2$ , and the dataset  $\mathbf{D}$ , the Bayes factor is the ratio of the posterior odds of the models to their prior odds. When the prior probabilities  $pr(\mathcal{M}_k), k = 1, 2$ , of the two models are equal, the Bayes factor coincides with the ratio of their marginal (or *integrated*) likelihoods  $pr(\mathbf{D} | \mathcal{M}_k), k = 1, 2$  obtained by integrating (9) over the parameter space with respect to their prior distributions.

The two versions of our model differ in the probability distribution of the *sub-ordinates* rupture time: in case of the flexible Weibull distribution (model  $\mathcal{M}_f$ ) the marginal likelihood is equal to -43.39 in  $\log_{10}$ -scale, whereas in case of the additive Weibull distribution (model  $\mathcal{M}_a$ ) the marginal likelihood is equal to -43.06. According to the interpretation of Jeffreys' scale, the Bayes factor  $\log_{10} B_{a,f} = \log_{10} pr(\mathbf{D} | \mathcal{M}_a) - \log_{10} pr(\mathbf{D} | \mathcal{M}_f) = -43.06 + 43.39 = 0.33$  indicates slight evidence in favor of the model  $\mathcal{M}_a$  which has a bathtub-shaped hazard function for the rupture time.

We note that aftershocks are lacking in the CPT104 catalogue according to the compilers; we hope for variations, possibly improvements, in the results from the application of the model to catalogues with a greater number of secondary events.

**Acknowledgements** This study was partially funded by the project PRIN 2015 - Prot. 20157PRZC4 "Complex space-time modeling and functional analysis for probabilistic forecast of seismic events".



**Fig. 2** (Left) Different estimates of the conditional intensity function  $\lambda_s(t)$  of the stress release model, and (Right) of the hazard function of the flexible Weibull distribution (top) and of the additive Weibull distribution (bottom).

## References

1. Bebbington, M., Harte, D.: The linked stress release model for spatio-temporal seismicity: formulations, procedures and applications, *Geophys. J. Int.*, **154**, 925-946 (2003)
2. Bebbington, M., Lai, C.D., Zitkikis, R.: A flexible Weibull extension, *Reliab. Eng. Syst. Safe.*, **92**, 719-726 (2007)
3. Berger, J.: The case for objective Bayesian analysis, *Bayesian Anal.*, **1**, 3, 385-402 (2006)
4. CPTI Working Group: Catalogo Parametrico dei Terremoti Italiani, version 2004 (CPTI04), INGV, Bologna (2004) available on <http://emidius.mi.ingv.it/CPTI04/>
5. DISS Working Group: Database of Individual Seismogenic Sources (DISS), Version 3.0.2: A compilation of potential sources for earthquakes larger than M 5.5 in Italy and surrounding areas, <http://diss.rm.ingv.it/diss/>, © INGV 2007 - Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy, DOI:10.6092/INGV.IT-DISS3.0.2
6. Lai, C.-D.: Generalized Weibull Distributions, *SpringerBriefs in Statistics* (2014) doi:10.1007/978-3-642-39106-4
7. Schoenberg, F., Bolt, B.: Short-term exciting, long-term correcting models for earthquake catalogs, *B. Seismol. Soc. Am.*, **90**, 4, 849-858 (2000) doi:10.1785/0119990090
8. Rotondi, R., Varini, E.: Bayesian inference of stress release models applied to some Italian seismogenic zones, *Geophys. J. Int.*, **169**, 1, 301-314 (2007)
9. Varini, E.: Sequential estimation methods in continuous-time state-space models, PhD Thesis, Institute of Quantitative Methods, Bocconi University, Milano, Italy (2005)
10. Varini, E., Rotondi, R., Basili, R., Barba, S.: Stress release model and proxy measures of earthquake size. Application to Italian seismogenic sources, *Tectonophysics*, **682**, 147-168 (2016) <http://dx.doi.org/10.1016/j.tecto.2016.05.017>
11. Zheng, X., Vere-Jones, D.: Application of stress release models to historical earthquakes from North China, *Pure Appl. Geophys.*, **135**, (4), 559-576 (1991) doi:10.1007/BF01772406

# Functional principal component analysis of quantile curves

## *Analisi in componenti principali funzionali di curve quantiliche*

M. Ruggieri, F. Di Salvo and A. Plaia

**Abstract** Literature on functional data analysis is mainly focused on estimation of individuals curves and characterization of average dynamics. The idea underlying this proposal is to focus attention on other particular features of the distribution of the observed data, moving from mean functions towards functional quantiles. The motivating examples are functional data sets that are collections of high frequency data recorded along time. As quantiles provide information on various aspects of a time series, we propose a modelling framework for the joint estimation of functional quantiles, varying along time, and functional principal components, summarizing some common dynamics shared by the functional quantiles.

**Abstract** *La letteratura sull'analisi di dati funzionali è prevalentemente rivolta alla modellazione e stima delle singole curve aleatorie e alla caratterizzazione del momento primo. L'idea di base di questo lavoro è considerare altri aspetti della distribuzione dei dati osservati, spostando l'attenzione verso i quantili funzionali. La tipologia di dati a cui questa analisi si rivolge è rappresentata da dati ad alta frequenza osservati nel tempo. Poiché i quantili sintetizzano informazioni sulle dinamiche temporali di una serie storica, si propone un approccio per la stima di quantili funzionali, in corrispondenza di diversi valori di probabilità, e per la derivazione di componenti principali funzionali che ne riassumano le dinamiche comuni.*

**Key words:** functional data, nonparametric quantile regression, penalized splines, functional principal components

---

M. Ruggieri e-mail: [mariantonietta.ruggieri@unipa.it](mailto:mariantonietta.ruggieri@unipa.it),  
F. Di Salvo, e-mail: [francesca.disalvo@unipa.it](mailto:francesca.disalvo@unipa.it),  
A. Plaia, e-mail: [antonella.plaia@unipa.it](mailto:antonella.plaia@unipa.it)

Department of Economics Business and Statistics, University of Palermo, viale delle Scienze, Palermo.

## 1 Introduction

Let consider high dimensional data observed at discrete times; although we observe a finite number of measured values, they are often analyzed as if they were defined in continuous time.

Traditional analysis concerns the conditional distribution at each time point, while in a functional data analysis (FDA) approach each time series is considered as a sample generated by a random curve, varying over a continuum; in both cases, more frequently the goal is the centre of the conditional distribution or a mean function describing the pattern of the set of functions.

In the univariate regression setting, quantile regression models the quantiles of the conditional distribution of the response variable; this is a valuable alternative to the conditional mean, when the interest is in the tails of the distribution or in presence of model mis-specification (see [4] and [5]). With the increasing demand of statistical tools for FDA is therefore natural to try to extend the definition and the estimation of quantile functions for infinite-dimensional data. However the extension of quantile function to a multivariate setting is not straightforward, because quantiles are basically defined by ordering values of a random variable. Since there is no natural order for  $R^n$  when  $n \geq 2$ , there is no obvious extension and a number of efforts has been devoted to this problem in the last years.

Our proposal explore the performance of the multi-way functional principal component analysis (FPCA) when functional quantiles of different order are simultaneously considered. There are some previous works motivating our idea and in particular [3] and [1]. An approach on generalized regression quantiles with their synthesis by means of a small number of principal components is proposed in [3] in a FDA framework.

Fraiman et al. [1] define directional quantiles and extend a projection-based definition of quantiles to infinite-dimensional Hilbert and Banach spaces; the authors develop a factor analysis based on principal directions and robust principal directions. The main results in [1] are based on an intuitive definition of directional quantiles, indexed by an order  $\alpha \in (0, 1)$  and a direction  $u$  in the unit sphere; the directional quantile describe the behavior of the probability distribution in finite and infinite-dimensional spaces; principal quantile directions are defined to summarize their information. Moreover, exploiting the idea of statistical depth, they generalize the definition of robust principal components for functional data.

In a previous paper [2], we estimate multivariate functional data by penalized B-spline; a working covariance matrix is also derived on the basis of coefficients of the splines, accounting for the main temporal effects; FPCA allow us to project data variations, observed in multidimensional space, into few dimensions. Due to dimensionality reduction and applying the Karhunen-Loève decomposition, this method is also useful for the representation of the random curves in terms of the factor functions.

In the present paper our main purpose is to investigate data by means of FPCA, capturing the tail behaviour of the distributions. The FDA approach is proposed for the simultaneous estimation of the functional regression quantiles; assuming that quan-

tiles, estimated at different values of probability, share some common features, they can be summarized by a small number of functional principal components, identifying the directions along which resuming the most interesting characteristics. The method is applied to air pollution data from a monitoring network.

## 2 The Methodology

The  $\alpha$ -quantile is defined as the inverse of a cumulative distribution function, given a real valued random variable  $X$ , with distribution  $F_X$ :

$$Q_X(\alpha) =: Q(F_X, \alpha), \inf\{x \in R : F_X(x) \geq \alpha\}.$$

We refer to the situation in which the interest is in the  $\alpha - th$  theoretical quantile of the conditional distribution of  $X$  at time  $t$ :

$$Q_{X|t}(\alpha|t) =: Q(F_{X|t}, \alpha). \tag{1}$$

In this setting the (1) is a time varying function:

$$Q_{X|t}(\alpha|t) = l_\alpha(t), \tag{2}$$

and the estimator is the minimizer of a expected (generally asymmetric) loss function.

Kato [4] is one of the earlier paper studying functional quantile regression; starting with a linear quantile regression, in which the response is scalar while the covariate is a function, and expanding the covariate and the slope function in terms of their principal components, the model is transformed into a quantile regression model with an infinite number of regressors. More recently, in [3] the functional quantiles  $l_{\alpha_i}(t)$  are estimated nonparametrically; on the basis of the Karhunen-Loève decomposition, they may be approximately represented by means of an Empirical Orthonormal Basis (EOF):

$$l_{\alpha_i}(t) = \mu(t) + \sum_{h=1}^H \psi^h(i) \xi_h^\alpha(t), \tag{3}$$

where  $\mu(t)$  is a mean function and  $\sum_{h=1}^H \psi^h(i) \xi_h^\alpha(t)$  is the reduced rank model obtained fixing the number  $H$  of bases; it is linear combination of principal components  $\psi^h(i)$  and eigenfunctions  $\xi_h^\alpha(t)$ .

The authors perform the analysis combining the representation (3) with the estimation procedure of the quantile function, after choosing a proper loss function.

We generalize this approach to the joint estimation of a collection of quantile functions, defined for a relevant set of probability values,  $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_q]$ , implementing a three-mode FPCA analysis together with a general smoothing approach.

A functional form is presented by means of multidimensional linear smooth functions:

$$l_\alpha(t) = \sum_{k=1}^K \theta_k^\alpha \phi_k(t), \quad (4)$$

where:  $\Phi(t) = \{\phi_k(t)\}$  is the set of  $K$  basis functions and  $\theta_1^\alpha = [\theta_{1,1}^\alpha, \dots, \theta_{1,K}^\alpha]$  is the vector of the coefficients.

In order to estimate the  $K \times N$  matrix  $\Theta^\alpha$  of parameters, the P-spline (penalized B-spline) approach is here considered, minimizing the penalized loss function:

$$PENRSS_\lambda(y) = \mathbf{w}(\alpha) \mathbf{I} |X - \Phi \Theta^\alpha|^T |X - \Phi \Theta^\alpha| + \lambda \Theta^\alpha \mathbf{H} \Theta^\alpha, \quad (5)$$

where the elements of vector  $w(\alpha) = [w_1(\alpha), \dots, w_n(\alpha)]$  are:

- $w_i(\alpha) = \alpha$ , if  $X_i > \alpha \Phi_i \Theta^\alpha$ ;
- $w_i(\alpha) = (1 - \alpha)$ , if  $X_i \leq \alpha \Phi_i \Theta^\alpha$ .

For details of penalty term  $\lambda \Theta^\alpha \mathbf{H} \Theta^\alpha$  in (5), as well as for the estimation procedure, we refer to [2]. In this framework, three-mode functional principal quantile are derived straightforward by decomposition of the variance function, estimated by a working variance array (referred to  $N$  curves,  $T$  time units and  $Q$  quantiles) defined in terms of the estimated coefficients (see also [2]). An interesting result is the decomposition of a random function into two sets: the set of factor scores, one for each curve, on the basis of all their quantiles, and the set of corresponding factor loadings, defining the mood of variations.

### 3 The application

We illustrate the proposed method with an example of  $PM_{10}$  daily time series registered in one year in  $N = 59$  stations of a monitoring network in California.

In Fig. 1 (a) the set of the  $N$  observed curves are represented (gray lines); a subset of seven curves are selected (coloured lines) in order to highlight the results of the procedure. In Fig. 1 (b) – (f) the estimated quantile functions for different probability values, from 0.1 to 0.9, synthesize the specific pattern of the respective curves. Fig. 2 show the projections of the quantile functions in the space of the first two principal components with proportion of variance explained 0.793 and 0.073; figures (a) – (e) are the partial scores for each quantile and (f) the total scores. We can observe that the functional principal components retain the most information of the original curves and curves with similar pattern have similar scores.



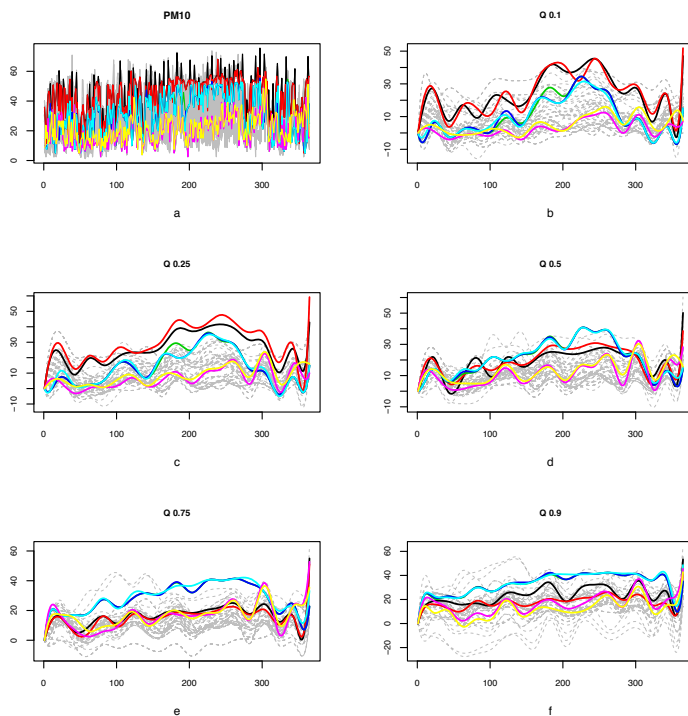
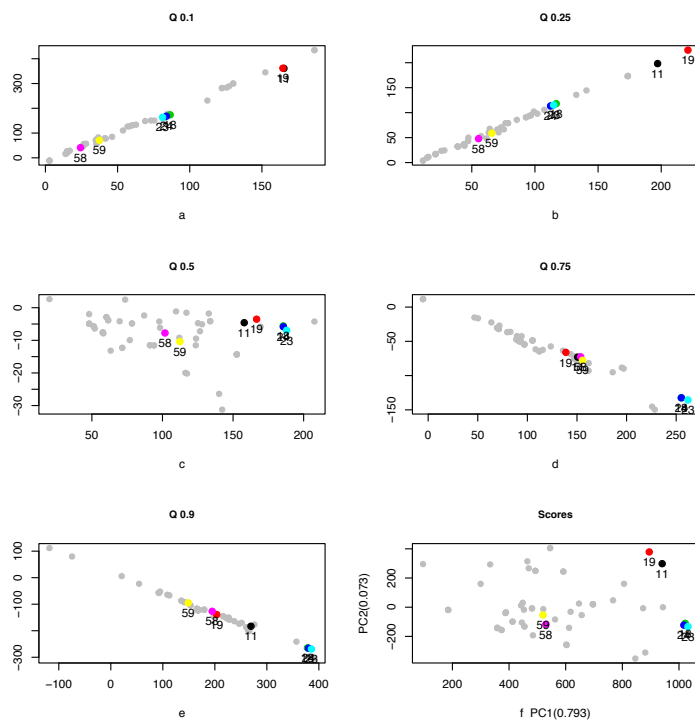


Fig. 1 Observed curves and estimated quantile functionals

## 4 Conclusion

The FDA approach is proposed for the simultaneous estimation of functional regression quantiles, when the main purpose is capturing the tail behaviour; assuming that quantiles estimated at different values of probability share some common features, they can be summarized by a small number of functional principal components, identifying the directions along which resuming the interesting characteristics. Some implication and appealing intuitions can be borrowed from approaches relied on depth measures, in order to construct basic tools for functional data. The approach has the advantage of further generalization, such as the inclusion of explanatory variables and distributional assumptions. Many consequent applications of the FPCA in quantile regression are motivated by the Karhunen-Loève theorem, by means of which the random curves find convenient representations in terms of empirical orthogonal functions.



**Fig. 2** Projection of the curves in the space of the first two partial (a)-(e) and total (f) principal components

## References

1. Fraiman, R., Pateiro-Lopez, B. Functional quantiles, in F. Ferraty and Y. Romain, eds, *Recent Advances in Functional Data Analysis and Related Topics. Contributions to Statistics*, Physica-Verlag HD, pp. 1231-129 (2011).
2. Di Salvo, F., Ruggieri, M., Plaia, A., Functional Principal Component analysis for multivariate multidimensional environmental data. *Environmental and Ecological Statistics*, 22 (4), 739-757 (2015).
3. Guo, M., Zhou, L., Huang, J.Z., Hardle, W.K., Functional data analysis of generalized regression quantiles. *Statistics and Computing*, 25 (2), 189-202 (2015). DOI 10.1007/s11222-013-9425-1
4. Kato, K., Estimation in functional linear quantile regression. *The Annals of Statistics*, 40 (6), 3108-3136 (2012). DOI: 10.1214/12-AOS1066
5. Koenker, R., *Quantile Regression*. Cambridge University Press, New York (2005).

# Detecting group differences in multivariate categorical data

## *Ricerca delle differenze tra gruppi in dati categoriali multivariati*

Massimiliano Russo

**Abstract** In several studies, a group indicator is collected together with a multivariate vector of categorical variables with main goal in assessing evidence of differences of the collected vector across these groups. Similar goals arise routinely, but very few general methods which can test for group differences in multivariate categorical data are discussed in literature. We address this goal proposing a Bayesian model which factorizes the joint probability mass function for the group variable and the multivariate categorical data as the product of the marginal probabilities for the groups and the conditional probability mass function of the multivariate categorical data given the group membership. To provide a flexible and computationally tractable model for the probability mass function of multivariate categorical vector we rely on a mixture of tensor factorizations, facilitating dimensionality reduction, while providing simple and accurate test procedures to assess global and local group differences.

**Abstract** *In molti studi, si osserva un vettore di dati qualitativi non ordinali associato con un indicatore di gruppo. In questo contesto uno degli obiettivi principali è quello di stabilire se esistono differenze significative nel vettore di variabili qualitative osservato al variare del gruppo. Simili obiettivi sono presenti in diverse applicazioni, ma la letteratura corrente deficit di metodologie generali che possano testare le differenze di gruppo in un vettore di dati qualitativi non ordinali a diversi livelli. Per perseguire tale obiettivo ci si è basati su un modello bayesiano, fattorizzando la probabilità congiunta per la variabile di gruppo ed il vettore di dati qualitativi come il prodotto della funzione di probabilità marginale della variabile di gruppo e la funzione di probabilità condizionata del vettore di variabili qualitative dato l'indicatore di gruppo. Al fine di ottenere un modello flessibile per la funzione di probabilità del tensore si è utilizzata una mistura di tensori che favorisce la riduzione della dimensionalità, portando a procedure accessibili per testare differenze globali e locali.*

---

Massimiliano Russo

Department of Statistical Sciences, University of Padova, 35121 Padova, Italy, e-mail: russo@stat.unipd.it

**Key words:** Categorical data; Endgame problem; Hypothesis testing; Tensor factorization.

## 1 Introduction

Categorical data frequently arise in different applications especially in such areas as clinical trials, psychology and social sciences in which many nominal data are required. In some applications, as medicine or social sciences, we have a group division — case/control, severity of a disease or social classes — with main goal being in testing how the whole set of measured variables changes according to the group structure. The main aim in our work is in presenting a model to test how the dependence structure of a whole contingency table varies across groups. We additionally propose local tests to establish which variables are responsible for such variation.

In accomplishing this goal a widely used approach consists in separately testing group difference in each marginal via chi-square test, accounting for multiplicity by false discovery rate control [2]. This approach does not incorporate dependence structure and hence is usually characterized by low power.

To take into account dependence underlying the data, a possible solution is given by nonparametric permutation tests [6], but although presenting a valid alternative these methods cannot detect changing that goes beyond marginals, giving inaccurate results when changes occurs in higher order structures.

To overcome this last issue, one possibility is to define a test based on a flexible representation for the probability mass function of the multivariate categorical data. Analysis of contingency tables is mostly based on log-linear models [1], but when the number of variable is even moderately high the set of possible interaction become huge, making successive inference a difficult task.

Recently Dunson and Xing [3] proposed a Bayesian nonparametric methodology which defines the probability measure over a tensor as a mixture of product of multinomial distributions, avoiding direct specification of the underlying dependence structure. The proposed model is computationally tractable, has theoretical justifications and has been recently generalized to different frameworks — e.g. [9] and [10].

While these lasts focus on modeling the conditional probability mass function of a univariate response with the categorical data acting as predictors, we consider instead the dual problem, assessing evidence of group differences in the entire probability mass function of a multivariate categorical random variable. In accomplishing this goal we factorize the joint probability mass function for the group variable and the multivariate categorical random variable as the product of the marginal probabilities for the groups and the conditional probability mass function of the multivariate categorical data. This last is defined via a mixture of tensor factorizations allowing a general and tractable formulation which facilitates global testing of group differences in the entire probability mass function for the multivariate categorical data.

## 2 Model formulation and testing

We propose a flexible model for the joint probability mass function  $p_{Y,X} = \{\text{pr}(Y = y, X = x) : y \in \mathcal{Y}, x \in \mathcal{X}\}$  underlying the observed data  $(y_i, x_i)$ , where  $y_i = (y_{i1}, \dots, y_{ip})^T \in \mathcal{Y} = (1, \dots, d_1) \times \dots \times (1, \dots, d_p)$  denote the observed vector of categorical data and  $x_i \in \mathcal{X} = (1, \dots, k)$  its corresponding group, for each subject  $i = 1, \dots, n$ .

Our main goal is to establish if the probability mass function varies across the level of  $X$ . This hypothesis can be formally stated as

$$H_0 : p_{Y,X}(y, x) = p_Y(y)p_X(x) \quad \text{v.s.} \quad H_1 : p_{Y,X}(y, x) \neq p_Y(y)p_X(x). \quad (1)$$

with  $p_Y(y) = \{\text{pr}(Y = y) : y \in \mathcal{Y}\}$  and  $p_X(x) = \{\text{pr}(X = x) : x \in \mathcal{X}\}$  the marginal probability mass functions of  $Y$  and  $X$ , respectively.

In order to develop an accurate test procedure for hypothesis (1), avoiding misspecification issues, a key step is to rely on a representation for  $p_{Y,X}$  which is sufficiently general to approximate any possible probability mass function in the  $|\mathcal{Y} \times \mathcal{X}| - 1$  dimensional simplex  $\Delta_{|\mathcal{Y} \times \mathcal{X}|}$ . We address this goal by expressing  $p_{Y,X}$  as

$$p_{Y,X}(y, x) = p_{Y|X=x}(y)p_X(x) \quad (y \in \mathcal{Y}, x \in \mathcal{X}), \quad (2)$$

with the conditional probability mass function of  $Y$  given  $X = x$  factorized as

$$p_{Y|X=x}(y) = \sum_{h=1}^H v_{hx} \prod_{j=1}^p \psi_{hy_j}^{(j)} \quad (y \in \mathcal{Y}, x \in \mathcal{X}), \quad (3)$$

where  $v_x = (v_{1x}, \dots, v_{Hx}) \in \Delta_H$  are vectors of mixing probabilities specific to each group  $x = 1, \dots, k$ , while  $\psi_{hy_j}^{(j)}$  is the probability that the categorical random variable  $Y_j$  assumes value  $y_j$  in mixture component  $h$ , for each  $y_j \in (1, \dots, d_j)$ ,  $j = 1, \dots, p$  and  $h = 1, \dots, H$ . Under factorization (2) the marginal probability  $p_X(x)$  is the probability mass function over a categorical vector having  $k$  levels and can be efficiently modeled via multinomial distribution.

Representation (3) for the conditional probability provides a parsimonious model, reducing dimensionality. Group-dependence is included in the model only in the mixture weights allowing for efficient borrowing of information. Additionally, it is easy to show that hypothesis (1) reduces to

$$H_0 : v_1 = \dots = v_k \quad \text{versus} \quad H_1 : v_x \neq v_{x'} \quad \text{for some } x, x'. \quad (4)$$

Hypothesis (4) is based on a representation of  $p_{Y,X}$  which is general and robust against model misspecification (refers to [7] for proofs and additional details) providing an accurate solution, moreover it can be directly included in the model via suitable prior [5]

$$\begin{aligned}
v_x &= (1 - T)u + Tu_x \\
u &\sim \text{Dirichlet}\{\gamma_1, \dots, \gamma_H\}, \quad u_x \sim \text{Dirichlet}\{\gamma_1, \dots, \gamma_H\}, \quad x = 1, \dots, k \quad (5) \\
T &\sim \text{Ber}\{\text{pr}(H_1)\}
\end{aligned}$$

where,  $T$  is a hypothesis indicator, with  $T = 0$  for  $H_0$  and  $T = 1$  for  $H_1$ . Under  $H_1$ , we generate group-specific mixing weights while under  $H_0$  we have equal weight vectors. In assessing evidence in favor of the alternative, we can rely on the posterior probability,  $\text{pr}(H_1 | (y_1, x_1), \dots, (y_n, x_n))$ , or the correspondent Bayes factor, easily obtained from the output of a Gibbs sampler. Specifically, under prior (5) the full conditional  $\text{pr}(T = 0 | -) = \text{pr}(H_0 | -) = 1 - \text{pr}(H_1 | -)$  is analytically available.

Although rejection of the global null (4) provides evidence of group differences in the multivariate categorical random variable  $Y$ , such changes may be due to several structures. To provide interpretable inference we additionally consider local analyses assessing evidence of group differences in each marginal  $Y_j$  of  $Y$ , for  $j = 1, \dots, p$ .

We address the above aim by adapting the model-based version of the Cramer's V coefficient [3] to our local tests. Specifically, we measure the association between each marginal  $Y_j$  and  $X$  for  $j = 1, \dots, p$  studying the posterior distributions of the coefficients.

$$\rho_j^2 = \frac{1}{\min\{k, d_j\} - 1} \sum_{x=1}^k p_X(x) \sum_{y_j=1}^{d_j} \frac{(p_{Y_j|x}(y_j) - p_{Y_j}(y_j))^2}{p_{Y_j}(y_j)}. \quad (6)$$

where  $p_{Y_j}(y_j)$  denotes  $\text{pr}(Y_j = y_j)$ , while  $p_{Y_j,x}(y_j, x) = \text{pr}(Y_j = y_j, X = x) = \text{pr}(Y_j = y_j | X = x)\text{pr}(X = x) = p_{Y_j|x}(y_j)p_X(x)$  for every  $y_j \in (1, \dots, d_j)$  and group  $x = 1, \dots, k$ .

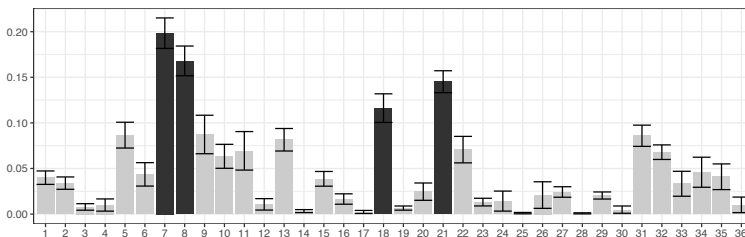
Relying on  $\rho_j \in [0, 1]$  to study variation of the marginal across groups provides a convenient choice for interpretation. In fact, according to (6), a value of  $\rho_j$  very close to 0 suggests low dependence between  $Y_j$  and  $X$ . A discussion on the choice of the prior distributions for the involved quantities and an efficient algorithm to perform posterior inference from the proposed model are described in [7].

### 3 Application to chess endgame data

Developing efficient strategies for the endgame part of a chess game presents many difficulties, especially when implementing a computer chess program. These difficulties arise since decisions to be adopted are different from the ones used for the first part of the game, strongly depending on which pieces are still on the board and on their position. Different ending scenarios can be considered but we focus on King-Rook vs. King-Pawn game. The data consists of  $n = 3196$  chess games where  $p = 36$  categorical variables indicating the chess board configuration are registered together with a variable indicating if the white can win the game or not (refer to [8] for a more detailed description of the data) Analysed data are publicly available at

the UCI machine learning repository [4]. Our interest is in establishing if the position of the pieces at the beginning of the endgame stage has an impact on the final result and if so targeting the piece and positions responsible for such variation. This last information might be used to develop refined game strategies, driving the game in the final position more suitable for victory or tie.

We consider 5000 Gibbs samples relying on the hyperparameter settings suggested in [7] except for the upper bound  $H = 20$ . Trace-plots suggest that convergence is reached after a burn-in of 1000. Results from posterior inference offer interesting insights on group differences in the considered endgame problem with a  $\text{pr}\{H_1 \mid (y_1, x_1), \dots, (y_n, x_n)\} > 0.95$  providing evidence that the starting position has a deep impact on the final result. Figure 1 shows the posterior mean and 0.9 credible interval for the Cramer V coefficient (6) for all the considered variables. We can notice how just few position of the pieces may impact on the final result, suggesting to drive the game in these lasts to build sophisticated strategies.



**Fig. 1** Posterior mean, 0.1 and 0.9 quantiles of  $\hat{\rho}_j$  for  $j = 1, \dots, 36$  considered chess piece/position. Dark grey bars are such that  $\text{pr}\{\hat{\rho}_j > 0.10 \mid (y_1, x_1), \dots, (y_n, x_n)\} > 0.95$ .

## References

1. Agresti, A.: Categorical Data Analysis. Wiley, (2013).
2. Benjamini, Y. and Hochberg, Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc: Series B* **57**, 289–300. (1995)
3. Dunson, D. B. and Xing, C: Nonparametric Bayes modeling of multivariate categorical data. *J. Am. Statist. Assoc.* (2009); **104**, 1042–1051.
4. Frank, A. and Asuncion, A. UCI machine learning repository, (2010).
5. Lock, E. F. and Dunson, D. B.: Shared kernel Bayesian screening. *Biometrika* **102**, 829–842. (2015)
6. Pesarin, F., and Salmaso L.: Permutation tests for complex data: theory, applications and software. John Wiley & Sons, (2010).
7. Russo, M., Durante .D., and Scarpa B: Bayesian inference on group differences in multivariate categorical data. arXiv preprint (arXiv:1606.09415), (2016).
8. Shapiro A. D: Structured induction in expert systems. Addison-Wesley Longman Publishing Co., Inc. (1987)

9. Yang, Y and Dunson, D. B.: Bayesian conditional tensor factorizations for high-dimensional classification. *J. Am. Statist. Assoc.*, (2015): **in publication**.
10. Zhou, J., Herring, A. H., Bhattacharya, A., Olshan, A. F. and Dunson, D. B.: Nonparametric Bayes modeling for case control studies with many predictors. *Biometrics*, (2015): **72**, 184–92.



# A Sequential Test for the $C_{pk}$ Index

## *Un test sequenziale per l'indice $C_{pk}$*

Michele Scagliarini

**Abstract** We propose a new sequential hypothesis test for the process capability index  $C_{pk}$ . We compare the statistical properties of the proposed test with the properties of non-sequential tests by performing a simulation study. The results indicate that the sequential test makes it possible to save a large amount of sample size, while type I and II error probabilities are maintained at their desired values.

**Abstract** *In questo lavoro si propone un nuovo test sequenziale per la verifica d'ipotesi sull'indice di capacità  $C_{pk}$ . Le proprietà statistiche del test sequenziale sono confrontate con le proprietà di test non sequenziali mediante simulazioni. I risultati indicano che il test sequenziale consente una notevole riduzione dell'ampiezza campionaria mantenendo le probabilità degli errori di primo e secondo tipo ai valori prefissati.*

**Key words:** Brownian motion, Monte Carlo Simulation, non-central  $t$  distribution, power function, process capability indices, sequential test.

## 1 Introduction

One of the process capability indices most widely used in industry today is  $C_{pk} = (d - |\mu - \frac{1}{2}(USL - LSL)|) / 3\sigma = (d - |\mu - m|) / 3\sigma$  where  $\mu$  is the process mean,  $\sigma$  is the process standard deviation,  $LSL$  and  $USL$  are the specification limits,  $d = (USL - LSL) / 2$  and  $m = (USL + LSL) / 2$  [4]. Often, as a part of contractual agreement, it is necessary to demonstrate that the process capability index  $C_{pk}$  meets or exceeds some particular target value, say  $c_{pk,0}$ . Such decision-making problem may be

---

<sup>1</sup> Michele Scagliarini, Department of Statistical Sciences, University of Bologna; michele.scagliarini@unibo.it

formulated as a hypothesis testing problem:  $H_0 : C_{pk} \leq c_{pk,0}$  *i.e.* the process is not capable versus  $H_1 : C_{pk} > c_{pk,0}$  *i.e.* the process is capable.

In this study, starting from some of the results obtained by [2], we propose a sequential test for the index  $C_{pk}$ . Firstly, we review two of the most used non-sequential tests for assessing whether a process is capable or not. Secondly, we analytically derive the test statistic of the sequential test. Thirdly, we describe in detail the testing procedure. Finally, we compare the sequential test properties with the performances of the non-sequential tests by performing an extensive simulation study. The results show that the proposed sequential test has good power behavior and makes it possible to save a large amount of sample size, which can be translated into reduced costs, time and resources.

## 2 Hypothesis testing on $C_{pk}$

Assuming a normally distributed quality characteristic,  $X \sim N(\mu, \sigma^2)$ , [5] proposed a statistical test (PC-test) based on the distribution of the estimator  $\tilde{C}_{pk} = b_f \hat{C}_{pk}^n$  where  $\hat{C}_{pk}^n = (d - (\bar{X} - m)I_A(\mu))/3S$ ,  $b_f = \sqrt{2}\Gamma[(n-1)/2]/\sqrt{n-1}\Gamma[(n-2)/2]$ ,  $n$  is the sample size,  $\bar{X} = \sum_{i=1}^n X_i/n$ ,  $S = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2/(n-1)}$ ,  $I_A(\mu) = 1$  if  $\mu \in A$  and  $I_A(\mu) = -1$  if  $\mu \notin A$  with  $A = \{\mu | \mu \geq m\}$ . Given the type I error probability  $\alpha$ , the critical value of the test is  $C_0 = b_f t_{n-1,\alpha}(\delta_c)/3\sqrt{n}$  where  $t_{n-1,\alpha}(\delta_c)$  is the upper  $\alpha$  quantile of a non-central  $t$  with  $n-1$  degrees of freedom and non-centrality parameter  $\delta_c = 3\sqrt{nc_{pk,0}}$ . The power function of the PC-test can be computed as  $\pi_{PC}(c_{pk,1}) = \Pr\{t_{n-1}(\delta) > 3\sqrt{n}C_0/b_f\}$  where  $\delta = 3\sqrt{nc_{pk,1}}$ .

Recently [3], for testing  $H_0$  versus  $H_1$ , discussed a test (LP-test) based on the estimator  $\hat{C}_{pk} = (1 - |\hat{\delta}|)/3\hat{\gamma}$ , where  $\hat{\gamma} = \hat{\sigma}/d$ ,  $\hat{\delta} = (\bar{X} - m)/d$  and  $\hat{\sigma} = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2/n}$ . Under the assumption of a normally distributed quality characteristic, the authors obtained the critical value for the test as  $c_{pk;\alpha} = t_{n-1,\alpha}(\delta_0)/3\sqrt{n-1}$  where  $\delta_0 = 3\sqrt{nc_{pk,0}}$ . The power function of the LP-test can be computed as  $\pi_{LP}(c_{pk,1}) = 1 - F_{\hat{C}_{pk}}(c_{pk;\alpha})$ , where  $F_{\hat{C}_{pk}}$  is the cumulative distribution function of  $\hat{C}_{pk}$  and is defined as  $F_{\hat{C}_{pk}}(x) = 1 - \Pr(t_{n-1}(\delta_1) \leq -t) + \Pr(t_{n-1}(\delta_2) \leq -t)$  if  $x \leq 0$  and  $F_{\hat{C}_{pk}}(x) = 1 - Q_{n-1}(-t, \delta_1; 0, R) + Q_{n-1}(t, \delta_2; 0, R)$  if  $x > 0$ . In the previous

equations  $t_{n-1}(\delta_1)$  and  $t_{n-1}(\delta_2)$  are non-central  $t$  variables with  $n-1$  degrees of freedom and non-centrality parameters  $\delta_1 = -3\sqrt{n}(1-\delta)C_{pk}/(1-|\delta|)$  and  $\delta_2 = 3\sqrt{n}(1+\delta)C_{pk}/(1+|\delta|)$  respectively,  $\delta = (\bar{X} - \mu)/d$ ,  $t = 3\sqrt{n-1}x$ ,  $R = \sqrt{n-1}(\delta_2 - \delta_1)/2t$ ,  $\Gamma$  is the gamma function,  $Q_f(t, \delta; 0, R) = \frac{\sqrt{2\pi}}{\Gamma(f/2)2^{(f-2)/2}} \int_0^R \Phi(tx/\sqrt{f} - \delta)x^{f-1}\phi(x)dx$ ,  $\Phi$  and  $\phi$  are respectively the normal cumulative distribution function and probability density function.

### 3 A sequential test for $C_{pk}$

Under the assumption that the data came from a multivariate distribution with density function  $f(x; \theta)$ , [2] proposed a general sequential testing procedure for testing  $H_0: h(\theta) = \mathbf{0}$  versus  $H_1: h(\theta) \neq \mathbf{0}$ , where  $h(\theta): \mathbb{R}^d \rightarrow \mathbb{R}^q$ , with  $q \leq d$ , is a function with first order derivative matrix denoted by  $H(\theta)$  with  $\theta$  unknown. Under the standard regularity conditions for the existence of the multivariate maximum likelihood estimators the author showed that the statistic  $W_k = kh(\hat{\theta}_k)[H'(\theta)I^{-1}(\theta)H(\theta)]^{-1}h(\hat{\theta}_k)^t$ , where  $k$  is the sample size,  $\hat{\theta}_k$  is a consistent estimator of  $\theta$  and  $I(\theta)$  is the Fisher information matrix, can be approximated by a functional of Brownian motions. Thus the author [2] proposed as test statistic  $W_k^* = kh(\hat{\theta}_k)[H'(\hat{\theta}_k)I^{-1}(\hat{\theta}_k)H(\hat{\theta}_k)]^{-1}h(\hat{\theta}_k)^t$  where  $\hat{\theta}_k$  is the maximum likelihood estimator of  $\theta$ . The  $\alpha$ -level sequential test procedure, truncated at the maximal allowable sample size  $n_0$ , is performed as follows:

1. for  $k = 2, 3, \dots, n_0$  compute of the statistic  $W_k^{*(1)} = \sqrt{k/n_0}\sqrt{W_k^*}$ ;
2. hypothesis  $H_0$  is rejected the first time that  $W_k^{*(1)}$  exceeds the critical value  $w_\alpha$ ;
3. if  $W_k^{*(1)}$  does not exceed  $w_\alpha$  by  $n_0$  then do not reject  $H_0$ .

The maximal sample size  $n_0$  can be decided on the basis of financial, ethical or statistical reasons as, for example, to achieve a desired power level. The critical value  $w_\alpha$ , given the type I error probability  $\alpha$ , can be obtained from [1].

Let us consider the hypothesis  $H_0: C_{pk} = c_{pk,0}$  versus  $H_1: C_{pk} \neq c_{pk,0}$  and assume that the quality characteristic is normally distributed:  $X \sim N(\mu, \sigma^2)$ . Let us define  $h(\theta)$  as  $h(\theta) = \ln\left(\left(\frac{C_{pk}}{c_{pk,0}}\right)^2\right) - \ln\left(\left(\frac{c_{pk,0}}{c_{pk,0}}\right)^2\right) = \ln\left[\frac{(d - |\mu - m|)^2}{9\sigma^2(c_{pk,0})^2}\right]$ , where

$\theta = (\mu, \sigma^2)$ . For  $C_{pk} \geq 0$ ,  $H_0$  is equivalent to  $H_0 : h(\theta) = 0$  and the alternative hypothesis is equivalent to  $H_1 : h(\theta) \neq 0$ . In the case at hand the statistic  $W_k^*$  can be written as  $W_k^* = kh^2(\hat{\theta}_k) [H^t(\hat{\theta}_k) I^{-1}(\hat{\theta}_k) H(\hat{\theta}_k)]^{-1}$  where  $\hat{\theta}_k = (\bar{X}_k, S_k^2)$  with  $\bar{X}_k = \sum_{i=1}^k X_i / k$  and  $S_k^2 = \sum_{i=1}^k (X_i - \bar{X}_k)^2 / k$ . The function  $h(\hat{\theta}_k)$  is therefore given by  $h(\hat{\theta}_k) = \ln \left( (d - |\bar{X}_k - m|)^2 / 9S_k^2 (c_{pk,0})^2 \right)$  and consequently  $W_k^*$  can be written as:

$$W_k^* = k \left[ \ln \left( \frac{(d - |\bar{X}_k - m|)^2}{9S_k^2 (c_{pk,0})^2} \right) \right]^2 \times \left[ \frac{4(\text{signum}[\bar{X}_k - m])^2 S_k^2}{[d - |\bar{X}_k - m|]^2} + 2 \right]^{-1} \tag{1}$$

where  $\text{signum}[a] = a/|a|$  if  $a \neq 0$ ;  $\text{signum}[a] = 0$  if  $a = 0$ . Therefore, given the value of  $\alpha$  and the maximal allowable sample size  $n_0$ , the test is performed by computing, for  $k=2,3,\dots, n_0$ , the statistic  $W_k^{*(1)} = \sqrt{k/n_0 W_k^*}$ .

Let  $n_{stop}$  be the first integer  $k=2,3,\dots, n_0$  for which  $W_k^{*(1)} > w_\alpha$ ; we reject  $H_0$  if  $W_{n_{stop}}^{*(1)} > w_\alpha$ ; we do not reject  $H_0$  if  $W_k^{*(1)}$  does not exceed  $w_\alpha$  by  $n_0$ . In this framework  $n_{stop}$  is the stopping sample size of the test.

#### 4 Simulation study and concluding remarks

We study the properties of the sequential procedure by comparing its performances with those of the LP and PC-tests. More precisely, we compare the tests in terms of the sample size required for achieving a given value of power. Note that the sequential test is two sided with composite alternative hypothesis  $H_1 : C_{pk} \neq c_{pk,0}$ , while the LP and PC-tests are unilateral. In order to correctly compare the statistical properties of the tests, we considered cases under  $H_1$  where  $C_{pk} = c_{pk,1}$  with  $c_{pk,1} > c_{pk,0}$ . In this manner the sequential bilateral test with Type I error probability  $\alpha$  can be compared with the non-sequential unilateral tests with Type I error probability equal to  $\alpha_u = \alpha/2$ . To study the properties of the sequential procedure we examined several scenarios where different values of  $C_{pk}$  under  $H_1$  ( $c_{pk,1}$ ) were considered for the unilateral test with  $\alpha_u = 0.01, 0.05$  and  $c_{pk,0} = 1.33, 1.67$ . For the LP and PC-tests we analytically determined the minimum sample size,  $n_{LP,0.80}$  and  $n_{PC,0.80}$  to achieve a power at least equal to or greater than 0.80: i.e.  $\pi_{LP}(c_{pk,1}) \geq 0.80$  and  $\pi_{PC}(c_{pk,1}) \geq 0.80$ . As far as the sequential test is concerned we used a set of simulation studies. For each value of  $\alpha$ ,

$c_{pk,0}$  and  $c_{pk,1}$  we generated  $10^4$  replicates from a normally distributed quality characteristic. The aim of these simulations was to determine the smallest maximal allowable sample size,  $n_{0;\hat{\pi}_s > 0.80}$ , which gives an empirical power  $\hat{\pi}_s$  greater than 0.80: i.e.  $\hat{\pi}_s > 0.80$ . The empirical power  $\hat{\pi}_s$  of the sequential test is estimated as the proportion of correctly rejected  $H_0$ . Note that, in order to obtain  $n_{0;\hat{\pi}_s > 0.80}$ , we implemented an iterative search algorithm which allows to determine  $n_{0;\hat{\pi}_s > 0.80}$  with a suitable precision.

The simulation results are summarized in Table 1 where, for each combination of  $\alpha$ ,  $c_{pk,0}$  and  $c_{pk,1}$ , the following quantities are reported:  $n_{0;\hat{\pi}_s > 0.80}$  the smallest maximal allowable sample size for the sequential test for achieving an empirical power  $\hat{\pi}_s > 0.80$ ;  $n_{avg}$  the average of the stopping sample sizes  $n_{stop}$  required for the sequential test with maximal allowable sample size  $n_{0;\hat{\pi}_s > 0.80}$  for concluding in favor of  $H_1$ ;  $S.D.(n_{stop})$  the standard deviation of the final sample sizes  $n_{stop}$ ;  $\hat{\pi}_s$  the estimated power of the sequential test. Table 1 contains also  $n_{LP;0.80}$  the minimum sample size required by the LP-test for achieving a power level  $\geq 0.80$  and  $n_{PC;0.80}$  the minimum sample size required by the PC-test for achieving a power level  $\geq 0.80$ .

**Table 1:** Simulation results under  $H_1$  with  $C_{pk} = c_{pk,1}$ , for  $c_{pk,0} = 1.33, 1.67$  and  $\alpha=0.02, 0.1$

$c_{pk,1}$	$n_{LP;0.80}$	$n_{PC;0.80}$	$n_{0;\hat{\pi}_s > 0.80}$	$n_{avg}$	$S.D.(n_{stop})$	$\hat{\pi}_s$
case $c_{pk,0} = 1.33$ and $\alpha=0.02$						
<b>1.60</b>	199	178	171	116.1	29.8	0.811
<b>1.70</b>	115	104	96	64.7	17.2	0.815
<b>1.80</b>	75	68	68	41.7	11.2	0.809
case $c_{pk,0} = 1.33$ and $\alpha=0.1$						
<b>1.60</b>	124	108	107	65.8	21.5	0.820
<b>1.70</b>	70	62	60	36.0	12.7	0.817
<b>1.80</b>	47	41	39	23.0	8.8	0.819
case $c_{pk,0} = 1.67$ and $\alpha=0.02$						
<b>2.00</b>	198	181	173	117.2	29.9	0.816
<b>2.10</b>	125	115	106	71.6	18.3	0.812
<b>2.20</b>	88	82	74	49.7	13.4	0.819
case $c_{pk,0} = 1.67$ and $\alpha=0.1$						
<b>2.00</b>	123	109	106	65.2	21.4	0.809
<b>2.10</b>	78	69	65	39.4	13.6	0.800
<b>2.20</b>	53	48	45	27.0	10.0	0.813

By examining the averages of the final sample sizes  $n_{avg}$  the results show that the sequential test, with the same power of the LP and PC-tests, saves a lot of sample size.

Furthermore, the maximum allowable sample size  $n_{0,\hat{\pi}_s > 0.80}$  required to achieve the desired power is always less than  $n_{LP,0.80}$  and  $n_{PC,0.80}$ . This indicates that even in the worst cases the sequential test needs a maximum allowable sample size not greater than the sample size of the non-sequential tests. As an example, under  $H_1 : C_{pk} = c_{pk,1}$  with  $c_{pk,1} = 1.60$ , when  $c_{pk,0} = 1.33$  and  $\alpha = 0.02$ , we have  $n_{LP,0.80} = 199$ ,  $n_{PC,0.80} = 178$ , while with a maximum allowable sample size equal to  $n_{0,\hat{\pi}_s > 0.80} = 171$  the power of the sequential test is  $\hat{\pi}_s > 0.80$  with an  $n_{avg} = 116.1 (= 117)$ . In this case the sequential procedure saves, on average, 41.2% of the sample size as to the LP-test and 34.3% as to the PC-test. Under  $H_1$  with  $c_{pk,1} = 2.0$ , when  $c_{pk,0} = 1.67$  and  $\alpha = 0.1$ , we have  $n_{LP,0.80} = 123$ ,  $n_{PC,0.80} = 109$ , while with a maximum allowable sample size equal to  $n_{0,\hat{\pi}_s > 0.80} = 106$  the power of the sequential test is  $\hat{\pi}_s > 0.80$  with an  $n_{avg} = 65.2 (= 66)$ . In this case the sequential procedure saves, on average, 46.3% of the sample size as to the LP-test and 39.4% as to the PC-test. A further simulation study, conducted under  $H_0$ , confirmed that the empirical type I error probability of the sequential test does not exceed the nominal  $\alpha$ -level.

The results show that the sequential test allows on average smaller stopping sample sizes as compared with the fixed sample size tests while maintaining the desired  $\alpha$ -level and power. Furthermore, the maximum allowable sample sizes required by the sequential test to achieve the desired power level are less than, or at most equal to, the sample sizes required by the non-sequential tests: this means that, even in the worst cases, the sequential procedure uses a sample size that does not exceed the sample size of the non-sequential tests with the same power level. Summarizing, the proposed sequential procedure has several interesting features: it offers a substantial decrease in sample size compared with the non-sequential tests, while type I and II error probabilities are correctly maintained at their desired values.

## References

1. Borodin, A.B., Salminen, P.: Handbook of Brownian Motion-Facts and Formulae. Birkhäuser Verlag, Basel, (1996).
2. Hussein, A., Ahmed, S.E., Bhatti, S.: Sequential testing of process capability indices. Journal of Statistical Computation and Simulation, 82(2), 279--282 (2012).
3. Lepore, A. and Palumbo, B.: New Insights into the Decisional Use of Process Capability Indices via Hypothesis Testing. Quality and Reliability Engineering International, 31(8), 1725--741 (2015).
4. Montgomery, D.: Introduction to Statistical Quality Control (6th ed). John Wiley & Sons, Hoboken, (2009).
5. Pearn, W.L., Chen, K.S.: Making decision in assessing process capability index Cpk. Quality and Reliability Engineering International, 15(4), 321--326 (1999).

# Industrial Applications of Bayesian Structural Time Series

## *Applicazioni industriali delle serie storiche strutturali bayesiane*

Steven L. Scott

**Abstract** Not every business problem involves time series, but every business has time series problems of one sort or another. Bayesian structural time series models are a flexible and powerful tool for modeling time series data. The models are additive, allowing the analyst to combine latent state components for handling trend, seasonal, regression, and other structural features. Additivity also makes it easy to place informative priors on individual components, like a sparsity-inducing spike and slab prior on a regression component when working with large numbers of contemporaneous predictors. These methods are encoded in the `bsts` R package [Scott(2011)], which was developed at Google to provide Bayesian time series modeling capabilities to non-experts in Bayesian modeling. The package has been used for a variety of purposes, including nowcasting economic time series, anomaly detection, forecasting, and causal inference.

**Abstract** *Non tutti i problemi aziendali hanno a che fare con le serie storiche ma ogni azienda ha qualche tipo di problema legato alle serie storiche. I modelli bayesiani strutturali sono uno strumento flessibile e potente per l'analisi di questo tipo di dati. Questi modelli sono additivi e permettono all'analista di combinare componenti allo stato latente per la modellare stagionalità, trend, regressione e altre caratteristiche strutturali. L'additività rende anche più facile specificare distribuzioni a priori informative sui singoli componenti, come una distribuzione spike-and-slab per indurre sparsità sui coefficienti di regressione quando si lavora con un gran numero di predittori. Questi metodi sono implementati nella libreria R `bsts` [Scott(2011)] sviluppata a Google per fornire uno strumento per l'analisi Bayesiana delle serie storiche a utenti non esperti di modellazione bayesiana. La libreria è stata usata per una serie di scopi tra cui il nowcasting di serie storiche economiche, l'anomaly detection, la previsione e l'inferenza causale.*

**Key words:** time series models, Kalman filter, spike and slab prior, variable selection, big data

---

Steven L. Scott  
Google, USA, e-mail: stevescott@google.com

## 1 Structural Time Series Models

A structural time series model is any model that can be written in the form

$$\begin{aligned} y_t &= Z_t^T \alpha_t + \varepsilon_t \\ \alpha_{t+1} &= T_t \alpha_t + R_t \eta_t, \end{aligned} \quad (1)$$

where  $\varepsilon_t \sim \mathcal{N}(0, H_t)$ ,  $\eta_t \sim \mathcal{N}(0, Q_t)$ , and  $\eta_t$  and  $\varepsilon_t$  are independent of all other quantities. The value  $y_t$  is the only observed data in equation (1). The vector of latent variables  $\alpha_t$  is known as the *state* of the model, and is used to encode the underlying time trend, seasonal pattern, and other desired structure. The remaining symbols in equation (1) are structural parameters. These values of  $T_t$ ,  $Z_t$ , and  $R_t$  may contain statistical parameters, but more often they consist of appropriately placed 0's and 1's. Equation (1) allows the modeler substantial flexibility in defining the state, by concatenating elements from commonly used state models as appropriate for the application at hand.

For example, a common trend model, known as the *local linear trend*, can be written

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t \\ \mu_{t+1} &= \mu_t + \delta_t + \eta_{1t} \\ \delta_{t+1} &= \delta_t + \eta_{2t}. \end{aligned} \quad (2)$$

The state for the local linear trend model is  $\alpha_t = (\mu_t, \delta_t)$ , the transition matrix is

$$T_t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

$R_t = I_2$ , the  $2 \times 2$  identity matrix, and  $Z_t^T = (1, 0)$ . This model captures the current level of the trend in  $\mu_t$ , and the slope of the trend (the extra amount of  $\mu$  as  $t$  increases by 1) in  $\delta_t$ . The trend is a random walk, with a drift term that is also a random walk.

The most commonly used seasonal model (assuming there are  $S$  seasons per cycle) is

$$\begin{aligned} y_t &= \gamma_t + \varepsilon_t \\ \gamma_{t+1} &= - \sum_{s=1}^{S-1} \gamma_{t+1-s} + w_t. \end{aligned} \quad (3)$$

Think of the seasonal model as a regression model where the coefficients of  $S$  seasonal dummy variables evolve over time. Rather than leave out one dummy variable, the regression enforces the constraint that the coefficients must sum to zero (in expectation) over the course of a full cycle. Thus the expected value of any one coefficient is the negative sum of the remaining coefficients.



The state for model (3) is  $\alpha_t = (\gamma_t, \gamma_{t-1}, \dots, \gamma_{t-S+2})$ , which stores the most recent  $S - 1$  seasonal coefficients. The transition matrix is

$$T_t = \begin{pmatrix} -\mathbf{1} \\ I_{S-2} \mathbf{0} \end{pmatrix}. \quad (4)$$

This matrix computes the mean  $\gamma_{t+1}$  and then shifts all current elements of state by one, so that the last one falls off the end. Notice that this model is not full rank. The state is  $S - 1$  dimensional, but only a one-dimensional error term is needed. That is the reason for the matrix  $R_t$ , which in this case is a column vector with a 1 in the first position and 0's everywhere else. It expands the one-dimensional error term  $\eta_t$  into an  $S - 1$  dimensional vector that can be added to  $\alpha_t$ .

A powerful feature of structural time series models is that state components can be combined modularly, by concatenating state vectors, concatenating the corresponding  $Z_t$  vectors, and combining structural matrices  $T_t$ ,  $R_t$ , and  $Q_t$  block-diagonally.

## 2 Bayesian priors and analysis

The parameters of a structural time series model are the residual variance  $H_t$  (often but not always taken to be a constant  $\sigma^2$ ) and any parameters of the state models. For many state models the parameters are simply the variances of the error terms in various forms of random walks. These can be modeled using inverse gamma or inverse Wishart priors.

Posterior inference for many state models would be trivial if the state  $\alpha_t$  were observed at every time point  $t$ . Thus structural time series models are obvious candidates for a data augmentation algorithm that alternates between drawing  $\alpha = (\alpha_1, \dots, \alpha_T)$  from  $p(\alpha|\mathbf{y}, \theta)$  and drawing  $\theta$  from  $p(\theta|\alpha, \mathbf{y})$ . Here  $\mathbf{y} = (y_1, \dots, y_T)$  and  $\theta$  denotes all model parameters. Several papers have been written explaining how to directly simulate from  $p(\alpha|\theta, \mathbf{y})$  using a technique known as “forward filtering backward sampling.” Important early papers include [Carter and Kohn(1994)], [Frühwirth-Schnatter(1994)], and [de Jong and Shepard(1995)]. The `bsts` package uses the technique from [Durbin and Koopman(2002)]. Assuming state models are combined using the concatenation procedure described at the end of Section 1, then state model parameters can be simulated conditionally independently given  $\alpha$  and  $\mathbf{y}$ .

In many applications it can be particularly useful to include a regression component as part of the model state. The `bsts` package offers two mechanisms for handling regression on contemporaneous predictors (which can of course include lags). The first is a dynamic regression, where

$$y_t = Z_t^T \alpha_t + \beta_t^T \mathbf{x}_t + \varepsilon_t. \quad (5)$$

In this model, each coefficient in  $\beta_t^T = (\beta_{t1}, \dots, \beta_{tp})$  obeys a random walk with  $\beta_{t+1,j} = \beta_{tj} + \eta_{tj}$  where  $\eta_{tj} \sim \mathcal{N}(0, \tau_j^2)$ . An alternative is a static regression model

$$y_t = Z_t^T \alpha_t + \beta^T \mathbf{x}_t + \varepsilon_t. \quad (6)$$

Both equations (5) and (6) can be rewritten in the form of equation (1), but it is convenient here to separate the regression from other state components. The equations differ based on whether  $\beta$  is subscripted by  $t$ . The dynamic model is more flexible, but the static model tends to perform better when there are many predictors.

When facing large numbers of predictors, a sparse “spike-and-slab” prior can be placed on the static regression coefficients. Let  $\gamma$  denote a binary vector of the same length as  $\beta$ , where  $\gamma_j = 1$  indicates  $\beta_j \neq 0$  and  $\gamma_j = 0$  indicates  $\beta_j = 0$ . Then the joint prior on  $\beta$  and the constant residual variance parameter  $H_t = \sigma^2$  can be written

$$p(\beta, \gamma, \sigma^2) = p(\gamma)p(\sigma^2|\gamma)p(\beta_\gamma|\gamma, \sigma^2), \quad (7)$$

where  $\beta_\gamma$  refers to the components of  $\beta$  where  $\gamma_j = 1$ .

A convenient choice is to model  $p(\gamma)$  as the product of independent Bernoulli distributions  $p(\gamma) = \prod_j \gamma_j^{\pi_j} (1 - \gamma_j)^{1 - \pi_j}$ . The  $\pi_j$  are prior quantities to be specified by the analyst. The prior can be elicited by asking the analyst for an “expected model size,” which is a guess at the number of nonzero coefficients. If the analyst expects  $k$  important coefficients, then set  $\pi_j = k/p$  where  $p$  is the dimension of  $\mathbf{x}_t$ . Specific values for individual  $\pi_j$  can also be set, for example if there are certain variables the analyst wishes to force in or out of the model.

A conjugate inverse gamma prior is a convenient choice for  $p(\sigma^2|\gamma)$ . The `bsts` package assumes this distribution is independent of  $\gamma$ . Finally, a Zellner prior is assumed for  $p(\beta_\gamma|\gamma, \sigma^2)$ .

From equation (6), notice that  $y - Z_t^T \alpha_t$  is just the equation of an ordinary regression model. Thus conditional on the draw of  $\alpha$ , we can integrate  $\beta_\gamma$  and  $\sigma^2$  from the model and simulate  $p(\gamma|y, \alpha, \theta)$  using a sequence of well understood Gibbs sampling steps (e.g. [George and McCulloch(1997)]). Then, conditional on the draw of  $\gamma$ , we can directly simulate from  $p(\beta_\gamma, \sigma^2|\alpha, y, \theta)$  through conjugacy.

## 3 Applications

### 3.1 Nowcasting economic time series

Structural time series models were used by [Scott and Varian(2014)] and [Scott and Varian(2015)] to make short term forecasts (or “nowcasts”) of economic time series using readily observed predictors that are released more rapidly than official government sanctioned numbers. The predictors in both cases were data from Google trends. Several hundred such predictors were available for any weekly or monthly time series, so the spike-and-slab prior discussed above was helpful.

### 3.2 Causal Modeling

Structural time series models were used by [Brodersen et al(2015)Brodersen, Gallusser, Koehler, R as a method of measuring the impact of a market intervention (e.g. an advertising campaign). The first step is to fit the models to data observed prior to the market intervention. The models are then used to forecast the counterfactual time series of metrics (e.g. sales) that would have occurred had the intervention not taken place. The difference between the observed series and the forecasted counterfactual is the estimated impact. A regression component is a critical feature of this model, because contemporaneous predictors unaffected by the intervention can be used to improve the counterfactual forecast. Examples of such predictors might include a competing firm’s sales, or counts of Google searches for relevant terms. The predictors are available during the forecast period because the “forecast” is done after the conclusion of the market intervention, after which time the relevant predictors are observed.

### 3.3 Long term forecasting

Structural time series models can be used for longer term forecasting, although analysts should think carefully about trend models based on random walks (like the local linear trend). The variance of a random walk increases linearly with the number of time steps into the future, which can lead to unrealistically large forecast errors. Replacing the slope in a local linear trend model with a stationary AR(1) process (centered on a long term global trend  $D$ ) provides additional stability.

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t \\ \mu_{t+1} &= \mu_t + \delta_t + \eta_{0t} \\ \delta_{t+1} &= D + \rho(\delta_t - D) + \eta_{1t} \end{aligned} \tag{8}$$

To illustrate the impact of this assumption, consider a forecast of the future Google stock price based on the historical data shown in Figure 1. There is clear daily volatility in this data, but there is also a clear “up and to the right” trend.

The top panel of Figure 2 shows a forecast 180 days into the future based on the local linear trend model. Notice that the scale of the plot is sufficiently wide to make the variation in the stock price over the last 3 years appear constant. The local linear trend model is extremely flexible. This flexibility is useful for short term forecasts, but for longer term forecasts the local linear trend allows for the possibility of hyper-explosive growth or catastrophic failure far beyond anything observed in the preceding 10-year period.

Contrast the bottom panel of Figure 2, which shows a forecast distribution that matches the empirical volatility of the historical data much more closely. At the end of 2016 the Google stock price was around 790.

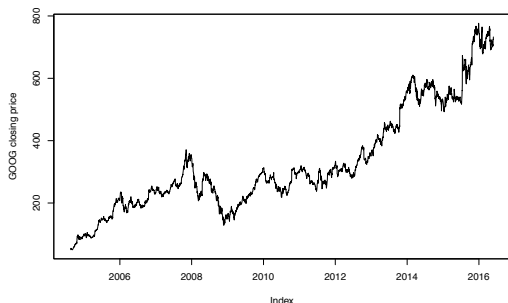


Fig. 1 Google stock price, adjusted for splits.

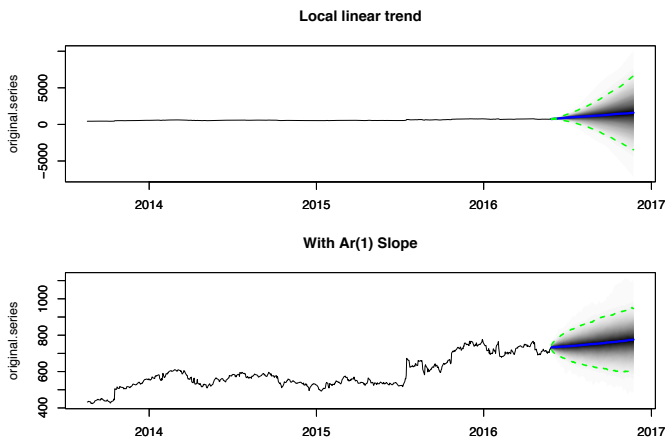


Fig. 2 Forecast distributions of Google stock price based on the local linear trend model (top panel) and the semi-local linear trend model from equation (8).

### 3.4 Handling non-Gaussian errors

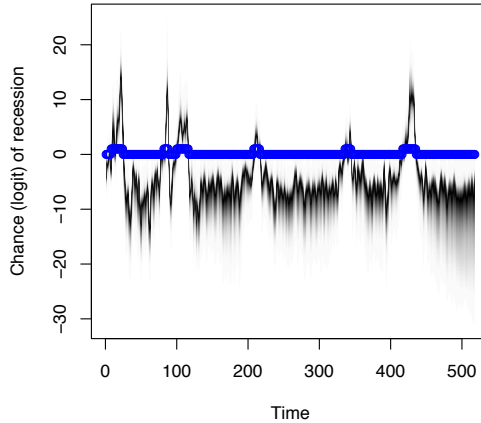
When non-Gaussian error distributions are required, structural time series models can often still be used by introducing a set of latent variables that render the model conditionally Gaussian. Well kWhen non-Gaussian error distributions are required, structural time series mod-els can often still be used by introducing a set of latent variables that render the model conditionally Gaussian. Well known methods exist for probit regression [Albert and Chib(1993)] and models with student T errors [Gelman et al(2014)Gelman, Carlin, Stern, Dunson, Vehtari, and Somewhat more complex methods exist for logistic regression [Frühwirth-Schnatter and Frühwirth(2005)], [Holmes and Held(2006)], [Gramacy and Polson(2012)] and Poisson regression [Frühwirth-Schnatter et al(2008)Frühwirthnown methods exist for probit regression

Additional methods exist for quantile regression, support vector machines, and multinomial logit regression, though they are not yet provided by the `bsts` package.

To see how non-Gaussian errors can be useful, consider the analysis done by [Berge et al(2016)Berge, Sinha, and Smolyansky] who used Bayesian model averaging (BMA) to investigate which of several economic indicators would best predict the presence or absence of a recession. The model they used was a probit regression, which took the presence or absence of a recession (as determined by the NBER recession determinations: <http://www.nber.org/cycles.html>) as a response variable. The model was highly predictive, but it ignored serial dependence in the data. The BMA done by [Berge et al(2016)Berge, Sinha, and Smolyansky] is essentially the same as fitting a logistic regression under a spike-and-slab prior like the one described in Section 2 with all  $\pi_j = 1/2$ . Running that analysis using the `Boom-SpikeSlab` R package [Scott(2010)] (similar to `bsts`, but without the time series) largely replicates their results (up to minor Monte Carlo error).

To capture serial dependence, consider the following dynamic logistic regression model with a *local level* trend.

$$\begin{aligned}\text{logit}(p_t) &= \mu_t + \beta^T \mathbf{x}_t \\ \mu_{t+1} &= \mu_t + \eta_t\end{aligned}\tag{9}$$



**Fig. 3** Distribution of state (on logit scale) for recession data. Blue dots show the true presence or absence of a recession, as determined by official statistics.

Here  $p_t$  is the probability of a recession at time  $t$ , and  $\mathbf{x}_t$  is the set of economic indicators used by [Berge et al(2016)Berge, Sinha, and Smolyansky] in their analysis. The distribution of  $\mu_t$  is plotted in Figure 3, which shows it going to very large values during a recession, and to very small values outside of a recession. This reflects the fact that recessions are rare, but once they occur they tend to persist. Assum-

ing independent time points is therefore unrealistic, and substantially overstates the amount of information available to identify logistic regression coefficients. The posterior distribution of the coefficients in model (9) gives fewer nonzero coefficients than the corresponding analysis assuming independent observations.

## References

- [Albert and Chib(1993)] Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88:669–679
- [Berge et al(2016)Berge, Sinha, and Smolyansky] Berge T, Sinha N, Smolyansky M (2016) Which market indicators best forecast recessions? Tech. rep., US Federal Reserve, <https://www.federalreserve.gov/econresdata/notes/feds-notes/2016/which-market-indicators-best-forecast-recessions-20160802.html>
- [Brodersen et al(2015)Brodersen, Gallusser, Koehler, Remy, and Scott] Brodersen KH, Gallusser F, Koehler J, Remy N, Scott SL (2015) Inferring causal impact using bayesian structural time-series models. *Ann Appl Stat* 9(1):247–274, DOI 10.1214/14-AOAS788, URL <http://dx.doi.org/10.1214/14-AOAS788>
- [Carter and Kohn(1994)] Carter CK, Kohn R (1994) On Gibbs sampling for state space models. *Biometrika* 81(3):541–553
- [Durbin and Koopman(2002)] Durbin J, Koopman SJ (2002) A simple and efficient simulation smoother for state space time series analysis. *Biometrika* 89(3):603–616
- [Frühwirth-Schnatter(1994)] Frühwirth-Schnatter S (1994) Data augmentation and dynamic linear models. *Journal of Time Series Analysis* 15(2):183–202
- [Frühwirth-Schnatter and Frühwirth(2005)] Frühwirth-Schnatter S, Frühwirth R (2005) Auxiliary mixture sampling with applications to logistic models. Tech. rep., IFAS Research Paper Series, Department of Applied Statistics, Johannes Kepler Univeristy Linz.
- [Frühwirth-Schnatter et al(2008)Frühwirth-Schnatter, Frühwirth, Held, and Rue] Frühwirth-Schnatter S, Frühwirth R, Held L, Rue H (2008) Improved auxiliary mixture sampling for hierarchical models of non-gaussian data. *Statistics and Computing* 19(4):479, DOI 10.1007/s11222-008-9109-4, URL <http://dx.doi.org/10.1007/s11222-008-9109-4>
- [Gelman et al(2014)Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014) *Bayesian Data Analysis*, 3rd edn. Chapman & Hall
- [George and McCulloch(1997)] George EI, McCulloch RE (1997) Approaches for Bayesian variable selection. *Statistica Sinica* 7:339–374
- [Gramacy and Polson(2012)] Gramacy RB, Polson NG (2012) Simulation-based regularized logistic regression. *Bayesian Analysis* 7(3):567–590
- [Holmes and Held(2006)] Holmes CC, Held L (2006) Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* 1(1):145–168
- [de Jong and Shepard(1995)] de Jong P, Shepard N (1995) The simulation smoother for time series models. *Biometrika* 82(2):339–350
- [Scott(2010)] Scott SL (2010) BoomSpikeSlab: MCMC for spike and slab regression. R package version 0.4.1
- [Scott(2011)] Scott SL (2011) bsts: Bayesian structural time series. R package version 0.6.2
- [Scott and Varian(2014)] Scott SL, Varian HR (2014) Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation* 5(1/2):4–23
- [Scott and Varian(2015)] Scott SL, Varian HR (2015) Bayesian variable selection for nowcasting economic time series. In: Goldfarb A, Greenstein S, Tucker C (eds) *Economics of Digitization*, NBER Press, London, pp 119 –136, URL <http://www.sims.berkeley.edu/hal/Papers/2012/fat.pdf>

# Asymptotically Efficient Estimation in Measurement Error Models

## *Stima Asintoticamente Efficiente in Modelli con Errori di Misurazione*

Catia Scricciolo

**Abstract** Inference on linear functionals of the latent distribution in measurement error models is considered. The issue about asymptotically efficient estimation by maximum likelihood in a convolution model with Laplace error distribution is settled in the affirmative: maximum likelihood estimators of certain linear functionals of the mixing distribution are  $\sqrt{n}$ -consistent, asymptotically normal and efficient. Asymptotic normality of a Studentized version of the maximum likelihood estimator allows to construct confidence intervals for linear functionals. Regarding maximum likelihood estimation of the mixing distribution as a data-driven choice of the *a priori* distribution on the mixing parameter in an empirical Bayes approach to the problem of estimating the single means, a sequence of estimators can be constructed such that it is asymptotically optimal in a decision-theoretic sense.

**Abstract** È d'interesse fare inferenza su funzionali lineari della distribuzione latente in modelli con errore di misurazione. Il problema della stima asintoticamente efficiente basata sul metodo della massima verosimiglianza in un modello miscuglio con distribuzione di Laplace degli errori è risolto in senso affermativo: gli stimatori di massima verosimiglianza di taluni funzionali lineari sono consistenti, asintoticamente normali ed efficienti. La normalità asintotica di una versione studentizzata dello stimatore di massima verosimiglianza consente di costruire intervalli di confidenza per funzionali lineari. Riguardando la stima di massima verosimiglianza della distribuzione misurante come la scelta guidata dai dati della legge iniziale sul parametro di mistura in un approccio empirico-bayesiano al problema di stima delle singole medie, è possibile costruire una successione di stimatori che sia asintoticamente ottimale in un inquadramento decisionale del problema.

**Key words:** asymptotic efficiency, asymptotic normality, empirical Bayes, Laplace mixture model, maximum likelihood estimate, mixing distribution.

Catia Scricciolo

Dipartimento di Scienze Economiche, Università di Verona, Polo Universitario Santa Marta, Via Cantarane 24, I-37129 Verona (VR), ITALY, e-mail: [catia.scricciolo@univr.it](mailto:catia.scricciolo@univr.it)

## 1 Introduction and Main Results

The problem of asymptotically efficient estimation by the maximum likelihood method of linear functionals in measurement error models is considered. The issue about asymptotic efficiency of the maximum likelihood estimator (MLE) for certain linear functionals in a convolution model with a Laplace error distribution is settled in the affirmative. Under regularity conditions, the MLE is  $\sqrt{n}$ -consistent, asymptotically normal and efficient, even though typically the unknown latent distribution can only be estimated at slower rates. Considered a consistent estimator of the efficient asymptotic variance, a Studentized version of the re-centered MLE also converges to a standard normal distribution. This allows to construct asymptotic confidence intervals for linear functionals and to make inference about them.

### *Model Description*

Let  $X$  be a real-valued random variable (r.v.) with unknown distribution  $P_0$ . Assume that  $P_0$  possesses density  $p_0$  with respect to Lebesgue measure  $\lambda$  on  $\mathbb{R}$ , that is,  $p_0 := dP_0/d\lambda$ . Assume that

$$X = Y + Z, \quad (1)$$

with  $Y$  and  $Z$  independent, unobservable random variables. The distribution  $G_0$  of  $Y$  is unknown, while  $Z$  has Laplace(0, 1) distribution with probability density function

$$f_Z(z) = \frac{1}{2}e^{-|z|}, \quad z \in \mathbb{R}.$$

Then,  $p_0$  is the convolution of  $f_Z$  and  $G_0$  or, in other terms, a location mixture of Laplace densities with mixing distribution  $G_0$  supported on some set  $\mathcal{Y} \subseteq \mathbb{R}$ ,

$$p_0(x) = \int_{\mathcal{Y}} f_Z(x-y) dG_0(y) = \frac{1}{2} \int_{\mathcal{Y}} e^{-|x-y|} dG_0(y), \quad x \in \mathbb{R}.$$

In what follows, we write  $p_{G_0}$  in place of  $p_0$  when we want to stress the dependence of  $p_0$  on  $G_0$ . We observe  $n$  independent and identically distributed (i.i.d.) copies  $X_1, \dots, X_n$  of  $X$  satisfying relationship (1),

$$X_i = Y_i + Z_i, \quad i = 1, \dots, n.$$

We observe the noisy data  $X_1, \dots, X_n$  instead of the uncorrupted r.v.'s  $Y_1, \dots, Y_n$ . The i.i.d. r.v.'s  $Z_1, \dots, Z_n$  represent additive errors and their known (Laplace) distribution is called the *error distribution*. In this classical additive measurement error model we have  $E(X|Y) = Y$  and  $X$  varies around  $Y$ . Interest in this model is motivated by the fact that measurement errors occur in nearly every discipline from medical statistics to astronomy and econometrics.

### *Asymptotic Efficiency of Linear Functionals of the MLE for the Mixing Distribution*

We are interested in estimating linear functionals of  $G_0$  of the form



$$\theta_0 \equiv \theta_{G_0} := \int_{\mathcal{Y}} a(y) dG_0(y) \tag{2}$$

with  $a: \mathcal{Y} \rightarrow \mathbb{R}$  a given function. Linear functionals of  $G_0$  describe different features of the unknown mixing distribution. For example, if, for any fixed  $y_0 \in \mathcal{Y} \subseteq \mathbb{R}$ , the function  $a(y) = 1_{(-\infty, y_0]}(y)$ , then the linear functional  $\theta_0 = P(Y \leq y_0) = G_0(y_0)$  is the cumulative distribution function of  $Y$  evaluated at the point  $y_0$ ; if  $a(y) = y$ , then the linear functional  $\theta_0 = EY$  is the mean of  $G_0$  which, for a zero-mean error density, coincides with the mean  $EX$  of the observations,

$$EX = \int_{\mathbb{R}} xp_{G_0}(x) dx = \int_{\mathcal{Y}} \int_{\mathbb{R}} x f_Z(x - y) dx dG_0(y) = \int_{\mathcal{Y}} y dG_0(y) = EY,$$

with  $a(y) = y = \int_{\mathbb{R}} x f_Z(x - y) dx$ . If, for a fixed real number  $s$  in a neighborhood of zero, the function  $a(y) = e^{sy}$ , then the linear functional  $\theta_0 = \int_{\mathcal{Y}} e^{sy} dG_0(y)$  is the moment generating function (m.g.f.) of  $G_0$  at the point  $s$ , denoted by  $M_Y(s)$ . Relevant aspects of the mixing distribution  $G_0$ , such as the mean and the variance, can be expressed as functionals of the m.g.f.  $M_Y(\cdot)$ . Hence, virtually all results in statistical estimation of characteristics of  $G_0$  can be obtained as by-products of the inference on the m.g.f.  $M_Y(\cdot)$ . In some cases, simple naïve estimators of  $\theta_0$  are available. For example, an estimator of the mean of  $Y$  is the sample mean of the observations  $\bar{X} = \sum_{i=1}^n X_i/n$ ; an estimator of the m.g.f.  $M_Y(s)$  is  $(1 - s^2)^{-1} \sum_{i=1}^n e^{sX_i}/n$  for  $|s| < 1$ , namely, the ratio between the empirical m.g.f. of the observations and the m.g.f. of the error r.v.  $Z$  which is equal to  $(1 - s^2)^{-1}$  for  $|s| < 1$ .

A principled method for estimating linear functionals of  $G_0$  as in (2) is that of estimating  $G_0$  by the MLE  $\hat{G}_n$  and then plugging  $\hat{G}_n$  into the expression of  $\theta_{G_0}$  to obtain the MLE  $\hat{\theta}_n \equiv \theta_{\hat{G}_n}$ . The (nonparametric) MLE  $\hat{G}_n$  of  $G_0$  is a measurable function of the observations  $X_1, \dots, X_n$  taking values in the collection  $\mathcal{G}$  of all probability measures on  $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ , with  $\mathcal{B}(\mathcal{Y})$  the Borel  $\sigma$ -field on  $\mathcal{Y}$ , such that

$$\hat{G}_n \in \arg \max_{G \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \log p_G(X_i) = \arg \max_{G \in \mathcal{G}} \int (\log p_G) d\mathbb{P}_n,$$

where  $p_G(\cdot) := \int_{\mathcal{Y}} f_Z(\cdot - y) dG(y)$  is the location mixture of Laplace densities with mixing distribution  $G \in \mathcal{G}$  and  $\mathbb{P}_n := n^{-1} \sum_{i=1}^n \delta_{X_i}$  is the empirical measure associated with the random sample  $X_1, \dots, X_n$ , namely, the discrete uniform distribution on the sample values that puts mass  $1/n$  on each one of the observations. We assume that the MLE exists, but do not require it to be unique. Lindsay [3] showed that the MLE  $\hat{G}_n$  is a discrete distribution supported on at most  $k \leq n$  support points,  $k$  being the number of distinct observed values or data points. Any linear functional as in (2) can then be estimated by the (plug-in) MLE

$$\hat{\theta}_n \equiv \theta_{\hat{G}_n} = \int_{\mathcal{Y}} a(y) d\hat{G}_n(y).$$

We study the behaviour of  $\hat{\theta}_n$  to answer the question of whether there are functionals of  $G_0$  that can be consistently estimated using the maximum likelihood method at

$\sqrt{n}$ -rate and for which the estimator  $\sqrt{n}(\hat{\theta}_n - \theta_0)$  is *asymptotically normal and efficient*, in the sense that it has asymptotically normal distribution with mean zero and minimum variance. In fact, in the theory of asymptotic efficiency of the MLE, see Chapter 11 in van de Geer [6], there may be linear functionals of  $G_0$  which can be estimated at  $\sqrt{n}$ -rate, even if  $G_0$  itself can only be estimated (at best) at a slower rate, which, in the present case, is  $O_{\mathbf{P}}(n^{-1/5})$  relative to the  $L^1$ -Wasserstein distance, see Dedecker *et al.* [2]. A related issue is the existence of estimators that are empirical means of certain transformations of the observations, like the ones previously considered, which may not be equal to the MLE, but are asymptotically efficient.

To establish asymptotic normality and efficiency of linear functionals of the MLE  $\hat{G}_n$ , we can appeal to Theorem 11.8 of van de Geer [6], pp. 217–220. We preliminarily introduce some more notation. For some  $\sigma_n \downarrow 0$ , let  $\tau_1^2(\sigma_n) := \int_{p_0 \leq \sigma_n} p_0 d\lambda$  and  $\tau_2^2(\sigma_n) := \int_{p_0 > \sigma_n} (1/p_0) d\lambda$ . Here below we state the assumptions we shall be using:

- (A1)  $|\hat{\theta}_n - \theta_0| = o_{\mathbf{P}}(1)$ ,
- (A2) for  $\sigma_n = n^{-3/8} \log^{1/8} n$ , we have  $\tau_1^2(\sigma_n) = O(\sigma_n^2)$  and  $\tau_2^2(\sigma_n) = O(|\log \sigma_n|)$ ,
- (A3) either  $a$  is bounded or  $G_0$  is compactly supported,
- (A4)  $\dot{a}(y) := da(y)/dy$  exists and  $\|\dot{a}\|_{\infty} < +\infty$ ,
- (A5) there exist constants  $0 < c_1, c_2 < +\infty$  such that, for every  $y \in \text{support}(G_0) \subseteq \mathbb{R}$ ,

$$\left| \frac{d(e^{-y}\dot{a}(y))}{dG_0(y)} \right| \leq c_1 \quad \text{and} \quad \left| \frac{d(e^y\dot{a}(y))}{dG_0(y)} \right| \leq c_2,$$

- (A6)  $M_Y(1) := \int_{\mathcal{Y}} e^y dG_0(y) < +\infty$  and  $M_Y(-1) := \int_{\mathcal{Y}} e^{-y} dG_0(y) < +\infty$ .

We now make some remarks on Assumptions (A1)–(A6). Assumption (A1) is satisfied if  $a$  is continuous, which is guaranteed by (A4), and either  $a$  is bounded or  $G_0$  is compactly supported, jointly with  $\hat{G}_n$  weakly converges to  $G_0$  almost surely, in other terms,  $\hat{G}_n$  is strongly consistent at  $G_0$ . Sufficient conditions for strong consistency of the MLE  $\hat{G}_n$  are stated in Theorem 2.3 of Chen [1]. Assumption (A2) proposes the same conditions employed by Scricciolo [5] in Proposition 4 to establish the rate of convergence in the Hellinger metric for the MLE of a Laplace mixture. The result is obtained adopting a convenient approach according to which it is the dimension of the class of kernels and the behaviour of the sampling density  $p_0$  near zero that jointly determine the rate of convergence for the MLE. Assumption (A4) is standard and is used to establish boundedness of subdirections and influence curves. The remaining Assumptions (A5)–(A6) are technical and are used for the same purpose. We are now in a position to state the main result.

**Proposition 1.** *Under Assumptions (A1)–(A6),*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow^{\mathcal{L}} \mathcal{N}(0, \tau_0^2),$$

where  $\tau_0^2$  is the efficient asymptotic variance.

If  $\hat{\tau}_n^2$  is a consistent estimator of the efficient asymptotic variance  $\tau_0^2$ , then the Studentized version of the recentered MLE is asymptotically distributed according to a standard normal,  $\sqrt{n}\hat{\tau}_n^{-1}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$  and asymptotic confidence intervals for  $\theta_0$  can be constructed as well as hypothesis testing be made.

*Asymptotic Optimality of the MLE in the Empirical Bayes Approach to the Decision Problem of Estimating the Single Means*

Maximum likelihood is a principled method for estimating the mixing distribution and leads to asymptotically efficient estimation of certain linear functionals. In a decision-theoretic framework, if the problem of estimating the mixing distribution in a mixture model is regarded as that of selecting the *a priori* distribution on the mixing parameter by a data-driven choice in an empirical Bayes approach to statistical decision problems when, to say it with Robbins [4], “the same decision problem presents itself repeatedly and independently with a fixed but unknown *a priori* distribution of the parameter”, then the MLE has also some optimality property. A description of the elements of the Bayesian decision problem is as follows:

- a parameter space  $\Theta$  with generic state of nature  $\theta$ ,
- an action space  $A$  with generic action  $a$ ,
- a prior distribution  $G$  on  $\Theta$ ,
- a r.v.  $X$  taking values in  $\mathcal{X} \subseteq \mathbb{R}$  such that  $X | (\Theta = \theta) \sim f_\theta(\cdot) := f_Z(\cdot - \theta)$ .

The statistical decision problem consists in choosing a decision function  $t : \mathcal{X} \rightarrow A$  that has conditioned expected loss  $R(t, \theta) := \int_{\mathcal{X}} L(t(x), \theta) f_\theta(x) dx$  when  $\theta$  is the parameter. By averaging over  $\theta$  when the *a priori* distribution is  $G$ , we get the overall expected loss or *Bayes risk*  $R(t, G) = \int_{\Theta} R(t, \theta) dG(\theta)$ .

Consider  $n$  independent repetitions of the same decision problem that give rise to

$$(\theta_1, x_1), (\theta_2, x_2), \dots, (\theta_n, x_n),$$

with  $x^{(n)} := (x_1, \dots, x_n)$  observed and  $\theta_1, \dots, \theta_n$  unknown. When a decision has to be made about  $\theta_{n+1}$ , the observations  $(x_1, \dots, x_{n+1})$  are available, whereas the values  $\theta_1, \dots, \theta_{n+1}$  remain unknown. One can use a function  $t$  of  $x^{(n)}$  evaluated at the point  $x_{n+1}$ , that is,  $t_n(x_{n+1}) \equiv t(x_{n+1}; x^{(n)})$ . An empirical decision procedure is then a sequence  $T = \{t_n\}$  with expected loss

$$R_n(T, G) := \int_{\mathcal{X}} \int_{\mathcal{X}^n} \int_{\Theta} L(t_n(x), \theta) f_\theta(x) dG(\theta) dP_0^n(x^{(n)}) dx.$$

The sequence  $T$  is said to be *asymptotically optimal (a.o.)* relative to every  $G$  in a class  $\mathcal{G}$  if, for every  $G \in \mathcal{G}$ , the expected loss  $R_n(T, G)$  is asymptotically equal to the minimum Bayes risk relative to  $G$ . In symbols,

$$\text{for every } G \in \mathcal{G}, \quad \lim_{n \rightarrow \infty} R_n(T, G) = \inf_t R(t, G).$$

It is known from Robbins [4] that an a.o. sequence  $T$  relative to some class  $\mathcal{G}$  exists if we can find a sequence  $G_n$  of distribution functions that converges in distribution

to  $G$  whichever is  $G \in \tilde{\mathcal{G}}$ . The MLE  $\hat{G}_n$  is one of such sequences if it is strongly consistent. Then, a sequence  $T$  can be constructed following the indications of Corollary 1 in Robbins [4]. Consider a finite action space  $A = \{a_0, a_1, \dots, a_m\}$  and any  $G \in \tilde{\mathcal{G}}$  such that  $\int_{\Theta} L(a_j, \theta) dG(\theta) < +\infty$  for  $j = 0, \dots, m$ . If, for every  $j = 0, \dots, m$  and  $\lambda$ -almost every  $x$ ,  $[L(a_j, \theta) - L(a_0, \theta)]f_{\theta}(x)$  is continuous and bounded as a function of  $\theta$ , then defined  $\Delta_{j,n}(x) := \int_{\Theta} [L(a_j, \theta) - L(a_0, \theta)]f_{\theta}(x) d\hat{G}_n(\theta)$  and

$$t_n(x) := a_k \quad \text{when} \quad \Delta_{k,n}(x) = \min_{0 \leq j \leq m} \Delta_{j,n}(x),$$

the sequence  $T = \{t_n\}$  is a.o. relative to  $G$ .

## 2 Final Remarks

We have considered the problem of asymptotically efficient estimation by the maximum likelihood method of linear functionals of the unknown mixing distribution in a standard additive Laplace measurement error model: the MLE of certain linear functionals is  $\sqrt{n}$ -consistent, asymptotically normal and efficient.

Maximum likelihood estimation of the mixing distribution can also be regarded as the selection of the *a priori* distribution on the mixing parameter by a data-driven choice in an empirical Bayes approach to the problem of estimating the single means. When the MLE  $\hat{G}_n$  is strongly consistent, a sequence of estimators  $T = \{t_n\}$  for the single means can be constructed based on  $\hat{G}_n$  such that it has expected loss asymptotically equal to the minimum Bayes risk. Since, however,  $\hat{G}_n$  is not explicitly known, it would be interesting to investigate when a sequence  $T = \{t_n\}$  can be constructed based on simple naïve efficient estimators  $\hat{\theta}_n$  of  $\theta_0$ .

## References

- [1] Chen, J.: Consistency of the MLE under mixture models. *Statist. Sci.* (*forthcoming*)
- [2] Dedecker, J., Fischer, A., Michel, B.: Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. *Electron. J. Statist.* **9**(1), 234–265 (2015)
- [3] Lindsay, B.G.: The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11**(1), 86–94 (1983)
- [4] Robbins, H.: The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35**(1), 1–20 (1964)
- [5] Scricciolo, C.: Bayes and maximum likelihood for  $L^1$ -Wasserstein deconvolution of Laplace mixtures. Working Paper Series of the Department of Economics, University of Verona, **wp2016n18** (2016)
- [6] van de Geer, S.A.: *Empirical Processes in M-Estimation*. Cambridge University Press, New York (2000)

# On the noisy high-dimensional gene expression data analysis

Angela Serra, Pietro Coretto and Roberto Tagliaferri

**Abstract** The main goal of microarray experiments is to identify, within thousands of genes, groups that show similar co-expression patterns. In most cases the analysis starts from the estimation of a sample correlation matrix used to construct the input dissimilarity. However, the sample correlation matrix is highly distorted by the presence of outlying experimental units, and the typical large ratio between the number of genes and the number of patients. We review the joint action of these issues, and we discuss some possible remedies. We consider real data from some well known microarray experiments, and we perform cluster analysis based on both the usual sample correlation, and some “cleaned” alternatives. Finally, we investigate on the differences between the obtained groups and we draw some conclusions.

**Key words:** Outliers, high-dimensional data, gene expression data, DNA microarrays.

## 1 Introduction

A major role of DNA microarrays is to find genes that behaves similarly across various experimental conditions. This biological co-expression concept translates into the technical notion of statistical similarity. Co-expression can be measured in

---

Angela Serra  
NeuRoNeLab, DISA-MIS, University of Salerno, Via Giovanni Paolo II, 84060 Fisciano, Salerno,  
Italy e-mail: [aserra@unisa.it](mailto:aserra@unisa.it)

Pietro Coretto  
DISES and STATLAB, University of Salerno, Via Giovanni Paolo II, 84060 Fisciano, Salerno,  
Italy, e-mail: [pcoretto@unisa.it](mailto:pcoretto@unisa.it)

Roberto Tagliaferri  
NeuRoNeLab, DISA-MIS, University of Salerno, Via Giovanni Paolo II, 84060 Fisciano, Salerno,  
Italy e-mail: [robttag@unisa.it](mailto:robttag@unisa.it)

several different ways, however the correlation-based similarity plays a crucial role in gene expression data analysis. The main practice is to perform clustering taking correlation-based dissimilarity matrix as input. The groups found by the following cluster analysis are highly dependent on the chosen similarity measure. In the next Section we discuss two major problems with these classical correlation measures: lack of robustness, and excessive variations due to the typical high-dimensional setting.

This paper reviews the impact of these issues in the context of DNA microarrays and possible alternatives already proposed in the literature (see Section 2) are also reviewed. As a major contribution, we investigate what happens to the routinely applied cluster analysis. We show that the obtained partitions differ substantially from those found based on classical correlation-based dissimilarities (see Section 3). Final remarks and an overview of future research projects are discussed in Section 4.

## 2 Issues and cures

In this section we will review some of the major issues in estimating correlation matrices in large scale genomic studies. Cures proposed in the literature are also reviewed.

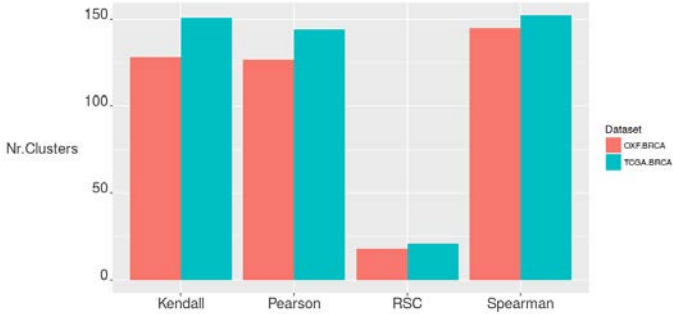
**Excessive noise.** “*Getting the noise out of gene arrays*” is the evocative title of the popular paper by Marshall (2004), where it was explained that microarrays data sampling is terribly noisy, and this undermines the possibility to reach scientific consensus on the empirical evidence. Biologists attribute the “excessive noise” to strong differences in experimental conditions (see Yang et al., 2002) and data acquisition platforms (see Wang et al., 2005). This “excessive noise” is essentially called “data contamination” in classical robust statistic. However, excessive noise is not always strictly related to data contamination. Often non regular data points, also known as outliers, are representative of a minority subpopulation as pointed in Coretto and Hennig (2016). This happens particularly in genomic studies where one cannot expect completely homogeneous populations. Whatever the cause is, it is well known that few outlying data points can completely break down the sampling correlation matrix. Even though this is a well-known problem within the statistical community, few efforts have been made in genetic studies. Hardin et al. (2007) proposed a robust metric based on the Tukey’s biweight statistics, while Bickel (2003) proposed classical rank based dissimilarity measures. Both contributions aim to solve the problem of pairwise correlation estimation. However, it is well known (see Maronna et al., 2006) that pairwise estimation does not necessarily lead to a well behaved correlation matrix estimator. The additional problem here is that DNA microarrays involve thousands of genes, and generally all high break-down covariance/correlation matrix estimators are not well defined in the high-dimensional setting. Serra et al. (2017) developed the  $\mathbf{R}_{\text{MAD}}$ , a robust correlation matrix estimator based on the robust pair-

wise correlation estimator developed in Pasman and Shevlyakov (1987). The  $\mathbf{R}_{\text{MAD}}$  is simple to compute and does not require tunings. Although  $\mathbf{R}_{\text{MAD}}$  performs better than classical estimators, Serra et al. (2017) showed that in high-dimensional setting the advantages of a robust procedure are counterbalanced by the additional issue described below.

**Effects of large concentration ratios.** Estimation of covariance matrices under the high-dimensional regime has been central in recent years. The high-dimensional regime is the situation where  $n =$  the number of sampling units is smaller than  $p =$  the number of genes. A large concentration ratio, that is  $p/n$ , drives the bias of the sampling correlation matrix to unacceptable levels. When  $n \ll p$  the spectral components of the sampling covariance/correlation matrix are dramatically distorted so that most classical dimensional reduction techniques (e.g. the PCA) would fail to recover appropriate and informative data subspaces. Large  $p/n$  also causes the emergence of many spurious correlations, and this has a strong impact in gene network reconstruction. Although in genomic studies there are several methods to filter out the too many small correlations, it seems that it is overlooked that the problem is mainly due to the additional noise introduced by the large  $p/n$ . Concentration ratios larger than  $p/n = 25$  are rather common in gene expression data sets. Since thousands of genes are sampled, it is generally thought that only a relatively small proportion of pairs are actually co-expressed. This translates into a sparsity assumption. Under sparsity assumptions two classes of statistical methods have recently emerged: (i) penalized methods, (ii) thresholding methods. One way to “clean” the effects of a large  $p/n$  is to estimate the precision matrix based on penalized likelihood-type estimators (Yuan and Lin, 2007; DAspremont et al., 2008; Friedman et al., 2008; Rothman et al., 2008; Cai et al., 2011). However, there are drawbacks: (i) these methods do not lead to scalable computations, (ii) they only provide sparse estimates of the inverse covariance matrix, while we are interested in the correlation matrix. Thresholding simply cuts off relatively small covariances/correlations (see Bickel and Levina, 2008; El Karoui, 2008; Rothman et al., 2009; Cai and Liu, 2011). Thresholding estimators are simple to compute and easy to interpret. Bickel and Levina (2008) also proposed random cross-validation to optimally choose the threshold parameter. In the gene expression context the problem has been addressed by Serra et al. (2017), where it is proposed to regularize the  $\mathbf{R}_{\text{MAD}}$  based on cross-validated hard thresholding. The resulting estimator, the Robust Sparse Correlation matrix ( $\mathbf{RSC}$ ) showed remarkable performances in both synthetic and real data. Perhaps the key finding of Serra et al. (2017) is that only jointly treating robustness and high-dimensionality produces the desired improvement.

### 3 Evidence from DNA microarray

Experiments have been performed on two real gene expression data sets related to Breast Cancer. The TCGA.BRCA data set was downloaded from the Cancer

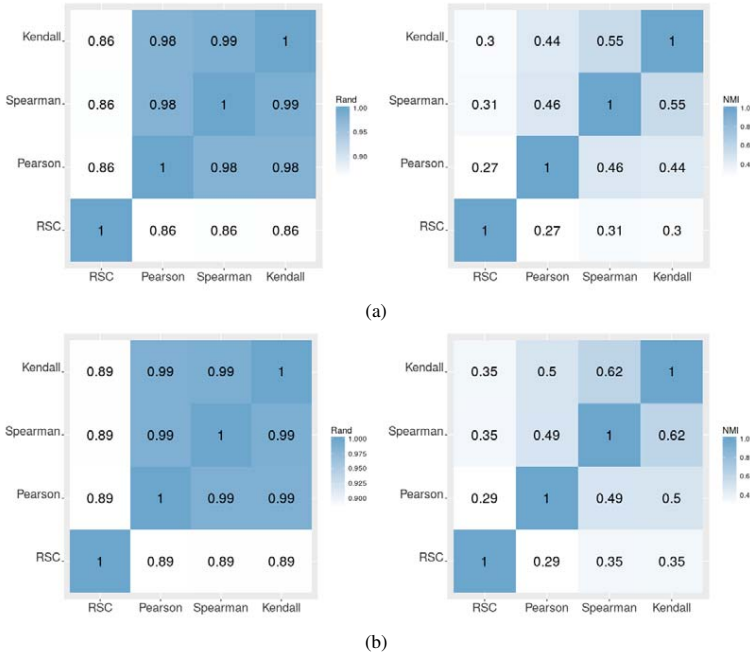


**Fig. 1** Number of clusters obtained with the Kendall, Pearson, RSC, and Spearman correlation-based dissimilarity for both for the Oxford (a) and TCGA (b) data sets.

Genome Atlas (TCGA) (<https://tcga-data.nci.nih.gov/tcga/>). It consists of 151 patients and 4100 genes. The OXF.BRCA data set, described in Buffa et al. (2011) was downloaded from NCBI GEO (<http://www.ncbi.nlm.nih.gov/geo/>) under the accession number GSE22219 and GSE22220. The two data sets were already preprocessed. As a further step, genes with low variance were eliminated and batch effect removal was performed with the `comBat` method available in the `sva` R package of Leek et al. (2011). For each data set correlation matrices have been estimated by using the RSC, Pearson, Spearman, and Kendall methods. These matrices were used as dissimilarity measures, and given as input to the hierarchical clustering method based on the complete linkage algorithm. Once the hierarchy between the genes is obtained, the optimal number of clusters is estimated with the dynamic tree cut approach of Langfelder et al. (2008, 2016). The latter detects the clusters in the dendrogram based on the shape of the branches. This method was proven by Langfelder et al. (2008) to have better performance compared to the classical fixed height dendrogram cut methodology when applied to protein-protein interaction network and gene expression data. In fact, the dynamic tree cut algorithm is able to identify nested clusters, and to identify outliers. Since tiny clusters of genes are to be avoided, the user can set the minimum number of genes needed to create a cluster. In our experiments this number is set to 15.

The differences between the obtained clusters were investigated. Figure 1 shows that the estimated number of clusters is significantly lower when using the RSC-based dissimilarity. The reason for this is that both the contamination and the large concentration ratio of these data sets cause the emergence of a huge number of spurious small correlations. The RSC correlation matrix can filter out these artifacts, so that the resulting number of connected genes is greatly reduced. The strong difference can also be seen in terms of dissimilarity between the optimal partitions. Partition dissimilarity is evaluated based on the Rand Index and the Normalized





**Fig. 2** Heatmaps of the Rand Index and the NMI comparing clusterings obtained based on RSC, Pearson, Spearman and Kendall correlation matrices. Panel (a) reports results for the Oxford data set, panel (b) reports results for the TCGA data set.

**Mutual Information (NMI).**

Let  $X = \{x_1, \dots, x_n\}$  be a set of objects, and consider two partitions to compare, let them be  $Cl_1 = \{X_1^1, \dots, X_r^1\}$ , and  $Cl_2 = \{X_1^2, \dots, X_m^2\}$ . The Rand index is defined as follow:

$$R = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}},$$

where  $a$  is the number of pairs of elements in  $X$  that are in the same subset in  $Cl_1$  and in the same subset in  $Cl_2$ ,  $b$  is the number of pairs of elements in  $X$  that are in different subsets in  $Cl_1$  and in different subsets in  $Cl_2$ ,  $c$  is the number of pairs of elements in  $X$  that are in the same subset in  $Cl_1$  and in different subsets in  $Cl_2$ ,  $d$  is the number of pairs of elements in  $X$  that are in different subsets in  $Cl_1$  and in the same subset in  $Cl_2$ . The Rand Index  $R \in [0, 1]$  with  $R = 0$  meaning no agreement between the partitions, and  $R = 1$  meaning complete agreement between them.

The normalized mutual information (NMI) between the two partitioning  $Cl_1$  and  $Cl_2$  is defined as follow:

$$NMI = \frac{I(Cl_1, Cl_2)}{H(Cl_1) + H(Cl_2)},$$

where  $I(Cl_1, Cl_2)$  is the mutual information between the clusterings  $Cl_1$  and  $Cl_2$ ,  $H(Cl_1)$  and  $H(Cl_2)$  are entropy of  $Cl_1$  and  $Cl_2$  respectively.  $NMI \in [0, 1]$ , where  $NMI = 0$  means that there is no dependence between the two clustering, and  $NMI = 1$  means that they are strongly dependent.

Again, figure 2 shows how the clustering obtained using the RSC matrix is dramatically different from those based on the competing methods. These strong differences may lead to interesting biological conclusions. Here we do not face the challenge to understand which method is better, in fact, this would need deep biological investigations. However, the analysis reported shows that taking into account the sources of the extra noise in an estimation step that is instrumental to the final clustering causes differences of unexpected magnitude.

## 4 Concluding remarks

In this work, the influence of the correlation matrix used as dissimilarity measure in the genes clustering is investigated. A hierarchical clustering algorithm with a dynamic tree cut approach was applied on two real gene expression data sets. It happened that the RSC correlation matrix produced clusters that are different from those obtained based on classical correlation measures. The study confirms that the effects of data contamination, and the high-dimensionality in gene expression data need to be considered carefully. We cannot argue which clustering is better, this needs to be investigated in future researches based on biological validation.

## References

- Bickel, D. R. (2003). Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically. *Bioinformatics* 19(7), 818–824.
- Bickel, P. J. and E. Levina (2008). Covariance regularization by thresholding. *Annals of Statistics* 36(6), 2577–2604.
- Buffa, F. M., C. Camps, L. Winchester, C. E. Snell, H. E. Gee, H. Sheldon, M. Taylor, A. L. Harris, and J. Ragoussis (2011, September). microrna-associated progression pathways and potential therapeutic targets identified by integrated mrna

- and microrna expression profiling in breast cancer. *Cancer research* 71(17), 5635–45.
- Cai, T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* 106(494), 672–684.
- Cai, T., W. Liu, and X. Luo (2011). A constrained  $l_1$  minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* 106(494), 594607.
- Coretto, P. and C. Hennig (2016). Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust gaussian clustering. *Journal of the American Statistical Association* 111(516), 1–12.
- DAspremont, A., O. Banerjee, and L. El Ghaoui (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications* 30(1), 5666.
- El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics* 36(6), 27172756.
- Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3), 432441.
- Hardin, J., A. Mitani, L. Hicks, and B. VanKoten (2007). A robust measure of correlation between two genes on a microarray. *BMC bioinformatics* 8(1), 1.
- Langfelder, P., B. Zhang, and S. Horvath (2008). Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics* 24(5), 719–720.
- Langfelder, P., B. Zhang, and with contributions from Steve Horvath (2016). *dynamicTreeCut: Methods for Detection of Clusters in Hierarchical Clustering Dendrograms*. R package version 1.63-1.
- Leek, J. T., W. E. Johnson, H. S. Parker, E. J. Fertig, A. E. Jaffe, and J. D. Storey (2011). *sva: Surrogate Variable Analysis*. R package version 3.18.0.
- Maronna, R. A., D. R. Martin, and V. J. Yohai (2006). *Robust Statistics: Theory and Methods*. Wiley.
- Marshall, E. (2004). Getting the noise out of gene arrays. *Science* 306(5696), 630–631.
- Pasman, V. and G. Shevlyakov (1987). Robust methods of estimation of correlation-coefficient. *Automation and Remote Control* 48(3), 332–340.
- Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* 2(0), 494515.
- Rothman, A. J., E. Levina, and J. Zhu (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* 104(485), 177186.
- Serra, A., P. Coretto, M. Fratello, and R. Tagliaferri (2017). Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data. The article is currently under peer-review. Preprint available on request (aserra@unisa.it).
- Wang, H., X. He, M. Band, C. Wilson, and L. Liu (2005). A study of inter-lab and inter-platform agreement of dna microarray data. *BMC Genomics* 6(1), 71.

- Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research* 30(4), e15.
- Yuan, M. and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* 94(1), 1935.

# Variable selection for (realistic) stochastic blockmodels

## *Selezione di variabili per modelli stocastici a blocchi*

Mirko Signorelli

**Abstract** Stochastic blockmodels provide a convenient representation of relations between communities of nodes in a network. However, they imply a notion of stochastic equivalence that is often unrealistic for real networks, and they comprise large number of parameters that can make them hardly interpretable. We discuss two extensions of stochastic blockmodels, and a recently proposed variable selection approach based on penalized inference, which allows to infer a sparse reduced graph summarizing relations between communities. We compare this approach with maximum likelihood estimation on two datasets on face-to-face interactions in a French primary school and on bill cosponsorships in the Italian Parliament.

**Abstract** *Sebbene i modelli stocastici a blocchi consentano di rappresentare convenientemente le relazioni fra gruppi di nodi in una rete, essi comportano una nozione di equivalenza stocastica spesso irrealistica in reti reali, e richiedono l'impiego di numerosi parametri che li rendono spesso difficilmente interpretabili. Oggetto di questo lavoro sono due estensioni di modelli stocastici a blocchi, ed un metodo di selezione di variabili fondato sulla penalizzazione della funzione di verosimiglianza che consente di derivare una rappresentazione grafica delle relazioni fra gruppi di nodi. Tale approccio è confrontato con lo stimatore di massima verosimiglianza tramite due applicazioni su una rete di interazioni in una scuola primaria francese e sulla cosponsorizzazione delle proposte di legge nella Camera dei Deputati.*

**Key words:** adaptive lasso; network; penalized inference; reduced graph; stochastic blockmodel; variable selection.

---

Mirko Signorelli

Johann Bernoulli Institute for Mathematics and Computer Sciences, University of Groningen  
Department of Statistical Sciences, University of Padova  
e-mail: signorelli@stat.unipd.it

## 1 Introduction

There is a long tradition in the study of graphs and relational data, whose origins can be arguably traced back to the seminal works of Moreno (1934) and Erdős and Rényi (1959). For decades, however, the study of real networks was limited by the difficulty to collect comprehensive data on large and complex systems. At the turn of the XX century, network science received a sudden boost from many technological advances that have facilitated the collection of relational data in a plentiful of fields. Examples include the advent of high throughput technologies in genetics and of functional magnetic resonance imaging in neuroscience, as well as the development of sensor-based measurements and the diffusion of social media in social network analysis.

The increasing availability of data on real networks has fostered research on their focal properties. These include the famous “small-world property”, encapsulated in the idea of “six degrees of separation” between any two inhabitants of the Earth, and the idea that networks are scale free, i.e., that a few nodes in a network account for most of the connections therein. A further commonly observed feature of real networks is the presence of groups of nodes (“communities”) that are highly connected to each other, and poorly connected to the rest of the network. This *community structure* may be induced by observed attributes of the nodes, or it could be thought as the result of an unknown latent factor. In this paper we focus on two extensions of stochastic blockmodels *a priori*, a class of network models that allow to relate such community structures to observed attributes of the nodes<sup>1</sup>. Although stochastic blockmodels are a convenient way to represent relations between groups of nodes in a network, they require a large number of parameters, which increases quadratically with the number of groups. As a consequence, when a large number of groups is considered, they typically yield cumbersome results that are hard to interpret. Signorelli and Wit (2016) proposed to address this issue by estimating stochastic blockmodels in a penalized likelihood setting. This allows to perform variable selection for stochastic blockmodels, to reduce model complexity and to derive a sparse reduced graph that summarizes the most important interactions within and between communities.

The paper is organized as follows. In Section 2 we discuss how the stochastic blockmodel can be extended so as to incorporate information on the degrees of nodes and on nodal or edge-specific covariates, and how to derive a reduced graph that summarizes relations between communities. Section 3 shortly introduces the variable selection approach proposed by Signorelli and Wit (2016). In Section 4, we illustrate the proposed methodology with two examples on face-to-face contacts in a French high school, and on bill cosponsorship in the Italian Parliament.

---

<sup>1</sup> A related class of blockmodels is that of *a posteriori* stochastic blockmodels (Wasserman and Anderson, 1987), whose aim is the detection of communities rather than the description of relations between known blocks of nodes.

## 2 Representing community structure with realistic stochastic blockmodels

We consider an undirected graph  $\mathcal{G} = (V, E)$ , which features a set of edges  $E \subseteq V \times V$  between a set of nodes or vertices  $V = \{1, \dots, n\}$ . We denote by  $A$  the (symmetric) adjacency matrix of the graph, whose entries  $a_{ij}$  are non-null if and only if an edge between nodes  $i$  and  $j$  is present, and we assume absence of self-loops, i.e.,  $a_{ii} = 0 \forall i \in V$ . Moreover, we distinguish binary graphs, where  $a_{ij} \in \{0, 1\}$ , from edge-valued graphs where  $a_{ij} \in \mathbb{N}$ , and we view each  $a_{ij}$  as a draw from the random variable  $Y_{ij}$ .

### 2.1 Stochastic blockmodel: definition and extensions

A stochastic blockmodel assumes that a partition  $\mathcal{P}$  of  $V$  into  $p$  blocks of nodes  $\{B_1, \dots, B_p\}$  is available. According to the definition proposed by Holland et al. (1983), a network model is a *stochastic blockmodel* if

- the random variables  $Y_{ij}$  are independent;
- $Y_{ij}$  and  $Y_{kl}$  are identically distributed if nodes  $i, k$  belong to the same block  $B_r$ , and nodes  $j, l$  to the same block  $B_s$ .

This definition implies that every node within the same block is *stochastically equivalent*, to wit, that it is possible to swap any two nodes that are members of the same block without affecting the probability distribution of the graph.

The assumption of stochastic equivalence within blocks represents a strong limitation of stochastic blockmodels. For example, it entails that the expected degree of nodes within a block is the same, whereas most real networks feature a strong heterogeneity in the distribution of node degrees. This was noted already by Wang and Wong (1987), who proposed to integrate the stochastic blockmodel with a set of nodal fixed effects. For undirected binary graphs, their *degree-corrected blockmodel* assumes that if  $i \in B_r$  and  $j \in B_s$ , then  $Y_{ij} \sim \text{Bern}(\pi_{ij})$  and

$$\text{logit } \pi_{ij} = \beta_0 + \alpha_i + \alpha_j + \phi_{rs}, \quad (1)$$

subject to the identifiability constraints  $\sum_i \alpha_i = 0$  and  $\sum_s \phi_{rs} = 0 \forall r \in \{1, \dots, p\}$ . Here, a positive block-interaction effect  $\phi_{rs}$  indicates that nodes in blocks  $B_r$  and  $B_s$  tend to interact preferentially with each other. Note that Equation (1) breaks the assumption of stochastic equivalence of nodes within a block and, thus, the model proposed by Wang and Wong (1987) is not a stochastic blockmodel in the sense of Holland et al. (1983). However, it allows a more realistic description of a network with known block-structure: as a matter of fact, it takes into account both nodal information on the popularity or productivity of each node ( $\alpha_i$  and  $\alpha_j$ ), and information on the extent of interaction between pairs of blocks ( $\phi_{rs}$ ).

A further limitation of stochastic blockmodels is that they postulate that the formation of edges depends only on block membership of the nodes. However, often it is reasonable to imagine that other factors besides block membership can affect the process of edge formation. Signorelli and Wit (2016) proposed an extension to stochastic blockmodels that allows the formation of an edge to depend both on block memberships, and on a set of nodal or edge-specific covariates  $x_{ij}$ . They considered the case of an undirected, edge-valued graph and viewed the formation of an edge between  $i \in B_r$  and  $j \in B_s$  as the result of a Poisson process whose rate depends both on blocks  $B_r$  and  $B_s$  and on the covariates  $x_{ij}$ . The resulting network model can be estimated with a generalized linear model where  $Y_{ij} \sim Poi(\mu_{ij})$  and

$$\log \mu_{ij} = \beta_0 + x_{ij}\beta + \gamma_r + \gamma_s + \phi_{rs}, \quad (2)$$

subject to the identifiability conditions  $\sum_r \gamma_r = 0$  and  $\sum_s \phi_{rs} = 0 \forall r \in \{1, \dots, p\}$ . Likewise model (1), also model (2) breaks the assumption of stochastic equivalence within blocks. However, it differs from model (1) in two aspects: it allows to account for factors other than group membership, and it replaces the nodal fixed effects  $\alpha_i$  with block effects  $\gamma_r$ .

## 2.2 How to derive a reduced graph

The focal point of a stochastic blockmodel and of its (more realistic) extensions outlined above is their capacity to summarize a (potentially large) network by making some statements on the relations that exist between the blocks (Anderson, 1992). In particular, stochastic blockmodels make it possible to infer from the observed graph  $\mathcal{G}$  a *reduced graph*  $\mathcal{G}_R = (\mathcal{P}, E_R)$  whose nodes are the blocks.

The reduced graph represents a synthetic way to visualize the relations that exist between blocks in the network. Typically, it is employed to show which blocks interact more with each other. For binary graphs, Anderson (1992) proposed to derive a reduced graph from a stochastic blockmodel by setting a threshold on the predicted interaction probability  $\hat{\pi}_{rs}$  to observe an edge between nodes in blocks  $B_r$  and  $B_s$ . However, the reduced graph obtained with this procedure arbitrarily depends on the choice of the threshold and, furthermore, it might display some blocks as connected to any other block, just because its nodes have, on average, high degrees. Moreover, this procedure does not directly generalize to the case of edge-valued graphs.

To overcome these problems, Signorelli and Wit (2016) derive the reduced graph in a different way, drawing an edge between two blocks  $B_r$  and  $B_s$  if the estimate  $\hat{\phi}_{rs}$  of the corresponding block-interaction parameter  $\phi_{rs}$  is positive. This approach is coherent with the parametrizations employed in models (1) and (2), where a positive  $\phi_{rs}$  entails evidence of attraction between  $B_r$  and  $B_s$ . Thus, the resulting reduced graph will display those pairs of blocks whose nodes tend to interact more with each other. The reduced graphs presented in Section 4 are obtained with this method.



### 3 Variable selection for stochastic blockmodels

The description of relations between pairs of blocks provided by stochastic blockmodels requires the use of a rather large number of parameters. This is necessary in order to model each interaction between blocks  $(B_r, B_s)$ ,  $s \geq r \in \{1, \dots, p\}$ . In particular, model (1) includes  $q_1 = n + p(p-1)/2$  parameters, and model (2)  $q_2 = \dim(\beta) + p(p+1)/2$ . As we will show in Section 4, when many blocks are considered ( $p \geq 10$ ) this often yields reduced graphs with a plentiful of links that are cumbersome to interpret.

In a study on collaborations between Italian political parties, Signorelli and Wit (2016) analysed bill cosponsorship networks in the Chamber of Deputies with model (2) and observed that although positive and negative estimates  $\hat{\phi}_{rs}$  respectively entail collaboration and repulsion between Deputies in parties  $B_r$  and  $B_s$ , it is also possible to imagine a situation of indifference between collaboration and repulsion for some pairs of parties. This indifference directly corresponds to  $\phi_{rs} = 0$  in models (1) and (2). However, with maximum likelihood estimation it is highly unlikely that any of the point estimates  $\hat{\phi}_{rs}$  will be exactly zero. For this reason, they advocated the penalization of the block-interaction terms  $\phi_{rs}$  (as well as of the covariate vector  $\beta$  in Equation (2)) and employed the adaptive lasso (Zou, 2006) to estimate their model.

This penalized inference approach yields two advantages: on the one hand, it allows to distinguish situations of indifference between blocks from collaborations or repulsions; on the other hand, it is capable to reduce the complexity of the inferred model by shrinking some of its parameters to 0. As a result, it enables to infer a sparse reduced graph, which is typically easier to interpret than the one based on maximum likelihood estimation.

We remark that because of the identifiability conditions that ought to be imposed in models (1) and (2),  $p$  block-interaction parameters do not directly appear in the models and, thus, they cannot be penalized. Given that the parameters for interactions within each block,  $\phi_{rr}$ , are anyway likely to be positive, we substitute  $\phi_{rr} = -\sum_{s \neq r} \phi_{rs}$  for every  $r \in \{1, \dots, p\}$  in (1) and (2). By doing so, we penalize each block-interaction parameter  $\phi_{rs}$  ( $r \neq s$ ), and derive each  $\phi_{rr}$  from the constraints.

In the next Section we consider two examples of penalized inference for stochastic blockmodels, and carry out a comparison of this approach with the one based on maximum likelihood.

## 4 Applications

### 4.1 Face-to-face contacts in a French primary school

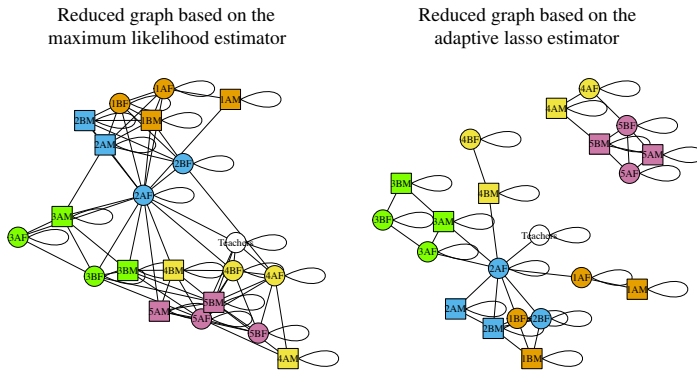
We consider data on face-to-face interactions in a French primary school collected by Stehlé et al. (2011). The study, which lasted 2 days, employed sensors to detect

face-to-face interactions between students and teachers that lasted at least 20 seconds. Here, we focus on the interactions measured in the first day and consider a binary graph whose nodes are students and teachers, and where an edge between two nodes indicates that at least an interaction between them was recorded during the day.

The school comprises 10 classes (2 for each level). The available information for each node is its status (student or teacher); furthermore, for students also class and gender are known. Thus, we partition the nodes into 21 blocks: 20 blocks partition students according to their class and gender, and the last one contains teachers.

We employ model (1) to study the pattern of interactions among the blocks, and compare the reduced graphs that can be derived by employing maximum likelihood, and the penalized likelihood estimation procedure described in Section 3.

Maximum likelihood estimation results into 86 positive, and 145 negative, estimates of the block-interaction parameters. As a result, the reduced graph in Figure 2 displays a large number of interactions between the blocks. Penalized likelihood estimation, instead, shrinks 88 block-interaction parameters to 0, resulting into 52 positive and 91 negative parameter estimate  $\hat{\phi}_{rs}$ . A direct consequence of this is that the reduced graph displaying interactions between blocks is now more readable. In particular, the presence of self-loops indicates that members within each block interact frequently with their peers. Furthermore, a link is present between male and female students within each class. Whereas students in their fifth grade also interact across classes in their same grade (5A and 5B) irrespective of gender, the pattern of interaction between the two third grade classes (3A and 3B) seems to be affected



**Fig. 1** Comparison of reduced graphs based on maximum likelihood and penalized likelihood inference, displaying interactions between groups of students (and teachers) in a French primary school. Node colors denote grades and their shapes distinguish female (circle) from male (square) students. The label of each block indicates the grade (1-5), the section (A or B) and the gender (F or M) of students. The white circular node indicates the block of teachers.

by gender identity (males in 3A interact with males in 3B, and females in 3A with females in 3B). Instead, we do not find any interaction between first, or fourth grade classes (1A-1B and 4A-4B, respectively).

### 4.2 Bill cosponsorship in the Italian Parliament

Signorelli and Wit (2016) employed data on bill cosponsorships to reconstruct the pattern of collaborations between Italian political parties in the Chamber of Deputies from 2001 to 2015. Here we focus our attention on the bill cosponsorship network for the first part of the XVII legislature (2013 - 2015) and make a comparison between maximum likelihood and penalized likelihood inference.

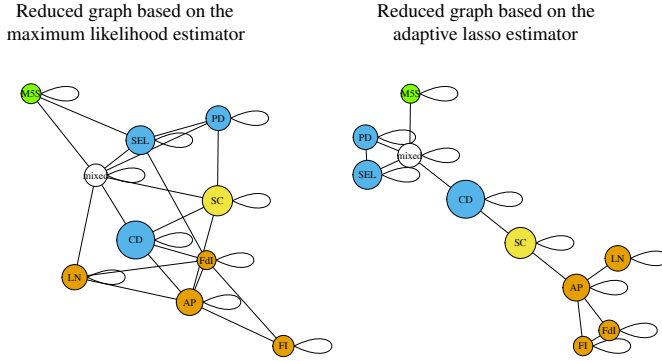
We define a bill cosponsorship network where a weighted undirected edge is present between two deputies if they have cosponsored together at least one bill. Edge weights represent the number of bills that each pair of deputies has cosponsored. During the XVII legislature, 10 parliamentary groups are represented in the Chamber. Those groups form the blocks in model (2), where we furthermore consider covariates for gender, age and seniority of deputies, besides a dummy variable that indicates whether two deputies have been elected in the same electoral constituency. In the penalized model, we penalize each of the covariates and the block-interaction terms, and we employ the adaptive lasso for estimation.

Table 1 compares the results for the intercept  $\beta_0$  and the parameter vector  $\beta$ . Here, the only (slight) difference is that the parameter for age difference is shrunk to 0 with the adaptive lasso. The other variables indicate that female and senior deputies are more active in cosponsorships, and that geographic proximity also increases the tendency to collaborate.

The main difference between the two approaches lies in the estimation of the block-interaction parameters  $\phi_{rs}$ . Maximum likelihood yields 29 positive, and 26 negative, estimates of the block-interaction parameters; the adaptive lasso, instead, shrinks 16 of those parameters to 0, resulting into 21 positive, 16 null and 18 neg-

**Table 1** Comparison of maximum likelihood and adaptive lasso estimators for the parameter vector  $\beta$  in model (2). The reference modes are interactions between two male deputies (male-male) for gender effects, and between two junior deputies (junior-junior) for seniority.

Covariate	Maximum likelihood estimate	Adaptive lasso estimate
Intercept ( $\beta_0$ )	-3.83	-3.86
Female-male interaction	0.233	0.210
Female-female interaction	0.659	0.634
Same electoral constituency	0.550	0.554
Age difference	-0.011	0
Junior-senior interaction	0.253	0.234
Senior-senior interaction	0.700	0.712



**Fig. 2** Comparison of reduced graphs based on maximum likelihood and penalized likelihood inference, displaying collaborations between Italian political parties. Node size is proportional to group productivity. The colour of nodes is lightblue for left-wing parties, orange for right-wing ones, yellow for “Scelta Civica”, green for “Movimento 5 Stelle” and white for the mixed group.

ative estimates. Once more, the reduced graph of collaborations based on maximum likelihood is rather cumbersome to interpret, whereas the one based on the adaptive lasso is more readable. In particular, the latter points out collaborations within each party, between the 4 right-wing parties (orange), between three parties (‘Centro Democratico’, ‘Scelta Civica’ and ‘Area Popolare’) that belong to different coalitions, between the two main left-wing parties, and that deputies in the ‘mixed group’ tend to collaborate with left-wing parties and with the ‘Movimento 5 Stelle’.

## References

- Anderson, C. J., Wasserman, S., Faust, K.: Building stochastic blockmodels. *Soc. Netw.*, **14**(1), 137-161 (1992).
- Erdős, P., Rényi, A.: On random graphs I. *Publ. Math. (Debr.)*, **6**, 290-297 (1959).
- Holland, P. W., Laskey, K. B., Leinhardt, S.: Stochastic blockmodels: First steps. *Soc. Netw.*, **5**(2), 109-137 (1983).
- Moreno, J. L.: Who shall survive? Nervous and mental disease monograph series, **58** (1934).
- Signorelli, M., Wit, E. C.: A penalized inference approach to stochastic blockmodelling of community structure in the Italian Parliament. arXiv:1607.08743 (2016).
- Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., et al.: High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS one*, **6**(8), e23176 (2011).
- Wang, Y. J., Wong, G. Y.: Stochastic blockmodels for directed graphs. *J. Am. Stat. Assoc.*, **82**(397), 8-19 (1987).
- Wasserman, S., Anderson, C.: Stochastic posteriori blockmodels: construction and assessment. *Soc. Netw.*, **9**, 1-36.
- Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.*, **101**(476), 1418-1429 (2006).

# Detection of spatio-temporal local structure on seismic data

## *Individuazione di strutture locali spazio-temporali su dati sismici*

Marianna Siino, Francisco J. Rodríguez-Cortés, Jorge Mateu and Giada Adelfio

**Abstract** For the description of the seismicity of an area, the comparison between local features of background and induced events could be a new perspective of research. In spatio-temporal point process, local second-order statistics provide information on the relationships of each event and its nearby events. In this paper, we use a test based on local indicators of spatio-temporal association (LISTA functions) for identifying different local structures comparing the two previous sets of events. We present a simulation study on the test and show the main results of the application on Greece earthquake data.

**Abstract** *In questo lavoro si propone una nuova prospettiva di analisi per la descrizione della sismicità di un'area considerando il confronto delle caratteristiche locali degli eventi di fondo e quelli indotti. Nell'ambito dell'analisi dei processi puntuali di tipo spazio-temporale, le statistiche del secondo ordine locali descrivono le relazioni esistenti tra ciascun evento e i suoi più vicini. Per identificare differenze tra i due insiemi di eventi individuati in precedenza, utilizziamo un test basato sugli indicatori locali di associazione spazio-temporale, chiamati funzioni LISTA. Presentiamo uno studio di simulazione e i principali risultati applicando la metodologia sui dati sismici della Grecia.*

**Key words:** earthquakes; local indicators of spatio-temporal association; second-order product density function

---

Marianna Siino

Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Palermo, Italy

Francisco J. Rodríguez-Cortés

Department of Mathematics, Universitat Jaume I, Castellón, Spain

Jorge Mateu

Department of Mathematics, Universitat Jaume I, Castellón, Spain,

Giada Adelfio

Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Palermo, Italy, e-mail: giada.adelfio@unipa.it

## 1 Introduction

In an observed area, earthquake events can be considered as a realization of a marked space-time point process, where the magnitude is the mark, and a point is identified by its geographical coordinates and time of occurrence (Illian et al, 2008). Generally, the description of seismic events requires the definition of more complex models than stationary Poisson process since clustering structure characterises these events. Therefore, spatio-temporal cluster analysis has a relevant role in the comprehension of seismic processes.

Commonly, global spatio-temporal second-order summary statistics (such as the  $K$ - and pair-correlation functions) are used to detect deviations from the Poisson assumption (Gabriel and Diggle, 2009). These tools play a fundamental role in the phase of descriptive analysis, in model validation and for testing procedures giving global information of a given point pattern. An interesting question may concern if the same conclusions are valid locally, and thus, for example, in testing procedures if in subregions of the spatio-temporal window the pattern behaves differently identifying specific regions where the null hypothesis is not accepted.

Anselin (1995) proposed the idea of considering individual contributions of a global estimator as a measure of clustering under the name of Local Indicators of Spatial Association (LISA). In spatial point processes, Cressie and Collins (2001) propose a local product density function developing theoretical properties, namely first- and second-order moments, of these functions. Some applications of LISA functions are in Mateu et al (2007) and Moraga and Montes (2011). Rodríguez-Cortés (2014) and Siino et al (2016b) extend the concept of LISA function to the spatio-temporal point pattern context defining the LISTA functions. A brief summary on this methodology is in Section 2. Moreover, Siino et al (2016b) develop a testing procedure for the local structure comparing spatio-temporal point patterns based on Local indicators of spatio-temporal association (LISTA) functions described in Section 3. A simulation study is performed to illustrate that the test proposed has the prescribed size (Section 4). For the analysis in Section 5, we consider earthquakes occurred in the Hellenic area between 2005 and 2014. We aim to detect which triggered events have a significant different local cluster structure with respect to the underlying process, represented by the background events, linking the results with the geological information available in the study area.

## 2 Methodology

We consider a spatio-temporal point process with no multiple points as a random countable subset  $\mathcal{X}$  of  $\mathbb{R}^2 \times \mathbb{R}$ , where for a point  $(\mathbf{u}, t) \in \mathcal{X}$ ,  $\mathbf{u} \in \mathbb{R}^2$  is the spatial location and  $t \in \mathbb{R}$  is the time of occurrence. In practice, an observed spatio-temporal pattern is a finite set  $\{(\mathbf{u}_i, t_i)\}_{i=1}^n$  of distinct points within a bounded spatio-temporal region  $W \times T \subset \mathbb{R}^2 \times \mathbb{R}$ , where usually  $W$  is a polygon with area  $|W| > 0$  and  $T$  a single closed interval with length  $|T| > 0$ . Considering a bounded spatio-temporal

region  $A \subset W \times T$ ,  $Y(A)$  denotes the number of the events of the process falling in  $A$ . The intensity of a process is defined as (Diggle, 2013)

$$\rho(\mathbf{u}, t) = \lim_{|\mathbf{du} \times dt| \rightarrow 0} \frac{\mathbb{E}[Y(\mathbf{du} \times dt)]}{|\mathbf{du} \times dt|}$$

where  $\mathbf{du} \times dt$  is a spatio-temporal region around the point  $(\mathbf{u}, t)$ ,  $|\mathbf{du}|$  is the area of the spatial region,  $|dt|$  is the length of the time interval and,  $\mathbb{E}(Y(\mathbf{du}, dt))$  denotes the expected number of events in the infinitesimal spatio-temporal region. The process is called homogeneous or stationary when the intensity is constant,  $\rho(\mathbf{u}, t) = \rho$  for all  $(\mathbf{u}, t) \in W \times T$ .

When the interest is in describing the spatio-temporal variability and correlations between points of a pattern, we have to consider second-order measures, such as the product density  $\rho^{(2)}(\cdot, \cdot)$ . This quantity provides an interpretable measure of the spatio-temporal dependence structure and it is defined as

$$\rho^{(2)}((\mathbf{u}_i, t_i); (\mathbf{u}_j, t_j)) = \lim_{|\mathbf{du}_i \times dt_i| |\mathbf{du}_j \times dt_j| \rightarrow 0} \frac{\mathbb{E}[Y(\mathbf{du}_i \times dt_i) Y(\mathbf{du}_j \times dt_j)]}{|\mathbf{du}_i \times dt_i| |\mathbf{du}_j \times dt_j|} \quad (1)$$

where  $\mathbf{du}_i \times dt_i$  and  $\mathbf{du}_j \times dt_j$  are small cylinders around two distinct points  $(\mathbf{u}_i, t_i)$  and  $(\mathbf{u}_j, t_j)$ .

Under the stationary case, and ignoring edge-effects, a global naive non-parametric kernel estimator for  $\rho^{(2)}(r, h)$  in (1) (Rodríguez-Cortés, 2014) is given by

$$\widehat{\rho}_{\varepsilon, \delta}^{(2)}(r, h) = \frac{1}{4\pi r |B|} \sum_{i=1}^n \sum_{j \neq i} \kappa_{\varepsilon, \delta}(\|\mathbf{u}_i - \mathbf{u}_j\| - r, |t_i - t_j| - h), \quad (2)$$

where the sum is over all pairs  $(\mathbf{u}_i, t_i) \neq (\mathbf{u}_j, t_j)$  of the data points,  $B = W \times T$ ,  $r > \varepsilon > 0$  and  $h > \delta > 0$ . The kernel function  $\kappa$  has a multiplicative form  $\kappa_{\varepsilon, \delta}(\|\mathbf{u}_i - \mathbf{u}_j\| - r, |t_i - t_j| - h) = \kappa_{1\varepsilon}(\|\mathbf{u}_i - \mathbf{u}_j\| - r) \kappa_{2\delta}(|t_i - t_j| - h)$ , where  $\kappa_{1\varepsilon}$  and  $\kappa_{2\delta}$  are kernel functions with bandwidths  $\varepsilon$  and  $\delta$ , respectively. For an approximately unbiased edge-corrected estimator for the spatio-temporal product density see Rodríguez-Cortés (2014). The R package `stpp` (Gabriel et al, 2013) implements the main code for the computation of the estimator in (2).

Considering the spatio-temporal product density in (1), its local version is denoted by  $\rho^{(2)i}(\cdot, \cdot)$ . Rodríguez-Cortés (2014) extends the operational definition of local indicator introduced by Anselin (1995), for fixed  $r$  and  $h$ , it holds that

$$\widehat{\rho}_{\varepsilon, \delta}^{(2)}(r, h) = \frac{1}{n-1} \sum_{i=1}^n \widehat{\rho}^{(2)i}_{\varepsilon, \delta}(r, h), \quad (3)$$

An unbiased edge-corrected kernel-based estimator for  $\widehat{\rho}^{(2)i}(r, h)$  is given by

$$\widehat{\rho}^{(2)i}_{\varepsilon, \delta}(r, h) = \frac{n-1}{4\pi r |B|} \sum_{j \neq i} \frac{\kappa_{\varepsilon, \delta}(\|\mathbf{u}_i - \mathbf{u}_j\| - r, |t_i - t_j| - h)}{w(\mathbf{u}_i, \mathbf{u}_j) w(t_i, t_j)}, \quad (4)$$

with  $r > \varepsilon > 0$ ,  $h > \delta > 0$ , for  $(\mathbf{u}_i, t_i) \in W \times T$ ,  $i = 1, \dots, n$ ,  $w(\mathbf{u}_i, \mathbf{u}_j)$  and  $w(t_i, t_j)$  are the edge-effect factors. For formal theoretical details on the LISTA functions see Rodríguez-Cortés (2014) and for the implementation of the function `LISTAfunc` in the GitHub repository `pdLISTA` (Rodríguez-Cortés, 2016).

### 3 Testing procedure

The test procedure is an extension of the test proposed in Moraga and Montes (2011) into the spatio-temporal context. It detects differences in the local structure of two given point spatio-temporal patterns  $X$  and  $Z$ . We test the null hypothesis of no difference in the spatio-temporal local structure of  $X$  and  $Z$  with respect to the  $i$ -th point  $(\mathbf{u}_i, t_i) \in X$ , where the number of points in the two patterns are respectively  $N(X) = n$  and  $N(Z) = m$ . The steps of the testing procedure are the following.

1. For each point  $(\mathbf{u}_i, t_i) \in X$ , for  $i = 1, \dots, n$  the LISTA function  $\widehat{\rho}^{(2)i}_{\varepsilon, \delta}(r, h)$  is estimated.
2. Secondly, for each fixed point  $(\mathbf{u}_i, t_i) \in X$ ,  $k$  point patterns are generated under the null hypothesis. For each fixed point  $(\mathbf{u}_i, t_i)$ ,  $k$  local spatio-temporal product density surfaces are estimated,  $\widehat{\rho}^{(2)iq}_{\varepsilon, \delta}(r, h)$  for  $q = 1, \dots, k$ . They are summarised in terms of the average surface, denoted by  $\bar{\rho}_{H_0}^i(r, h)$ .
3. Based on the previous quantities, the following statistic is considered

$$T^i = \int_0^{h_0} \int_0^{r_0} \left( \widehat{\rho}^{(2)i}_{\varepsilon, \delta}(r, h) - \bar{\rho}_{H_0}^i(r, h) \right)^2 dr dh, \quad (5)$$

where  $r_0$  and  $h_0$  are chosen using the Diggle's rule (Diggle, 2013).

4. The theoretical distribution of our statistics under the null hypothesis is not known, so we rely on simulation-based empirical distributions. Fixing a point  $(\mathbf{u}_i, t_i) \in X$ , the estimated value of the statistic, is compared with the empirical distribution of the  $k$  values of  $T_{H_0}^{iq}$  with  $q = 1, \dots, k$  that are obtained computing the test between the  $q$ -th generated LISTA surfaces under the null hypothesis and their sample mean function. The  $p$ -value of  $T^i$  is the following ratio  $p^i = \sum_{q=1}^k \mathbf{I}(T_{H_0}^{iq} \geq T^i) / k$ . The null hypothesis is rejected if  $p^i \leq \alpha$ , where  $\alpha$  is the type I error.

### 4 Simulation study

A simulation study with some scenarios is carried out to assess the performances in terms of type I error of the test introduced in the previous section.



The patterns are generated in the unit cube,  $W \times T = [0, 1]^2 \times [0, 1]$  and varying the type of process (Poisson, Poisson cluster). We also consider  $\mathbb{E}[N(W \times T)] = n + m = \{150, 300\}$  for  $X \cup Z$ . Under the null hypothesis, a pattern is generated with expected number of points equal to  $n + m$ , and the points are randomly associated to the pattern  $X$  or  $Z$  such that the number of points for the two sets is equal, and the test is computed for all the points belonging in  $X$ . For each point, the number of permutations is equal to  $k = 99$ .

The spatio-temporal Poisson point patterns are generated using the function `rpp` in the package `stpp` of **R**. The Poisson cluster processes are simulated using `rpcp` in `stpp`. Given the values of  $n + m$  and the dispersion parameters, we control the degree of clustering by changing the expected number of parents ( $n_p = \{5, 10\}$ ) and the number of offspring points with respect to each parent.

For each of the resulting scenarios under the null hypothesis, 100 pairs of patterns of  $(X, Z)$  are generated, and the type I error probability is defined as the proportion of points belonging to  $X$  for which the null hypothesis is rejected considering a fixed nominal value of  $\alpha$ . Table 1 presents the average and the variance of the  $p$ -values under  $H_0$  with the rejection rate for  $\alpha = 0.05$ . The statistical test exhibits acceptable empirical rejection rates for the several scenarios. There are no remarkable differences in the results when changing the intensity, the type of the process and the degree of clustering.

Scenarios		Dispersion	$n + m$	$n_p$	$\varepsilon$	$\delta$	$T_1$			
$X$	$Z$	parameters					Rej.	Mean	Var	
$P$	$P$	-	150	-	0.134	0.080	0.057	0.490	0.087	
			300	-	0.111	0.069	0.055	0.487	0.086	
$PC$	$PC$	$\{h, r\} = (0.26; 0.13)$	150	5	0.094	0.069	0.053	0.500	0.086	
				15	0.114	0.074	0.048	0.493	0.084	
				300	5	0.082	0.061	0.047	0.500	0.085
				15	0.095	0.065	0.047	0.511	0.085	

**Table 1:** Rejection rates (Rej.) at  $\alpha = 0.05$ , the mean and the variance (Var.) of the  $p$ -values for the statistic. The spatio-temporal models considered are homogenous Poisson point processes ( $P$ ) and Poisson cluster point processes ( $PC$ ). Dispersion parameters for the  $PC$  model are given in the table,  $n + m$  is the total expected number of points for  $X \cup Z$ , and  $n_p$  is the expected number of parents for the  $PC$  model. For each scenario, 100 simulations are considered,  $\varepsilon$  and  $\delta$  are the bandwidths in space and time, respectively.

## 5 Application

In the seismological context, a background event refers to an earthquake that has not been triggered by another and that might be related to changes in the tectonic field. On the other hand, triggered events are thought to have been caused by a pre-

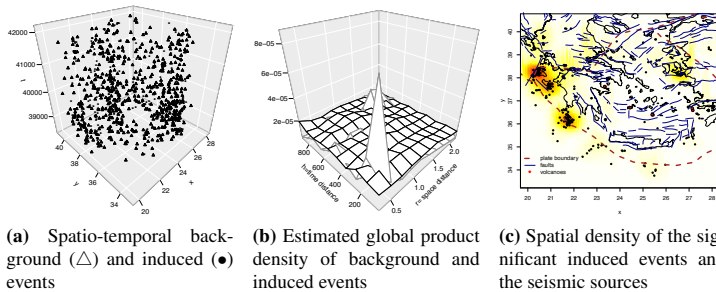
vious earthquake. Globally, these two set of events present different spatio-temporal global interaction structure, however it can be of interest to compare them focusing on a local scale.

In this application, we consider earthquake events occurred in the Greek area between 2005 and 2014 with a magnitude greater than 4 (Figure 1a), for a total number of 1105 events. Its complex spatial multiscale structure has been analysed in Siino et al (2016a).

The earthquakes are classified into background and induced events using a declustering procedure: a probability of being independent events is assigned to each one and it comes from an algorithm for the estimation of Epidemic Type Aftershocks-Sequences (ETAS) model (Ogata, 1988). We fitted the model using the R package `etasFLP` (Chiodi and Adelfio, 2014) based on the method developed in Adelfio and Chiodi (2015). We use the final probabilities provided in the last step of the iteration procedure to classify the events with a magnitude greater than 4 into the two groups, obtaining 580 background events and 525 triggered events (Figure 1a).

Considering the two clusters of independent and induced earthquakes, we would answer the following research questions: Is there a different global structure between the two point patterns? Which triggered events have a significant different local structure with respect to the underlying process (background events)? Is there any geological justification for the identified clusters?

As expected, the estimated spatio-temporal product density of the spontaneous events does not show any particular behaviour. On the other hand, for the induced seismicity, there is a spatio-temporal clustering around at  $t < 300$  days and  $r < 65$  kilometres, in terms of temporal and spatial distances, respectively (Figure 1b). However, we further aim to detect if we can obtain different conclusions focusing on a local scale detecting the spatio-temporal clusters.



**Fig. 1:** (1a) Scatterplot of the spatio-temporal earthquake data classified in background events (triangles) and induced events (points) according to the procedure of declustering using the ETAS model. (1b) Estimated global product density for the background events (black surface), and induced ones (grey surface) with bandwidths  $\varepsilon = 30.44$  km and  $\delta = 44.14$  days. (1c) Image plot in space of the significant induced events and seismogenetic sources.

We apply the testing procedure of Section 3. The point pattern  $Z$  is represented by the background events and the events in  $X$  are the triggered ones. The representation of the results of the significant points in space allows to interpret them in relation to the geological information available in the study area (Figure 1c). We can identify some areas in which the induced events (with a magnitude greater than 4) are different in terms of spatio-temporal local structure than the background seismicity: islands of Kefalonia and Zakynthos and the Samos area (East Aegean Sea). The different behaviour is due to their specific geological characteristics and, in particular, to a higher fracturation degree of their seismogenetic volumes. These results confirm our idea that the observed seismicity is generated by a complex model, characterised by spatial-temporal interaction, with events happening at several scales, and with spatial inhomogeneity related to the geological information available in the study area.

## 6 Final remarks

We deal with a non-parametric testing approach for spatio-temporal point processes, in order to compare the local structure of two spatio-temporal point patterns (say  $X$  and  $Z$ ). The used statistic leads to approximately valid test and the results in terms of type I error are reasonably good. Using the aforementioned test, we compare background and induced seismicity with a magnitude greater than 4 in the Greek area. It seems that the sequences of events that are strongly different to the underlying process, are placed in specific regions of the study window.

As a possible future development, we may consider further simulation scenarios to assess the power of the test. Moreover, it could be interesting to define other local tests based on the LISTA surfaces, changing what is postulated under the null hypothesis. With the analysis of the LISTA surfaces for a given point pattern, we can explore how individual points are related to their neighbouring events, clustering surfaces in order to classify points with similar spatio-temporal local structure. Moreover, we could develop a diagnostic tool based on the LISTA functions computing a weighted version of them by the inverse of the intensity function, looking for points with a more relevant contribution to the global summary statistics.

## Acknowledgments

This paper has been partially supported by the national grant of the Italian Ministry of Education University and Research (MIUR) for the PRIN-2015 program (Progetti di ricerca di Rilevante Interesse Nazionale), “Prot. 20157PRZC4 - Research Project Title Complex space-time modeling and functional analysis for probabilistic forecast of seismic events. PI: Giada Adelfio”

## References

- Adelfio G, Chiodi M (2015) Alternated estimation in semi-parametric space-time branching-type point processes with application to seismic catalogs. *Stochastic Environmental Research and Risk Assessment* 29(2):443–450
- Anselin L (1995) Local indicators of spatial association-lisa. *Geographical analysis* 27(2):93–115
- Chiodi M, Adelfio G (2014) *etasflp*: Estimation of an etas model. mixed flp (forward likelihood predictive) and ml estimation of non-parametric and parametric components of the etas model for earthquake description. R package version 10
- Cressie N, Collins LB (2001) Analysis of spatial point patterns using bundles of product density lisa functions. *Journal of Agricultural, Biological, and Environmental Statistics* 6(1):118–135
- Diggle PJ (2013) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press
- Gabriel E, Diggle PJ (2009) Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica* 63(1):43–51
- Gabriel E, Rowlingson BS, Diggle PJ (2013) *stpp*: An R package for plotting, simulating and analyzing Spatio-Temporal Point Patterns. *Journal of Statistical Software* 53(2):1–29
- Illian J, Penttinen A, Stoyan H, Stoyan D (2008) *Statistical Analysis and Modelling of Spatial Point Patterns*, vol 70. John Wiley & Sons
- Mateu J, Lorenzo G, Porcu E (2007) Detecting features in spatial point processes with clutter via local indicators of spatial association. *Journal of Computational and Graphical Statistics* 16(4):968–990
- Moraga P, Montes F (2011) Detection of spatial disease clusters with lisa functions. *Statistics in Medicine* 30(10):1057–1071
- Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* 83(401):9–27
- Rodríguez-Cortés FJ (2014) *Modelling, estimation and applications of second-order spatio-temporal characteristics of point processes*. PhD thesis, Departament de Matemàtiques; Universitat Jaume I
- Rodríguez-Cortés FJ (2016) *pdLISTA*: Second-order product density local indicator of spatio-temporal association function. URL <https://github.com/frajaroco/pdLISTA>
- Siino M, Adelfio G, Mateu J, Chiodi M, D'Alessandro A (2016a) Spatial pattern analysis using hybrid models: an application to the hellenic seismicity. *Stochastic Environmental Research and Risk Assessment* pp 1–16
- Siino M, Rodríguez-Cortés FJ, Mateu J, Adelfio G (2016b) An approach to hypothesis testing based on local indicators of spatio-temporal association. In: *CM-Statistics 2016*. University of Seville; Seville Spain. Blanco-Fernandez, A. and Gonzalez-Rodriguez, G.

# Bayesian Mixture Models for the Detection of High-Energy Astronomical Sources

## *Modelli Mistura Bayesiani per la Rilevazione di Sorgenti Astronomiche ad Alta Energia*

A. Sottosanti, D. Bastieri, A. R. Brazzale

**Abstract** The search of gamma-ray sources in the extra-galactic space is one of the main targets of the *Fermi* telescope project, which aims to identify and study the nature of high energy phenomena in the universe. Starting from a collection of photons, we perform an unsupervised analysis using a Bayesian mixture model with an unknown number of components to determine the number of gamma ray sources in the map. The parameters of the model are estimated using a reversible jump MCMC algorithm. We finally propose a new method which exploits the distributions of both the weights of the mixture components and the energy spectra to qualify the nature of each cluster.

**Abstract** *La rilevazione di sorgenti gamma nello spazio extra-galattico è uno dei principali obiettivi del telescopio Fermi, nato con l'intento di studiare la diversa natura dei fenomeni ad alta energia. Partendo da un insieme di fotoni, viene proposta un'analisi non supervisionata tramite un modello mistura bayesiano con un numero finito e non noto di componenti, al fine di individuare il numero di sorgenti luminose in una mappa. Per stimare i parametri del modello, viene proposto un algoritmo reversible jump MCMC. Viene infine discussa una nuova procedura per individuare la vera natura dei gruppi attraverso l'analisi dei pesi di mistura e degli spettri di energia.*

**Key words:** Astrostatistics, Bayesian Statistics, Finite Mixture Model, MCMC

---

Andrea Sottosanti  
University of Padua, Department of Statistical Sciences, via Cesare Battisti, 241, Padova, e-mail: [sottosanti@stat.unipd.it](mailto:sottosanti@stat.unipd.it)

Denis Bastieri  
University of Padua, Department of Physics and Astronomy *Galileo Galilei*, via Belzoni, 7, Padova, e-mail: [denis.bastieri@unipd.it](mailto:denis.bastieri@unipd.it)

Alessandra R. Brazzale  
University of Padua, Department of Statistical Sciences, via Cesare Battisti, 241, Padova, e-mail: [alessandra.brazzale@unipd.it](mailto:alessandra.brazzale@unipd.it)

## 1 Introduction

The detection of astronomical sources is an interdisciplinary field which includes both statistical and astronomical methods. This work analyzes the gamma rays detected by *Fermi* LAT in the energy window 1 to 300 *GeV*, grouped into rectangular bins using galactic coordinates. The principal goals are: *i*) to determine the number of overlapping sources; *ii*) to measure their intensities; *iii*) to pool the individual counts into clusters.

In this paper, we present a statistical model which uses the direction of photons to identify the coordinates of sources in the analyzed extra-galactic space. Instead of using the *Pixel-by-Pixel* approach discussed in [3], we implement a Bayesian algorithm which simultaneously estimates the number of sources in the map, their coordinates and their intensities. We also extend [2] when the background contamination can not be assumed to be uniformly distributed over the entire map.

### 1.1 *Fermi* LAT data

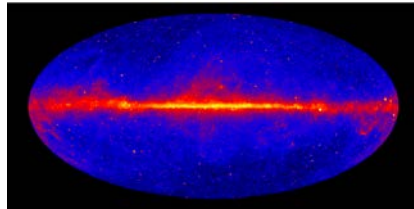
We start from a collection of  $\gamma$ -ray photons detected by the LAT telescope on-board the *Fermi* satellite, which is designed to record high energy particles. The map represented in Figure 1 is plotted in galactic coordinates. The red bright band in the middle part corresponds to the Milky Way, and the flare particular in the center indicates the presence of a black hole. This region is called *galactic space*, while the blue part of the map is the *extra-galactic space*.

One of the main goals of the *Fermi* satellite is to detect new  $\gamma$ -ray sources in the extra-galactic space, such as *active galactic nuclei (AGN)*, or in our own galaxy such as *supernova remnants (SNR)* and *pulsar wind nebula (PWN)*.

These sources are typically point-like, but there are also diffuse sources like the so-called *isotropic diffuse gamma-ray background (IGRB)* that, as its name implies, is uniformly distributed in the plane of directions, and it will not be taken into account in this analysis.

In addition, since our analysis will initially deal with extragalactic sources, we remove all photons with a galactic latitude value in  $[-10^\circ, 10^\circ]$ , presumably belonging to our galaxy.

**Fig. 1** Whole sky map at  $\gamma$ -ray wavelengths accumulated over six years of operations; the minimum observed value of energy is 1 *GeV* (<http://fermi.gsfc.nasa.gov/>)



## 2 Bayesian Finite Mixture Modelling

### 2.1 Model Specification

For a generic photon  $i = 1, \dots, N$ , we consider its galactic coordinates  $(X_i, Y_i)$ , where  $X$  represents the longitude and is defined in the interval  $[-180^\circ, 180^\circ]$ , while  $Y$  is the galactic latitude and is defined in  $[-90^\circ, 90^\circ]$ . We want to study how particles are scattered in the space, and then infer the position of the sources.

Consider a photon generated by a source  $j$ : its direction is randomly distributed as

$$(X_i, Y_i) | \mu_j \sim PSF(\mu_j), \tag{1}$$

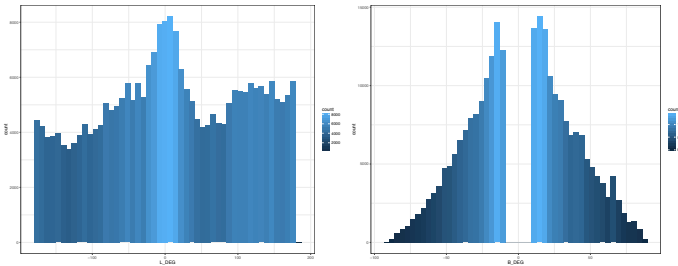
where  $\mu_j = (\mu_{jx}, \mu_{jy})$  represents the unknown coordinates of the source  $j$ . Here we use a 2-dimensional *King profile distribution* (see [2], Appendix C).

If instead we know that a certain photon was not emitted from a source, we assume this particle comes from the *background contamination*; we hence specify a distribution function also for this component.

For instance, [2] propose to assume a uniform distribution over the entire map. If we take a look at Figure 2, we can easily see that this assumption is not verified. The histogram of photon counts related to longitude shows a peak around  $0^\circ$ , while latitude presents a descending trend while moving away from the center of the galaxy. We extend the model specification given in [2] for *Fermi* LAT data by considering the bi-dimensional distribution

$$(X_i, Y_i) | \sigma_b \sim Unif(-x_{min}, x_{max}) \times Lap(0, \sigma_b). \tag{2}$$

The longitude of a photon emitted by the background contamination is then modeled as a uniform over the observed range, without taking into account the peak at the center of the galaxy, while its latitude is modeled as a Laplace distribution with mean 0 and scale parameter  $\sigma_b$ . The two components can be taken as independent because of the isotropy of the background.



**Fig. 2:** Left: histogram of longitude values of *Fermi* LAT photons. Right: histogram of latitude values.

In practice, we have no information about the real number of sources and their coordinates in space. The origin of each photon is unknown. We further need to take into account the background contamination, which has a masking effect on signal sources. We thus translate these assumptions into a statistical model, which uses a mixture of components, one for each source in the map, and one for the background contamination, that is

$$(X_i, Y_i) \sim \omega_0 g_b(X_i, Y_i | \theta_0) + \sum_{j=1}^K \omega_j f_j(X_i, Y_i | \theta_j). \quad (3)$$

In (3),  $g_b(\cdot|\cdot)$  represents the distribution of photons from the background,  $f_j(\cdot|\cdot)$  represents the source  $j$  and  $\omega = (\omega_0, \dots, \omega_K)$  is a vector of weights, which can be viewed as the intensity of each component.

Note that the number of sources  $K$  is assumed to be unknown, and will be estimated as the other parameters of the model.

In particular, for  $f_j$  we consider the density function defined in (1), while  $g_b$  assumes the form described in (2). We also have  $\theta_0 = \{\sigma_b\}$  and  $\theta_j = \{\mu_j\}$ . The total set of unknown model parameters is then  $\Theta_K = \{\omega, \theta_0, \dots, \theta_K\}$ . The goal is to make inference on  $(\Theta_K, K)$ .

From the Bayesian point of view, each unknown parameter is a random variable. Since the number of sources  $K$  is itself unknown and must be estimated, we attach to it a probability distribution. Here we put

$$K \sim Poi_t(\kappa, \kappa_{min}, \kappa_{max}), \quad (4)$$

where  $Poi_t$  is a truncated Poisson, defined in the interval  $[\kappa_{min}, \kappa_{max}]$ . The a priori distributions for  $\omega$  and  $\mu_j$  are chosen as in [2], while for  $\sigma_b$  we choose an Inverse Gamma distribution, which is the conjugate prior of the Laplace distribution.

## 2.2 Simulation Algorithm

The main challenge of our model lies in the fact that the dimension of the parameter space is itself unknown.

*Reversible Jump Markov Chain Monte Carlo* was introduced for the first time in [1] to infer model parameters when the model specification is uncertain. We propose a two step simulation algorithm.

In the first step, we fix the dimension of the parameter space and we simulate from the posterior distribution of  $\Theta_K$  using *Gibbs Sampling* and the *Metropolis-Hastings* algorithm. The updating sequence starts with an allocation of photons into the identified sources and the background. We then update the posterior distribution of the weight vector  $\omega$ , the coordinates of each source  $\mu_j$  and the scale parameter of the background  $\sigma_b$ .



**Algorithm 1** Reversible Jump MCMC - split move

---

```

1: procedure SPLIT  $j$  INTO  $j_1, j_2$  WITH PROBABILITY  $b_k = 0.25$  (from  $k$  to  $k + 1$  sources)
2:    $u_1, u_2, u_3 \sim \text{Beta}(2, 2)$ ,  $v \sim \text{Unif}(0, 1)$ 
3:    $\omega_{j_1} \leftarrow u_1 \omega_j$  and  $\omega_{j_2} \leftarrow (1 - u_1) \omega_j$ 
4:   compute  $\mu_{j_1}, \mu_{j_2}$  from  $\mu_j, \omega_{j_1}, \omega_{j_2}$  (see [2])
5:   compute quantities  $p_{k+1}, p_k, g(u_1, u_2, u_3)$  and  $J$  (see [2])
6:    $q_k \leftarrow b_k/k$  and  $q_{k+1} \leftarrow d_{k+1}/(k+1)$ 
7:   if  $\text{argmin}_j \|\mu_{j_1}, \mu_{j_2}\| = j_2$  and  $\text{argmin}_j \|\mu_{j_2}, \mu_{j_1}\| = j_1$  then
8:      $q_{k+1} \leftarrow 2q_{k+1}$ 
9:    $A \leftarrow p_{k+1}q_{k+1}J/p_kq_kg(u_1, u_2, u_3)$ 
10:  if  $v \leq \min(1, A)$  then accept split.

```

---

In the second step a trans-dimensional jump is proposed, which evaluates whether to increase or decrease the number of sources by one. The choice is made randomly; *split*, *combining*, *birth* and *death* moves are used with equal probability to jump among different dimensions of the space, according to [4]. The pseudocode for the *split* move is given in Code Box (1).

Each step of the algorithm adds a source to or removes it from the mixture. The background contamination is thus left unchanged and therefore it will never be excluded from the model.

The final estimation of  $(\Theta_K, K)$  depends on the number of sources included in the model; we fix  $K$  equal to the mode of its posterior distribution.

### 3 Application to *Fermi* LAT data

We now apply the proposed technique to the *Fermi* LAT data; in particular, we focus on the region of the sky map represented in Figure 1 with longitude values less than  $-10^\circ$  and latitude values larger than  $10^\circ$ . This choice excludes the contamination coming from the center of the galaxy.

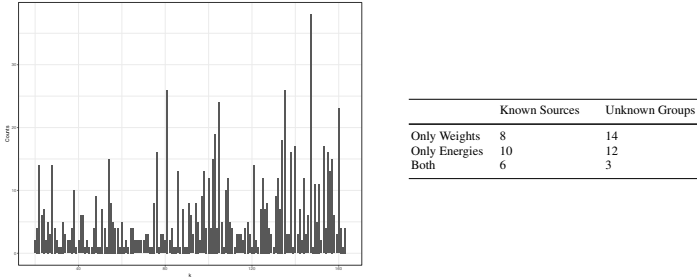
Previous analyses identified in this sector 54 sources<sup>1</sup>. We now want to evaluate the behaviour and the clustering performance of our method when applied to a map characterized by an high value of background contamination, as it is the case for the *Fermi* LAT data. The list of discovered sources is used as a benchmark for the clustering performance of our model.

We start the procedure running 10,000 iterations of the reversible jump MCMC algorithm from different starting points to explore the entire map; Figure 3 shows the posterior distribution of  $K$  produced by the first chain.

Although it emerges from this graph that the most visited value of  $K$  is 147, this number most likely does not represent the real number of sources in the analyzed map. This result seems instead to be the side-effect of the rather strong background contamination, which both masks the signal and thus leads to the detection of false

---

<sup>1</sup> <https://fermi.gsfc.nasa.gov/ssc/data/access/lat/2FHL/>



**Fig. 3:** Left: posterior distribution of  $K$ . Right: results obtained from classification with three proposed methods.

positives. A statistical method to discriminate between real and fake clusters is thus necessary.

We propose to use an empirical analysis of the posterior distributions of the weights  $\omega$  and of the energy spectra of the clusters. From our simulation studies (results not shown here), it emerged that at times the reversible jump MCMC overestimates the number of components of the mixture in the presence of background contamination; however, the posterior distributions of weights associated to these false groups are very close to 0. This empirical result leads us to select as sources all those clusters with a median value of the weights higher than a defined threshold, which we fix to 0.01.

A similar empirical approach can be applied to the energy spectra of the clusters, and selects those groups with high levels of energy. In particular, we selected 125 GeV as the threshold for classification with energy spectra.

Table in Figure 3 compares our results after discrimination with the list of published sources. It emerges that 6 clusters selected after both weight and energy discrimination coincide with known sources. If instead a single classification method is applied, both, known sources and unknown groups, emerge from the results.

## References

1. Green, Peter J. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika* (1995): 711-732.
2. Jones, David E., Vinay L. Kashyap, and David A. Van Dyk. "Disentangling Overlapping Astronomical Sources using Spatial and Spectral Information." *The Astrophysical Journal* 808.2 (2015): 137.
3. Mattox, James R., et al. "The likelihood analysis of EGRET data." *The Astrophysical Journal* 461 (1996): 396.
4. Richardson, Sylvia, and Peter J. Green. "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." *Journal of the Royal Statistical Society: series B (statistical methodology)* 59.4 (1997): 731-792.

# Causal analysis of Cell Transformation Assays

## *Analisi causale dei Cell Transformation Assay*

Federico Mattia Stefanini

**Abstract** A Cell Transformation Assay (CTA) is an *in vitro* method to test a chemical for carcinogenicity. In a recent contribution from an international expert group created to improve the analysis of BALB/c 3T3 CTA data, two classes of models in the frequentist paradigm were recommended. Here a Bayesian model for potential outcomes is developed to estimate the causal effect of some concentrations of a candidate carcinogen on counts of *foci* growing within Petri dishes. The reanalysis of an actual case study is performed to illustrate some limitations of current models and the main features of the proposed approach.

**Abstract (in Italian)** Il Cell Transformation Assay (saggio di trasformazione cellulare) è un metodo *in vitro* per saggiare la carcinogenicità di una sostanza chimica. In un recente contributo di un gruppo di esperti internazionali, creato per migliorare l'analisi dei dati provenienti dal saggio con la linea cellulare BALB/c 3T3, sono stati raccomandati due classi di modelli nel paradigma frequentista. In questo lavoro viene sviluppato un modello Bayesiano per i risultati potenziali allo scopo di stimare l'effetto causale di alcune concentrazioni di un candidato cancerogeno sul conteggio dei *foci* che crescono entro capsule Petri. La rianalisi di un caso di studio reale viene realizzata per illustrare alcune limitazioni dei metodi correnti e le principali caratteristiche del modello proposto.

**Key words:** Bayesian causal model, potential outcomes

## 1 Introduction

It has been estimated that annual cancer incidence will rise from 14 million in year 2012 to 22 within the next 2 decades [1]. Chemical carcinogenicity is defined as

---

Department of Statistics, Computer Science, Applications  
University of Florence  
e-mail: stefanini@disia.unifi.it

the ability of a chemical substance, or a mixture of chemical substances, to induce cancer or to increase its incidence. Given the role played by environmental and chemical exposures, no wonder that the evaluation of chemical carcinogenicity has become a leading task in public health risk assessment during the last decades.

Cell Transformation Assays (CTAs) are a family of *in vitro* methods for the identification of potential chemical carcinogens. It has been shown that CTAs nicely correlate with rodent bioassay, which is considered the standard approach for carcinogenicity testing [2]. The endpoint assessed in CTAs is the progression of cultured cells from immortality to tumorigenicity, as evidenced by formation of *foci* of multilayered and disorganised cells, growing over the surrounding regular cell monolayer. Therefore, the number of fully transformed cell colonies, called type III *foci*, grown within a Petri dish (experimental unit) after 4 weeks from treatment with the chemical under testing is the outcome of primary interest.

In the following, a Bayesian model for potential outcomes is developed to estimate the causal effect of different concentrations of a chemical in a CTA.

## 2 A Bayesian model

The case study here considered is a CTA experiment performed to test o-toluidine (CAS chemical registry number # 636-21-5). A total of eight different concentrations and the negative control were considered, and they are ( $\mu\text{g/ml}$ ): 0 (negative control,  $i = 0$ ), 20 ( $i = 1$ ), 100 ( $i = 2$ ), 200 ( $i = 3$ ), 500 ( $i = 4$ ), 800 ( $i = 5$ ), 1000 ( $i = 6$ ), 1200 ( $i = 7$ ), 1750 ( $i = 8$ ). A total of 90 Petri dishes containing BALB/c 3T3 cells sampled at the same passage from the original cell culture were treated after random assignment of each concentration to  $n_i = 10$  dishes (replicates) for each  $i$ . All experimental units received protocol ingredients taken from the same batch, including medium and serum. After 4 weeks from treatment, Petri dishes were visually scored under a light microscope and the number of type III *foci* within each dish counted.

Following Rubin's framework for causal inference [3], potential outcomes are introduced for every treatment and experimental unit under consideration. Let  $Y_k^{<i>}$  be random variables representing the potential number of *foci* within Petri dish  $k = 1, \dots, n$  under treatment (concentration)  $i = 0, 1, \dots, L$ . A plausible size for the sample space  $\Omega_{Y^{<i>}}$  is around 30, thus  $\Omega_{Y^{<i>}} = \{0, 1, \dots, 30\}$ , because the available physical space on a Petri dish is limited.

The potential outcomes referred to concentration  $i$  define the vector  $Y^{<i>}$ . Let  $W = (W_1, \dots, W_n)^T$  be the vector of indicators of treatment assignment, with sample space  $\Omega_{W_i} = \{0, \dots, L\}$ . Given that CTAs belong to the class of randomized experiments, the assignment mechanism is ignorable and characterized by unit-level probability of treatment assignment in the interval  $(0, 1)$ , in particular the probabilistic mass function of vector  $W$  that represents the assignment mechanism is:

$$p(W | Y^{<0>}, \dots, Y^{<L>}) = \binom{n}{n_1 n_2 \dots n_L}^{-1} \tag{1}$$

for all  $W$  satisfying  $\sum_{k=1}^n I_i(W_k) = n_i$  for each  $i$ . Under row (unit) exchangeability of matrix  $(Y^{<0>}, \dots, Y^{<L>})$  the joint distribution of potential outcomes is:

$$p(Y^{<0>}, \dots, Y^{<L>}) = \int \prod_{k=1}^n p(Y_k^{<0>}, \dots, Y_k^{<L>} | \theta) p(\theta) d\theta \tag{2}$$

where  $\theta$  is a vector of model parameters belonging to the parameter space  $\Theta$ .

The elicitation of conditional distributions for potential outcomes given model parameters (eq. 2), the so called science, should take into account the main processes driving the emergence of *foci*. Even if it is not carcinogenic, a chemical may exert a toxic effect on cultured cells, thus causing a reduction in the final number of type III *foci*. If a chemical is carcinogenic then it is expected to stimulate the emergence of *foci*, but this driving force also depends on concentration: too low doses are ineffective, too high doses are often cytotoxic. Despite that concentrations are selected to be within a convenient range, it is quite difficult to anticipate any correlation between potential outcomes. For these reasons, conditional independence among potential outcomes is here assumed:

$$p(Y_1^{<0>}, \dots, Y_k^{<L>}, \theta) = \prod_{i=0}^L p(Y_k^{<i>} | \theta_i) p(\theta_i) \tag{3}$$

where the joint distribution of model parameters is factorized into marginally independent subvectors,  $\theta = (\theta_0, \dots, \theta_i, \dots, \theta_L)$ .

At the end of the experiment, the vector of observed potential outcomes is:

$$Y^{<obs>} = \left\{ \left( \dots, \sum_{i=0}^L Y_k^{<i>} I_i(W_k), \dots \right)^T : k = 1, \dots, n \right\}$$

while the collection of vectors  $C^{<mis>} = \{Y^{<mis,0>}, \dots, Y^{<mis,L>}\}$  with

$$Y^{<mis,i>} = \left\{ (\dots, Y_k^{<i>}, \dots)^T : W_k \neq i, k = 1, \dots, n, \right\}$$

and  $i = 0, \dots, L$ , contains missing potential outcomes. The conditional predictive distribution  $p(C^{<mis>} | Y^{<obs>}, W)$  is exploited to impute missing values.

Three causal estimands of particular interest are finite sample averages of indicators, with  $i \geq 1$ : the probability of positive effect on the treated (PPET) units,

$$\tau_{PPET}^{<i>} = \sum_{k=1}^n I_{\{y^{<i>} - y^{<0>} > 0 \wedge W_k = i\}} (Y_k^{<i>}, Y_k^{<0>}, W_k) / \sum_{k=1}^n I_i(W_k),$$

the probability of null effect on the treated (PNET) units, where the indicator of the event is  $I_{\{y^{<i>} - y^{<0>} = 0 \wedge W_k = i\}}(Y_k^{<i>}, Y_k^{<0>}, W_k)$ , the probability of cytotoxic effect on the treated (PCET) units, where  $I_{\{y^{<i>} - y^{<0>} < 0 \wedge W_k = i\}}(Y_k^{<i>}, Y_k^{<0>}, W_k)$ .

Further structure may be imposed on the model after borrowing context and assumptions settled by an international expert group (European Centre for the Validation of Alternative Methods, ECVAM), in particular:

1. the number of Petri dishes at each concentration is typically 10 (never below 9);
2. the number of levels for the concentration typically ranges from 3 to 7;
3. *focus*-inducing chemicals are expected to show non-monotone dose-concentration relationships, mostly due to cytotoxicity at higher concentrations;
4. positive controls are not informative;
5. at small concentrations the empirical distribution of counts may be degenerate, typically at zero;
6. concentrations have to be considered as levels of a qualitative factor, although originally on a quantitative scale ( $\mu\text{g/ml}$ ).

ECVAM's experts recommended two tentative classes of models, the first one is a Normal model for Nishiyama-transformed counts,  $x = \sqrt{y} + \sqrt{1+y}$ , and the second one is a Negative Binomial model for original counts. Unfortunately, the two recommended classes did not seem suited to our case study (Section 3).

ECVAM's committee proposed two family of distributions allowing asymmetry (original scale) and smooth changes of probability value in contiguous (transformed) counts. By introducing latent variables  $X_k^{<i>} \sim \text{Beta}(x \mid \alpha_i, \beta_i)$  in the Beta family, we essentially maintained the original belief while gaining in flexibility: values of variance smaller than the mean became possible. The probability of observing count  $y_{i,j}$  is thus  $P[Y_j^{<i>} = y] = \int_{y/31}^{(y+1)/31} \text{Beta}(x \mid \alpha_i, \beta_i) dx$ . A weakly informative initial distribution was elicited for marginally independent model parameters, with  $\alpha_i \sim \text{Uniform}(1, 1000)$  and  $\beta_i \sim \text{Exponential}(0.01)$ ,  $i = \{0, \dots, L\}$ .

### 3 Results and discussion

Computations were performed in R<sup>1</sup> using RStudio<sup>2</sup> and the following packages<sup>3</sup>: *MASS*, *fitdistrplus*, *rjags*, *coda*, *knitr*.

In the first step of the analysis, Normal models for Nishiyama-transformed counts were considered. Note that after transformation, null counts are mapped to 1. From unbiased point estimates of model parameters at each concentration  $i$ , plug-in estimates of probability values  $\hat{P}[X_i < 1 \mid \hat{\mu}_i, \hat{\sigma}_i^2]$  were calculated, and for concentrations from 0 to 200 they resulted well above 0.15, that is: 0.1904, 0.2265, 0.2673, 0.2734.

<sup>1</sup> <https://www.R-project.org/>

<sup>2</sup> <http://www.rstudio.com>

<sup>3</sup> <https://cran.r-project.org/web/packages/>

For concentrations equal to 500 and 800 estimates were 0.1470 and 0.0550 respectively. Only for the last 3 concentrations the point estimates were below 0.02. Fitting distributions by maximum likelihood always reduced the the above estimated probability values of about 0.01. Quantile-quantile plots (not shown), even if based on just 10 observations, detected clear departures of Nishiyama-transformed counts from normality for all concentrations smaller than 800.

Note that no concentration showed counts all equal to zero (or to one), an event of appreciable probability in many case studies, therefore it was possible to obtain the unbiased point estimate of the variance. For this reason, we did not consider the recommended artificial increase of sample size by one observation equal 1, an action that would have determined an increase of sample size of about 10% at each concentration. Furthermore, in case all observations are equal to one, such artificial change of observed counts is not even uniquely defined. Given the role played by the predictive distribution in the Bayesian causal model, and therefore by the model for observations, the Normal model for Nishiyama-transformed counts seemed unsatisfactory and therefore was not considered further.

In the second step, we considered the class of Negative Binomial models. The optimization of likelihood functions at each concentration often failed due to divergence of the scale parameter towards infinity. Even upon termination, it was sometimes impossible to calculate standard errors, or in other cases estimated values were huge. Similar failures were observed using other algorithms, for example iterated moment matching. Indeed, a small sample size at each concentration ( $n_i = 10 \forall i$ ) and the presence of sample variances often smaller than sample averages made the optimization hard, given that the Negative Binomial family is not suited to under-dispersed count data. All things considered, the class of Negative Binomial models seemed unsatisfactory for the case study at hand and therefore it was not considered further.

The proposed Beta latent model was fitted by MCMC: a sample of  $1 \times 10^5$  realizations from the final distribution of model parameters was collected after thinning one chain by 4. The initial burn-in consisted of 10'000 iterations. Values of difference between pair of counterfactuals in the numerator of  $\tau_{PPET}^{<i>$ ,  $i = 1, \dots, 8$  were saved and further processed to obtain their distribution conditioned to observed outcomes at each concentration. One-chain output diagnostics did not suggest lack of convergence. In Table (1), estimated probability of carcinogenic/null/cytotoxic effects are shown, based on a sample of  $1 \times 10^6$  imputed counterfactuals for each concentration. The odds of a carcinogenic effect of o-toluene at concentrations 1000 and 1200 is about ten. The 5% quantile of the distribution of odds for a carcinogenic effect of o-toluene is equal to 2.3333 at both concentrations 1000 and 1250, thus they are both well above 1.

**Table 1** Estimated probability values of causal effects.

Concentration <math>i</math>	20	100	200	500	800	1000	1200	1750
$\tau_{PCET}^{<i>i</i>}$	0.403	0.454	0.441	0.324	0.194	0.032	0.024	0.107
$\tau_{PNET}^{<i>i</i>}$	0.386	0.417	0.391	0.365	0.254	0.067	0.053	0.147
$\tau_{PPET}^{<i>i</i>}$	0.211	0.129	0.168	0.311	0.552	0.901	0.923	0.746
odds $\frac{\tau_{PPET}^{<i>i</i>}}{(1-\tau_{PPET}^{<i>i</i>})}$	0.266	0.148	0.202	0.451	1.232	9.101	11.987	2.937

## 4 Conclusions

We developed a Bayesian causal model based on latent Beta distributions to overcome limitations found in alternative proposals when applied to the o-toluene case study. Instead of exploiting beliefs by adding virtual observations, prior information was formally introduced. Causal estimands like PPET were formulated to restrain the assessment by excluding the magnitude of differences, but motivations of this choice will be detailed elsewhere. Further similar estimands might address effects due to the increase of concentration, an important issue for applications not reported in this work.

The limited sample size and the small number of concentrations characterizing a typical CTA make the assessment of model assumptions hard. It is clear that alternative, and more general, latent models might be formulated, for example one that allows for rare but very extreme counts. These outliers were absent in our case study but they are not so rarely observed in CTAs. Replicates of the same experiment performed on a chemical in different laboratories represent an opportunity to increase sample size, at least after properly considering transportability.

Finally, a battery of positive controls made by known genotoxic and nongenotoxic carcinogens, for example 3-methylcolantrene, could be introduced to study the variability in the response of experimental units to treatments.

**Acknowledgements** The author thanks Chiara Urani e Giulia Callegaro for their valuable expertise and many fruitful discussions on CTAs. This work is partially supported by University of Florence, frame ‘Disegno e analisi di studi sperimentali e osservazionali per le decisioni’.

## References

1. Stewart, B. W., C. P. Wild: World Cancer Report 2014. OCLC: 908606220. Lyon: International Agency for Research on Cancer/World Health Organization (2014).
2. Creton, S., Aardema, M.J., Carmichael, P.L., Harvey, J.S., Martin, F.L., Newbold, R.F., O’Donovan, M.R., Pant, K., Poth, A., Sakai, A., Sasaki, K., Scott, A.D., Schechtman, L.M., Shen, R.R., Tanaka, N., Yasaei, H.: Cell transformation assays for prediction of carcinogenic potential: state of the science and future research needs. *Mutagenesis* **27**, 93–101 (2012).
3. Rubin, D.: Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics* **6**, 34–58, (1978).



# Estimation and Inference of Skew–Stable distributions using the Multivariate Method of Simulated Quantiles

*Stima e inferenza per i parametri delle distribuzioni Stabili asimmetriche utilizzando il metodo dei quantili simulati*

Stolfi Paola and Bernardi Mauro and Petrella Lea

**Abstract** The multivariate method of simulated quantiles (MMSQ) is proposed as a likelihood–free alternative to indirect inference procedures that does not rely on an auxiliary model specification and its asymptotic properties are established. As a further improvement we introduce the Smoothly clipped absolute deviation (SCAD)  $\ell_1$ –penalty into the MMSQ objective function in order to achieve sparse estimation of the scaling matrix. We extend the asymptotic theory and we show that the sparse–MMSQ estimator enjoys the oracle properties under mild regularity conditions. The method is applied to estimate the parameters of the Skew Elliptical Stable distribution.

**Abstract** *In questo lavoro viene proposto il metodo dei quantili simulati multivariati che rappresenta una valida alternativa alle procedure di inferenza indiretta e che non richiede la specificazione di un modello ausiliario e vengono dimostrate le proprietà asintotiche dello stimatore. Allo scopo di indurre una stima sparsa della matrice di scala introduciamo inoltre la funzione di penalità SCAD all'interno della funzione obiettivo del metodo. Un ulteriore contributo è rappresentato dall'estensione della teoria asintotica nel caso sparso e dalla dimostrazione che lo stimatore soddisfa le proprietà ORACLE. Il metodo è applicato alla stima dei parametri della distribuzione Stabile asimmetrica.*

**Key words:** Directional quantiles; Sparse regularisation; Skew Elliptical Stable distribution.

---

Stolfi Paola

Department of Economics, University of Rome Tre and Istituto per le Applicazioni del Calcolo “Mauro Picone” - CNR, e-mail: paola.stolfi@uniroma3.it.

Bernardi Mauro

Department of Statistical Sciences, University of Padova and Istituto per le Applicazioni del Calcolo “Mauro Picone” - CNR, e-mail: mauro.bernardi@unipd.it.

Petrella Lea

MEMOTEF Department, Sapienza University of Rome, e-mail: lea.petrella@uniroma1.it.

### 1 Introduction

In this paper we extend the method of simulated quantiles (MSQ) of Dominicy and Veredas (2013) to a multivariate framework (MMSQ). The method of simulated quantiles like alternative likelihood-free procedures is based on the minimisation of the distance between appropriate quantile-based statistics evaluated on the true and simulated data. The MMSQ effectively deals with distributions that do not admit moments of any order, like the  $\alpha$ -Stable or the Tukey lambda, without relying on the choice of a misspecified auxiliary model. The lack of a natural ordering in the multivariate setting requires a careful definition of the concept of quantile. Here, we rely on the notion of projectional quantile recently introduced by Hallin et al. (2010) and Kong and Mizera (2012). This notion of multivariate quantile makes the estimator flexible and it allows us to deal with non-elliptically contoured distributions. As a further improvement we introduce the smoothly clipped absolute deviation (SCAD)  $\ell_1$ -penalty of Fan and Li (2001) into the MMSQ objective function in order to achieve sparse estimation of the scaling matrix. The method is illustrated using several synthetic datasets from distributions for which alternative procedures are recognised to perform poorly, such as the Skew Elliptical Stable distribution (SESD) firstly mentioned by Branco and Dey (2001).

The remainder of the paper is structured as follows. Section 2 introduces the sparse MMSQ estimator. Section 3 defines the Skew-Elliptical distribution of Branco and Dey (2001) while Section 4 presents simulated-data experiments to assess the effectiveness of the proposed method. Section 5 concludes.

### 2 The Multivariate Method of Simulated Quantiles

Let:

- (i)  $\mathbf{Y} \in \mathbb{R}^d$  be a random variable with distribution function  $F_{\mathbf{Y}}(\cdot, \vartheta)$ , which depends on a vector of unknown parameters  $\vartheta \subset \Theta \in \mathbb{R}^k$ , and  $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)^T$  be a vector of  $n$  independent realisations of  $\mathbf{Y}$ ;
- (ii)  $\mathbf{q}_{\vartheta}^{\tau, \mathbf{u}} = (q_{\vartheta}^{\tau_1 \mathbf{u}}, q_{\vartheta}^{\tau_2 \mathbf{u}}, \dots, q_{\vartheta}^{\tau_s \mathbf{u}})$  be a  $s \times 1$  vector of projectional quantiles at given confidence levels  $\tau_i \in (0, 1)$  with  $i = 1, 2, \dots, s$ , and  $\mathbf{u} \in \mathbb{S}^{d-1}$ ;
- (iii)  $\Phi_{\mathbf{u}, \vartheta} = \Phi(\mathbf{q}_{\vartheta}^{\tau, \mathbf{u}})$  be a  $b \times 1$  vector of quantile functions assumed to be continuously differentiable with respect to  $\vartheta$  for all  $\mathbf{Y}$  and measurable for  $\mathbf{Y}$  and for all  $\vartheta \subset \Theta$ ;
- (iv)  $\hat{\mathbf{q}}^{\tau, \mathbf{u}} = (\hat{q}^{\tau_1 \mathbf{u}}, \hat{q}^{\tau_2 \mathbf{u}}, \dots, \hat{q}^{\tau_s \mathbf{u}})$  and  $\hat{\Phi}_{\mathbf{u}} = \Phi(\hat{\mathbf{q}}^{\tau, \mathbf{u}})$  be the corresponding sample counterparts;

and assume that  $\Phi_{\mathbf{u}, \vartheta}$  cannot be computed analytically but it can be empirically calculated on simulated data. At each iteration  $j = 1, 2, \dots$  the MMSQ compute  $\tilde{\Phi}_{\mathbf{u}, \vartheta_j}^R = \frac{1}{R} \sum_{r=1}^R \tilde{\Phi}_{\mathbf{u}, \vartheta_j}^r$ , where  $\tilde{\Phi}_{\mathbf{u}, \vartheta_j}^r$  is the function  $\Phi_{\mathbf{u}, \vartheta}$  computed at the  $r$ -th simulation path from  $F_{\mathbf{Y}}(\cdot, \vartheta^{(j)})$ . The parameters are subsequently updated by min-

imising the distance between the vector of quantile measures calculated on the true observations  $\hat{\Phi}_{\mathbf{u}}$  and that calculated on simulated realisations  $\tilde{\Phi}_{\mathbf{u},\vartheta}^R$ . The subscript  $\mathbf{u}$  denotes that those quantities depend on a set of directions that should be properly chosen. We establish consistency and asymptotic normality of the proposed estimator. The MMSQ estimator is then extended in order to achieve sparse estimation of a scaling matrix  $\Sigma$ . Specifically, the SCAD  $\ell_1$ -penalty of Fan and Li (2001) is introduced into the MMSQ objective function as follows

$$\hat{\vartheta} = \arg \min_{\vartheta} \left( \hat{\Phi}_{\mathbf{u}} - \tilde{\Phi}_{\mathbf{u},\vartheta}^R \right)' \mathbf{W}_{\vartheta} \left( \hat{\Phi}_{\mathbf{u}} - \tilde{\Phi}_{\mathbf{u},\vartheta}^R \right) + n \sum_{i < j} p_{\lambda}(|\sigma_{i,j}|), \tag{1}$$

where  $\mathbf{W}_{\vartheta}$  is a  $b \times b$  symmetric positive definite weighting matrix,  $\Sigma = (\sigma_{i,j})_{i,j=1}^n$  is the scale matrix and  $p_{\lambda}(\cdot)$  is the SCAD  $\ell_1$ -penalty. By setting the tuning parameter  $\lambda = 0$ , equation (1) reduces to non sparse MMSQ estimator. We extend the asymptotic theory and we show that the sparse–MMSQ estimator enjoys the oracle properties under mild regularity conditions.

### 3 Skew Elliptical Stable distribution

In this Section we define the quantile–based measures and the optimal directions  $\mathbf{u} \in \mathbb{S}^{m-1}$  for the parameters of the SESD distribution  $\mathbf{Y} \sim \text{SESD}_m(\alpha, \xi, \Omega, \delta)$  introduced by Branco and Dey (2001). For the shape parameter  $\alpha$ , the locations  $\xi_i$ , the skewness parameters  $\delta_i$  and scale parameters  $\omega_{ii}$ ,  $i = 1, 2, \dots, m$  we consider

$$\begin{aligned} \kappa_{\mathbf{u}} &= \frac{q_{0.95\mathbf{u}} - q_{0.05\mathbf{u}}}{q_{0.75\mathbf{u}} - q_{0.25\mathbf{u}}} \\ m_{\mathbf{u}} &= q_{0.5\mathbf{u}} \\ \gamma_{\mathbf{u}} &= \frac{q_{0.95\mathbf{u}} + q_{0.05\mathbf{u}} - 2q_{0.5\mathbf{u}}}{q_{0.95\mathbf{u}} - q_{0.05\mathbf{u}}} \\ \zeta_{\mathbf{u}} &= q_{0.75\mathbf{u}} - q_{0.25\mathbf{u}}, \end{aligned}$$

where  $\mathbf{u} \in \mathbb{S}^{m-1}$  defines a relevant direction. Once the quantile–based measures have been selected, we need to identify the optimal directions for each parameter. Let us consider the locations first. Because of the presence of skewness, the median computed along the canonical directions is not a good quantile measure for the locations. Therefore, we consider a transformation of the data in order to remove the skewness. The properties of the Skew Elliptical Stable distribution imply that  $\mathbf{Y}^- = -\mathbf{Y} \sim \text{SESD}_m(\alpha, \xi, \Omega, -\delta)$  independent of  $\mathbf{Y}$ , therefore it holds

$$\mathbf{Z} = \frac{\mathbf{Y} + \mathbf{Y}^-}{\sqrt{2}} \sim \text{SESD}_m\left(\alpha, \sqrt{2}\xi, \Omega, \mathbf{0}\right), \tag{2}$$

which means that the variable  $\mathbf{Z}$  is symmetric and, up to a constant, it has the same location parameter of  $\mathbf{Y}$ . Therefore, we choose, as informative measure for the locations, the median of the transformed variable  $\mathbf{Z}$  in equation (2). In order to estimate the remaining parameters, we consider univariate marginals that have Skew Elliptical Stable distribution, i.e.,  $Y_i \sim \text{SESD}_1(\alpha, \xi_i, \omega_i, \delta_i)$ , by construction. The quantile-based measures for the shape, skewness and for the diagonal elements of the scale matrix  $\omega_i$  are then computed along the canonical directions.

Now we need to identify the optimal directions for the off-diagonal elements of the scale matrix  $\Omega$ . To this end, we consider the bivariate marginal variables  $\mathbf{Y}_{ij} = (Y_i, Y_j)'$  for  $1 \leq i < j \leq m$ . It holds  $\mathbf{Y}_{ij} \sim \text{SESD}_2(\alpha, \xi_{ij}, \Omega_{ij}, \delta_{ij})$ , where  $\xi_{ij} = (\xi_i, \xi_j)'$  and  $\Omega_{ij} = \begin{bmatrix} \omega_{ii} & \omega_{ij} \\ \omega_{ij} & \omega_{jj} \end{bmatrix}$ , while  $\delta_{ij} = (\delta_i, \delta_j)'$ . Moreover, let  $\mathbf{Y}_{ij}^- \sim \text{SESD}_2(\alpha, \xi_{ij}, \Omega_{ij}, -\delta_{ij})$  independent of  $\mathbf{Y}_{ij}$  and let us consider the same construction introduced for the locations, that is the random variable  $\mathbf{Z}_{ij} = \frac{\mathbf{Y}_{ij} + \mathbf{Y}_{ij}^-}{\sqrt{2}}$ , having distribution  $\mathbf{Z}_{ij} \sim \text{SESD}_2(\alpha, \sqrt{2}\xi_{ij}, \Omega_{ij}, \mathbf{0})$ . Since  $\mathbf{Z}_{ij}$  is a symmetric variable we choose the optimal direction  $\mathbf{u}^* \in \mathbb{S}^1$  such that

$$\mathbf{u}^* = \arg \max_{\mathbf{u} \in \mathbb{S}^1} \sqrt{\mathbf{u}' \Omega_{ij} \mathbf{u}}. \tag{3}$$

### 4 Simulated-data experiment

To illustrate the effectiveness of the MMSQ in dealing with parameters estimation of the SEDS we consider a simulation example where we fix the dimension  $m = 5$  and  $\alpha = 1.70$ , while the location, shape and scale parameters are  $\xi = \mathbf{0}$ ,  $\delta = (0, 0, 0, 0.9, 0.9)$  and

$$\Sigma_5^s = \begin{pmatrix} 0.25 & 0.25 & 0.4 & 0 & 0 \\ 0.25 & 0.5 & 0.4 & 0 & 0 \\ 0.4 & 0.4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2.55 \\ 0 & 0 & 0 & 2.55 & 4 \end{pmatrix}, \tag{4}$$

respectively. We consider two different sample sizes  $n = 500, 2000$  and we fix the number of simulated paths  $R = 5$ . Simulation results over 100 replications are reported in Table 1. Table 1 reports the bias (BIAS), the standard deviation (SSD) and the empirical coverage probabilities (ECP) obtained over 100 replications of the simulation experiment. Our results show that the MMSQ estimator is always unbiased, indeed the BIAS is always less 0.15. The SSDs are always small, in particular for  $n = 500$  it is always less than 0.5. Moreover, the empirical coverages are always in line with their expected values. In order to apply the sparse-MMSQ we consider a simulation example of dimension  $m = 12$ , with  $n = 200$  and  $R = 5$  where the location parameters are equal to zero, the shape parameters  $\delta = (0, 0, 0.6, 0, 0, 0, 0, 0, 0.6, 0.6, 0)'$ , while we consider the same scale matrix as

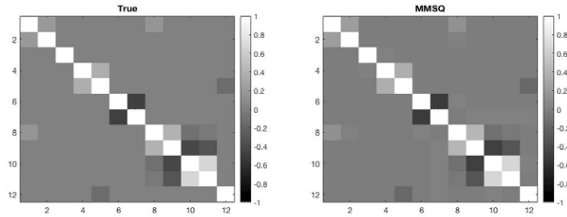
Par.	True	n = 500			n = 2000		
		BIAS	SSD	ECP	BIAS	SSD	ECP
$\alpha$	1.70	-0.0068	0.0690	0.9500	0.0013	0.0320	0.9500
$\delta_1$	0.00	0.0048	0.0067	0.9400	0.0022	0.0010	0.1100
$\delta_2$	0.00	0.0048	0.0063	0.9500	0.0082	0.0016	0.0100
$\delta_3$	0.00	0.0040	0.0038	0.9100	0.0013	0.0005	0.0600
$\delta_4$	0.90	-0.0116	0.1648	0.9700	0.0180	0.0185	0.8100
$\delta_5$	0.90	-0.0179	0.1649	0.9700	0.0167	0.0234	0.9200
$\xi_1$	0.00	0.0016	0.0365	0.9600	0.0032	0.0218	0.9400
$\xi_2$	0.00	-0.0029	0.0534	0.9700	0.0023	0.0286	0.9400
$\xi_3$	0.00	0.0093	0.0757	0.9400	0.0065	0.0393	0.9500
$\xi_4$	0.00	-0.0051	0.0703	0.9700	0.0041	0.0356	0.9500
$\xi_5$	0.00	-0.0059	0.1089	0.9200	-0.0040	0.0618	0.9400
$\omega_{11}^s$	0.2500	-0.0126	0.0259	0.9400	-0.0027	0.0140	0.9800
$\omega_{22}^s$	0.5000	0.0184	0.0596	0.9200	0.0003	0.0261	0.9400
$\omega_{33}^s$	1.0000	0.0038	0.0998	0.9700	0.0166	0.0538	0.9500
$\omega_{44}^s$	2.0000	-0.1397	0.3571	0.9300	-0.1571	0.1700	0.8800
$\omega_{55}^s$	4.0000	-0.4342	0.6637	0.9100	-0.1142	0.3980	0.9600
$\omega_{12}$	0.7071	-0.0438	0.1336	0.9400	-0.0345	0.1055	0.9100
$\omega_{13}$	0.8000	-0.1043	0.1487	0.9200	-0.0173	0.1050	0.9800
$\omega_{14}$	0.00	0.0075	0.0256	0.9300	0.0018	0.0148	0.9400
$\omega_{15}$	0.00	0.0085	0.0445	0.9700	0.0040	0.0170	0.9400
$\omega_{23}$	0.5657	-0.0851	0.1680	0.9300	-0.0323	0.1255	0.9700
$\omega_{24}$	0.00	0.0049	0.0306	0.9600	0.0032	0.0154	0.9200
$\omega_{25}$	0.00	0.0076	0.0414	0.9300	0.0053	0.0172	0.9600
$\omega_{34}$	0.00	0.0047	0.0277	0.9100	0.0022	0.0151	0.9300
$\omega_{35}$	0.00	0.0100	0.0332	0.9500	0.0032	0.0151	0.9300
$\omega_{45}$	0.9016	-0.0727	0.0785	0.8200	-0.0552	0.0573	0.9300

**Table 1** Bias (BIAS), sample standard deviation (SSD), and empirical coverage probability (ECP) at the 95% confidence level for the locations  $\mu = (\mu_1, \mu_2, \dots, \mu_d)$ , scale matrix  $\Omega = \{\omega_{ij}\}$ , with  $i, j = 1, 2, \dots, d$  and  $i \leq j$ , tail parameter  $\alpha = 1.70$  and skewness parameter  $\delta_i, i = 1, 2, \dots, d$  of the Skew Elliptical Stable distribution in dimension 5. The results reported above are obtained using 100 replications.

in Wang (2015) and reported below

$$\Sigma_{12}^s = \begin{pmatrix} 0.239 & 0.117 & 0 & 0 & 0 & 0 & 0 & 0.031 & 0 & 0 & 0 & 0 \\ 0.117 & 1.554 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.362 & 0.002 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.002 & 0.199 & 0.094 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.094 & 0.349 & 0 & 0 & 0 & 0 & 0 & 0 & -0.036 \\ 0 & 0 & 0 & 0 & 0 & 0.295 & -0.229 & 0.002 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.229 & 0.715 & 0 & 0 & 0 & 0 & 0 \\ 0.031 & 0 & 0 & 0 & 0 & 0.002 & 0 & 0.164 & 0.112 & -0.028 & -0.008 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.112 & 0.518 & -0.193 & -0.09 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.028 & -0.193 & 0.379 & 0.167 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.008 & -0.09 & 0.167 & 0.159 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.036 & 0 & 0 & 0 & 0 & 0 & 0.207 \end{pmatrix} \quad (5)$$

As regards the simulated example in dimension  $m = 12$  we plot in Figure 1 the images displaying the band structure of the true estimated scale matrices are very close.



**Fig. 1** Images displaying the band structure of the true (*left*) and estimated (*right*) scale matrices of the simulated example in dimension  $m = 12$ .

## 5 Conclusion

In this paper the problem of parameter estimation and inference of Skew–Stable distributions has been approached using the multivariate method of simulated quantiles. Moreover, since as the number of dimensions increases the curse of dimensionality problem prevents any effective inferential procedure we introduce the sparse–MMSQ estimator and we prove that the estimator enjoys the oracle properties under mild regularity conditions. The MMSQ and the sparse–MMSQ have been applied to the problem of estimating the parameters of the multivariate Skew–Stable distribution introduced by Branco and Dey (2001). Our simulation results show that the proposed methodology effectively achieves sparse estimation of the scale parameter.

## References

1. Branco, M. D. and Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *J. Multivariate Anal.*, 79(1):99–113.
2. Dominicy, Y. and Veredas, D. (2013). The method of simulated quantiles. *J. Econometrics*, 172(2):235–247.
3. Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.
4. Hallin, M., Paindaveine, D., and Šíman, M. (2010a). Multivariate quantiles and multiple-output regression quantiles: from  $L_1$  optimization to halfspace depth. *Ann. Statist.*, 38(2):635–669.
5. Kong, L. and Mizera, I. (2012). Quantile tomography: using quantiles with multivariate data. *Statist. Sinica*, 22(4):1589–1610.
6. Wang, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Anal.*, 10(2):351–377.

# Sparse Indirect Inference

## *Inferenza indiretta sparsa*

Stolfi Paola and Bernardi Mauro and Petrella Lea

**Abstract** In this paper we propose a sparse indirect inference estimator. In order to achieve sparse estimation of the parameters, the Smoothly Clipped Absolute Deviation (SCAD)  $\ell_1$ -penalty of Fan and Li (2001) is added into the indirect inference objective function introduced by Gouriéroux et al. (1993). We derive the asymptotic theory and we show that the sparse-Indirect Inference estimator enjoys the oracle properties under mild regularity conditions. The method is applied to estimate the parameters of large dimensional non-Gaussian regression models.

**Abstract** *In questo lavoro si propone un metodo di stima indiretta sparsa. A tal fine la funzione di penalità SCAD- $\ell_1$  di Fan and Li (2001) è introdotta nella funzione obiettivo del metodo di inferenza indiretta di Gouriéroux et al. (1993). Sotto usuali condizioni di regolarità vengono inoltre dimostrate la consistenza e la Normalità asintotica unitamente alle proprietà di stimatore ORACLE. Il metodo è illustrato con l'applicazione alla stima di modelli di regressione lineare con distribuzione non-Gaussiana del termine di errore.*

**Key words:** Indirect inference; sparse regularisation; SCAD penalty, stable non-Gaussian models.

## 1 Introduction

Indirect inference (II) methods (Gouriéroux et al. 1993, Gallant and Tauchen, 1996) are likelihood-free alternatives to maximum likelihood or moment-based estimation methods for parametric inference when a closed-form expression for the density is not available. Throughout the paper we consider the following dynamic model

---

Stolfi Paola

Istituto per le Applicazioni del Calcolo “Mauro Picone” - CNR, e-mail: p.stolfi@iac.cnr.it

Bernardi Mauro

Department of Statistical Sciences, University of Padova e-mail: mauro.bernardi@unipd.it

Petrella Lea

MEMOTEF Department, Sapienza University of Rome, e-mail: lea.petrella@uniroma1.it.

$$y_t = r(y_{t-1}, \mathbf{x}_t, u_t, \vartheta) \tag{1}$$

$$u_t = \phi(u_{t-1}, \varepsilon_t, \vartheta), \tag{2}$$

where  $\mathbf{x}_t$  are exogenous variables whereas  $u_t$  and  $\varepsilon_t$  are latent variables. We assume that: (i)  $\mathbf{x}_t$  is an homogeneous Markov process with transition distribution  $F_0$  independent of  $\varepsilon_t$  and  $u_t$ ; (ii) the process  $\varepsilon_t$  is a white noise whose distribution  $G_0$  is known, and (iii) the process  $\{y_t, \mathbf{x}_t\}$  is weekly stationary. We further assume that the joint density function of the observations  $\{y_t, \mathbf{x}_t\}_{t=1}^T$  is not known analytically. The II method replaces the maximum likelihood estimator of the parameter  $\vartheta$  in equations (1)–(2) with a quasi–maximum likelihood estimator which relies on an alternative auxiliary model and then exploits simulations from the original model to correct for inconsistency. Specifically, let  $Q_T(y_T, \mathbf{x}_T, \beta)$  the auxiliary criterion function, which depends on the observations  $\{y_t, \mathbf{x}_t\}_{t=1}^T$  and on the auxiliary parameter  $\beta \in \mathbf{B} \subset \mathbb{R}^q$ , such that  $\lim_{T \rightarrow \infty} Q_T(y_T, \mathbf{x}_T, \beta) = Q_\infty(F_0, G_0, \vartheta_0, \beta)$ , a.s., where  $\vartheta_0$  is the true parameter of interest, then

$$\hat{\beta}_T = \arg \max_{\beta \in \mathbf{B}} Q_T(y_T, \mathbf{x}_T, \beta). \tag{3}$$

Under the additional assumptions that the limit criterion is continuous in  $\beta$  and has a unique maximum  $\beta_0$ , then the estimator  $\hat{\beta}_T$  is a consistent estimator of  $\beta_0$ , that is unknown since it depends on  $F_0$  and  $\vartheta_0$  that are unknown. To overcome this problem, the II method simulates, for each value of  $\vartheta$ ,  $H$  paths  $\tilde{y}_T^h$  for  $h = 1, 2, \dots, H$  and computes the QML estimate  $\tilde{\beta}_T^h$  for the auxiliary model in equation (3) and subsequently minimises the following objective function

$$\hat{\vartheta}_T = \arg \min_{\vartheta} \left( \hat{\beta}_T - \frac{1}{H} \sum_{h=1}^H \tilde{\beta}_T^h \right)' \hat{\Omega}_T \left( \hat{\beta}_T - \frac{1}{H} \sum_{h=1}^H \tilde{\beta}_T^h \right), \tag{4}$$

for an appropriately chosen positive–definite square symmetric matrix  $\hat{\Omega}_T$ . Indirect estimator are consistent and asymptotically Normal under mild regularity conditions, see Gourieroux et al. (1993). The most important condition concerns the binding function that maps the parameter space of the auxiliary model onto the parameter space of the true model

$$b(F, G, \theta) = \arg \max_{\beta \in \mathbf{B}} Q_T(F, G, \theta, \beta), \tag{5}$$

must be one–to–one. We further assume that  $\frac{\partial b}{\partial \vartheta}(F_0, G_0, \cdot)$  is of full–column rank. In the following Section we introduce the Sparse–II estimator.



### 2 Sparse indirect inference

In order to achieve sparse estimation of the parameter  $\vartheta$  we introduce the Smoothly Clipped Absolute Deviation (SCAD)  $\ell_1$ -penalty of Fan and Li (2001) into the II objective function. The SCAD function is a non-convex penalty function with the following form

$$p_\lambda (|\gamma|) = \begin{cases} \lambda |\gamma| & \text{if } |\gamma| \leq \lambda \\ \frac{1}{a-1} \left( a\lambda |\gamma| - \frac{\gamma^2}{2} \right) - \frac{\lambda^2}{2(a-1)} & \text{if } \lambda < \gamma \leq a\lambda \\ \frac{\lambda^2(a+1)}{2} & \text{if } a\lambda < |\gamma|, \end{cases} \tag{6}$$

which corresponds to quadratic spline function with knots at  $\lambda$  and  $a\lambda$ . The SCAD penalty is continuously differentiable on  $(-\infty; 0) \cup (0; \infty)$  but singular at 0 with its derivatives zero outside the range  $[-a\lambda; a\lambda]$ . This results in small coefficients being set to zero, a few other coefficients being shrunk towards zero while retaining the large coefficients as they are. The Sparse II estimator minimises the penalised II objective function, as follows

$$\hat{\vartheta}^* = \arg \min_{\vartheta} D^* (\vartheta), \tag{7}$$

where

$$D^* (\vartheta) = \left( \hat{\beta}_T - \frac{1}{H} \sum_{h=1}^H \tilde{\beta}_T^h \right)' \hat{\Omega}_T \left( \hat{\beta}_T - \frac{1}{H} \sum_{h=1}^H \tilde{\beta}_T^h \right) + n \sum_i p_\lambda (|\vartheta_i|), \tag{8}$$

where  $\hat{\Omega}_T$  is a positive-definite square symmetric matrix. A similar approach in a different context has been recently proposed by Blasques and Duplinskiy (2015).

### 3 Asymptotic theory

As shown in Fan and Li (2001), the SCAD estimator, with appropriate choice of the regularisation (tuning) parameter, possesses a sparsity property, i.e., it estimates zero components of the true parameter vector exactly as zero with probability approaching one as sample size increases while still being consistent for the non-zero components. An immediate consequence of the sparsity property of the SCAD estimator is the, so called, oracle property, i.e., the asymptotic distribution of the estimator remains the same whether or not the correct zero restrictions are imposed in the course of the SCAD estimation procedure. More specifically, let  $\vartheta_0 = (\vartheta_0^1, \vartheta_0^0)$  be the true value of the unknown parameter  $\vartheta$ , where  $\vartheta_0^1 \in \mathbb{R}^s$  is the subset of non-zero parameters and  $\vartheta_0^0 = 0 \in \mathbb{R}^{k-s}$  and let  $A = \{i : \vartheta_i \in \vartheta_0^1\}$ , we consider the following definition of oracle estimator given by Zou (2006).

**Definition 1.** An oracle estimator  $\hat{\vartheta}_{\text{oracle}}^1$  has the following properties:

- (i) consistent variable selection:  $\lim_{n \rightarrow \infty} \mathbb{P}(A_n = A) = 1$ , where  $A_n = \{i : \hat{\vartheta}_i \in \hat{\vartheta}_{\text{oracle}}^1\}$ ;
- (ii) asymptotic normality:  $\sqrt{n}(\hat{\vartheta}_{\text{oracle}}^1 - \vartheta_0^1) \xrightarrow{d} N(\mathbf{0}, \Sigma)$ , as  $n \rightarrow \infty$ , where  $\Sigma$  is the variance covariance matrix of  $\vartheta_0^1$ .

In the remainder the Section we establish the oracle properties of the penalised SCAD II estimator. To this end, the following set of assumptions are needed:

(i)

$$\xi_T = \sqrt{T} \left( \frac{\partial Q_T}{\partial \beta}(y_T, x_T, \beta_0) - \frac{1}{H} \sum_{h=1}^H \frac{\partial Q_T}{\partial \beta}(\tilde{y}_T^h, x_T, \beta_0) \right), \tag{9}$$

is asymptotically normal with mean zero, and asymptotic variance–covariance matrix given by  $W = \lim_{T \rightarrow \infty} V(\xi_T)$ ;

(ii)

$$\lim_{T \rightarrow \infty} V \left( \sqrt{T} \frac{\partial Q_T}{\partial \beta}(\tilde{y}_T^h, x_T, \beta_0) \right) = I_0, \tag{10}$$

and the limit is independent of the initial values  $z_0^h$ , for  $h = 1, 2, \dots, H$ ;

(iii)

$$\lim_{T \rightarrow \infty} \text{Cov} \left( \sqrt{T} \frac{\partial Q_T}{\partial \beta}(\tilde{y}_T^h, x_T, \beta_0), \sqrt{T} \frac{\partial Q_T}{\partial \beta}(\tilde{y}_T^l, x_T, \beta_0) \right) = K_0, \tag{11}$$

and the limit is independent of  $z_0^h$  and  $z_0^l$  for  $h \neq l$ ;

(iv)

$$\text{plim} - \frac{\partial^2 Q_T}{\partial \beta \partial \beta'}(\tilde{y}_T^h, x_T, \beta_0) = - \frac{\partial^2 Q_\infty}{\partial \beta \partial \beta'}(F_0, G_0, \vartheta_0, \beta_0), \tag{12}$$

and the limit is independent of  $z_0^h$ .

The next Theorem states that the estimator defined in equation (7) satisfies the sparsity property.

**Theorem 1.** *Given the SCAD penalty function  $p_\lambda(\cdot)$ , for a sequence of  $\lambda_n$  such that  $\lambda_n \rightarrow 0$ , and  $\sqrt{n}\lambda_n \rightarrow \infty$ , as  $n \rightarrow \infty$ , there exists a local minimiser  $\hat{\vartheta}$  of  $D^*(\vartheta)$  in (7) with  $\|\hat{\vartheta} - \vartheta_0\| = \mathcal{O}_p(n^{-\frac{1}{2}})$ . Furthermore, we have*

$$\lim_{n \rightarrow \infty} \mathbb{P}(\hat{\vartheta}^0 = 0) = 1. \tag{13}$$

The following theorem establishes the asymptotic normality of the penalised SCAD II estimator; we denote by  $\vartheta^1$  the subvector of  $\vartheta$  that does not contain zero elements and by  $\hat{\vartheta}^1$  the corresponding penalised II estimator.

**Theorem 2.** *Given the SCAD penalty function  $p_\lambda (|\vartheta_i|)$ , for a sequence  $\lambda_n \rightarrow 0$  and  $\sqrt{n}\lambda_n \rightarrow \infty$  as  $n \rightarrow \infty$ , then  $\hat{\vartheta}^1$  has the following asymptotic distribution:*

$$\sqrt{n} \left( \hat{\vartheta}^1 - \vartheta_0^1 \right) \xrightarrow{d} \mathbf{N} \left( \mathbf{0}, \left( 1 + \frac{1}{H} \right) \mathbf{W} \right), \tag{14}$$

as  $n \rightarrow \infty$ , where

$$\mathbf{W} = \left( b' (F_0, G_0, \vartheta_0)' \Omega b' (F_0, G_0, \vartheta_0) \right)^{-1} \mathbf{W}_1 \left( b' (F_0, G_0, \vartheta_0)' \Omega b' (F_0, G_0, \vartheta_0) \right)^{-1},$$

and

$$\mathbf{W}_1 = b' (F_0, G_0, \vartheta_0)' \Omega \mathbf{J}_0^{-1} (\mathbf{I}_0 - \mathbf{K}_0) \mathbf{J}_0^{-1} \Omega b' (F_0, G_0, \vartheta_0), \tag{15}$$

where  $b' (F_0, G_0, \vartheta_0) = \frac{\partial b(F_0, G_0, \vartheta_0)}{\partial \vartheta'}$  is the first derivative of the binding function  $b(F_0, G_0, \vartheta_0)$ .

### 4 Sparse II algorithm

The objective function of the sparse estimator is the sum of a convex function and a non convex function which complicates the minimisation procedure. Here, we adapt the algorithms proposed by Fan and Li (2001) and Hunter and Li (2005) to our objective function in order to allow a fast procedure for the minimisation problem. To this aim we consider the first order Taylor expansion of the penalty, for  $\vartheta_i \approx \vartheta_{i0}$

$$p_\lambda (|\vartheta_i|) \approx p_\lambda (|\vartheta_{i0}|) + \frac{1}{2} \frac{p'_\lambda (|\vartheta_{i0}|)}{|\vartheta_{i0}|} (\vartheta_i^2 - \vartheta_{i0}^2), \tag{16}$$

where the first derivative of the penalty function has been approximated as follows:

$$\left[ p_\lambda (|\vartheta_i|) \right]' = p'_\lambda (|\vartheta_i|) \text{sgn} (\vartheta_i) \approx \frac{p'_\lambda (|\vartheta_{i0}|)}{|\vartheta_{i0}|} \vartheta_i, \tag{17}$$

when  $\vartheta_i \neq 0$ . The objective function  $D^*$  in equation (7) can be locally approximated, except for a constant term by

$$\begin{aligned} D^* (\vartheta) \approx & \left( \hat{\beta} - \tilde{\beta}_{\vartheta_0}^h \right)' \hat{\Omega} \left( \hat{\beta} - \tilde{\beta}_{\vartheta_0}^h \right) - \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} \hat{\Omega} \left( \hat{\beta} - \tilde{\beta}_{\vartheta_0}^h \right) (\vartheta - \vartheta_0) \\ & + \frac{1}{2} (\vartheta - \vartheta_0)' \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} \hat{\Omega} \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} (\vartheta - \vartheta_0) + \frac{n}{2} \vartheta' \tilde{\mathbf{P}}_{\lambda_n} (\vartheta_0) \vartheta, \end{aligned} \tag{18}$$

where  $\bar{\beta}_{\vartheta_0}^h = \frac{1}{H} \sum_{h=1}^H \tilde{\beta}_{\vartheta_0}^h$  and  $\bar{\mathbf{P}}_{\lambda_n}(\vartheta) = \text{diag} \left\{ \frac{p'_{\lambda_n}(|\vartheta_i|)}{|\vartheta_i|}; \vartheta_i \in \vartheta^1 \right\}$ . Then the first order condition becomes

$$\begin{aligned} \frac{\partial D^*(\vartheta)}{\partial \vartheta} &\approx -\frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} \hat{\Omega} \left( \hat{\beta} - \bar{\beta}_{\vartheta_0}^h \right) \\ &\quad + \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} \hat{\Omega} \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} (\vartheta - \vartheta_0) + n \bar{\mathbf{P}}_{\lambda_n}(\vartheta_0) \vartheta \\ &= -\frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} \hat{\Omega} \left( \hat{\beta} - \bar{\beta}_{\vartheta_0}^h \right) + \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} \hat{\Omega} \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} (\vartheta - \vartheta_0) \\ &\quad + n \bar{\mathbf{P}}_{\lambda_n}(\vartheta_0) (\vartheta - \vartheta_0) + n \bar{\mathbf{P}}_{\lambda_n}(\vartheta_0) \vartheta_0 \\ &= 0, \end{aligned} \tag{19}$$

therefore

$$\begin{aligned} \vartheta = \vartheta_0 - &\left[ \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} \hat{\Omega} \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} + n \bar{\mathbf{P}}_{\lambda_n}(\vartheta_0) \right]^{-1} \\ &\times \left[ \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} \hat{\Omega} \left( \hat{\beta} - \bar{\beta}_{\vartheta_0}^h \right) - n \bar{\mathbf{P}}_{\lambda_n}(\vartheta_0) \vartheta_0 \right]. \end{aligned} \tag{20}$$

The optimal solution can be found iteratively, as follows

$$\begin{aligned} \vartheta^{(k+1)} = \vartheta^{(k)} - &\left[ \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta^{(k)}}^h}{\partial \vartheta} \hat{\Omega} \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta^{(k)}}^h}{\partial \vartheta} + n \bar{\mathbf{P}}_{\lambda_n}(\vartheta^{(k)}) \right]^{-1} \\ &\times \left[ \frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta^{(k)}}^h}{\partial \vartheta} \hat{\Omega} \left( \hat{\beta} - \bar{\beta}_{\vartheta_0}^h \right) - n \bar{\mathbf{P}}_{\lambda_n}(\vartheta^{(k)}) \vartheta^{(k)} \right], \end{aligned} \tag{21}$$

and if  $\vartheta_i^{(k+1)} \approx 0$ , then  $\vartheta_i^{(k+1)}$  is set equal zero. When the algorithm converges the estimator satisfies the following equation

$$\frac{1}{H} \sum_{h=1}^H \frac{\partial \tilde{\beta}_{\vartheta_0}^h}{\partial \vartheta} \hat{\Omega} \left( \hat{\beta} - \bar{\beta}_{\vartheta_0}^h \right) - n \bar{\mathbf{P}}_{\lambda_n} \vartheta_0 = 0, \tag{22}$$

that is the first order condition of the minimisation problem of the Sparse-II estimator.

### 5 Tuning parameter selection

The SCAD penalty requires the selection of two tuning parameters  $(a, \lambda)$ . The first tuning parameter is fixed at  $a = 3.7$  as suggested in Fan and Li (2001), while the parameter  $\lambda$  is selected using the cross validation function

$$CV(\lambda) = \sum_{k=1}^K \frac{1}{n_k} \left( \hat{\beta} - \frac{1}{H} \sum_{h=1}^H \tilde{\beta}_{\hat{\vartheta}_{\lambda,k}^h} \right)' \hat{\Omega} \left( \hat{\beta} - \frac{1}{H} \sum_{h=1}^H \tilde{\beta}_{\hat{\vartheta}_{\lambda,k}^h} \right), \tag{23}$$

where  $\hat{\vartheta}_{\lambda,k}$  denotes the parameters estimate over the sample  $(\cup_{i=1}^K T_k) \setminus T_k$  with  $\lambda$  as tuning parameter. Then the optimal value is chosen as  $\lambda^* = \arg \min_{\lambda} CV(\lambda)$ , where again the minimisation is performed over a grid of values for  $\lambda$ .

### 6 Application

Let  $\mathbf{y} = (y_1, y_2, \dots, y_T)'$  be the vector of observations on the scalar response variable  $Y$ ,  $\mathbf{X} = (\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_T)'$  is the  $(n \times p)$  matrix of observations on the  $p$  covariates, i.e.,  $\mathbf{x}_{j,t} = (x_{j,1}, x_{j,2}, \dots, x_{j,p})$  and consider the following regression model

$$\mathbf{y} = \iota_T \delta + \mathbf{X}\gamma + \varepsilon, \quad \varepsilon \sim S_{\alpha}(0, \sigma), \tag{24}$$

where  $\iota_T$  is the  $T \times 1$  vector of unit elements,  $\delta \in \mathbb{R}$  denotes the parameter related to the intercept of the model,  $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_p)'$  is the  $p \times 1$  vector of regression parameters and  $S_{\alpha}(0, \sigma)$  denotes the symmetric  $\alpha$ -Stable distribution (Samorodnitsky et al. 1994) centred at zero with characteristic exponent  $\alpha \in (0, 2)$  and scale parameter  $\sigma > 0$ . We further assume that the element of the vector of innovations  $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T)$  are independent i.e.  $\varepsilon_j \perp\!\!\!\perp \varepsilon_k$ , for any  $j \neq k$  and they are independent of  $\mathbf{x}_l$ , for  $l = 1, 2, \dots, p$ . Indirect inference for Stable distributions has been previously considered by Lombardi and Veredas (2009). The Sparse-II method requires the definition of the auxiliary model as well as the metric used to compare the synthetic data generated by the method  $\tilde{\mathbf{y}}$  with the true data  $\mathbf{y}$ . As auxiliary distribution we consider the Student-t regression model defined in equation (24), with the only difference that the error term follows a Student-t distribution  $\varepsilon \sim T(0, \sigma^2, \nu)$ . As regards the metric, we consider the  $L_2$  distance between the scores of the auxiliary model evaluated at the true  $\mathbf{y}$  and simulated  $\tilde{\mathbf{y}}$ , i.e.,  $\|\nabla(\hat{\beta}, \tilde{\mathbf{y}}) - \nabla(\hat{\beta}, \mathbf{y})\|_2^2$ . In Table 1, we report the empirical inclusion probabilities of the regression parameters obtained over 1,00 replications of the  $\alpha$ -Stable regression model defined in equation (24), for two values of  $\alpha = (1.70, 1.95)$  with  $n = 250$ . The true parameters are defined in the column (Par.) of Table (24), while the scale parameter of the Stable distribution is held fixed at  $\sigma = 0.05$ . Our simulation results confirm that the sparse Indirect estimator perform well in detecting zeros in linear non-Gaussian regression models.

Par.	True	EIP		Par.	True	EIP	
		$\alpha = 1.70$	$\alpha = 1.95$			$\alpha = 1.70$	$\alpha = 1.95$
$\delta$	1	0	0	$\gamma_{11}$	0	0.6591	0.8919
$\gamma_1$	2	0	0	$\gamma_{12}$	0	0.7500	0.8378
$\gamma_2$	2	0	0	$\gamma_{13}$	0	0.8182	0.9459
$\gamma_3$	3	0	0	$\gamma_{14}$	0	0.7273	0.9189
$\gamma_4$	1	0	0	$\gamma_{15}$	0	0.7955	0.9730
$\gamma_5$	2	0	0	$\gamma_{16}$	0	0.7273	0.8378
$\gamma_6$	3	0	0	$\gamma_{17}$	0	0.7727	0.8919
$\gamma_7$	1	0	0	$\gamma_{18}$	0	0.8182	0.9189
$\gamma_8$	2	0	0	$\gamma_{19}$	0	0.8636	0.9459
$\gamma_9$	3	0	0	$\gamma_{20}$	0	0.8636	1.0000
$\gamma_{10}$	0	0	0	$\gamma_{21}$	0	0.8636	0.9459

**Table 1** Empirical inclusion probabilities (EIP) evaluated over 1,00 replications for the regression parameters ( $\delta, \gamma$ ) of the  $\alpha$ -Stable regression model defined in equation (24).

## 7 Conclusion

In this paper we introduce the sparse indirect inference (SII) estimator and we extend the asymptotic theory. Empirical properties of the estimator are evaluated by means of a simulation study where a moderately large linear regression model with non-Gaussian innovations is considered. Our results confirm that the SII estimator performs well in detecting non zero regressor parameters.

## References

- Blasques, F. and Duplinskiy, A. (2015). Penalized Indirect Inference. *Tinbergen Institute, WorkingPaper*, 15-09/III.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.
- Gouriéroux, C. and Monfort, A. (1996). *Simulation-based econometric methods*. CORE lectures. Oxford University Press, 1996.
- Gouriéroux, C., Monfort, A. and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8(S1):S85–S118, 1993.
- Hunter, D.R. and Li, R. (2005). Variable selection using mm algorithms. *Ann. Statist.*, 33(4):1617–1642.
- Lombardi, M. J. and Veredas, D. (2009). Indirect estimation of elliptical stable distributions. *Comput. Statist. Data Anal.*, 6(53):2309–2324.
- Samorodnitsky, G. and Taqqu, M. S. (1994). Stable non-Gaussian random processes. *Chapman & Hall, New York*, xxii+632.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 476(101):1418–1429.

# The ESSnet Big Data: Experimental Results

## *Gli ESSnet Big Data: risultati sperimentali*

Peter Struijs<sup>1</sup>, Anke Consten, Piet Daas, Marc Debusschere, Maiki Ilves, Boro Nikic, Anna Nowicka, David Salgado, Monica Scannapieco, and Nigel Swier

**Abstract:** In the ESSnet Big Data, 22 partners from 20 countries of the European Statistical System, the ESS, collaborate in exploring the possibilities of using big data as a source for official statistics. The research focuses on seven areas: (1) web scraping for statistics on job vacancies, (2) web scraping for obtaining enterprise characteristics, (3) use of smart meter data, (4) use of AIS data (maritime data), (5) use of mobile phone data, (6) use of big data for early estimates, and (7) combining data sources for statistics on the domains of population, tourism and border crossing, and agriculture. The paper presents the main results of the ESSnet after its first year, highlighting the opportunities and challenges of using big data for official statistics.

**Key words:** Big Data, European Statistical System, ESSnet Big Data, Official Statistics

## 1 Big Data and the European Statistical System

The emergence of big data is having a big impact on organisations for which the production and analysis of data and information is core business. National Statistical Institutes (NSIs), which are responsible for official statistics, are such organisations. Official statistics are heavily used by policy makers and society as a whole, and the way NSIs take up big data will eventually have implications for all of society.

The relevance of big data for official statistics stems from the exponential increase of data registered through networks of sensors, camera's, public

---

<sup>1</sup> Contact: Peter Struijs, Statistics Netherlands, [p.struijs@cbs.nl](mailto:p.struijs@cbs.nl). The co-authors are all work package leaders of the ESSnet.

administrations, banks, enterprises, mobile networks, satellites, drones, social networks, internet sites, etc. This not only creates many opportunities for improving official statistics, such as reporting on phenomena whose measurement used to be out of reach, but also profoundly influences the context in which statistics are produced, for better or worse. And there are many issues with big data that may have an impact on NSIs, such as on the required statistical methodology, the way data is obtained, privacy and ethical considerations, the need for an appropriate IT infrastructure, the skills needed to deal with big data, the quality of statistics based on big data, and the positioning of NSIs in the emerging data society. Official statistics are generally based on established, validated methods, but for big data new approaches are clearly needed [5].

The possible strategic impact of big data for official statistics was recognised by several NSIs and international organisations some years ago. In September 2013 the Directors-General of the NSIs of the European Statistical System (ESS), adopted the so-called Scheveningen Memorandum on Big Data and Official Statistics [1], in which a course of action was set out, including the drafting of an ESS action plan and a roadmap. The action plan, which has been worked out in the context of the ESS Vision 2020 [2], identifies nine themes: policy; quality; skills; experience sharing; legislation; IT infrastructure; methods; ethics/communication; and partnerships. It also recognised the need for carrying out concrete pilots. For the pilots a so-called ESSnet was created, the results of which are the subject of this paper.

But what is, in fact, big data in the context of official statistics? This question was also considered at the international level. In official statistics, big data is generally considered as a data source. An attempt to define big data for statistical purposes was made by UNECE, the UN Economic Commission for Europe. Building on a definition by Gartner [4] it defined big data as follows [3]: “Big data are data sources that can be – generally – described as: high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making.”

## **2 The ESSnet Big Data**

An ESSnet is a project in which a number of ESS partners collaborate in order to pursue an ESS goal, with partial EU-funding. Usually, a so-called Framework Partnership Agreement (FPA) is established, after which one or more so-called Specific Grant Agreements (SGAs) are concluded. In the case of the ESSnet Big Data (henceforth: the ESSnet), the FPA is linked to a consortium of 22 partners, consisting of 20 National Statistical Institutes (NSIs) and two Statistical Authorities, and covers the period from January 2016 to May 2018. There are two SGAs, with an overlap in time: SGA-1 extends from February 2016 to July 2017, and SGA-2 from January 2017 to May 2018. For each of them, the grant has a value of 1 million euro, and funding is limited to 90%.



The overall objective of the project is to prepare the ESS for integration of big data sources into the production of official statistics. The consortium has organised the core of its work around a number of work packages (WPs), each WP dealing with one pilot and a concrete output. In SGA-1 there are seven of them: WP 1 Web Scraping / Job Vacancies; WP 2 Web Scraping / Enterprise Characteristics; WP 3 Smart Meters; WP 4 AIS Data; WP 5 Mobile Phone Data; WP 6 Early Estimates; WP 7 Multi Domains

The outputs of these pilots so far are described in the remainder of this paper. They have one thing in common: they cover the complete statistical process, from data acquisition to the production of statistical output. In addition, and in accordance with the general objective of the FPA, the pilots also consider future perspectives. Thus, all pilots comprise the following five phases:

1. Data access
2. Data handling
3. Methodology and technology
4. Statistical output
5. Future perspectives

SGA-1 covers only some of the five phases for each of the WPs, the rest being covered by SGA-2. And the phases covered by SGA-1 are not the same for each pilot (WP), as for some areas it was possible to plan ahead further (in time and phases) than for other areas. In particular, WP 5 concentrated on data access issues in SGA-1 and could not plan further ahead, as data processing would depend on the results of the efforts to realise data access. Therefore, WP 5 was planned to end in December 2016, whereas the other WPs would continue into 2017. This explains the overlap in time of SGA-1 and SGA-2.

Given the overall objective, the findings need to be generalised. This is done in SGA-2, for which a new WP is added, WP 8. This WP covers overarching aspects, in particular methodology, quality and IT infrastructure.

The ESSnet has organised support in several ways. In addition to the IT infrastructure of NSIs, common IT facilities were ensured by subscribing to the UNECE Sandbox, a facility maintained by the Central Statistics Office (CSO) of Ireland and the Irish Centre for High-End Computing (ICHEC)<sup>2</sup>. In order to ensure that professional standards are met, a Review Board was created that systematically reviews the products of the ESSnet. Concerning communication and dissemination, the ESSnet uses a Mediawiki website<sup>3</sup>. CROS Portal, the general ESS dissemination site, also links to the products of the ESSnet<sup>4</sup>. The products described in the sections below can be found there. After the first year of the ESSnet, a dissemination workshop<sup>5</sup> was held in Sofia, Bulgaria.

---

<sup>2</sup> <http://www1.unece.org/stat/platform/display/bigdata/Sandbox>

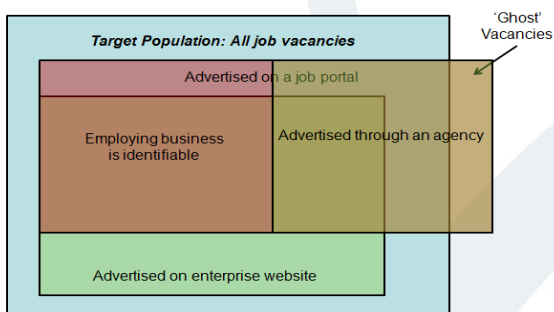
<sup>3</sup> <https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/>

<sup>4</sup> [https://ec.europa.eu/eurostat/cros/content/essnetbigdata\\_en](https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en)

<sup>5</sup> [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Project\\_meetings](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Project_meetings)

### 3 Web Scraping: Job Vacancies

Since this pilot involves each country taking its own specific approach, there are a lot of intermediate country specific results. However, general selection criteria have been identified for targeting portals for scraping. Taking into account the distinction between job vacancy and job advertisement, a conceptual model is proposed of how online job advertisements correspond to the target population (see Figure 1). In practical terms this may be defined as all vacancies that are available to be measured by existing job vacancy surveys. As well as providing a conceptual framework for understanding the coverage of job vacancies from online sources and how these relate to the measurement of all job vacancies, this approach may also provide the conceptual basis for an estimation framework.



**Figure 1:** Conceptual model for measuring job vacancies from on-line sources

### 4 Web Scraping: Enterprise Characteristics

Six use cases have been identified in the pilot: (1) enterprise URLs inventory, (2) e-commerce in enterprises (about predicting whether or not an enterprise provides web sales facilities on its website), (3) job vacancies ads on enterprises' websites, (4) social media presence on enterprises webpages, (5) sustainability reporting on enterprises' websites (linked to the UN Sustainability Development Goals), and (6) relevant categories of enterprises' activity sector (NACE) aimed at checking or completing statistical business registers. A common use case template was developed and has been used. For each use case, a pilot definition was performed and all of them were mapped to a general "logical architecture" (see Figure 2). Also, a report was produced on legal aspects related to web scraping of enterprise websites, aimed at showing the real possibilities for the NSIs to perform activities of web scraping. These appear to be generally favourable.

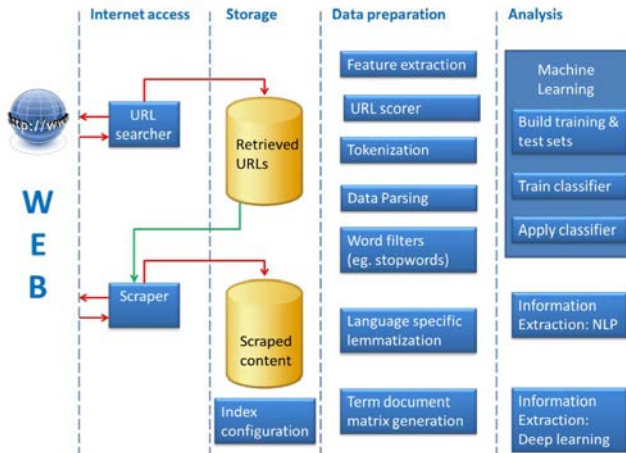


Figure 2: Logical architecture for software for web scraping of enterprise websites

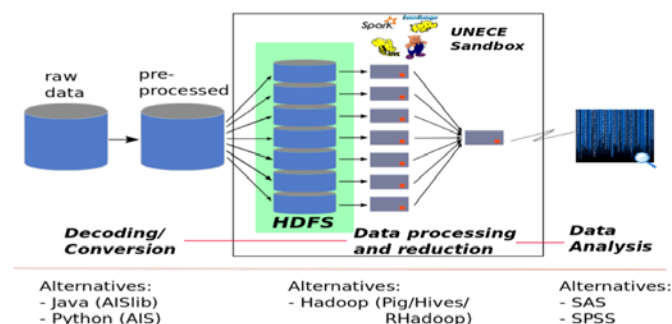
## 5 Smart Meters

This pilot has produced a report on data access and data handling of smart meter electricity data. The report includes the results of a literature study and a survey on access to smart meter data, which was sent to the NSIs of all EU member countries in the spring of 2016, with 18 responses. It appeared that only two countries currently have access to data: Denmark and Estonia. Several countries were aware of substantial legal barriers, and it was unclear if market participants could even share data with each other. Some countries such as Poland are in the process of drawing up legislation that will enable smart meter data use. For two countries, Estonia and Denmark, the pilot has defined and assessed the quality of smart meter electricity data, and a synthetic dataset was analysed as well, aimed at generating demo output and developing and testing statistics and algorithms for situations where linkage to enterprise or household characteristics is necessary. The assessment of quality indicators comprised: under- and overcoverage; percentage of units that fail checks; percentage of units that are adjusted; percentage of units that are imputed; data periodicity; delay in data provision

## 6 AIS Data

This pilot investigates whether real-time measurement data of ship positions (measured by the so-called AIS system) can be used to improve the quality and

internal comparability of existing statistics and for new statistical products relevant for the ESS. Reports were produced on (1) the possibilities and pitfalls of creating a database with AIS data for official statistics, (2) deriving harbour visits and linking data from maritime statistics with AIS data, and (3) sea traffic analyses using AIS data. While the possibility of using AIS data from EMSA, the European Maritime Safety Authority, is still being investigated, AIS data from Dirkzwager, a private company, was used, and the data quality analysed. Visualisations were made, showing the coverage of the ships by the data, and showing the path of a ship through time. A method to build a reference frame of maritime ships was developed and software options considered (see Figure 3).



**Figure 3:** pre-processing, processing and storing the AIS data

First results show that AIS data can be used as a backbone for maritime statistics. This is important, since the added value of running a pilot with AIS data at European level is linked to the fact that the source data is generic worldwide and data can be obtained at European level.

## 7 Mobile Phone Data

This pilot has focussed on data access, which will be needed for SGA-2. A preliminary analysis of the issues regarding the access to mobile phone data was made, which was the basis for the design of a questionnaire surveying the status of this access across the ESS. Belgium, Finland, France, and Italy were found to have succeeded in their negotiations to have access to a concrete mobile phone data set that can be used for SGA-2. Spain and Romania are still under contact pursuing this goal. A workshop was held in Luxembourg to bring together mobile network operators (MNOs), NSIs, Eurostat (the statistical office of the EU) and other stakeholders, including some other international organizations (UN, OECD, ITU, DG Connect, DG Digit). Basically five main groups of issues were identified regarding the access to mobile phone data, namely (i) the characteristics of the

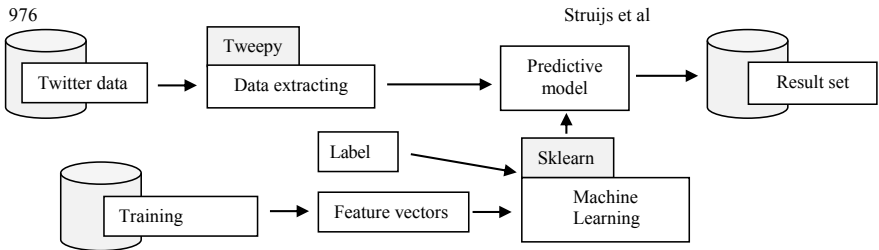
MNO, (ii) the legal requirements, (iii) the access conditions, (iv) the data characteristics and (v) other aspects.

## **8 Early Estimates**

The aim of this pilot is to investigate how a combination of multiple big data sources and existing official statistical data can be used in order to create existing or new early estimates for statistics. Two pilots were chosen. The first was web-based sales inquiries for the aim of nowcasts of turnover indices. In this context, Statistics Finland has examined the nowcasting performance of large dimensional models for the year-on-year growth rate of the turnover indexes, considering the main sectors of the Finnish economy. The Slovenian NSI has prepared and tested nowcasting methods on their data, and has also prepared the application in R environment (together with a manual), which will enable also other countries to test their data on the nowcasting model. The second pilot concerned social media data, newsfeeds and survey data for the aim of a Consumer Confidence Index (CCI). Furthermore, for eight statistical domains the potential of combining multiple big data sources and existing official statistical data was investigated in order to create existing or new early estimates for statistics. It was concluded that the most promising statistics for which early estimates could be produced are statistic related to early economic estimates. A proposal along these lines was prepared for SGA2.

## **9 Multi Domains**

The aim of this pilot is to find out how a combination of big data sources, administrative data and statistical data may enrich current statistical output. Three statistical domains are being investigated: (1) population, (2) tourism/border crossings, and (3) agriculture. For population, three areas are looked at: daily (life) satisfaction, the moods of population associated with public events (e.g., Brexit, voting), and morbidity areas (e.g., flu). For tourism/border crossings, a number of possible data sources have been identified and investigated, for instance with regard to traffic intensity information. For agriculture, the focus is on satellite data applications, in particular monitoring of crop conditions, seasonal changes, soil properties and mapping tillage activities. For each (sub)domain a big data framework is developed (see Figure 4).



**Figure 4:** Big data framework for daily (life) satisfaction

## 10 Outlook

In the remaining time of the ESSnet, the pilots will deal with the phases not covered so far, which implies less focus on data access and more focus on statistical outputs. In particular, attention will be paid to the meaning of the results for the future of ESS statistics. The pilot for mobile phone data is the only one that did not involve actual data handling in SGA-1, as data access issues had to be solved first. This is generally considered to be one of the most promising new data sources for statistics.

Apart from the results of the individual pilots, the ESSnet is also going to draw conclusions about methodology, quality and IT infrastructure for big data in general, building on the results of the seven pilots as well as on experiences and results described in the literature. Fascinating questions are to what extent the more traditional, established corpus of statistical methods can be deployed or needs to be supplemented with new approaches when dealing with big data, and to what extent the findings on methodology, quality and IT infrastructure are source dependent or can be generalised.

## References

1. DGINS: Scheveningen Memorandum on Big Data and Official Statistics (2013) Link: <http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13>
2. ESS Committee: ESS Vision 2020, Building the Future of European Statistics (2014) Link: <http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020>
3. Glasson, M., Trepanier, J., Patrino, V., Daas, P., Skaliotis, M., Khan, A.: What does "Big Data" mean for Official Statistics? Paper for the High-Level Group for the Modernization of Statistical Production and Services (2013) Link: <http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614>
4. Laney, D.: The importance of big data: a definition. Gartner Inc.(2012)
5. Struijs, P.: Big Data for Official Statistics, Game-Changer for National Statistical Institutes? Paper prepared for the 29th International Statistical Seminar, Vitoria, Spain (2016) Link: [http://www.eustat.eus/productosServicios/datos/58\\_Big\\_Data\\_for\\_Official\\_Statistics\\_Peter\\_Struijs.pdf](http://www.eustat.eus/productosServicios/datos/58_Big_Data_for_Official_Statistics_Peter_Struijs.pdf)

# Smart view selection in multi-view clustering

## *Selezione intelligente di viste per clustering multi-vista*

J r mie Sublime

**Abstract** Multi-view clustering is a data mining task in which a data set is processed by several algorithms observing different features of the same data. The main difficulty of this task is to detect whether or not sharing informations between the views may be beneficial: Some views contain mostly noisy features, while others simply contain features which lead to different clusters. One of the challenges of multi-view clustering is therefore to find which views should work together or not. Within this context, in this article we propose an optimisation method which sets the exchange weights between the different algorithms based on the maximization of the global likelihood function.

**Abstract** *Il clustering multi vista   una tecnica di data mining in cui un insieme di dati viene elaborato da diversi algoritmi analizzando diverse caratteristiche (o "viste") degli stessi dati. La difficolt  principale di questa tecnica   la capacit  di rilevare quali informazioni possono essere utili da condividere tra i diversi algoritmi: Alcune viste contengono per lo pi  rumore, mentre altre contengono semplicemente le caratteristiche specifiche dei diversi cluster. Una delle sfide del clustering multi-vista  , quindi, quella di trovare quali viste dovrebbero essere messe in comune e quali no. Il presente articolo propone un metodo di ottimizzazione basato sulla massimizzazione della funzione di verosimiglianza globale, ottenuto assegnando il peso di scambio tra i vari algoritmi.*

**Key words:** Multi-view clustering, optimization

## 1 Introduction

Data clustering is an exploratory data mining task that aims at discovering hidden intrinsic structures in a data set by forming groups (clusters) of objects that share similar features. Data clustering is usually considered more difficult than supervised classification because of its exploratory nature which makes the results difficult to rate. Nowadays, with data being abundant, most data sets tend to exist based on different representations. This gave birth to the field of multi-view learning, a re-

---

J r mie Sublime

LISITE Laboratory, RDI Team - ISEP Paris, France, e-mail: jeremie.sublime@isep.fr

cent paradigm which consists in learning and analyzing data using several views with redundant features. While this increased number of information has proved very beneficial in the context of supervised learning, multi-view clustering remains problematic in the sense that since it is difficult to assess the quality of a clustering result in an unsupervised context, it is equally difficult to rate the quality of a view and therefore to know whether or not an exchange of information between different views will be beneficial or detrimental.

Several unsupervised multi-view methods are available in the literature [13, 11, 2] with applications ranging from the clustering of distributed data [9, 3] to collaborative clustering [12, 6]. All of these models have in common that they mention the importance of properly weighting the exchange links between the different views in order to avoid "negative collaborations" [4, 10, 6, 5]. These so called negative collaborations may come from different reasons such as the lack of common structures in the different views, or a too large number of noisy features in some of them.

Within this context, the aim of this article is to propose an optimization method to detect which views should or should not exchange their information with the goal of minimizing the risk of negative collaborations.

## 2 General form for multi-view clustering likelihood functions

Let us define  $J$  algorithms  $\{\mathcal{A}^1, \dots, \mathcal{A}^J\}$ , with each of them processing a different view among  $\{X^1, \dots, X^J\}$ . These views describe the same  $N$  elements with different real features that may be redundant from one view to another:  $\forall i \quad X^i \subseteq X, X^i = \{x_1^i, \dots, x_N^i\}, x_i \in \mathbb{R}^{d_i}$ . From there, each algorithm  $\mathcal{A}^i$  tries to find a local solution  $S^i$  (hard or fuzzy) based on explicit or implicit distribution parameters  $\Theta^i$  and using information exchanged with the other views. Each algorithm is therefore defined by its subset, its solution and its distribution parameters. To this simple model, we add the exchange weights  $\tau_{j,i}$  that define how much information will be transferred from algorithm  $\mathcal{A}^j$  to algorithm  $\mathcal{A}^i$ . Therefore, we have:  $\mathcal{A}^i = \{X^i, S^i, \Theta^i, \tau_{\cdot,i}\}$ .

Using this model, most multi-view model found in the literature [11, 2, 9, 12, 6] try to optimize a variant of the fitness function given in Equation (1), where  $\mathcal{L}(X^i, S^i, \Theta^i)$  is a local term usually derived from a log-likelihood or a quality index specific to each algorithm  $\mathcal{A}^i$ , and  $\Delta_{i,j}$  is a divergence term assessing the pairwise difference or divergence between the models or partitions of two algorithms  $\mathcal{A}^i$  and  $\mathcal{A}^j$ .

$$\{S_{opt}, \Theta_{opt}\} = \operatorname{argmax}_{S, \Theta} \sum_{i=1}^J \left( \mathcal{L}(X^i, S^i, \Theta^i) - \sum_{j \neq i} \tau_{j,i} \cdot \Delta_{i,j} \right) \quad (1)$$

Depending on the multi-view framework, the divergence term  $\Delta_{i,j}$  may be based on the partitions [12]:  $\Delta_{i,j} = \Delta(S^i, S^j)$ , or on the distribution parameters and prototypes:  $\Delta_{i,j} = \Delta(\Theta^i, \Theta^j)$ ; and may regardless be concretely computed as an entropy, a Kullback-Leibler divergence, or a custom distance function between prototypes.



Please note that depending on the function used, the quality  $\Delta_{i,j} = \Delta_{j,i}$  is not always true.

As stated in the introduction, in this article we will mostly be interested in the exchange weights  $\tau_{j,i}$  and how to set them up to ensure the best possible results. The available literature on this subject is rather slim when it comes to unsupervised multi-view or collaborative learning: In some works, the problem is simply ignored and the weights set to 1 by default [2, 12], while other works describe setting up the weights manually [6]. Finally, in the work closest to this article, a method applicable to several collaborative clustering algorithms is proposed relying on the quality and diversity of the partitions to set up the weights based on regression computed on points cloud [10].

The weakness of this later approach is that it relies on clustering internal indexes that may be biased toward certain views or algorithms depending on the index and distance function used. Furthermore, the specific adjustment of the parameters depends on the data themselves and is subject to trials and errors. To solve this problem, in the next section we will propose a sound mathematical-based approach relying solely on the fitness function described in Equation (1) to optimize the exchange links. Our proposed approach therefore has the advantage to be more generic and unbiased.

### 3 Optimization under KKT conditions

The method that we propose consists in optimizing Equation (1) under the Karush-Kuhn-Tucker conditions (KKT) [7], with the goal of finding the ideal  $\tau_{ji}$ . In a second step, we will interpret the expression found with the aforementioned method.

We first want to make the hypothesis that all divergence terms  $\Delta_{i,j}$  are normalized between 0 and 1. Therefore, for any divergence term, there is an opposite consensus term  $C_{i,j}$  such that  $\Delta_{i,j} = 1 - C_{i,j}$ . If we inject this term into Equation (1), we get Equation (2) bellow.

$$\{S_{opt}, \Theta_{opt}\} = \underset{S, \Theta}{\operatorname{argmax}} \sum_{i=1}^J \left( \mathcal{L}(X^i, S^i, \Theta^i) - \sum_{j \neq i} \tau_{j,i} + \sum_{j \neq i} \tau_{j,i} \cdot C_{i,j} \right) \quad (2)$$

From there, finding the  $\tau_{ji}$  that maximize this equation with the constraint  $\sum_{j \neq i}^J \tau_{j,i} = 1$  gives us the following system:

$$\begin{cases} T = \operatorname{argmax}_T \sum_{i=1}^J \sum_{j \neq i} \tau_{j,i} \cdot C_{i,j} \\ \forall i \quad \sum_{j \neq i}^J \tau_{j,i} = 1 \\ \forall (i, j) \quad \tau_{j,i} \geq 0 \end{cases} \quad (3)$$

Note that the middle term from Equation (2) has been discarded because it does not depend on the data or partitions, and is constant under the second constraint on the weights. In the same way, the first term does not contain any  $\tau_{j,i}$  and is therefore

also removed from the problem. From this system, by using the Lagrange multipliers we get the following KKT conditions :

$$\forall (i, j), i \neq j \begin{cases} (a) & \tau_{j,i} \geq 0 \\ (b) & \sum_{j \neq i}^J \tau_{j,i} = 1 \\ (c) & \lambda_{j,i} \geq 0 \\ (d) & \tau_{j,i} \cdot \lambda_{j,i} = 0 \\ (e) & C_{i,j} - \lambda_{j,i} - v_i = 0 \end{cases} \quad (4)$$

With (c) and (e), we have:

$$\lambda_{j,i} = C_{i,j} - v_i \geq 0 \quad (5)$$

Let us suppose that there is a  $k$  so that  $\tau_{k,i} > 0$ . Then with (d), we have:  $\lambda_{k,i} = 0$ . And with Equation (5), we have:  $v_i = C_{i,k}$ . Then, using this information we can say that:

$$\forall j \neq k \begin{cases} \tau_{j,i} \neq 0 \implies C_{i,j} = C_{i,k} \implies \lambda_{j,i} = 0 \\ \tau_{j,i} = 0 \implies \lambda_{j,i} = C_{i,j} - C_{i,k} \geq 0 \end{cases} \quad (6)$$

From the second line of Equation (6), we can conclude the following:

$$\tau_{j,i} \neq 0 \implies C_{i,j} = \max_k C_{i,k} \quad (7)$$

Then, if we use (d) and (e), we have:

$$\tau_{j,i}(C_{i,j} - v_i) = 0 \quad (8)$$

If we sum Equation (8) over  $j$  and use (b), we have:

$$v_i = \sum_{j \neq i}^J \tau_{j,i} \cdot C_{i,j} \quad (9)$$

For Equation (9) to be correct while respecting the constraints given in Equations (6) and (7), the only solution is:

$$\forall j \neq i, \tau_{j,i} = \begin{cases} \frac{1}{\text{Card}(C_{i,j=\max_k}(C_{i,k}))} & \text{if } C_{i,j} = \max_k(C_{i,k}) \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

It is possible to get a relaxed version of this result by modifying the system from Equation (3) as follows:

$$\begin{cases} T = \operatorname{argmax}_T \sum_{i=1}^J \sum_{j \neq i} \tau_{j,i} \cdot C_{i,j} \\ \forall i \quad \sum_{j \neq i}^J (\tau_{j,i})^p = 1, \quad p \in \mathbb{N}^*, \quad p > 1 \\ \forall (i, j) \quad \tau_{j,i} \geq 0 \end{cases} \quad (11)$$

It is worth noting that using this new system does not exactly optimize Equation (2) since the middle term cannot be ignored using the new constraints. Instead, we are trying to optimize Equation (12), which can be seen as a "conjugate problem in spirit" since we want to favor the consensus instead of penalizing the divergences between models.

$$\{S_{opt}, \Theta_{opt}\} = \operatorname{argmax}_{S, \Theta} \sum_{i=1}^J \left( \mathcal{L}(X^i, S^i, \Theta^i) + \sum_{j \neq i} \tau_{j,i} \cdot C_{ij} \right) \quad (12)$$

However, mathematically speaking the  $\tau_{j,i}$  of this system are not the same ones as in Equations (1) and (2). For the sake of completeness we will compute them and show that they can still be used in the original equation. With this system, we get the following new KKT conditions:

$$\forall (i, j), i \neq j \begin{cases} (a) & \tau_{j,i} \geq 0 \\ (b) & \sum_{j \neq i}^J (\tau_{j,i})^p = 1, \quad p > 1 \\ (c) & \lambda_{j,i} \geq 0 \\ (d) & \tau_{j,i} \cdot \lambda_{j,i} = 0 \\ (e) & C_{i,j} - \lambda_{j,i} - v_i \cdot (p \cdot (\tau_{j,i})^{p-1}) = 0 \end{cases} \quad (13)$$

If we consider the case  $\tau_{j,i} \neq 0$  and  $\lambda_{j,i} = 0$  in Equation (d), then with (e) we have:

$$\tau_{j,i} = \left( \frac{C_{i,j}}{p \cdot v_i} \right)^{\frac{1}{p-1}} \quad (14)$$

From Equation (14) and (b), we have:

$$1 = (p \cdot v_i)^{\frac{-p}{p-1}} \sum_{j \neq i} (C_{i,j})^{\frac{p}{p-1}} = (v_i)^{\frac{-p}{p-1}} \sum_{j \neq i} \left( \frac{C_{i,j}}{p} \right)^{\frac{p}{p-1}} \quad (15)$$

Then we can write:

$$v_i = \frac{1}{p} \left( \sum_{j \neq i} (C_{i,j})^{\frac{p}{p-1}} \right)^{\frac{p-1}{p}} \quad (16)$$

Then by injecting the expression of  $v_i$  into Equation (14),  $\forall (i, j), i \neq j, p > 1$  we have:

$$\tau_{j,i} = \frac{(C_{i,j})^{\frac{1}{p-1}}}{\left( \sum_{k \neq i}^J (C_{i,k})^{\frac{p}{p-1}} \right)^{\frac{1}{p}}} \quad (17)$$

## 4 Interpretation

In this section, we will interpret and explain the results from Equations (10) and (17).

We begin by analyzing the first result which basically says that in the most constrained case, each view of a multi-view framework should only acquire information from the view that has the closest model and not at all from the other views. The model specifies that in case several views have equally close models, information should be acquired from all of them with an equal exchange link. We can reasonably say that mostly pairwise exchanges between most similar views is very restricting, hence our proposal of a relaxed version of the weights.

In Equation (17), we have a second result based on the conjugate problem favorising the consensus rather than penalizing the differences. The idea is that while the exchange link should still be primarily stronger between views that have similar models, depending on the value of  $p$  the information coming from divergent models should be given some importance too based on their degree of divergence. In fact, when  $p$  grows towards infinity all weights would become equal.

- For  $p > 1$ :

$$\forall j \neq i, \tau_{j,i} = \frac{|C_{i,j}|^{\frac{1}{p-1}}}{\left(\sum_{k \neq i}^J |C_{i,k}|^{\frac{p}{p-1}}\right)^{\frac{1}{p}}} \quad (18)$$

- When  $p \rightarrow \infty$ :

$$\forall (i, j), \tau_{j,i} = Cte \quad (19)$$

While this interpretation seems to be in the continuity of the result for  $p = 1$  in Equation (10), as stated earlier, these weights were computed for the conjugate problem shown in Equation (12) which is consensus based instead of divergence based, and we shall therefore prove that these weights are still applicable and make sense for the original model from Equation (1) which is the one used in most multi-view algorithms.

Under the hypothesis that  $0 \leq \Delta_{i,j} = 1 - C_{i,j} \leq 1$ , when we inject these weights into Equation (1), we have the following interpretation: The exchange link should be stronger when acquiring information from views that have models with the lowest divergences. Therefore the optimization process used to find the partitions and in fine the model would have two interesting properties:

1. It would attempt to reduce the divergences in model between similar partitions.
2. It would reduce the exchanges between too dissimilar partitions or models.

The first property is interesting because it relates with the concept of stability in clustering partitions [1, 8]. Indeed, by encouraging views with similar results to work together in an unsupervised context, it increases the chances of achieving better results, since structures found in several views can be considered stable which is a good and unbiased indication for exploratory data mining.

The second and first properties together have direct applications to tackle the problem of feature selection. If we consider multi-view data where the same objects

are represented using different redundant views, we know that the two main problems are: views that contain mostly noise, views or groups of views that contains different and incompatibles data structures. Using our proposed weighting system, the noisy views would not hinder the others by sending wrong information because they are too different and will have low exchange links toward every other views. As for groups of views containing different structures, our weighting system would mostly foster exchanges between similar views, thus creating meta-clusters of similar views. These meta-clusters of views would still exchange information through the views that are compatible with several meta-clusters, and noisy views would remain isolated as outliers.

Using the graph of the weights  $\tau_{j,i}$ , it would be very simple to detect views containing mostly noisy features, but also to remove redundant attributes by detecting communities of hyper-connected views forming clusters in the graph and removing some of them.

If we had attempted the same optimization based on the original Equation (1) using  $\Delta_{i,j}$  instead of  $C_{i,j}$ , it would have led to  $p = 2$  being the only valid constraint on the weights, and on the solution given in Equation (20).

$$\tau_{j,i}^* = \frac{-\Delta_{i,j}}{\sqrt{\sum_{k \neq i}^J \Delta_{i,k}^2}} \quad (20)$$

While this result is mathematically correct, it leads to negative weights the absolute value of which is the highest between views that have most divergent models. Compared with the result from Equation (17), this solution offers little interest from a multi-view perspective because a negative exchange of information cannot practically translated in a multi-view framework. Furthermore, this model is difficult to interpret and leads to close to weak exchange weights between similar and mildly similar solutions. This second point is also problematic from a multi-view perspective where the views are supposed to exchange with a goal of mutual improvement.

## 5 Conclusion

In this article, we have proposed an optimization method to set up the exchange weights between views in the context of unsupervised multi-view clustering. Our method relies on the optimization of a conjugate fitness function under the Karush-Kuhn-Tucker conditions and provides an analytic expression for the weights.

Our results are interesting in the sense that the resulting weights rely on the unbiased notion of clustering stability instead of regular clustering indexes, and they are applicable to most multi-view framework unlike others weighting methods in the literature. Furthermore, our method creates meta-clusters of views which makes it possible to regroup views containing similar features that lead to different structures

thus enabling the detection of redundant attributes, and more importantly it makes it easy to detect views with mostly noisy features.

Since this work is only a preliminary theoretical background, in our future works we look forward to apply it to various multi-view methods and see how it fares in practice.

## References

1. Ben-David, S., von Luxburg, U., Pal, D.: A sober look at clustering stability. In: G. Lugosi, H. Simon (eds.) *Learning Theory, Lecture Notes in Computer Science*, vol. 4005, pp. 5–19. Springer Berlin Heidelberg (2006)
2. Bickel, S., Scheffer, T.: Estimation of mixture models using co-em. In: J. Gama, R. Camacho, P. Brazdil, A. Jorge, L. Torgo (eds.) *Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings, Lecture Notes in Computer Science*, vol. 3720, pp. 35–46. Springer (2005)
3. Depaire, B., Falcon, R., Vanhoof, K., Wets, G.: PSO Driven Collaborative Clustering: a Clustering Algorithm for Ubiquitous Environments. *Intelligent Data Analysis* **15**, 49–68 (2011)
4. Falcon, R., Jeon, G., Bello, R., Jeong, J.: Learning collaboration links in a collaborative fuzzy clustering environment. In: *Proceedings of the artificial intelligence 6th Mexican international conference on Advances in artificial intelligence, MICAI'07*, pp. 483–495. Springer-Verlag, Berlin, Heidelberg (2007). URL <http://portal.acm.org/citation.cfm?id=1775967.1776018>
5. Grozavu, N., Cabanes, G., Bennani, Y.: Diversity analysis in collaborative clustering. In: *2014 International Joint Conference on Neural Networks, IJCNN 2014, Beijing, China, July 6-11, 2014*, pp. 1754–1761 (2014)
6. Grozavu, N., Ghassany, M., Bennani, Y.: Learning confidence exchange in collaborative clustering. In: *IJCNN*, pp. 872–879 (2011)
7. Kuhn, H.W., Tucker, A.W.: Nonlinear programming. In: B.U. of California Press (ed.) *Proceedings of 2nd Berkeley Symposium*, pp. 481–492 (1951)
8. von Luxburg, U.: Clustering stability: An overview. *Foundations and Trends in Machine Learning* **2**(3), 235–274 (2010)
9. Pedrycz, W.: Collaborative fuzzy clustering. *Pattern Recognition Letters* **23**(14), 1675–1686 (2002)
10. Rastin, P., Cabanes, G., Grozavu, N., Bennani, Y.: Collaborative clustering: How to select the optimal collaborators? In: *IEEE Symposium Series on Computational Intelligence, SSCI 2015, Cape Town, South Africa, December 7-10, 2015*, pp. 787–794. IEEE (2015). DOI 10.1109/SSCI.2015.117. URL <http://dx.doi.org/10.1109/SSCI.2015.117>
11. Sublemontier, J.: Unsupervised collaborative boosting of clustering: An unifying framework for multi-view clustering, multiple consensus clusterings and alternative clustering. In: *The 2013 International Joint Conference on Neural Networks, IJCNN 2013, Dallas, TX, USA, August 4-9, 2013*, pp. 1–8. IEEE (2013). DOI 10.1109/IJCNN.2013.6706911. URL <http://dx.doi.org/10.1109/IJCNN.2013.6706911>
12. Sublime, J., Grozavu, N., Bennani, Y., Cornu ejols, A.: Collaborative clustering with heterogeneous algorithms. In: *2015 International Joint Conference on Neural Networks, IJCNN 2015, Killarney, Ireland, July 12-18, 2015* (2015)
13. Zimek, A., Vreeken, J.: The blind men and the elephant: on meeting the problem of multiple truths in data from clustering and pattern mining perspectives. *Machine Learning* **98**(1-2), 121–155 (2015). DOI 10.1007/s10994-013-5334-y. URL <http://dx.doi.org/10.1007/s10994-013-5334-y>

# Social Sensing and Official Statistics: call data records and social media sentiment analysis

## *Social Sensing e Official Statistics: analisi di dati telefonici e misure di sentiment dai social media*

Emilio Sulis

**Abstract** This contribution explores the relationship between indicators from official statistics and measures derived from new data sources. In particular, phone call data from Italy's leading phone company offer suggestions on patterns of behavior, as well as on the demographic composition of municipalities. A classification experiment based on ego-network measures reaches an F-measure accuracy of 0.68 (baseline: 0.63) in distinguishing high and low presence of migrants in Italian municipalities. In addition, this paper explores Sentiment Analysis to investigate the content of online social media communications. Measures about peoples' sentiment from a microblogging platform show some correlations with economic data. These results provide interesting arguments about the usefulness of integrating new kinds of data to estimate subjective well-being and official statistics.

**Abstract** *Il presente contributo affronta il rapporto tra indicatori provenienti dalla statistica ufficiale e misure derivate da nuove fonti di dati digitali. In particolare, l'esame di dati telefonici permette di rilevare informazioni sul comportamento delle persone. Tali dati hanno dimostrato di essere utili anche per valutare la composizione demografica dei comuni italiani. Un esperimento di classificazione basato su misure di ego-network ha permesso di migliorare la precisione (F-measure 0.68, baseline 0.63) nel distinguere tra comuni italiani con alta o bassa presenza di migranti. Inoltre, è possibile esaminare il contenuto delle comunicazioni dei social media, applicando tecniche di Sentiment Analysis. Le misure di polarità dei messaggi, aggregate a livello provinciale, hanno mostrato una correlazione positiva con alcuni indicatori sociali. Questi risultati offrono argomenti circa l'utilità di integrare nuovi tipi di dati alle stime del subjective well-being e alla statistica ufficiale.*

**Key words:** Official Statistics, Big Data, Sentiment Analysis, Call Data Record

---

Emilio Sulis  
Computer Science Department, Corso Svizzera 185 - Torino, e-mail: sulis@di.unito.it

## 1 Introduction

Several studies combine computational methods with theories and techniques of social sciences. Recently, Computational Social Science (CSS) investigates individuals and groups in order to understand social phenomena, organizations and companies exploiting the so-called Big Data [6]. The data-driven computational analysis of large datasets offers new types of insights complementary to classical methods such as surveys, self-reported data and direct observations. In particular, some information extraction techniques include Social Network Analysis (SNA) and machine learning algorithms combined together. In this kind of studies based on user generated content on a big scale, humans can be considered as social sensors [3, 4]. This contribution focus on a set of different experiments on new and traditional data: section 2 describes data and methodology related to two studies exploiting Call Data Records (CDR). In section 3, an experiment of text content analysis of social media information sheds light on the relationship between social media and social well-being. Finally, the last section contains some concluding remarks and ideas for future work.

## 2 Exploiting phone calls to assess behavioral patterns, demographic and immigration data

This section details two different kinds of experiment based on mobile phone call datasets from Italy’s telecommunication leading company “Telecom Italia”<sup>1</sup>. *Experiment 1* explores SNA measures from a communication network of one week in Italy to give an insight into patterns of behavior and demographics. *Experiment 2* concerns Ego-Network analysis investigating immigration data. In the following, we briefly describe related dataset, methodology and main results for each experiment.

### 2.1 Patterns of behavior and demography

*Data and methodology.* In *Experiment 1* the dataset *D1* includes about 300 millions of mobile phone calls made in Italy during a representative week of the year. The network of phone calls considers the geographical references of antennas as nodes, as well as the sum of the duration of each single call as the weight of the edges. In addition, we use also measures about population, real estate market and enterprises<sup>2</sup>.

*Calling patterns.* Dataset *D1* has face validity: as Figure 1 clearly states, the distribution over time of phone calls data exhibit a weekly cycle and a daily cy-

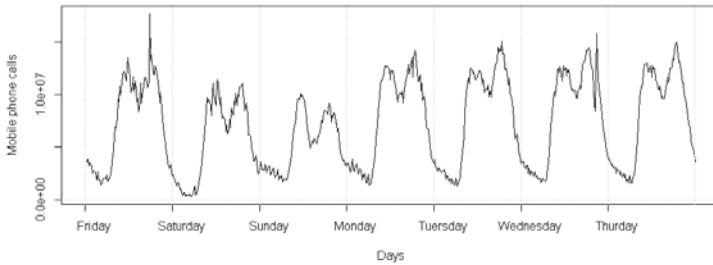
---

<sup>1</sup> Thanks to Trusted Digital Life Innovation group of Telecom Italia SpA, Turin.

<sup>2</sup> Our data sources are ISTAT, the Italian Revenue Agency (*Agenzia delle Entrate*), and the Chambers of Commerce (CCIAA)



cle. In particular, the distribution of each day involves two daily peaks, one in the late morning and another one in the afternoon. As expected, Sunday presents slight



**Fig. 1** Frequency of Italian mobile phone calls in one week using Dataset *D1*.

differences: first, we noticed a lower number of calls; moreover, the morning peak comes later than the other days (at around 11, instead of 10), consistently with the sleep and wake cycles already detected in similar studies.

*Degree and centrality measures.* In the network of phone calls from *D1*, arcs with higher weights connect the most relevant nodes. As degree indicates the number of ingoing and outgoing phone calls for each node, the lowest degree correspond to a site in a mountain alpine valley (Alta Valsesia), while the highest degree corresponds to an antenna of the “Termini” railway station in Rome. Centrality measures offer important suggestions about the connectivity and the position of nodes in the graph. Regarding the measure of betweenness, the highest value corresponds to a CDR in the Italian capital, consistently with the central geographical position of the city.

*Demographics and network measures.* Looking at population by age, the duration of phone calls highly correlates with the presence of young people ( $r=0.90$ ), while the values are lower for adult (0.87) and elderly (0.79). Similarly, phone calls negatively correlate both with the aging index ( $r=-0.35$ ) and with the incidence of elderly ( $r=-0.36$ ). In addition, network measures correlate with the number of enterprises (0.78), which is an indicator of the economic well-being of a province. A less evident relationship is observed with respect to the number of residential property sales ( $r=0.48$ ). These results assess the usefulness of phone call datasets and open the way to further investigation in this subject, as detailed in the next section.

## 2.2 Call data records and immigration

*Data and methodology.* To investigate phone calls to abroad, a *D2* dataset refers to calls made in a not-working day. This dataset includes about 67 millions of single calls, including information about the duration and the direction (incoming or outgoing) aggregated on the single cell by their latitude and longitude values. The

attention here is on Ego-Network (EN), where Ego nodes are the Italian municipalities and Alter nodes are the foreign states. Arcs represent the communication links, weighted on the basis of calls duration. The representation of a first-order EN is a star, where an Italian municipality is Ego, while second-order EN includes also the arcs sorting from the related foreign states to other Italian municipalities<sup>3</sup>. This kind of EN takes into account the impact of foreign states over a municipality and over Italy as a whole.

In order to obtain a single measure concerning migration movements in one year, a Demographic Index of Migrations (DIM) is computed for each of the 8k Italian municipalities as the combination of four single indicators: *i.* one-year variation of immigrant residents; *ii.* register office movements with foreign states; *iii.* register office movements with foreign states; *iv.* net migration rate. Each indicator is normalized between 0 and 1 and the sum of the four is the value of the index. Accordingly to DIM, the list of municipalities is labeled with *HighDIM* (values over the mean), while *LowDIM* includes the lower ones. A binary classification experiment includes a set of 37 features from phone calls ego-networks (i.e., degree, strength, constraint, centrality), using different machine learning algorithms in Weka<sup>4</sup>.

*Classification results.* As smaller municipalities values are less reliable, we finally considered the 538 municipalities with more than 10k families. Then, the classification task includes 254 municipalities with *LowDIM* and 282 with *highDIM*. The baseline measures were computed on the basis of the number of foreign states linked to each municipality (63.6%) and the strength of the network (62.8%). Adopting Logistic Regression as a machine learning algorithm, the confusion matrix shows good results, with a number of correct predictions higher for *LowDIM* class (see Table 1). As the classification result obtain an F-measure score of 67.9, we finally assess the usefulness of network measures derived from phone calls.

	<i>LowDIM</i>	<i>HighDIM</i>
<i>LowDIM</i>	201	81
<i>HighDIM</i>	91	163

**Table 1** Confusion matrix for high and low DIM. The *LowDIM* class obtains the best results.

Computing the Information Gain for each feature, the most relevant ones for this classification task are network metrics. In particular, these features are: the duration of calls (or the sum of the weights) in the complete second-order EN, and the number of arcs of second-order EN. Comparing the results with our baseline, consistently with the role of social networks in migrations, we state the relevance of phone-calls ego-network measures in this prediction task.

<sup>3</sup> As this dataset comes from a single Italian company, it does not consider the links between Alter nodes (the call between foreign states, i.e. from Spain to France and so on), as well as other phone companies data

<sup>4</sup> We adopt Naive Bayes, Support Vector Machines, as well as Logistic Regression in Weka toolkit: <http://www.cs.waikato.ac.nz/ml/weka/>

### 3 Sentiment Analysis of Social Media messages

The content of communication can be investigated by a lexical-based approach to count the polarity of each word in a sentence [13]. Dictionaries containing the semantic orientation of terms have been developed and used in different tasks, as in distinguishing emotions [7] or sarcasm detection [11]. Machine learning techniques automatically classify labeled instances of texts or sentences [8] in supervised classification tasks. Furthermore, a NLP approach already detected happiness in Twitter Italian messages [2]. The current section introduces a general framework architecture, implemented on a large corpus of tweets.

#### 3.1 A general framework

A whole framework architecture to manage social media data and official statistics, detailed in [12], consists of five main parts: providers, data gathering, data analysis and visualization. *Providers* are the data sources (i.e. Twitter, as well as demographic data). A *data gathering* module manages three particular tasks: first, a submodule collects information from different providers; a filtering step removes noisy data (i.e. duplicate records, empty voices, formatting errors); finally, results are standardized in a unified format. Then, a *data analysis* module combines several data processing: first of all, a sentiment analysis submodule returns a mood value (positive, negative, neutral) for each geolocated tweet; a *mash-up* submodule aggregates results by regions, provinces and municipalities; finally, data are stored in a database to be further processed for elaboration and *visualization*.

#### 3.2 Sentiment analysis and social indicators

On the basis of the general framework presented in previous section, data concerning moods and social indicators can be compared as they belongs to the same period and to the same administrative level. A correlation analysis across moods and, in turn, several social indicators is performed, in order to quantify the strength of the relationship between the variables.

*Data and methodology.* The current study considers a large dataset<sup>5</sup>, which includes 259,886,462 Italian tweets from January to December 2014. Tweets having geo-location information (coordinates) are 4,686,251. The line plots in Figure 2 show the monthly trends, both for georeferenced tweets (line plot *a*) and for all tweets (line plot *b*). As expected, the distribution is well-balanced in each month, with the exception of August which corresponds in Italy to the holiday time.

---

<sup>5</sup> Cfr. TWITA project [1].

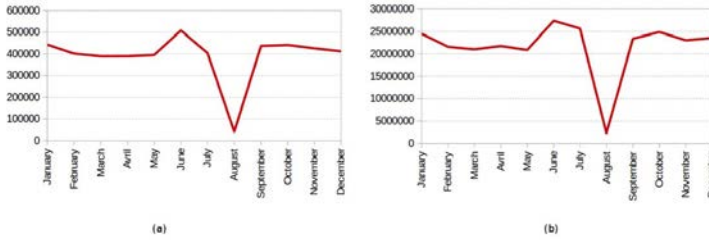


Fig. 2 The line plot with the monthly trend of georeferenced tweets (a) and all tweets (b) in Italy by months, year 2014.

Data of tweets aggregated at the level of the Italian regions are described on the left of Figure 3. The number of tweets by region is consistent with the size of the population, as presented in the plot on the right. The highest number of tweets comes from Lombardia, the most important industrialized area in Italy.

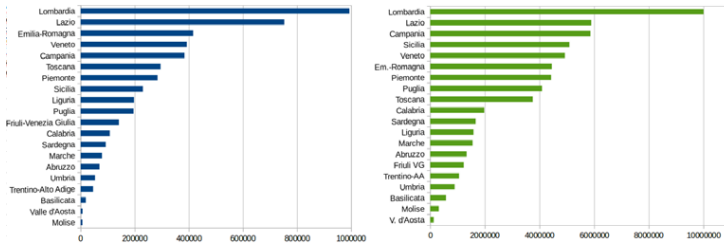


Fig. 3 The distributions both of Tweets (left) and of Italian population by Regions, year 2014 (source: ISTAT).

Once confirmed the validity of the dataset, the polarity of messages is computed by applying a lexical-based approach. Polarity is computed with the formula presented by Kramer in [5] and used also by Quercia [10], which is a normalized count of the occurrences of positive and negative terms from lexical resource LIWC<sup>6</sup>. For  $n$  tweets including a number of positive (Pos) and negative (Neg) terms collected in one month in a territorial unit having a certain mean (MeanPos and MeanNeg) and standard deviation (SDPos and SDNeg) for positive and negative contents, the polarity  $p$  is:

$$p = \sum_{i=1}^n \frac{Pos - MeanPos}{SDPos} - \frac{Neg - MeanNeg}{SDNeg}$$

<sup>6</sup> Cfr. LIWC is presented in [9]. We adopt an own version translated in Italian.

A single polarity value is computed for each month and for each Italian province in 2014, obtaining 1,320 distinct values (12 months for 110 provinces). The five most positive provinces are Napoli, Milano, Savona, Torino and Cremona, while the 5 most negative ones are: Roma, Pordenone, Forlì, Udine and Lodi.

*Correlation between sentiment and socio-economic data.* The aggregated results for each month in 2014 and the 110 Italian Provinces are firstly correlated with the number of hours in which workers were laid off collecting unemployment benefits, and the demographic data concerning the number of births. These two are fine-grained measures available in Italy at the level both of provinces and months. Other data consider all provinces for the whole year, ranging from unemployment to bank deposits. The most promising Pearson's correlation results between sentiment values and socio-economic data ( $p$ -value  $< 0.05$ ) are number of active companies (0.50), real estate transactions (0.38) and bank deposits (0.33). While a weak correlation is observed with respect to population (0.24), registered companies rate (0.21), active companies rate (0.20), births (0.15) and bank loans (0.12), some indicators are not correlated at all, such as the employment or the activity rates. Finally, the correlation between sentiment and Social Security measures is slightly negative, even if very weak. In this experiment, the main idea was to analyze social media content in comparison to a set of more extensive data coming from official statistics. Even considering the problem of selection bias in non-representative samples [15], some results seem quite promising, while others are not, similarly to Wang [14] where the social media sentiment overall seems weakly correlated with official statistics. Nevertheless, the focus here was on describing the situation in fine-grained temporal and geographical subdivisions.

## 4 Concluding remarks

This contribution assesses the utility of digital traces in relation to official statistics. Social signals from phone calls gave interesting insights over socio-economic indicators, as well as Ego-Network measures have been demonstrated useful to distinguishing high or low presence of immigrants. Future works will improve such analysis in a wider range of time, including a larger phone call dataset of one month. In addition, we plan to extend the number of network features, e.g. including diversity. By considering diurnal phone calls on working days, we will test the hypothesis that more industrialized provinces have wider network connections with richer countries. Text content analysis offers different suggestions. A general framework to compare socio-economic data and social media content is presented. More detailed analysis would include the filtering out of messages concerning different topics (sports, television etc.) from the social media corpus. We suppose it would improve the accuracy in the computation of sentiment polarity, as well as the correlation with economic data. Nevertheless, identifying a smaller set of tweets would also reduce the reliability of the correlation results.

## References

1. V. Basile and M. Nissim. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, 2013.
2. C. Bosco, L. Allisio, V. Mussa, V. Patti, G. Ruffo, M. Sanguinetti, and E. Sulis. Detecting happiness in italian tweets: Towards an evaluation dataset for sentiment analysis in felicità. In B. Schuller, P. Buitelaar, L. Devillers, C. Pelachaud, T. Declerck, A. Batliner, P. Rosso, and S. Gaines, editors, *Proc. of the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data, ESSSLOD 2014, LREC*, pages 56–63, Reykjavik, Iceland, 2014. European Language Resources Association.
3. N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.
4. S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, Sept. 2011.
5. A. Kramer. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 287–290. ACM, 2010.
6. D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstynne. Social science: Computational social science. *Science*, 323(5915):721–723, February 2009.
7. R. Meo and E. Sulis. Processing affect in social media: A comparison of methods to distinguish emotions in tweets. *ACM Trans. Internet Technol.*, 17(1):7:1–7:25, Jan. 2017.
8. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.
9. J. W. Pennebaker, M. E. Francis, and R. J. Booth. *Linguistic Inquiry and Word Count*. Lawrence Erlbaum Associates, Mahwah, NJ, 2001.
10. D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. In the mood for being influential on twitter. In *Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom)*, pages 307–314. IEEE, IEEE, 2011.
11. E. Sulis, D. I. H. Fariás, P. Rosso, V. Patti, and G. Ruffo. Figurative messages and affect in twitter: Differences between# irony,# sarcasm and# not. *Knowledge-Based Systems*, 108:132–143, 2016.
12. E. Sulis, M. Lai, M. Vinai, and M. Sanguinetti. Exploring sentiment in social media and official statistics: a general framework. In *Proceedings of the 2nd International Workshop on Emotion and Sentiment in Social and Expressive Media: Opportunities and Challenges for Emotion-aware Multiagent Systems co-located with 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Istanbul, Turkey, May 5, 2015*, pages 96–105, 2015.
13. P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
14. C. Wang, Z. Xiao, Y. Liu, Y. Xu, A. Zhou, and K. Zhang. Sentiview: Sentiment analysis and visualization for internet popular topics. *Human-Machine Systems, IEEE Transactions on*, 43(6):620–630, 2013.
15. E. Zagheni and I. Weber. Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1):13–25, 2015.

# Knowledge mapping by a functional data analysis of scientific articles databases

## *Mappare la conoscenza attraverso un'analisi di dati funzionali di basi di dati di articoli scientifici*

Matilde Trevisani and Arjuna Tuzzi

**Abstract** Scientometrics studies in quantitative fashion the evolution of science focusing on the analysis of publications. One of its objectives is the development of information systems that can help to explore the enormous amount of scientific articles unceasingly published. Our study proposes an information system to reconstruct a dynamical knowledge mapping from a functional data analysis perspective. The source database is a diachronic corpus which originates a words $\times$ time-points contingency table displaying the frequencies of each keyword in the set of texts grouped by time-points in the observed time span. The information system consists of an information retrieval procedure for keywords' selection and a two-stage functional clustering to reconstruct the historical evolution of the knowledge field under investigation.

**Abstract** *La scientometria studia con un approccio quantitativo l'evoluzione della scienza attraverso l'analisi delle pubblicazioni. Uno degli obiettivi è lo sviluppo di sistemi di informazione di ausilio nell'esplorare l'enorme mole di articoli scientifici pubblicati incessantemente. Il nostro studio propone un sistema di informazione atto a ricostruire una mappatura dinamica della conoscenza secondo una prospettiva di analisi di dati funzionali. Il database di partenza è un corpus diacronico che dà origine a una tabella di contingenza parole $\times$ punti temporali contenente le frequenze di ogni parola chiave nell'insieme dei testi raggruppati per punti temporali lungo l'arco di tempo osservato. Il sistema informativo è costituito da una procedura di recupero delle informazioni per la selezione delle parole chiave e un clustering funzionale a due stadi per ricostruire l'evoluzione storica del campo di conoscenza in esame.*

---

Matilde Trevisani

Department of Economics, Business, Mathematics and Statistics, University of Trieste, Via Tigor 22, 34124 Trieste (Italy), e-mail: matilde.trevisani@deams.units.it

Arjuna Tuzzi

Department of Philosophy, Sociology, Education and Applied Psychology, Via M. Cesarotti 10/12, 35123 Padova (Italy) e-mail: arjuna.tuzzi@unipd.it

**Key words:** scientometrics, diachronic corpus, functional data analysis, cluster validation

## 1 Introduction

Scientometrics studies in quantitative fashion the evolution of science focusing on the analysis of publications. One of its major objectives is the development of information systems that can help to explore the enormous amount of scientific articles unceasingly published. The two main methods for automatically designing lexical maps are *citation-based analysis* and *co-word analysis*. Co-citation analysis maps the literature under consideration from the interaction of document citations whereas co-word analysis deals directly with the interaction of key terms shared by documents. Dynamical science mapping is another challenge that aims at describing dynamical patterns in science evolution.

In our study a dynamical knowledge mapping is reconstructed from a functional data analysis (FDA) perspective. The source database is a diachronic corpus which is a collection of texts including information on the time period to which they relate. In *bag-of-words* approaches, a diachronic corpus originates a words  $\times$  time-points contingency table displaying the frequencies of each keyword in the set of texts grouped by time-points in the observed time span. Diachronic corpora represent the ideal ground for studying the history of linguistic phenomena, e.g., when a corpus is able to reflect the relevant features of a text genre in a well-defined time period, the temporal evolution of word occurrences mirrors the historical development of the corresponding concepts [3].

This study proposes an information system consisting of (1) an information retrieval procedure for keywords' selection and (2) a functional clustering two-stage approach to identify words showing prototypical temporal patterns and cluster words portraying similar temporal patterns.

The procedure has been and is being applied to corpora of scientific papers published by leading journals of several disciplines, namely, statistics, social psychology, sociology and philosophy. This work connects to the project *Tracing the History of Words. A Portrait of a Discipline Through Analyses of Keyword Counts in Large Corpora of Scientific Literature* (University of Padova, CPDA145940, 2015-2017), involving an interdisciplinary research group whose aim is to construct chronological corpora, and, hence, to investigate whether a discipline history can be traced from analyzing the keywords' temporal pattern. Several analyses are performed to reconstruct a dynamical evolution: correspondence analysis, topic latent Dirichlet allocation, similarity analysis (using co-occurrences), and—the object of the present work—curve clustering.



## 2 Material: the corpus

Our databases are collections of articles published by a selection of premier journals of the disciplines of interest over a long time period. Text under consideration consist of titles and/or abstracts and/or full texts of the scientific papers. Time is typically discretized by years according to the cadence of volume publication.

As an example, consider one of the corpora analyzed for exploring the historical evolution of Statistics. The database is the collection of papers published by the *Journal of the American Statistical Association* (JASA, 1922-) and its predecessors, *Publications of the ASA* (1888-1912) and *Quarterly Publications of the ASA* (1912-1921). Taking into account only the texts of titles including content words and disregarding items that not refer to research papers (e.g., *List of publications*, *News*, *Comment*, *Rejoinder*), the corpus includes 10,077 titles of articles published in the period 1888-2012 (125 years, from Volume No. 1, Issue No. 1 to Volume No. 107, Issue No. 500, since at the very beginning the volumes were biennial). The corpus is composed of 87,060 word-tokens and 7,746 word-types. To solve the problem of identifying a set of keywords that prove relevant for the study of the history of Statistics, we adopt a stepwise procedure:

1. to overcome some of the limitations of analyses based on simple word-types, we replace words with stems by means of the popular Porter's stemming algorithm;
2. to take into account compounds, multi-words and sequences of words which have different meanings when they are considered in their context of use and together with adjacent words, we identify n-stem-grams;
3. to identify the most relevant statistical keywords, we match the vocabulary with popular statistics glossaries available on-line: ISI-International Statistical Institute; OECD-Organisation for Economic Cooperation and Development; Statistics.com-Institute for Statistics Education; StatSoft Inc.; University of California, Berkeley; University of Glasgow.
4. to reduce low frequency keywords we select keywords with frequencies  $\geq 10$ .

The final contingency table includes the frequencies of 900 keywords over 107 time-points.

## 3 Method: a functional clustering two-stage approach

From a FDA perspective, discrete observations  $\mathbf{y}_i = \{y_{ij}\}$  of the frequency of a keyword  $i (= 1, \dots, N)$  in the volumes  $j = 1, \dots, T$  are viewed as a realization of an underlying continuous function  $x_i(t)$ . As  $\mathbf{y}_i$  is a noisy observation of  $x_i(t)$ , an adequate model is  $\mathbf{y}_i = x_i(\mathbf{t}) + \boldsymbol{\varepsilon}_i$ , where  $\mathbf{t} = \{t_j\}$  is the finite set of time-points and  $\boldsymbol{\varepsilon}_i = \{\varepsilon_{ij}\}$  is a zero mean vector with dispersion matrix  $\Sigma_{\boldsymbol{\varepsilon}_i}$ .

For representing functional data (FD) as smooth functions one method is the basis function approach where  $x_i(t)$  is represented by a finite-dimensional linear combination  $x_i(t) = \sum_{k=1}^K c_{ik} \phi_k(t)$ ,  $c_{ik} \in \mathfrak{R}$ , for sufficiently large  $K$ , of real-valued func-

tions  $\phi_k$  called basis functions. In this study we consider B-splines as they consist in a very flexible basis for non-periodic FD. As regards the positioning of breakpoints, a direct and reasonable choice is placing knots at each time-point  $t_j$ .

We adopt the *roughness penalty* approach for estimation under which the estimate  $\hat{x}_i$  is that function minimizing the penalized residual sum of squares,  $\text{PENSSE} = \text{SSE} + \lambda \cdot \text{PEN}_r$ , where  $\text{SSE} = \{\mathbf{y}_i - x_i(\mathbf{t})\}^T W \{\mathbf{y}_i - x_i(\mathbf{t})\}$  ( $W = \Sigma_e^{-1}$ ) is the residual sum of squares,  $\text{PEN}_r = \int [D^r x_i(s)]^2 ds$  is the penalty term and  $\lambda$  is a smoothing parameter.

A standard practice for choosing  $\lambda$  is to use the generalized cross validation,  $\text{GCV}(\lambda) = T / (T - df(\lambda))^2 \text{SSE}(\hat{x}_i)$ , which provides a convenient approximation to leave-one-out CV.  $df(\lambda)$  is the effective degrees of freedom, which is monotone decreasing in  $\lambda$  with maximum equal to  $K$  when  $\lambda = 0$ .

We smooth the data by trying different spline orders combined with various roughness penalties and varying the smoothing parameter over an opportune range of values.

We adopted a distance-based approach, in particular the  $k$ -means algorithm combined with the  $L_2$  metric to measure distance between curves. Besides the  $L_2$  metric other measures of proximity can be considered, such as the  $L_1$  metric, the adaptive dissimilarity index, and the correlation-based dissimilarity [2].

Cluster validation is an essential step in the cluster analysis process. Within the approaches to cluster validation [1], the use of external information is a valuable and ultimately necessary tool. Here, external information consists of an informal assessment of subject matter experts. On the other side, a large number of indexes has been proposed in the literature for a validation based on the clustered data alone. In this study we combine a large number of internal validation indexes without integrating subject matter knowledge, so as to let the data bring out the best rated groupings. Our clustering procedure is thought of as a tool of thorough investigation before submitting the results to experts who possibly will guide towards other analyses.

## 4 Theory: corpus data transformation

The decision about what data to use is an important part of the clustering process, and often has a fundamental impact on the resulting clusters.

If we consider the keywords $\times$ time-points table by row, a typical feature of a word trajectory is a sharp peak-and-valley trend, mainly due to the sparsity affecting frequency data of a corpus. On the other hand, if we look at data by column they appear strongly asymmetrical, in particular for the marked disparity of frequency classes between the most popular words and all of the others. This is a typical feature of word-type frequency distributions aka *large number rare events* property. Lastly, the size of time-point subcorpora may vary greatly over time.

In our research, we envision several transformations which address two different objectives: whether, in assessing two curves as similar, we should consider height (word popularity) and timing (synchrony) jointly, or timing only. In the first case

we just need to normalize data by column, in the other case we need to normalize by row, or better still, since a sort of column-normalization should be regarded as preliminary, to resort to some double normalization.

The normalization step (Table 1) of our procedure provides several transformations by column ( $c_1$ - $c_5$ ) and by row ( $r_1$ - $r_5$ ).

**Table 1** Normalization plan

		normalized by column		(corpus logic)	("table" logic)	(LNRE)		
		subcorpus		#titles	#tokens	sum ( $\sqrt{\cdot}$ )	max. freq.	dynamic density
normalized by row				$d_1$	$d_2$	$d_3$	$d_4$	
Strong asymmetry	row sum	$d$	$d$	$d_1$	$\chi^2$	$d$	$d_{1b}$	$r_1$
	z-score by row	$d$	$d_2$	$d$	$d$	$d$	$d$	$r_2$
	maximum row frequency	$d$	$d_3$	$d$	$d$	$d$	$d$	$r_3$
	nonlinear transformation: $p_{x(1)}$	$d$	$d_4$	$d$	$d$	$d$	$d$	$r_4$
	nonlinear transformation: $p_{x(2)}$	$d$	$d_{4b}$	$d$	$d$	$d$	$d$	$r_{4b}$
	relative difference	$d$	$d_5$	$d$	$d$	$d$	$d$	$r_5$
		$c_1$	$c_2$	$c_3$	$c_4$	$c_5$		

Crossing a column- by a row-normalization generates a double normalization. Our comprehensive study examines all the transformations specifically indicated in the table. Here we present a small subset:  $c_2, d_1, d_3$ .

### 5 Results and conclusions

Optimal smoothing for  $c_2$  normalized data is achieved with  $m = 5$  and  $\lambda = 10^3$  ( $df = 7.7$ ) after setting a  $PEN_2$  roughness penalty, whereas for both  $d_1$  and  $d_3$  normalized data the criterion lead to  $m = 3$  and  $\lambda = 10^{1.75}$  ( $df = 7.375$ ) under a  $PEN_1$  roughness penalty. Curves are then partitioned by the  $k$ -means algorithm on the basis of the Euclidean distance. The algorithm is re-run, for each  $k$  from 2 to 26, 20 times from different initial configurations set through the  $k$ -means++ seeding method.

A set of 49 quality criteria are then computed in order to identify the best partition/number(s) of clusters. By pooling rankings from all the quality indices, the frequency of being in the top-1 up to the top-4 is calculated for each cluster number  $k$ . In general partitions into two/three clusters are the best rated. This reflects the substantial bifurcation of the historical period around the sixties at which the birth of Statistics as an autonomous and established discipline can be placed. Moreover, partitions with a number of clusters close to the maximum of the considered range have also been frequently selected. This result may be a failure due to the standard assumption of data normally distributed. From the foregoing, once discarded the solutions picking the extremes, the most selected cluster numbers are: 5 for  $c_2$ , 6 for  $d_1$  and 5 for  $d_2$ . To compare some aspects of how the three transformations affect

clustering, we consider the best partition found with the above numbers of clusters (conclusions are below).

Let us now examine some aspects of clustering, in the three cases of normalization, by varying the number of clusters (Table 2): how much groups are balanced; how many groups are singletons; how much groups are heterogeneous in being composed of words of different frequency class or popularity.

**Table 2** Balance, presence of singletons and heterogeneity of frequency classes

normalization	cluster#	2	3	4	5	6	7	8	9	10	15	20	25
$c_2$		.00	.12	.26	.29	.44	.49	.56	.59	.63	.80	.86	.91
Quality of	$d_1$	.72	.93	.90	.90	.94	.96	.95	.97	.97	.98	.98	.99
balancing	$d_3$	.84	.88	.92	.93	.93	.95	.95	.96	.97	.97	.98	.99
	$c_2$	1	1	1	2	2	3	3	3	3	7	10	11
Number of	$d_1$	0	0	0	0	0	0	0	0	0	1	1	1
singletons	$d_3$	0	0	0	0	1	0	0	1	0	3	5	5
Heterogeneity	$c_2$	1	.50	.06	.09	.02	.02	.05	.09	.09	.11	.05	.12
of frequency	$d_1$	1	1	1	.99	.99	.99	.98	.98	.97	.96	.95	.94
classes	$d_3$	.90	.95	.95	.93	.81	.85	.80	.82	.80	.80	.78	.77

A summary of conclusions follows.

1. Normalization by column maintains the level of word popularity differentiated and produces a dominance of high frequency words on the clustering results. Significant imbalance in cluster size, large presence of singletons, lack of heterogeneity of frequency classes in group composition and, finally, the presence of one or more “amorphous” groups, made up almost exclusively of low frequency words, are some of the most evident effects of this type of transformation.
2. Conversely, the double normalization produces groups normally well balanced both in cluster size and frequency classes, rare singletons, and almost never amorphous groups, but does lose the information on word popularity.
3. In specific, type- $d_1$  normalization is better able to recognize any group of words having “sparse” trajectories, i.e., which have experienced birth and/or death over the period considered, while the  $d_3$  variant, which more properly “normalizes” the frequency, builds the groups primarily looking at the curve shape, i.e., at if the “relative popularity” of a word has been constant over time or has fluctuated (and how) during its life cycle.

## References

1. Hennig, C., Meila, M., Murtagh, F., Rocci, R. E.: Handbook of cluster analysis. Chapman & Hall (2016).
2. Montero, P., Vilar, J.: Tslust: An R package for time series clustering, J. Stat. Softw. **62** (1), 1–43 (2014)
3. Trevisani, M., Tuzzi, A.: A portrait of JASA: the history of Statistics through analysis of keyword counts in an early scientific journal, Quality and Quantity **49** (3), 1287–1304 (2015)

# Characterizing the extent of rater agreement via a non-parametric benchmarking procedure

## *Caratterizzazione del grado di accordo intra/inter-valutatore mediante una procedura non parametrica di benchmark*

Amalia Vanacore<sup>1</sup> and Maria Sole Pellegrino<sup>2</sup>

**Abstract** In several context ranging from medical to social sciences, rater reliability is assessed in terms of intra (-inter) rater agreement. The extent of rater agreement is commonly characterized by comparing the value of the adopted agreement coefficient against a benchmark scale. This *deterministic* approach has been widely criticized since it neglects the influence of experimental conditions on the estimated agreement coefficient. In order to overcome this criticism, in this paper a statistical procedure for benchmarking is presented. The proposed procedure is based on non parametric bootstrap confidence intervals. The statistical properties of the proposed procedure have been studied via a Monte Carlo simulation.

**Abstract** *In numerosi contesti applicativi, dal medico al sociale, l'affidabilità di un valutatore è valutata in funzione del grado di accordo intra (-inter) valutatore. La caratterizzazione del grado di accordo è tipicamente effettuata confrontando la stima del coefficiente di accordo adottato con una scala di riferimento (benchmark). Questo approccio "deterministico" è stato spesso criticato in letteratura in quanto non tiene in conto l'influenza delle condizioni sperimentali sul processo di stima. In questo lavoro è presentata una procedura di benchmark basata su intervalli di confidenza bootstrap. Le proprietà statistiche della procedura proposta sono state studiate mediante simulazione Monte Carlo.*

**Key words:** rater reliability, kappa-type agreement index, statistical power, Monte Carlo simulation

---

<sup>1</sup> Amalia Vanacore, Department of Industrial Engineering, University of Naples Federico II; email: amalia.vanacore@unina.it

<sup>2</sup> Maria Sole Pellegrino, Department of Industrial Engineering, University of Naples Federico II; email: mariasole.pellegrino@unina.it

## 1. Introduction

In many context of research (*e.g.*, cognitive and behavioural science, quality science, clinical epidemiology, diagnostic imaging, content analysis), there is frequently a need to assess the performance of human instruments (*i.e.*, raters) providing subjective measurements, expressed on a dichotomous, nominal or ordinal rating scale. Rater reliability is often evaluated in terms of the extent of agreement between two or more series of ratings provided by two or more raters (inter-rater agreement) or by the same rater in two or more occasions (intra-rater agreement). Specifically, inter-rater agreement is concerned about the reproducibility of measurements provided by different raters, whereas intra-rater agreement is concerned about self-reproducibility (also known as repeatability).

The easiest way of measuring agreement between ratings is to calculate the overall percentage of agreement; nevertheless, this measure does not take into account the agreement that would be expected by chance alone [11]. A reasonable alternative is to adopt the widespread kappa-type index that was introduced by Cohen in 1960 as a rescaled measure of the probability of observed agreement corrected with the probability of agreement expected by chance alone. A main issue for the correct definition of a kappa-type index regards the notion of expected proportion of agreement: chance measurements are conceived as blind (that is, uninformative about the rated items) and any distributional assumption for them is likely to be arbitrary. A solution is to adopt the notion of uniform chance measurement [2] that — given a certain rating scale — can be assumed as a reasonable model for the maximally non-informative measurement system. This uniform version of Kappa is often referred to as Brennan-Prediger coefficient [3].

The extent of a kappa-type index is generally qualified through a benchmark scale [*e.g.* 1, 8, 10]: threshold values against which compare the estimated agreement coefficient for deciding whether the extent of agreement is good or poor. Although commonly adopted, this deterministic benchmarking approach does not consider that the value of the information provided by an agreement coefficient is unknown since, being computed on a sample of items, its estimate is subject to sampling error. In order to identify a suitable neighbourhood of the truth (*i.e.*, the true population value), sampling error has always to be considered.

In this paper a benchmarking procedure based on bootstrap resampling is proposed in order to take into account the sampling uncertainty when characterizing the extent of rater agreement. The main statistical properties of the proposed procedure have been assessed via a Monte Carlo simulation study.

The remainder of this paper is organized as follows: in Section 2 the weighted Brennan-Prediger coefficient is introduced; Sections 3 is devoted to coefficient estimation and inference; in Section 4 the simulation design is described and the main results are discussed; finally, conclusions are summarized in Section 5.

## 2. Weighted Brennan-Prediger Coefficient

Let  $n$  be the number of items rated twice (*i.e.*, two replications) on an ordinal  $k$ -points rating scale (with  $k > 2$ ),  $n_{ij}$  the number of items classified into  $i^{\text{th}}$  category in the first replication and into  $j^{\text{th}}$  category in the second replication and  $w_{ij}$  the corresponding symmetrically weight (*i.e.*,  $w_{ij} = w_{ji}$ ) introduced in order to consider that on an ordinal rating scale, some disagreements are more serious than others (*i.e.*, disagreement on two distant categories are more relevant than disagreement on neighbouring categories).

The weighted Brennan-Prediger coefficient [9] is defined as:

$$\hat{K}_w^U = (\hat{p}_a - p_{alc}^U) / (1 - p_{alc}^U)$$

where  $\hat{p}_a = \sum_{i=1}^k \sum_{j=1}^k w_{ij} (n_{ij} / n)$  and  $p_{alc}^U = (1/k^2) \sum_{l=1}^k \sum_{j=1}^k w_{lj}$ .

The  $\hat{K}_w^U$  coefficient ranges from -1 to +1 and it can be assumed asymptotically normal distributed [9] with mean  $K_w^U$  and variance  $\hat{\sigma}_{K_w^U}^2$  given by:

$$\hat{\sigma}_{K_w^U}^2 = \frac{1}{n(n-1)} \sum_{l=1}^n a_h^2 / (1 - p_{alc})^2 \quad (2)$$

where  $h$  refers to the generic rated item and  $a_h = \sum_{i,j=1}^k w_{ij} (\delta_{ij}^{(h)} - p_{ij})$  with  $\delta_{ij}^{(h)} = 1$  if the rater rated item  $h$  into  $i^{\text{th}}$  and  $j^{\text{th}}$  category in the first and second replication, respectively, and  $\delta_{ij}^{(h)} = 0$  otherwise.

## 3. Characterization of rater agreement

The approach currently adopted to characterize the extent of agreement is based upon a straight comparison between the estimated coefficient and an adopted benchmark scale. The most widespread benchmark scale for interpreting the magnitude of agreement coefficients was proposed by Landis and Koch [10]. According to this scale, there are 5 categories of agreement corresponding to as many ranges of coefficient values: slight, fair, moderate, substantial and almost perfect agreement for coefficient values ranging between 0 and 0.2, 0.21 and 0.4, 0.41 and 0.6, 0.61 and 0.8 and 0.81 and 1.0, respectively.

Although benchmark scales are widely adopted for relating the magnitude of the coefficient to the notion of extent of agreement, some researchers question their validity and give advice that their uncritical application may lead to practically questionable decisions [11]. Actually, as argued in [9] the choice of the benchmark scale is less important than the way it is used for characterizing the extent of agreement.

A deterministic approach to benchmarking does not account for the influence of experimental conditions on the estimated coefficient and, thus, it does not allow for a statistical characterization of the extent of rater agreement. This criticism may be overcome by benchmarking the lower bound of the confidence interval of the agreement coefficient rather than its point estimate.

Assuming the asymptotic normal approximation, the lower and upper bound of a two-sided  $(1 - \alpha)\%$  confidence interval for  $K_W^U$  are given by:

$$\hat{K}_W^U \pm z_{\alpha/2} \hat{\sigma}_{K_W^U}$$

The accuracy of the above confidence interval depends on the asymptotic normality of  $\hat{K}_W^U$  and on the asymptotic solution for  $\hat{\sigma}_{K_W^U}^2$  which are both questionable for small sample sizes.

Resampling, which is generally considered the approach of choice when the assumptions of classical statistical methods are not met, may yield more accurate confidence limits and thus it can be usefully adopted to characterize the extent of rater agreement.

Among the available methods to build bootstrap confidence intervals, the percentile bootstrap (hereafter,  $p$ ) is the simplest and the most popular one. The lower and upper bounds of the  $(1 - \alpha)\%$  two-sided  $p$  confidence interval are, respectively, the  $(\alpha/2)$  and  $(1 - \alpha/2)$  percentiles of the cumulative distribution function  $G$  of the bootstrap replications of  $\hat{K}_W^U$ . On the other hand the Bias-Corrected and Accelerated bootstrap (hereafter, BCa) confidence interval is recommended for severely non normal data [4, 6]. Despite the high computational complexity needed, BCa confidence intervals have generally smaller coverage errors than the others. The lower and upper bounds of the  $(1 - \alpha)\%$  two-sided BCa confidence interval are defined as:

$$G^{-1}\left(\Phi\left(b \pm (z_{\alpha/2} \pm b) / [1 + a(\mp z_{\alpha/2} - b)]\right)\right) \quad (4)$$

being  $\Phi$  the cumulative distribution function of the normal distribution,  $z_{\alpha/2}$  the  $\alpha/2$  percentile of the standard normal distribution,  $b$  the bias correction parameter and  $a$  the acceleration parameter.

#### 4. Simulation study

In order to analyse the statistical properties of the proposed benchmarking procedure in terms of Type I error and statistical power, a Monte Carlo simulation study has been developed considering one rater who classifies  $n$  items into one of  $k$  possible ordinal rating categories. The simulation has been conducted by sampling  $r = 2000$  Monte Carlo data sets from a multinomial distribution with parameters  $n$  and  $\mathbf{p} = (\pi_{11}, \dots, \pi_{ij}, \dots, \pi_{kk})$ ; the  $\pi_{ij}$  values have been chosen so as to obtain nine true population values of  $K_W^U$  (viz., 0, 0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 1.0),





As foreseen, the statistical properties of the proposed benchmarking procedure improve as the sample size increases being satisfactory even for relatively small sample size. Specifically, the significance level is generally slightly better for  $BC_a$  bootstrap confidence interval; it decreases with increasing sample size but it grows up for increasing true population value of  $K_w^U$ ; it is close to the nominal level ( $\alpha = 0.025$ ) only in the case of null rater agreement for  $n \geq 40$ ; however, it is always less than 0.10 except for  $n = 20$  when testing an high rater agreement level. The statistical power, instead, is generally slightly better for  $p$  bootstrap confidence interval; for  $n \geq 30$ , it is less than 80% only in testing hypotheses referring to adjacent agreement categories (*e.g.*, poor vs slight, moderate vs substantial).

## 5. Conclusions

The proposed benchmarking procedure can be suitably applied for the characterization of the extent of agreement over a small or moderate number of subjective ratings provided by one or more raters. The procedure shows satisfactory statistical properties in testing both null and non-null cases of rater agreement, being adequately powered in detecting differences in the extent of rater agreement that are of practical interest for agreement studies (*i.e.*, differences of at least 0.2).

It is worthwhile to note that the proposed benchmarking procedure can be also adopted to characterize the extent of inter-rater agreement which, in the case of more than two raters, could be estimated using a suitable variant of kappa coefficient, such as the Fleiss' kappa.

## References

1. Altman, D. G.: Practical Statistics for Medical Research. Chapman and Hall (1991)
2. Bennett, E. M., Alpert, R., Goldstein, A. C.: Communications through limited response questioning. Public Opinion Quarterly. 18.3, 303--308 (1954). DOI: 10.1086/266520
3. Brennan, R. L., Prediger, D. J.: Coefficient Kappa: Some Uses, Misuses, and Alternatives. EPM (1981), 41, 687--699
4. Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statist. Med. 19, 1141--1164. (2000),
5. Cicchetti, D. V., Allison, T.: A new procedure for assessing reliability of scoring EEG sleep recordings. Amer. J. EEG Technol. 11.3, 101--110. (1971)
6. Cohen, J.: A coefficient of agreement for nominal scale. EPM. 20.1, 37-46 (1960).
7. Efron, B., Tibshirani, R. J.: An introduction to the bootstrap. CRC press (1994)
8. Fleiss, J. L.: Statistical Methods for Rates and Proportions. John Wiley & Sons (1981)
9. Gwet, K. L.: Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC (2014)
10. Landis, J. R., Koch, G. G.: The measurement of observer agreement for categorical data. Biometrics. 33.1, 159--174 (1977)
11. Sim, J., Wright, C. C.: The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys. Ther. 85.3, 257--268 (2005).

# Mining Mobile Phone Data to Detect Urban Areas

## *Analisi di dati di telefonia mobile per l'individuazione di aree urbane*

Maarten Vanhoof, Stephanie Combes and Marie-Pierre de Bellefon

**Abstract** The production of Urban Areas zonings at national level is characterized by long delays between consecutive updates. As mobile phone data has recently shown promising results for automated land use classification, we investigate the possibility to reproduce the French Urban Area Zoning (ZAUER). We exploit a dataset of hourly mobile phone activity profiles at cell-tower level, discuss methodological challenges, and find Urban Centers to be most correctly classified. Our findings frame the possibilities and limits for using mobile phone data to automatically, and continuously produce urban zonings

**Abstract** *In questo articolo esaminiamo labilt dei dati del telefono cellulare nel riprodurre la zonizzazione dellarea urbana francese. Partendo un dataset di profili di attiviti di telefonia oraria registrati dalle antenne di uno dei maggiori operatori telefonici francesi, analizziamo le sfide metodologiche coinvolte, e identifichiamo i centri urbani pi facilmente predicibili. I risultati proposti mostrano alcune delle possibilit e dei limiti legati allutilizzo dei dati del telefono cellulare per produrre zonizzazioni automaticamente e continuamente.*

**Key words:** Supervised classification, Mobile phone data, Spatial autocorrelation, Urban areas, Map comparison

---

Maarten Vanhoof  
Open Lab, Newcastle University, Newcastle-Upon-Tyne, UK and Orange Labs, Paris, FR  
e-mail: M.Vanhoof1@newcastle.ac.uk

Stephanie Combes  
INSEE, 18 boulevard Adolphe PINARD, Paris, France  
e-mail: stephanie.combes@gmail.com

Marie-Pierre de Bellefon  
INSEE, 18 boulevard Adolphe PINARD, Paris, France  
e-mail: marie-pierre.de-bellefon@insee.fr

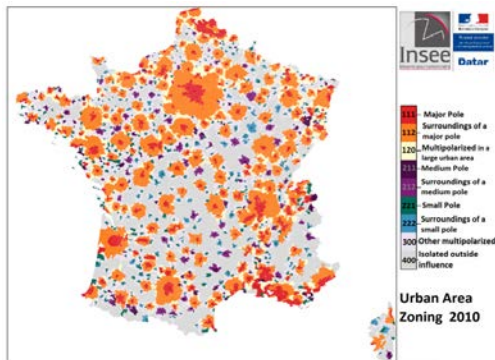
## 1 Introduction

The growth of cities, and with it the extension of urban agglomerations, has become characteristic for contemporary times [Galster et al., 2001]. In this context, the identification of economically integrated areas is crucial for the adequate implementation of policy measures [Duranton and Puga, 2014] and thus calls for the definition of other typologies than purely administrative regions. As a mean of defining integrated (urban) areas, [Berry et al., 1969] suggested to rely on commuting patterns toward a predefined urban core to delineate metropolitan areas. In the same spirit, the National Statistics Office of France (INSEE) nowadays produces a zoning (ZAUER: Urban Area and Rural Employment Areas Zoning) that identifies cities areas of influence at a national level. Producing urban area zonings is a complex task involving multiple actors and methods. As a consequence, long delays are observed between consecutive publications (between five and ten years in France) which contrast with the fast pace of change in territories. [Floch and Levy, 2011]. In this context, alternative sources of more timely, high-resolution data associated with simpler procedures could contribute in a meaningful way to the production of more recurrent releases of such typologies.

In this paper we investigate the capabilities of French mobile phone data to reproduce the ZAUER zoning and explore a procedure that could lead towards a data-driven and recurrent production of the typology between official releases. Our contribution is twofold. First we demonstrate how mobile phone data can be mobilized to develop a nationwide typology of urban areas. Secondly, we elaborate a case that demonstrates how supervised classification tools can be of interests to official statistics. In addition to their ease-of-use, supervised classification techniques provide for both classification outputs and a quality evaluation. The latter being key in official statistics, we deem that a wider investment in these techniques could be profitable.

## 2 Urban Areas in France

The official french Urban Areas classification (ZAUER), as produced by the French National Statistical Institute (INSEE), consists of 9 classes, being distinguished mainly by the size of the employment pole in the 'central' Urban Unit. Major, medium and small centers are Urban Units offering respectively more than 10,000, between 5,000 and 10,000 and between 1,500 and 5,000 jobs that are inhabited by at least 2,000 people and cover a continuous build-up area with no more than 200 meters between buildings. Surroundings of urban centers are municipalities for which more than 40 % of the working population commute daily to an adjacent urban center. Special cases are being recognized for municipalities that have several urban centers to commute to (multi-polarized municipalities) or municipalities that are not 'influenced' by any urban centers.



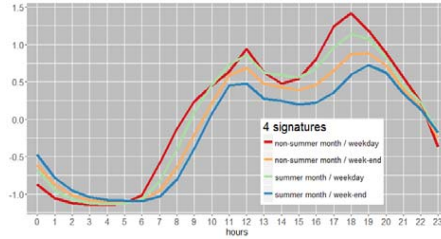
**Fig. 1** Spatial distribution of ZAUER classes. The 2010 ZAUER classification is founded on data collected between 2008 to 2010 in the national census survey

### 3 Activity profiles from Mobile Phone Data

Call Detail Record (CDR) data are collected by mobile phone service providers for billing and network maintenance purposes. CDR data gather locational, temporal, and interactional information (who contacts whom) every time a phone call or text message is initiated by a user. The spatial resolution of observations is restricted to the locations of cell-towers, which are not uniform in space because of demand-driven placement.

In this study we use an aggregated CDR dataset from France provided by Orange and based on the activities of 18 million subscribers during a period of 154 consecutive days in 2007 (May 13 to October 14). Anonymisation at individual level was complemented with aggregation at the cell-tower level (see next paragraph) hereby ensuring full privacy of individual users as demanded by the French Data Protection Agency (CNIL) in light of the EU General Data Protection Regulation.

Literature validates the hypothesis that cell-tower profile activity (amounts of events measured at an antenna over time) can be informative for territory qualification [Soto and Martinez, 2011]. Therefore, we construct antenna activity profiles as a time series of the amount of activities registered each day at every hour and standardize the series for comparative purposes. Next, the obtained relative hourly profiles are averaged per hour of the day for the entire six-months window resulting in activity profiles for each antenna ('signatures'). In total we build four distinct signatures per antenna averaging each hour on i) weekdays in non-summer months, ii) weekends in non-summer months, iii) weekdays in summer months and iv) weekends in summer months. This results in  $24 \times 4$  features per antenna (figure 2).



**Fig. 2** Average relative activity profile for all antennas grouped by summer/non-summer months and week-/week-end days.

Because standardizing the activity profiles implies a loss of information about the absolute amount of activities, we add the circumference of the Vorono polygon for each antenna as a complementary feature. Lower circumferences indicate locally higher antenna densities and, given demand-driven placement, higher expected amounts of activities by the operator.

## 4 Methodology

### 4.1 Classification methods

For our classification task, we consider each antenna as an observation that needs classification in one ZAUER class. The output of the procedure will form a zoning that can be compared to the official one. Multiple algorithms are available to carry out multiclass classification procedures. We implement the random forest approach described by [Breiman, 2001], Boosting Trees [Schapire, 2003] and the Elastic-Net penalized Logistic Regression [Zou and Hastie, 2005].

The Logistic Regression with Elastic Net penalty consists of maximizing the likelihood under a constraint expressed on the coefficients' amplitude [Zou and Hastie, 2005]. Specifically, this approach is better in accounting for multicollinearity between features (which is likely to happen here since our features are extracted from a temporal profile) than the initial LASSO [Tibshirani, 1996]. In contrast to the Logistic Regression, Random Forests and Boosting Trees do not assume linear interactions between variables. Random forests [Breiman, 2001] aggregate classification trees built on bootstrap samples, but introduce randomness by sampling a set of regressors from the initial set of variables at each separation step of each tree. Boosting Trees [Schapire, 2003] is rather different. It is an additive adaptive procedure which takes into account the biggest forecasting errors at a given iteration when calibrating the next iteration, by actualizing observations' weights.

## 4.2 Challenges

Mobilizing mobile phone data for urban areas classification at a national level raises several challenges. First the official ZAUER classification consists of 9 imbalanced classes, meaning that both municipalities and antennas are heterogeneously distributed among the classes and with respect to the urban tissue. In anticipation of this problem, we regroup the existing 9 classes into 6 by merging medium urban centers, small urban centers, and their respective surroundings. More importantly however, we apply downsampling, which consists of removing instances from the majority class, to minimize the effect of imbalanced classes on our classifiers.

Secondly, the extended area of investigation implicates high degrees of spatial autocorrelation and high volatility of antenna activity profiles within the different classes. As such we pay special attention to spatial autocorrelation. To de-correlate testing and training sets, we first operate stratification sampling while segmenting the map of France in four comparable quadrants to produce an initial test set (guaranteeing some minimal representativeness among regions). Next, the nearest neighbours of each selected observation are added to the test set so that we can evaluate the algorithms on their ability to reproduce a zoning and not only a punctual classification of one antenna. Finally, once the test set is built, we consider every left antenna not located in a buffer region around the test observations as a training sample. This last step ensures spatial de-correlation. We use the same approach to build the data partitions mobilized for cross validation.

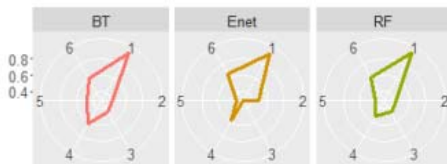
Thirdly, Urban Areas are characterized by various degrees of similarity and spatial entanglement (especially at the borders of areas where the validity of urban area typologies may be less reliable). We address this issue by recouring to the use of the Fuzzy Kappa metric ([Hagen, 2003] and the improved Fuzzy Kappa [Hagen-Zanker, 2009]) allowing us to evaluate (and calibrate) our models while taking into account both location and category fuzziness.

## 5 Results

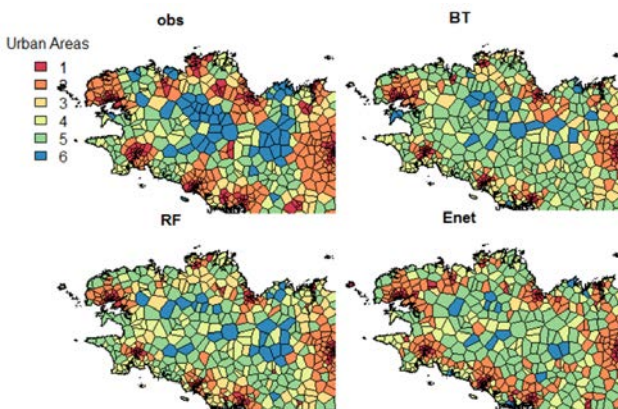
Following the procedures outlined before, we applied three classifiers in order to predict urban areas from signatures of mobile phone activity. Kappa and Fuzzy Kappa computed on the test sets are reported in table 1. Fuzzy Kappas values range from 0.66 to 0.70 whereas Kappas stand between 0.59 and 0.65. According to magnitude guidelines in literature (for example [Landis and Koch, 1977]), values between 0.61 and 0.80 are considered substantial (1 being the perfect agreement). Detection rates per class are represented in figure 3 for the different classifiers. We can see that synthetic measures like Kappa or Fuzzy Kappa mask an heterogeneity in the detection rates par class, highlighting the fact that some classes are more difficult to detect than others.

**Table 1** Kappa and Fuzzy Kappa for the different classifier

Method	Kappa	Fuzzy Kappa
Random Forests (RF)	0.62	0.67
Boosting Trees (BT)	0.63	0.69
Elastic Net (ENet)	0.59	0.67

**Fig. 3** Classification rate (in %) per class for each classifier

We observe small differences in the capabilities of the algorithms to detect classes. In general, major urban centers (class 1) present an excellent rate of detection (about 95 %) for every approach. Correct rates (50 to 70 %) can be achieved for classes 4,5, and 6 (medium and small urban centers and their surroundings, multipolarized municipalities, and isolated municipalities). Classes 2 and 3 are more difficult to discern. Especially Class 3 (multipolarized municipalities in a large urban area) whose detection rate varies from 30 % to 60 % in the best scenario. Class 2 (surroundings of major urban center) get properly detected for only 40 to 50 % of the cases. The results of our classification for Normandy, a region in the west of France that mixes all urban classes are displayed in figure 4.

**Fig. 4** Observed (obs) and predicted (based on the different classifiers) Urban areas for Normandy



## 6 Discussion

Our most remarkable findings are the difference between the accuracy of the prediction for major urban centers (class 1) and the heterogeneous performance in predicting the other classes (ranging between 30 and 70 %). Still, different classifiers show consistencies in results (with slight variations observed for one class or the other), which urges us to find explanations for these results based on the characteristics of our validation data (ZAUER) and mobilized data (mobile phone).

A first remark can be made by reflecting on municipalities in border areas between different zauer classes. In this perspective, antenna signatures of two distinct urban areas may sometimes be very similar and thus hard to distinguish (a case rather common in border areas). The difference of about 0.1 between Fuzzy Kappa and Kappa, however, denotes that our algorithms are able to, at least, partly cope with this difficulty by predicting sometimes wrongful but close classes (in terms of similarity of the classes or of spatial location).

A second remark urges us to consider the limitations of mobile phone as a data source for urban areas recognition. CDR data is, by design, subordinate to users' usage and the extracted activity profiles are subordinate to user's movement patterns, both of which might differ between regions and urban areas. In addition, local market shares of single operators are unknown making it impossible to correctly control for representativeness. Other uncertainties stem from using spatial resolution of the cellular tower network. This resolution does, of course, not collide with administrative borders which renders a discrepancy between information gathering and the proposed classification task. Ultimately, antenna positioning may hinder the antenna signature to be characterized by the presence of (local) populations. Some antennas might capture only specific behavior of local subscribers when, for instance, being positioned along transport axes.

The choice of the methods seems less at stake. One alternative would have been to recourse to unsupervised techniques, which is often done in land use literature [Aguilra et al., 2014, Soto and Martinez, 2011]. Yet differences in antennas signatures can be interpreted in multiple ways. Exploring supervised classification therefore appears as a useful preliminary step for relevant features extraction. In this context, classification algorithms with feature selection designs like penalized logistic regression and random forests are extremely useful as they allow to identify the features contributing most to the discrimination of the classes of interest (amplitude of the coefficients in the penalized logistic regression, importance measure of the variables in random forests), hence leveraging interesting insights.

## 7 Conclusion

Concluding, we would like to consider improvements and applications. In terms of applications, our results encourage us to promote the use of mobile phone data as an alternative source for producing recurrent urban area zoning between official but less frequent releases. Specifically, We are quite optimistic on using supervised classification to, for example, show patterns on the emergence of urban centers or the progression of urban areas. We reckon, however, that assessments regarding the urbanization of rural and isolated areas should remain cautious as our classification tasks underperformed there. Thinking improvement, we hope that more recent sources of CDR data, or other sources like DDR (data detail record data) could provide for more dense and frequent observations in remote areas, improving our automated classification. Ultimately, we hope that the results of our classification case can encourage a more widespread use of machine learning techniques in official statistics. In our work we have shown that the application of such techniques is rather straightforward and can be instructive for both future and past work.

## References

- [Aguilra et al., 2014] Aguilra, V., Milion, C., and Allio, S. (2014). Territory analysis using cell-phone data. *Transport Research Arena 2014, Paris*.
- [Berry et al., 1969] Berry, B. J. L., Goheen, P. G., and Goldstein, H. (1969). *Metropolitan area definition: A re-evaluation of concept and statistical practice*, volume 28. [Washington]: US Bureau of the Census.
- [Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- [Duranton and Puga, 2014] Duranton, G. and Puga, D. (2014). Urban land use.
- [Floch and Levy, 2011] Floch, J. and Levy, D. (2011). Poursuite de la priurbanisation et croissance des grandes aires urbaines. *INSEE Premire*, 1375.
- [Galster et al., 2001] Galster, G., Hanson, R., Ratcliffe, M. R., Wolman, H., Coleman, S., and Freihage, J. (2001). Wrestling sprawl to the ground: defining and measuring an elusive concept. *Housing policy debate*, 12(4):681–717.
- [Hagen, 2003] Hagen, A. (2003). Fuzzy set approach to assessing similarity of categorical maps. *International Journal of Geographical Information Science*, 17(3):235–249.
- [Hagen-Zanker, 2009] Hagen-Zanker, A. (2009). An improved fuzzy kappa statistic that accounts for spatial autocorrelation. *International Journal of Geographical Information Science*, 23(1):61–73.
- [Landis and Koch, 1977] Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- [Schapire, 2003] Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer.
- [Soto and Martinez, 2011] Soto, V. and Martinez, E. F. (2011). Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch, HotPlanet, New York, NY, USA*, 11:17–22.
- [Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.
- [Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, page 301320.

# Statistical methods in assessing the equality of income distribution, case study of Poland

## *Metodi statistici per valutare l'uguaglianza nella distribuzione del reddito, caso di studio della Polonia*

<sup>1</sup>Viktoriya Voytsekhovska, <sup>2</sup>Olivier Butzbach

**Abstract** The development of gross wages for employees in Poland for a different time periods by social justice criteria is made. Based on gross wages the interval differentiation approach is considered for correlation to gross remuneration in adjacent time period. The use of dependencies is carried out to determine the value, that increases the gross wages and to determine the rate of remuneration's growth rate for individual categories of employees. To assess the adequacy of the developed approach the results of theoretical calculations are compared with statistical data. The conclusions regarding the availability of sustainable patterns of equal distribution by achieving appropriate dynamics of wages' growth in certain categories of workers were done.

**Key words:** social justice, gross wages, income distribution, equality

**Abstract** *In Polonia negli ultimi anni si è perseguito l'obiettivo di sviluppare il salario lordo dei lavoratori dipendenti sulla base di principi di equità sociale.*

*In questo lavoro, con riferimento agli anni dal 2010 al 2014, ci proponiamo di analizzare le relazioni degli indici di concentrazione dei redditi con il livello dei salari lordi dell'anno precedente.*

*Nelle conclusioni del lavoro vengono discusse le possibilità di impiego di modelli sostenibili di equidistribuzione per raggiungere appropriate dinamiche di crescita dei salari in determinate categorie di lavoratori per ammontare di reddito.*

**Parole chiave:** giustizia sociale, salari lordi, distribuzione del reddito, uguaglianza

### **Introduction**

During the past decades the social and income inequality questions, besides evident recession from global financial crisis and GDP growth, raise extended interest of number of international organisations, such as United Nations, European Union, OECD, politicians and scientists globally [2, 13, 14, 16].

---

<sup>1</sup> Viktoriya Voytsekhovska, Lviv polytechnic National University; email:

[viktoriia.v.voytsekhovska@lpnu.ua](mailto:viktoriia.v.voytsekhovska@lpnu.ua)

<sup>2</sup>Olivier Butzbach, Università degli studi della Campania "Luigi Vanvitelli"; email:

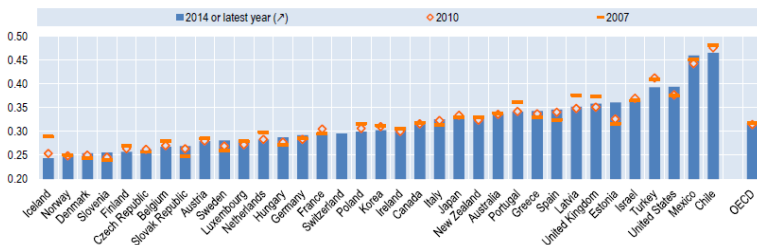
[OlivierKarlEmmanuel.BUTZBACH@unina2.it](mailto:OlivierKarlEmmanuel.BUTZBACH@unina2.it)

Recent study conducted by OECD confirms the hypothesis from number of scientists, that indeed besides the GDP growth, inequality increases, therefore 10% of richest population receives 9.5 times the income of 10% of the poorest, unlike in 80's this relation was 7.1[5, 10, 11]. However, Atkinson and Tinbergen, studying inequality for 40 years highlight the importance of investment in human capital and its education, that would respond to increasing demand of highly skilled employees in technologically growing world [1, 13]. Also Italian economists M.Raitano and E.Granaglia see the roots of inequality in globalization and routinization by means of examining the good and the bad effects of certain policies on inequalities [3, 10]. The magnitude of inequality differs from country to country and has variety of factors from tax and educational policies to decision-making.

A Structural Perspective brought by Van de Sande and Byvelds argues, that social work research, including statistics, should be taught from a structural perspective and must follow anti-oppressive principles, which view the problems experienced by people as rooted in the social, political and economic structures of society [15].

## 1.1 Main study

On a macro-level Gini coefficient among countries increases despite the fact, that during past years, economic growth leads to new jobs creation and lower unemployment, which in turn shall decrease income inequality. As can be observed from Fig.1 in OECD countries the Gini coefficient of household income, that varies from 0 to 1, where 0 is 100% equality, it can be observed, that in some countries it decreased since 2007 in such countries as Turkey, Iceland and Latvia, while increased in Slovak Republic, Spain and Sweden [5,9].

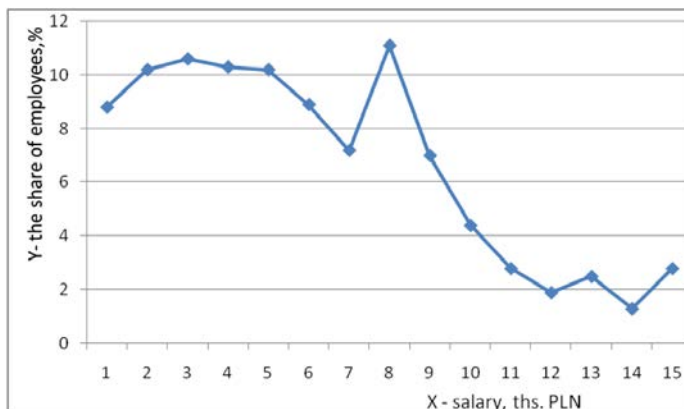


**Figure 1:** Gini coefficient of disposable income inequalities 2007-2014, OECD countries [9].

However, economic growth doesn't effect all countries equally. Poland overall has little decrease of Gini coefficient since 2007 (from 0,32 to 0,30 points), while overall inequalities increased since 80's due to the change from socialism to market economy [11, 13].

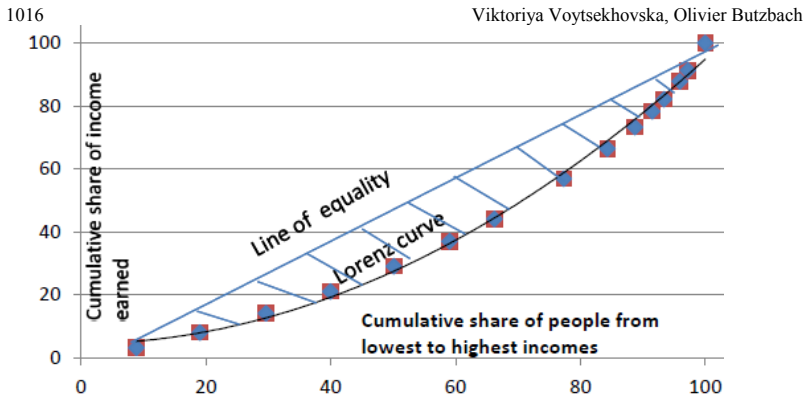
Listed above we can observe the historical data of Gini coefficient from OECD countries, inequalities can come both from economic or social nature and it's important to understand the dynamics for concrete country in order to be able to make adequate decisions and policies.

The goal of this work was to analyze the dynamics of employees' wages growth in Poland over the period of 2010-2014. The feature of this study was that according to statistical groupings in Poland there are 15 slots for wages, which we grouped into 5 in order of their value increase [6,7,8]. In each of these groups the smoothing determined the average level of wages. Thus for each group we have obtained 5 year average salary in accordance with the law of their distribution. The initial data can be described by dome shape distribution law (Figure 2).



**Figure 2.** Wage distribution in Poland, 2012 [7].

The Fig.2 illustrates the income distribution, which is asymmetrical, it determines the result that 50% of employees in the range of lower wages obtain only 29% of total wages. By means of appropriate accumulated values to determine the Lorenz curve and the Gini coefficient is presented on graphic interpretation of Fig. 3 according to the 2012 data. According to the empirical data, we can observe the peak in 8<sup>th</sup> interval of salaries, that is the same for each studied year, that in our opinion, means, that the share of employees with such salary is maximal.



**Figure 3.** Lorenz curve for the Polish data on gross wages (2012) [7].

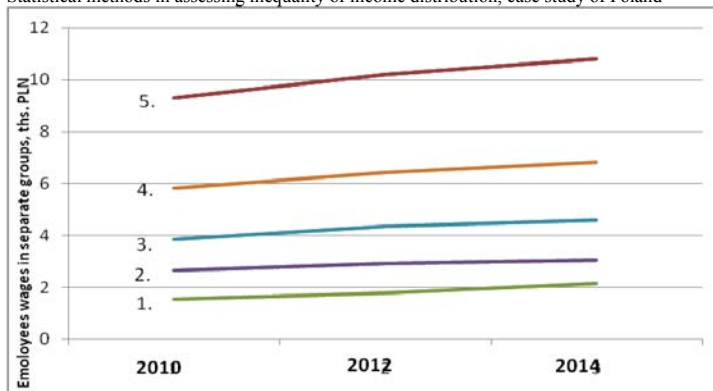
To simplify the determination of Gini coefficient, the Lorenz curve is approximated by the following polynomial function:  $y = 0,008x^2 + 0,017x + 4,397$ . This enables to simplify the determination of the area under the Lorenz curve by integration  $\int_0^{100} y(x)dx$ , which equals to 3191,367. Then the Gini coefficient:  $G=36,2\%$ .

It should be noted, that function of parabolic type is one of the approach and can be described by other type of function, such as spline and others.

It should be noted that this figure is found for gross salaries only and in general it is calculated for the entire population, taking into account other factors. Also it should be highlighted, that the distribution and G coefficient are approximately identical for all the period. Its value is determined for the labour market may differ due to additional redistribution of similar size defined for the entire population.

The sustainability of Gini coefficient in the target interval of time leads to the conclusion that the general practice in business reached approximately the same level of fairness of the distribution of wages. But it is important to analyse the dynamics of salaries for certain categories of workers by their wage.

Directly, the dynamics of wages growth, as we found differs in some selected groups of employees. These selected groups have the feature, that number of employees is almost constant in each studied period. For the selected 5 groups, this dynamic is shown on Fig. 4.



**Figure.4.** *The dynamics of wages in 5 major groups of employees*[6,7,8].

According to the actual statistics the wages of employees in 4 years increased from 1,564 to 2.148 ths. PLN, which equals 37%. The salaries of employees in the group with highest wages over the same period increased from 9.283 to 10.806 ths. PLN, so by 16%. When oriented in general on all employees, the average salary of 3.373 ths PLN in 2010, rose in 2014 to 3.981 ths.PLN, 18%. Thus, we have different dynamics of growth of wages for individual groups of employees. And the biggest difference in growth is among the groups with the lowest and highest earnings.

The research allowed to identify some patterns of non-standard quantitative growth of employees' wages on their teams. The following linear dependences were obtained by means of correlation with almost functional dependence ( $R^2= 0,995$ ).

Between the salaries of employees in 2010 and 2012 the following relationship takes place :  $x_{2012}=1,092x_{2010}+0,069$ .

The important here is the fact, that this relationship is one and the same for all 5 groups of employees. So we have the pattern, according to which the distribution of wages in 2010, transformed in their division in 2012 for one and the same linear dependence. A similar transformation is made for salaries in 2014 and the corresponding dependence is as follows:  $x_{2014}=1,039 x_{2012}+0,174$ .

Here this relationship is valid for all 5 groups of employees. The presence of a linear relationship can turn to determine the appropriate rate of wage growth:

$$1) \frac{x_{2012}}{x_{2010}} = 1,092 + \frac{0,069}{x_{2010}}; \quad 2) \frac{x_{2014}}{x_{2012}} = 1,039 + \frac{0,174}{x_{2012}}$$

## Conclusions

According to obtained dependencies, the rate of wage growth is inversely proportional to their basic values - higher wage increases to a lesser extent. Therefore, the growth rate of wages contains two components, including a constant and a variable that inversely proportional to the basic salary.

So we can assume that some justice is achieved in such a distribution of total payroll. It should be noted, that a constant linear dependence undergoes some changes with time influenced by the process of economic growth and changes in the average wage of workers. In subsequent studies, in our view, it is important to consider the factors and rationale availability transformation formulas on wages and their existence in the economies of other countries. In subsequent studies, in our opinion, the factors and rationale availability of transformation formula regarding costs shall be considered along with comparative studies for other countries. These results are important for economic decision and social policy making and requires future analysis.

## References

1. A. Atkinson : Inequality. What can be done?, Cambridge, Harvard University Press (2015);
2. Barro, R. Inequality and growth in a panel of countries. *Journal of Economic Growth* 5: 5-32. (2000),
3. E.Granaglia, M.Raitano:Le tante face della disuguaglianza economica.La rivista delle politiche sociali. PP.11-15 (2016)
4. Eurostat:Income distribution statistics (2017), electronic resource [[http://ec.europa.eu/eurostat/statistics-explained/index.php/Income\\_distribution\\_statistics](http://ec.europa.eu/eurostat/statistics-explained/index.php/Income_distribution_statistics)]
5. Focus on inequality and growth, OECD, (2014)
6. GUS, Rocznik statystyczny Rzeczypospolitej Polski/Statistic yearbook of Republic of Poland (2010)
7. GUS, Rocznik statystyczny Rzeczypospolitej Polski/ Statistic yearbook of Republic of Poland (2012)
8. GUS, Rocznik statystyczny Rzeczypospolitej Polski/ Statistic yearbook of Republic of Poland (2014)
9. Income inequality remains high in the face of weak recovery, OECD statistics (2016)
10. M.Raitano. Income inequality in Europe since the crisis/Leibniz information center for economics.PP.67-72, ZBW and Springer-Verlag Berlin Heidelberg (2016)
11. Maciej Kropiwnicki, Michał Polakowski, Dorota Szelewa and others. *Globalna polityka sprawidliwości społecznej*. Friedrich-Ebert-Stiftung. Warszawa (2015)
12. P.Wróbel. Wyznaczniki ekonomiczne sprawidliwości społecznej w Polsce/ *Studia Socialia Cracoviensia* 5, nr 2 (9), Krakow (2013)
13. Pathways, a magazine on poverty, inequality and social policy, Stanford Center on Poverty&Inequality (2016)
14. Tinbergen, J. *Income distribution*, North-Holland, Amsterdam (1975)
15. United Nations, Economic and Social Council/Progress towards Sustainable Development goals E/2016/75\*, New York (2016)
16. Van de Sande, A., &Byvelds, . *Statistics for social justice: A structural perspective*. Halifax, NS: Fernwood Publishing.(2015)
17. Wren-Lewis, S. "Ready for take-off: the role of helicopter money in supporting wage growth in future recessions" in G Kelly and C D'Arcy, editors, *Securing a pay rise*, Resolution Foundation, London (2015).



# Network inference in Genomics

Ernst C. Wit

**Abstract** The whole concept of network inference in genomics has multiple meanings and interpretations. It can refer to “causal” or “topological” considerations, i.e., learning about functional relationships in the genomic system or to considerations about the structure of the overall genomic network. Moreover, the genomic network does not really exist and can refer to gene regulatory networks, cell signalling networks, metabolic networks etc.

In this manuscript I aim to clarify the underlying genomics in order to motivate a hierarchy of four network inference strategies, starting at the single cell level and finishing at global structural network inference. It will involve stochastic and ordinary differential equation models, causal inference and graphical modelling.

**Key words:** networks, stochastic differential equations, ordinal differential equations, causal inference, graphical models

## 1 Introduction

Networks have become an important paradigm to describe genomic systems: from describing the physical, molecular interactions between proteins to the abstract interactions functional genetic units, the jargon of networks has been adopted eagerly by biologists tasked with study of this complex system. In this paper, we outline four modelling strategies, that are useful in various aspects of this enterprise. We start with a system of stochastic differential equations to describe single cell interactions. Often, however, data is observed at either a more agglomerated or across a number of cells that are destructively sampled. In those case, temporal models are more appropriately described by means of ordinary differential equations. Both these models

---

Ernst C. Wit

Johann Bernoulli Institute of Mathematics and Computer Science, University of Groningen

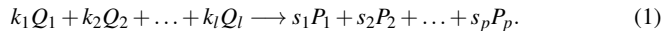
e-mail: e.c.wit@rug.nl

are inherently dynamic. Nevertheless, it is sometimes more appropriate to describe the genomic interactions by means of a single directed network, whereby the arrows have an inherently causal interpretation. We will introduce Pearl's causal framework and show how we can extend this beyond the usual Gaussian assumptions. Finally, we will drop the causal assumptions to merely describe the relationships between genomic components. This can give hints about the existence of certain interactions that may be responsible for particular phenotypes.

## 2 Stochastic differential equations for single cell interactions

The process of carrying over of signal (information) in the cell's environment is regulated by various signal transduction pathways. This signalling process is typically started by an external stimulus of the pathway leading to a binding of the signal to a receptor, i.e. hormones or growth factors, and ends up by a binding of a target protein. All cellular decisions such as cell proliferation, which refers a frequent and repeated reproduction of the cell, differentiation, which is the development of cells with specialized structure, or apoptosis, which implies cell death as a result of an intracellular suicide programme, are directed by different levels of transductions Hornberg (2005).

A general biochemical reaction can be defined as



where the terms on the left side, denoted as  $Q$ , are called the *reactants* and ones on the right side, denoted as  $P$ , are named as the *products*. The coefficients  $k_i$  ( $i = 1, \dots, l$ ) and  $s_j$  ( $j = 1, \dots, p$ ) represent the *stoichiometric coefficients* associated with the  $i$ th reactant  $Q_i$  and the  $j$ th product  $P_j$ , respectively.  $l$  refers the number of required reactants and  $p$  stands for the number of resulting products. So the chemical interpretation of this equation is that  $k_i$  molecules of type  $Q_i$  collide with each other and produce  $s_j$  molecule of type  $P_j$  while molecules move around randomly by Brownian motion Wilkinson (2006). Therefore under thermal equilibrium and fixed volume a biochemical reaction shows which species and in what proportions react together and what they produce Bower & Bolouri (2001).

For a set of  $r$  reaction and  $n$  species, accordingly, we can show the molecular transfer from reactant to product species as a net change of  $V = S - K$  where  $V$  is called the  $n \times r$  dimensional *net effect* matrix when  $S$  denotes the  $n \times r$  dimensional matrix of stoichiometry of products and  $K$  is the  $n \times r$  dimensional matrix of stoichiometry of reactants.

The master equation is defined as a differential equation for the process transition probability and can be written as:

$$\frac{dP(X;t)}{dt} = \sum_{k=1}^r \{h_k(X - V_k, \theta)P(X - V_k, t) - h_k(X, \theta)P(X, t)\} \quad (2)$$

By means of a multivariate Taylor expansion, is possible to derive an equivalent and alternative formulation of any master equation, named the Kramers-Moyal expansion Kampen (1981):

$$\frac{dP(X;t)}{dt} = \sum_{m=1}^{\infty} \frac{(-1)^m}{m!} \sum_{j_1, \dots, j_m=1}^N \frac{d^m}{dX_{j_1}, \dots, dX_{j_m}} [a_m(X, \theta)P(X, t)] \quad (3)$$

where  $a_m(X)$  are  $m$ -order symmetric tensors commonly called *jump moments* Moyal (1949) or *propagator moment functions* Gillespie (1992).

We will derive a methods of moments type estimator to infer the parameters of interest. This involves matching for each observation  $X_t$  with its expected value given the previous observation,

$$X_t = m(t; \theta) + \varepsilon_t \quad (4)$$

where  $m(t; \theta)$  is a known non-linear function of the process state at time step  $t - 1$  and  $\varepsilon_t$  is an  $N$ -dimensional mismatch variable with  $E[\varepsilon_t] = 0_N$  and  $\text{Var}(\varepsilon_t) = W_t$  and  $W_t = g(X_{t-1}; \theta)$  is a  $N \times N$  matrix, for some conveniently defined expectation and variance operators.

The conditional expectation of the process at time  $t$  given the previous time point  $t - 1$

$$m(t; \theta) = E[X_t | X_{t-1}, \theta]$$

is non-linear. It is an  $N$ -dimensional vector corresponding to the  $N$  predicted values for  $X_t$  at time  $t$  given the previous observed process state,  $X_{t-1}$ . Then we can derive a system of differential equations for the evolution of process first moments. We define the function  $m$  as the solution of the  $N$ -dimensional system of ordinary differential equations,

$$\frac{dm_i(t)}{dt} = E[a_1^i(X_t; \theta) | X_{t-1}], \quad i = 1, \dots, N, \quad (5)$$

with initial conditions  $m(t - 1) = X_{t-1}$ ,  $i = 1, \dots, N$ , where

$$a_1^i(x, \theta) = \sum_{k=1}^r V_{ik} h_k(x; \theta), \quad i = 1, \dots, N. \quad (6)$$

If the hazard functions  $h_k$  are linear in  $x$ , then (5) simplifies to  $\frac{dm_i(t)}{dt} = a_1^i(m; \theta)$ ,  $i = 1, \dots, N$ .

### 3 Ordinal differential equations for genomic interactions

Consider the gene regulatory or signaling network, described by a system of ordinary differential equations of the form

$$\begin{cases} \mathbf{x}'(t) = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}), & t \in [0, T], \\ \mathbf{x}(0) = \boldsymbol{\xi}, \end{cases} \quad (7)$$

where  $\mathbf{x}(t)$  takes values in  $\mathbb{R}^d$ ,  $\boldsymbol{\xi}$  in  $\Xi \subset \mathbb{R}^d$ ,  $\boldsymbol{\theta}$  in  $\Theta \subset \mathbb{R}^p$  and  $\mathbf{f} = (f_1, \dots, f_d)^\top$  is a known function. Given the values of  $\boldsymbol{\xi}$  and  $\boldsymbol{\theta}$ , we denote the solution of (7) by  $\mathbf{x}(t) = \mathbf{x}(t, \boldsymbol{\theta}, \boldsymbol{\xi})$ . For simplicity, assume that we have noisy observations  $Y_i(t_j)$ ,  $j = 1, \dots, n$  of the first  $1 \leq d_1 \leq d$  states  $x_i(t, \boldsymbol{\theta}, \boldsymbol{\xi})$ ,  $i = 1, \dots, d_1$  at time points  $t_j \in [0, T]$ ,  $j = 1, \dots, n$ :

$$Y_i(t_j) = x_i(t_j, \boldsymbol{\theta}, \boldsymbol{\xi}) + \varepsilon_i(t_j), \quad i = 1, \dots, d_1; j = 1, \dots, n, \quad (8)$$

where  $0 \leq t_1 < \dots < t_n = T < \infty$  and  $\varepsilon_i(t_j)$  is the unobserved measurement error for  $x_i$  at time  $t_j$ . The problem is to estimate  $\boldsymbol{\theta}$  from the data  $\mathbf{Y}$ , where  $\mathbf{Y} = \{Y_i(t_j)\}_{ij}$  denotes the matrix that contains all the observations.

We define an estimator  $\hat{\boldsymbol{\theta}}_n$  of  $\boldsymbol{\theta}$  as follows:

$$\hat{\boldsymbol{\theta}}_n = \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} M_n(\boldsymbol{\theta}, \hat{\mathbf{x}}(\boldsymbol{\theta}) | \mathbf{Y}), \quad (9)$$

where for every  $\boldsymbol{\theta} \in \Theta$  the approximation  $\hat{\mathbf{x}}(\cdot, \boldsymbol{\theta})$  of the ODE solution  $\mathbf{x}(\cdot, \boldsymbol{\theta})$  is defined by

$$\hat{\mathbf{x}}(\boldsymbol{\theta}) = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}_n} \mathcal{I}_{\alpha, \gamma}(\mathbf{x} | \boldsymbol{\theta}, \mathbf{Y}). \quad (10)$$

Here  $\mathcal{I}_{\alpha, \gamma}$  is a functional with tuning parameters  $\alpha \geq 0$  and  $\gamma \geq 0$ , which is optimized over a certain finite-dimensional function space  $\mathcal{X}_n$ .  $M_n$  is a criterion function to be minimized, for example the negative log-likelihood criterion. Minimization of  $M_n$  involves starting from some initial guess  $\boldsymbol{\theta}_0$  and iterating over the parameter  $\boldsymbol{\theta}$ , where at every iteration minimization problem (10) is solved.

## 4 Causality in genomic networks

Pearl (2009) defined causality through intervention, whereby variables were externally manipulated to take certain values. This intervention changes the underlying distribution  $P$  and can be expressed by adapting the direct effect diagram. The new distribution is called the *intervention distribution* and we say that the variables, whose structural equations we have replaced have been “intervened on.” The intervention distribution of  $Y$  when doing an intervention and setting the variable  $X_i$  to a value  $x'_i$  is denoted by  $P(Y | \operatorname{do}(X_i = x'_i))$ . The intervention on variable  $X_i$  is characterized by a *truncated factorization*, in which an intervention DAG  $G'$ , arising from the non-intervention DAG  $G$  can be defined by deleting all edges which point into the node  $X_i$ . Whereas most theory is derived using a underlying multivariate normal distribution, Mahmoudi & Wit (2017) derived a way to extend the causal effect calculus to the class of nonparanormal distributions.

## 5 Graphical models for genomic networks

Graphical models are a general class of probabilistic models that interpret the network as a conditional independence graph. This simple assumption is both quite sensible from an applied biological point of view and powerful from a computational point of view. The absence of a link in a genetic network typically means that the associated two nodes do not interact or regulate each other *directly*. This means that if any intermediate regulator or metabolic state were to be kept fixed, then the pair of nodes would either not vary at all or the variation would be unrelated to each other. Computationally, this conditional independence structure by means of the Hammersley-Clifford theorem directly translates into simple factorization of the probabilistic model in more easily manageable components.

Most texts on graphical models start with the general theory and slowly build towards more practical models, such as Gaussian graphical models. We have decided to turn this around. We begin with the very specific Gaussian graphical model, which has the advantage of being estimable even in large networks. Then we slowly peel away the restrictions of such models to be able to deal also with more complicated structures.

We will describe a network inference method for sparse high dimensional biological networks. In the Gaussian graphical model setting, this means specifically inferring a multivariate normal with many zeros in its associated precision matrix. As well as making the problem more tractable by reducing the number of parameters to estimate, sparsity of the network is also something expected from the underlying biology. A sparse estimate of the precision matrix  $\Theta$  can be obtained by imposing the  $L_1$ -penalty constraint directly on the entries of the precision matrix, rather than on the regression coefficients associated to each node. The *graphical lasso*, as it is called, is defined as the following solution,

$$\hat{\Theta}_C = \arg \max_{\|\Theta\|_1 \leq C} [\log |\Theta| - \text{Trace}(S\Theta)]$$

where  $\|\Theta\|_1 = \sum_{i,j} |\theta_{ij}|$  and  $C$  is a non-negative tuning parameter.

## 6 Discussion

In this manuscript, we have introduced the idea that depending of the underlying datastructure and the question of interest, it is crucial to adjust the modelling framework. Although networks have become an important modelling paradigm in genomics, there currently is no single network model to describe the genomic interaction structures, and in many ways, it will be unlikely that there will ever be one. In fact, the fact that the underlying generative model in biology is extremely complicated, we will always rely on convenient parametrizations to answer specific questions that arise in system biology.

## References

- BOWER, J. M., & BOLOURI, H. 2001. *Computational Modelling of Genetic and Biochemical Networks*. Second edn. Massachusetts Institute of Technology.
- GILLESPIE, D. 1992. *Markov processes: An introduction for physical scientists*. Academic Press.
- HORNBERG, J. J. 2005. *Towards integrative tumor cell biology control of MAP kinase signalling*. Ph.D. thesis, Vrije Universiteit, Amsterdam.
- KAMPEN, N.G.V. 1981. *Stochastic Processes in Physics and Chemistry*. Amsterdam:North-Holland.
- MAHMOUDI, M., & WIT, E. C. 2017. *Statistical inference of causal and ordinary differential equation models*. Ph.D. thesis, University of Groningen.
- MOYAL, J. 1949. Stochastic processes and statistical physics. *Journal of the Royal Statistical Society. Series B*, **11**, 150â210.
- PEARL, J. 2009. *Causality*. Cambridge university press.
- WILKINSON, D. J. 2006. *Stochastic Modelling for Systems Biology*. Chapman and Hall/CRC.

# Using Twitter data for Population Estimates

## *Usare dati Twitter per Stime di Popolazione*

Dilek Yildiz, Jo Munson, Agnese Vitali, Ramine Tinati and Jennifer Holland

**Abstract** Twitter is increasingly being used as a source of data for the Social Sciences. However, deriving the demographic characteristics of users and dealing with the non-random non-representative populations from which they are drawn represent challenges for social scientists. This paper has two objectives: first, it compares different methods for estimating demographic information from Twitter data based on the crowd-sourcing platform CrowdFlower and the image-recognition software Face++. Second, it proposes a method for calibrating the non-representative sample of Twitter users with auxiliary information from official statistics, hence allowing to generalize findings based on Twitter to the general population.

**Abstract** *Twitter è sempre più usato come fonte di dati per la ricerca sociale. Derivare le caratteristiche demografiche degli utenti di Twitter e la natura non-random e non rappresentativa del campione, però, rappresentano una sfida. Questo lavoro si propone due obiettivi: il primo è di confrontare due diversi metodi per derivare le caratteristiche demografiche degli utenti di Twitter, uno basato sulla piattaforma di crowd-sourcing CrowdFlower, l'altro sul software di riconoscimento di immagini Face++. Il secondo obiettivo propone un metodo per calibrare il campione non rappresentativo di Twitter con informazioni sulla popolazione ottenute da fonti di statistica ufficiale, in modo da poter fare inferenza sulla popolazione di interesse, partendo dal campione non rappresentativo di Twitter.*

**Key words:** Calibration, Population, Social Media, Twitter

---

<sup>1</sup> Dilek Yildiz, Wittgenstein Centre (IIASA, VID/ÖAW, WU), VID/ÖAW; email: [Dilek.Yildiz@oeaw.ac.at](mailto:Dilek.Yildiz@oeaw.ac.at)

Jo Munson, University of Southampton; email: [J.Munson@soton.ac.uk](mailto:J.Munson@soton.ac.uk)

Agnese Vitali, University of Southampton; email: [A.Vitali@soton.ac.uk](mailto:A.Vitali@soton.ac.uk)

Ramine Tinati, University of Southampton; email: [R.Tinati@soton.ac.uk](mailto:R.Tinati@soton.ac.uk)

Jennifer Holland, University of Rotterdam; email: [j.a.holland@fsw.eur.nl](mailto:j.a.holland@fsw.eur.nl)

## 1 Introduction

Twitter data, just like data from other social media, are increasingly being used for Social Science research. However, such data are not representative of the total population: the inference made using social media is hence invalid. Also, the basic demographic characteristics of the Twitter users are not readily available and hence need to be estimated. This paper proposes a method based on calibration for reducing the existing bias between the Twitter population and the total population and it compares the results obtained using two different methods for estimating the demographic characteristics of Twitter users with the aim of establishing best practices which were used in previous research (e.g. McCormick et al. 2015; Zagheni et al. 2014).

## 2 Data

We collected Twitter data between 23 June and 4 July 2014 using DataSift's Twitter Firehose connection. This one-week period straddles the mid-year population estimates (MYE) for the usual resident population of England and Wales on 30 June 2014 which are produced annually by the Office for National Statistics (2015). Our Twitter sample consists of users who tweeted at least once during the reference week. In addition, we restrict our sample to those Twitter users who have at least one geo-located tweet in South-East England during the week of observation. The final sample comprises 22,356 unique users.

## 3 Estimating Age and Sex of Twitter Users

We estimate age and gender of the Twitter users using two distinct methodologies: crowdsourcing, via the CrowdFlower Crowdsourcing platform, and the image-recognition software Face++. By restricting our sample to all geo-located tweets, we further have information on the location of the users.

CrowdFlower provides access to a large pool of crowd-workers who will execute a specific task in exchange of a monetary reward. We designed a task which presented crowd-workers with a user's profile description and picture (if available) and random tweet, and asked them two questions: "Would you say this Twitter user is female; male; don't know; the Tweeter is a company/organization/not a person" and "Take the best guess at the user's age in years: 0-19; 20-29; 30-39; 40-49; 50+". Given the cost of such experiment, we restrict the sample to be analysed by the crowd-workers to Twitter users in the South-East England.

Face++ is an automated face-detection algorithm developed by Megvii Inc. (2013). Face++ takes links to image files as its input variable and outputs an age and



gender estimate. Face++ demands that there are one or more distinguishable faces in the image provided in order to return a valid result, hence images showing non-human entities, or where the algorithm is unable to identify a face return a null result.

Figure 1 reports the population pyramids from the 2014 Twitter population with demographic information estimated via CrowdFlower and Face++. For both Crowdflower and Face++, males outnumber females in all age groups, with the exception of ages 0-19 in Face++. According to the gender estimates based on CrowdFlower and Face++, we find that the average number of males per 100 females in the Twitter sample to be equal to 149 and 138.6 males, respectively, whereas there are 96.8 males per 100 females according to the 2014 MYE.

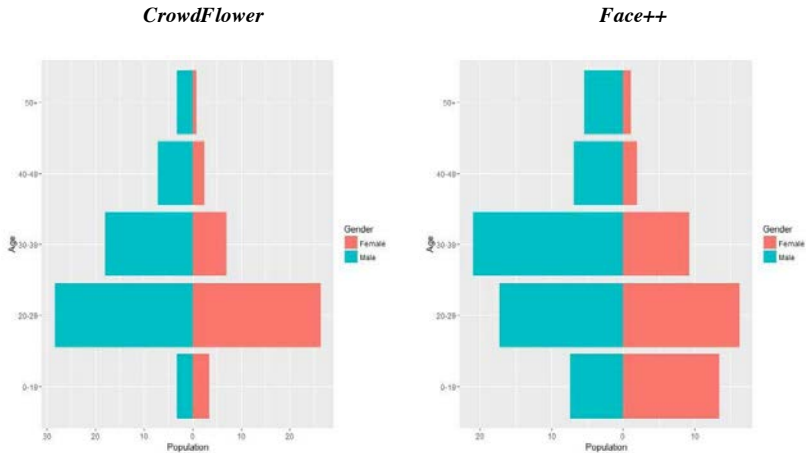
According to CrowdFlower, the age group 20-29 represents the modal age for both males and females, followed by the age group 30-39. The age groups 0-19 and 50+ are, as expected, the least represented age groups in the Twitter sample. For Face++, the most frequent group in the Twitter sample is the males aged 30-39, followed by both males and females aged 20-29. The youngest age group represents a higher proportion of the total Twitter population compared to the CrowdFlower estimates, especially among females.

In order to compare CrowdFlower and Face++, we compute a measure of performance for algorithms which attempt to assign data points to one of two or more categories, i.e. the Total Accuracy, as follows:

$$\text{Total Accuracy} = (TN + TP) / (FN + FP + TN + TP) \quad (1)$$

where T and F stands for True and False and N and P stands for Negative and Positive, respectively. In order to compute the Total Accuracy, we refer to a gold standard set of 123 randomly selected users with a valid profile picture, for whom we know the true age and sex as these were manually verified using LinkedIn profiles, Electoral Roll listings, personal websites. As Table 1 shows, the accuracy is higher with CrowdFlower. The gender matching when there is a valid profile picture is nearly 92% accurate for Face++ and 97% for CrowdFlower, but the age matching is only 35% accurate for Face++ vs. 79% for CrowdFlower.

**Figure 1:** Population pyramids based on Twitter data, demographic variables estimated with Crowdflower and Face++



**Table 1:** Accuracy of Face++ and CrowdFlower

	<i>Total Accuracy, valid images (N.=123)</i>	
	<i>Age</i>	<i>Gender</i>
<i>Face++</i>	35.8%	91.9%
<i>CrowdFlower</i>	73.2%	97.6%

## 4 Calibration Methodology

We propose a calibration approach for correcting the selection bias in a non-representative internet population. This approach relies on a regression framework for calibrating the non-representative sample of Twitter users with the auxiliary marginal information from the ‘ground truth’ data source, using log-linear models with offsets. We extend a calibration methodology developed by Yildiz and Smith (2015) to the framework proposed by Zagheni and Weber (2015).

If an auxiliary data source exists which can be assumed to measure the ‘true’ population, it can be combined with the dataset containing the counts from the Twitter population. This approach proposes to compare the ‘true’ counts of specific population subgroups by age and sex in each geographical location obtained from the ‘ground truth’ data source, with those obtained from the non-representative sample. In this example, we compare the Twitter population to the usual resident population of South-East England using the 2014 mid-year population estimates.

This source is assumed to represent the ‘true’ population count in each of the 67 local authorities in South East England, by gender and age group.

We fit a sets of log-linear models with offsets which takes into account the fact that the Twitter sample differs from the ‘ground truth’ data in terms of association structures between age and/or gender and/or location. Such models are estimated by an iterative process are similar to multiplicative weighting, raking or raking ratio estimation. We employ the IPF algorithm to fit the log-linear models with offsets and produce maximum likelihood estimates. We evaluate the capability of each model of calibrating the Twitter users’ data.

The best model, i.e. the model which reduces the bias between the Twitter sample and the total population the most, is the AS, AL model (the best model was chosen according to the mean percentage differences –see below–; results for other models are not shown). This model calibrates the Twitter population counts so that the marginal age-sex and sex-local authority marginal totals are equal to the ‘ground truth’ marginal totals. Instead, the three-way age-sex-local authority association structure is different from the ‘ground truth’ data source. The AS,SL Model can be written as follows:

$$\log(\mu_{asl}) = \lambda + \lambda_a^A + \lambda_s^S + \lambda_l^L + \lambda_{as}^{AS} + \lambda_{sl}^{SL} + \log(T_{asl}) \quad (2)$$

We denote the Census estimates and the MYE for age group  $a$ , sex  $s$ , and local authority  $l$  by  $C_{asl}$  where  $a$  denotes age groups “0-19”, “20-29”, “30-39”, “40-49” and “50+”; and  $s = 1, 2$  for males and females respectively. We assume that  $C_{asl}$  comes from a super population model and has Poisson distribution with mean  $\mu_{asl}$ . In this application we focus on the South-East region of England which consists of 67 local authorities, i.e.  $l = 1, 2, \dots, 67$ .  $T_{asl}$  is the ‘offset’ term and denotes the count of Twitter users in local authority  $l$  who are estimated to be in age group  $a$  and sex  $s$ . The factor  $\lambda$  calibrates the Twitter sample to match the South-East total population count;  $\lambda_a^A$  calibrates its age distribution, irrespectively of sex and location;  $\lambda_{as}^{AS}$  calibrates its age-sex distribution, irrespectively of location; etc.

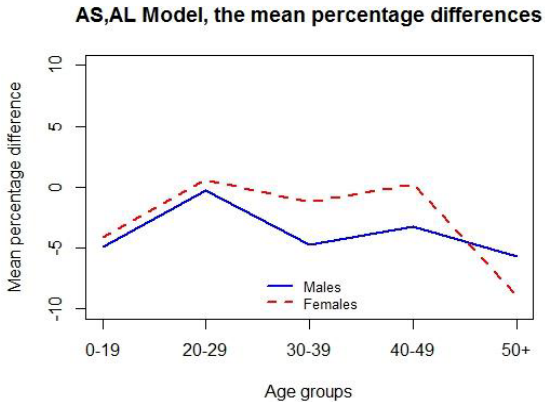
In order to ease the interpretation of results, the models are evaluated using percentage differences between the Twitter population and the population estimates in the ‘ground truth’ data source, defined as follows:

$$D_{asl} = 100 \times (P_{asl} - C_{asl}) / C_{asl} \quad (3)$$

where  $C_{asl}$  denotes the population counts estimated by the MYE for age group  $a$ , sex  $s$  and local authority  $l$  and  $P_{asl}$  the corresponding population counts. Figure 2 plots the mean percentage differences by age group and sex for the AS,AL model, using the CrowdFower estimates of age and sex. This figure shows that combining the Twitter sample with auxiliary age-sex and age-location association structures indeed decreases the bias in the Twitter sample substantially: the mean percentage differences decrease to reach the 0-5% range. We conclude that adjusting the Twitter sample by both the age group-gender association and the age group-region

Figure 2 also shows that overall our model slightly underestimates the populations of both sexes. The age category which is underestimated the most is the 50+ for both sexes.

**Figure 2:** Mean percentage difference between the MYE and calibrated models based on the Twitter users' population according to age groups, 2014 CrowdFlower



## 5 Conclusions

This paper proposed a modelling approach based on log-linear models with offsets for reducing the selection bias in the Twitter population. The population estimates derived from the model allows a considerable improvement towards the correction of the bias between the Twitter population and the real population, allowing researchers to make inference from the non-representative Twitter sample to the population of interest.

Moreover, this contribution has compared the accuracy of the age and gender estimates produced by the crowd-sourcing and image-recognition approaches. One of the major drawbacks of the Face++ approach is that it takes only an image as its input variable. If there is no image available for a user, or if the image does not clearly display a human user, the Face++ algorithm fails. In contrast, CrowdFlower users are able to utilise the username, tweet content and description as well as the image to guess the demographics of the user. Whilst the CrowdFlower results are clearly the most accurate, Crowd-sourcing assignment is not free and can be time consuming. Face++ is free and comparatively quick and could thus be

considered the best approach for gender matching where there is an identifiable user in the profile image, whereas Face++ is not an effective tool for the measurement of age.

## References

1. Megvii Inc. (2013) Face++ Research Toolkit. Available at: <http://www.faceplusplus.com>
2. McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., and Spiro, E. S. Using Twitter for Demographic and Social Science Research Tools for Data Collection and Processing. *Sociological Methods & Research*, doi: 0049124115605339 (2015).
3. Office For National Statistics (2015). Annual Mid-year Population Estimates, 2014
4. Yildiz, D., and Smith, P.W.F. Models for Combining Aggregate-Level Administrative Data in the Absence of a Traditional Census. *Journal of Official Statistics*, 31(3):431-451 (2015).
5. Zagheni, E. and Weber, I. Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1): 13-25 (2015)
6. Zagheni, E., Garimella, V.R.K., Weber, I. and State, B. Inferring international and internal migration patterns from twitter data. *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web (WWW)*: 439-444 (2014)



# Structured Approaches for High-Dimensional Predictive Modeling

Marco Seabra dos Reis<sup>1</sup>

**Abstract** Current predictive analytics approaches are strongly focused on optimizing accuracy metrics, leaving little room to incorporate a priori knowledge about the processes under analysis and relegating to a secondary concern the interpretation of results (Hastie, Tibshirani, & Friedman, 2001; Reis & Saraiva, 2005; Rendall, Pereira, & Reis, 2017). However, in the analysis of complex systems, one of the main interests is precisely the induction of relevant associations, in order to understand or clarify the way systems operate. On the other hand, there is often information available regarding the structure of the processes, which could be used in benefit of the analysis and to enhance the interpretation of results. The importance of this issue is not new and has motivated the development of multiblock approaches that try to improve the interpretation of results, while maintaining the quality of predictions (Naes, Tomic, Afseth, Segtnan, & Måge, 2013; Tenenhaus & Tenenhaus, 2014; Trygg & Wold, 1998; Westerhuis, Kourti, & MacGregor, 1998).

In this paper, two classes of multiblock frameworks are addressed, that present interpretational-oriented features, while allowing some system structure to be incorporated. One class is based on the existence of a priori knowledge for building the blocks of variables, while the other is able to extract the system structure in a data-driven way (Reis, 2013a, 2013b). The introduction of such block structures in the predictive platforms constraint their predictive spaces, for the sake of enabling interpretable elements in the final model. These constraints do not usually compromise the methods' performance when compared to their unconstrained counterparts, and sometimes even led to improvements in prediction ability, due to the use of more parsimonious and robust models.

**Key words:** Multiblock methods; Network-Induced Supervised Learning; Concatenated PLS; Multiblock PLS; Hierarchical PLS; Sequential Orthogonalised PLS

---

<sup>1</sup>Marco Seabra dos Reis  
CIEPQPF – Department of Chemical Engineering, University of Coimbra, Rua Sílvio Lima, 3030-790,  
Coimbra, Portugal, email: marco@eq.uc.pt

## 1 Introduction

Modeling is a fundamental piece of most workflows for process improvement, monitoring and control. With the increasing availability of data in the Big Data era and with the emergence of Industry 4.0, there is a strong emphasis in developing data-driven modeling approaches to address these tasks. In the domain of data-driven predictive frameworks, the mainstream methods tend to treat all variables, a priori in the same way, and the different methodologies available provide different solutions to the way variables are selected and/or combined in order to compound the final model. We call these methods “single-block”, as they treat all regressors equally in a first stage.

Taking a closer look to the recurrent structures found in data and to the way systems generating them actually work, one can notice that, most often, not all variables are actively contributing to the phenomenon under study (as assumed in multivariate methods) nor are they bringing independent and isolated pieces of information to the model (for which variable selection schemes would be adequate). Rather, the prevailing structure seems to present the form of clusters of variables – modules – composed by sets of variables, where each cluster is relative to a given functional mechanism. Variables falling in the same cluster exhibit some degree of mutual correlation, and may be aggregated in a higher level in an hierarchical way at an higher level of abstraction (Clauset, Moore, & Newman, 2008; Guimerà & Amaral, 2005; Newman, 2006; Ravasz, Somera, Mongru, Oltvai, & Barabási, 2002).

In this context, instead of multivariate or variable selection schemes for handling high-dimensional systems, methods should enable the definition and selection of modules of variables that better reflect the system structure, which can then be selected for integrating the model according to their predictive power (Reis, 2013a, 2013b). This setting calls for methods that are able to handle simultaneously several heterogeneous blocks of variables in their formulation, called hereafter as multiblock methods.

Two classes of multiblock methods can be identified upon a close analysis of the available literature. On one hand, there are methods where the blocks of variables are defined based on a priori knowledge about the system’s structure, e.g., when variables regard different process units, arise from different analytical measurement sources or are related to distinct and identifiable functional modules of the system. On the other hand, one can find methods where such knowledge is not explicitly known, but a modular structure is believed to exist, that must be inferred and extracted from the available existing data.

It is the purpose of this work to briefly provide an overview of both classes of multiblock methods and to illustrate their application with resort to a real world case study. This article is organized as follows. In the next section, the main



representatives of the two classes of methods, that will also integrated the present work, are briefly presented. Then, in Section 3, the results obtained are presented and discussed. Section 4 concludes this paper, with an overview of its contents and some concluding remarks about the relevance of considering multiblock methods in the analysis of high-dimensional systems.

## 2 Multiblock methods for predictive data analysis

In this section, a short presentation of the methods addressed in this work is provided. For more details on the implementation and use of these methods, we refer the interested readers to the references cited.

### 2.1 Class 1 – Composition of the blocks is known a priori

Belong to this class all multiblock methods that assume the composition of the different blocks to be known a priori, i.e., the following mapping can be established using background knowledge: variable<sub>*i*</sub> → block<sub>*j*</sub>. This attribution is often possible to be made when there is sufficient knowledge about the system, and the way variables are naturally organized regarding how they contribute to the final outcome. Several multi- and megavariate methods fall in this category, and we will address the mainstream ones, namely, Concatenated PLS (CPLS), Hierarchical PLS (HPLS), Multiblock PLS (MPLS), as well as recent advances in this field, such as Sequential Orthogonalised PLS (SO-PLS).

In brief terms, Concatenated PLS (CPLS) consists in concatenating all blocks of variables in a single augmented matrix and perform the classical PLS method over the entire data array. The different blocks should be weighted before being used in the model in order to give equal importance to all or to increase or decrease the importance of a given block in the model. Typically two block-scaling methods are described in the literature: soft block scaling and hard block scaling (Eriksson et al., 2006).

In Hierarchical PLS (HPLS), each data block is considered as a separate source of information and the multiblock model extracts the common structure for all the different blocks. This common structure forms the so-called super level of the model, combining information from all blocks of predictors at the lower levels. This means that block scores, loadings and weights for each separate block are available for interpretation in the lower level and super scores, loadings and weights are available in the super level for the interpretation of the global model.

Multiblock PLS (MBPLS) was proposed by Wold et al. (Wold, Martens, & Wold, 1983) and later by Wangen and Kowalski (Wangen & Kowalski, 1988). Similarly to the HPLS method, this method also presents two levels: the super level with global information and the lower level with information from each block. The main difference between this method and HPLS is that the Y block is regressed on all descriptor X blocks, whereas in HPLS the Y block is only regressed on the super block, which means that the block scores are calculated in an unsupervised way. This causes the block scores to be different in the two methods.

Orthogonalized Partial Least Squares (SO-PLS) was proposed by Naes et al. (2013) and is a methodology that incorporates the several blocks of variables in the model, one at a time, while evaluating/interpreting the incremental or additional contribution of the different blocks for improving the model predictions. This capability is relevant when one wants to assess the gain of introducing an additional source of information. The sequential nature of SO-PLS implies that the order chosen for including the blocks in the model can influence the final result.

## **2.2 Class 2 – Composition of the blocks is unknown**

When the composition of the blocks is unknown, it must be induced from data. Network Induced Supervised Learning Regression method (NI-SL) is a method proposed by Reis (Reis, 2013a, 2013b), aiming at bringing interpretation features to the forefront of the analysis goals. The method was divided in two stages. Stage 1 (Network-Induced Clustering) aims at finding functionally related groups of variables (clusters), which will form meaningful X blocks with predictive power for Y. The second stage consists in developing a predictive model, based on variates computed for the blocks induced in the first stage. For such, classical PLS models are developed separately between each X block and the Y response, and a predefined number of latent variables are retrieved from each block (in the present study five latent variables were retrieved from each block). These latent variables are gathered into a super block and a forward stepwise regression procedure is used to select the subgroup of latent variables that lead to the best fit.

NI-SL can also belong to Class 1 if the cluster compositions are known beforehand, in which case the NI-clustering stage is bypassed and the method starts immediately in the second stage of modelling. This is the case for the example addressed in this work, which will be described in the next section.

### 3 Results

We illustrate the application of multi-block methodologies with resort to an example from the wine production industry. More specifically, this example is focused on the prediction of ageing time in Madeira wine, based on different analytical measurement sources: volatile profile (1<sup>st</sup> block), the polyphenols and two furanic compounds (2<sup>nd</sup> block), the organic acids quantification (3<sup>rd</sup> block) and the ultraviolet-visible spectra (4<sup>th</sup> block). The volatile profile was analysed by gas chromatography coupled to mass spectrometry (GC-MS), preceded by solid phase extraction; the second block of data was obtained by High-Performance Liquid Chromatography combined with Photodiode Array Detection (HPLC-DAD; direct injection); organic acids (the 3<sup>rd</sup> block of variables) were also quantified by Liquid Chromatography combined with Photodiode Array Detection; UV-Vis absorbance spectra (4<sup>th</sup> block of variables) was done in a Perkin-Elmer Lambda 2 spectrophotometer (Waltham, MA, USA). More information about this case study and results is available elsewhere (Campos, Sousa, Pereira, & Reis, 2017).

A total of 26 samples were analysed, covering a range of 20 years (2-3 wine samples were taken per ageing year, with 2 year intervals). All samples correspond to wines produced from the same grape variety (Malvasia) and were supplied from the same Madeira wine producer.

In this paper, and due to space constraints, the analysis will be focused on the predictive capabilities of the methods, which was assessed according to the procedure described next. For each multiblock algorithm, 50 models were estimated using the datasets described above and Monte Carlo assignment of samples. In each Monte Carlo assignment, the dataset is randomly divided into a test set (20%) and a training set (80%). The training set is used to calibrate the model and to determine the respective hyper-parameters based on 10-fold cross validation method. This procedure is repeated 50 times originating 50 models for each multiblock method. The test sets are used for prediction based on which one computes the root mean square error of prediction (RMSEP), for each Monte Carlo run and each method, using equation 1.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_{test}} (y_{pred,i} - y_{obs,i})^2}{n_{test}}} \quad (1)$$

The distribution of the RMSEP over the 50 trials characterizes the method performance in terms of prediction accuracy and robustness. Moreover it can be used to compare different methods by evaluating the statistical difference in the prediction errors obtained by both, under similar testing conditions.

Table 1 presents the mean RMSEPs obtained for the several multiblock methods studies in this article. The methods leading to better performance are CPLS (with a new scaling methodology developed by the authors; see (Campos et al., 2017)) and SO-PLS, followed by NI-SL. The first two methods present superior predictive performances than the best linear multivariate methodologies applied to each block separately – see results for Principal Component Regression (PCR) and Partial Least Squares (PLS) in Table 2 (Rendall et al., 2017). These results indicate that it is possible to synergistically combine different sources of information for improving the predictive performance of the methods, even though the single-block methods based on the Polyphenol Content already lead to interesting predictive results. Moreover, the multiblock methodologies bring other interpretational dimensions to the analysis, namely regarding the importance of the different blocks for predicting the response and their redundancy or overlap, which are not addressed in this short article.

**Table 1.** Average root mean square error of prediction and the respective interquartil range (IQR) obtained in the Monte Carlo Cross-Validation procedure for the different multiblock methods tested in this paper.

<b>Method</b>	$\overline{RMSEP}$	<b>IQR</b> (75%-25%)
Concatenated PLS (CPLS)	0.93	0.61
Hierarchical PLS (HPLS)	1.48	0.44
Multiblock PLS (MBPLS) - Block Scores deflation	1.36	0.58
Multiblock PLS (MBPLS) - Super Scores deflation	1.34	0.54
Network Induced Supervised Learning (NI-SL)	1.17	0.85
Sequential Orthogonal-Partial Least Squares (SO-PLS)	0.97	0.49

**Table 2.** Average root mean square error of prediction and the respective interquartil range (IQR) obtained in the Monte Carlo Cross-Validation for single block approaches.

<b>Chemical Data</b>	<b>Method</b>	$\overline{RMSEP}$
----------------------	---------------	--------------------

Polyphenol Content	PCR	1.18
	PLS	1.17
Volatile Composition	PCR	1.55
	PLS	1.43
UV-Vis	PCR	2.23
	PLS	2.86
Organic Acids	PCR	2.93
	PLS	2.86

## 4 Conclusions

In this work, we illustrate the potential of using multiblock predictive methods in datasets composed by natural blocks of variables. The case study illustrates the advantage of using all blocks of variables simultaneously, in a structured way, rather than in an isolated fashion.

Even though multiblock methods represent constraint versions of their single-block counterparts, the predictive ability found may not be inferior. On the contrary, it was often found to be superior, which is due to their more parsimonious nature that leads to a more stable parameter estimation and finally to more accurate predictions (Reis, 2013a, 2013b). If, on top of this, one considers the expected higher interpretability of the multiblock methods, one can easily conclude about the increasing interest in adopting this modeling formalism to address the analysis of data collected from complex processes and phenomena.

## Acknowledgements

Marco Reis acknowledges financial support through project 016658 (references PTDC/QEQ-EPS/1323/2014, POCI-01-0145-FEDER-016658) financed by Project 3599-PPCDT (Promover a Produção Científica e Desenvolvimento Tecnológico e a Constituição de Redes Temáticas) and co-financed by the European Union's FEDER.

## References

- Campos, M. P., Sousa, R., Pereira, A. C., & Reis, M. S. (2017). Advanced predictive methods for wine age prediction: Part II - a comparison study of multiblock regression approaches. *Talanta*, *Accepted*.
- Clauset, A., Moore, C., & Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, *453*, 98-101.
- Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C., & Wold, S. (2006). *Multi- and Megavariate Data Analysis Part I – Basic Principles and Applications*. Umeå, Sweden: Umetrics Inc.
- Guimerà, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, *433*(24), 895-900.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. NY: Springer.
- Naes, T., Tomic, O., Afseth, N. K., Segtnan, V., & Måge, I. (2013). Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis. *Chemometrics and Intelligent Laboratory Systems*, *124*, 32-42.
- Newman, J. A. S. (2006). Modularity and Community Structure in Networks. *Proc. Natl. Acad. Sci. USA*, *103*(23), 8577-8582.
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A.-L. (2002). Hierarchical Organization of Modularity in Metabolic Networks. *Science*, *297*, 1551-1555.
- Reis, M. S. (2013a). Applications of a new empirical modelling framework for balancing model interpretation and prediction accuracy through the incorporation of clusters of functionally related variables. *Chemometrics and Intelligent Laboratory Systems*, *127*, 7-16.
- Reis, M. S. (2013b). Network-Induced Supervised Learning: Network-Induced Classification (NI-C) and Network-Induced Regression (NI-R). *AIChE Journal*, *59*(5), 1570-1587.
- Reis, M. S., & Saraiva, P. M. (2005). Integration of Data Uncertainty in Linear Regression and Process Optimization. *AIChE Journal*, *51*(11), 3007-3019.
- Rendall, R., Pereira, A. C., & Reis, M. S. (2017). Advanced predictive methods for wine age prediction: Part I - a comparison study of single-block regression approaches based on variable selection, penalized regression, latent variables and tree-based ensemble methods. *Talanta*, *Accepted* (in press).
- Tenenhaus, A., & Tenenhaus, M. (2014). Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of Operational Research*, *238*, 391-403.
- Trygg, J., & Wold, S. (1998). PLS Regression on Wavelet Compressed NIR Spectra. *Chemometrics and Intelligent Laboratory Systems*, *42*, 209-220.
- Wangen, L. E., & Kowalski, B. R. (1988). A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics*, *3*, 3-20.
- Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1998). Analysis of Multiblock and Hierarchical PCA and PLS Models. *Journal of Chemometrics*, *12*, 301-321.
- Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In B. Kågström & A. Ruhe (Eds.), *Matrix Pencils. Lecture Notes in Mathematics* (Vol. 973). Berlin, Heidelberg: Springer.