New perspectives for the quality of sub-municipal data with the Italian permanent population and housing census

Giancarlo Carbonetti, Stefano Daddi, Giampaolo De Matteis, Marco Di Zio, Davide Fardelli, Raffaele Ferrara, Fabio Lipizzi, Enrico Orsini

1. Introduction

Over the years, official statistics have shown increasing attention to the strong need for statistical information referring to sub-municipal territorial levels and, in this sense, the Population and Housing Census has always ensured the availability of sub-municipal data useful for territorial analyses, for business objectives and for social, economic and environmental decision-making processes.

Istat modernisation programme introduced the Permanent Census that, differently from the traditional decennial census essentially based on collecting data from people, is strongly based on the integration of administrative and sample data, and planned for providing yearly census results (Falorsi, 2017). This change required the adoption of new methodological and IT architectures with the aim of providing accurate and consistent figures at the various territorial levels.

In this framework, sub-municipal data derives from the integration of the Base Register of Individuals (BRI) and the Base Register of Places (BRP) (Crescenzi and Lipizzi, 2020; Fardelli et al., 2021). The quality of data depends on the quality of the registers and the procedures adopted to integrate and elaborate input data. In this regard, Istat is working to improve the result of the linkage task between the two registers to allocate individuals that, for various reasons, could not be geocoded.

This paper describes the strategy for the Permanent Census of Population and Households (PC) in Italy, with particular reference to the process of determining data at the sub-municipal level, the main criticalities and the solutions proposed for the production of quality information. The results of an experimental study conducted for the imputation of the enumeration area to non-geocoded units and for the production of the first sub-municipal census data are also reported.

2. The permanent census strategy and the production of sub-municipal data

Since 2018, ISTAT has been conducting the Permanent Census of Population and Housing. The traditional census has been replaced by a census based on a system of registers supported by sample surveys. Every year, counts at municipal level are disseminated according to the BRI, the BRP and a Population Coverage Survey (PCS). BRI contains information on some demographic variables such as gender, place and date of birth, citizenship, place of residence, derived by administrative data. BRP contains addresses, Enumeration Areas (EAs) and if possible, geographical coordinates.

All other census variables not present in the registers are collected with the traditional census questionnaire each year on household samples on representative sets of municipalities. From the integration of the data in the registers and the data collected on the sample households, census results are produced for different information details down to the municipal level.

The production of sub-municipal data in the Permanent Census is based on the integration of BRI and BRP (henceforth FRAME) which allows to locate individuals and households on the territory and enumeration areas. From the FRAME corrected for coverage errors, population

Giancarlo Carbonetti, ISTAT, Italian National Institute of Statistics, Italy, carbonet@istat.it, 0000-0003-1073-9813
Stefano Daddi, ISTAT, Italian National Institute of Statistics, Italy, daddi@istat.it
Giampaolo De Matteis, ISTAT, Italian National Institute of Statistics, Italy, dematteis@istat.it
Marco Di Zio, ISTAT, Italian National Institute of Statistics, Italy, dizio@istat.it, 0000-0002-6648-6934
Davide Fardelli, ISTAT, Italian National Institute of Statistics, Italy, fardelli@istat.it
Raffaele Ferrara, ISTAT, Italian National Institute of Statistics, Italy, rferrara@istat.it, 0000-0001-7777-3835
Fabio Lipizzi, ISTAT, Italian National Institute of Statistics, Italy, lipizzi@istat.it
Enrico Orsini, ISTAT, Italian National Institute of Statistics, Italy, eorsini@istat.it, 0000-0002-3472-4344

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giancarlo Carbonetti, Stefano Daddi, Giampaolo De Matteis, Marco Di Zio, Davide Fardelli, Raffaele Ferrara, Fabio Lipizzi, Enrico Orsini, *New perspectives for the quality of sub-municipal data with the Italian permanent population and housing census*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0106-3.20, in Enrico di Bella, Luigi Fabbris, Corrado Lagazio (edited by), *ASA 2022 Data-Driven Decision Making. Book of short papers*, pp. 113-118, 2023, published by Firenze University Press and Genova University Press, ISBN 979-12-215-0106-3, DOI 10.36253/979-12-215-0106-3

counts for sub-municipal domains can be produced. Such integration may fall giving rise to units without an enumeration area. This is mainly due to the quality of address information in administrative sources, problems in identifying and classifying addresses and to linkage errors between the BRI and BRP[1].

3. Process improvement actions

In order to overcome the criticalities of the archives and to make the calculation of sub-municipal data, ISTAT is working on different methodological solutions to improve the FRAME in order to deal with mismatches due to problems in the archives:

- *ex-ante: actions to improve the recognition and linkage of addresses between the BRI and BRP components;*
- *ex-post: deterministic and probabilistic procedures for recovering the enumeration area code of non-geocoded units.*

Ex-ante solutions will be implemented in the FRAME definition process, while ex-post solutions will be used for estimation purposes.

4. Procedures for improving address recognition and linkage

The following paragraph describes the techniques of processing addresses not recognized in the BRP entities as part of the construction of the integrated system of registers. The goal is to improve the quality and coverage of the geocoding of the resident population in Italy starting from the administrative archives of the Municipal Registry Lists (MRL).

To this end, new processing processes have been applied based on the use of different address recognition algorithms. Algorithms (normalizers) process input addresses by providing their output recognition according to their own normalized form. The address is characterized by four attributes: location; street; house number; address exponent. Failure to recognize an address is due either to under-coverage of the database on which the comparison is made or to systematic errors in the address string. In particular, systematic errors are treated according to two independent methodologies:

- *Machine learning algorithm for the deterministic parsing of systematic errors;*
- *Probabilistic Record linkage algorithm for matching the street.*

4.1 Machine learning algorithm for the deterministic parsing of systematic errors

The machine learning algorithm is used to have a tool capable of predicting the address string in its locality, street, house number and address exponent in order to identify the systematic error and then, where possible, clean the address.

The probabilistic record linkage algorithm is applied to have a tool to allow the recognition of addresses even in cases where the deterministic process with parsing fails, or to recognize addresses regardless of systematic error.

In detail, address parsing is performed using Conditional Random Fields (CRF), a probabilistic algorithm that allows the construction of a model for the segmentation and labelling of data sequences (Comber and Arribas-Bel, 2019). In the specific case reported here, it is a question of predicting the constituent parts of the address by assigning the corresponding labels to locality, street, house number and address exponent, dividing the individual words of the address into tokens. For labelling, the IOB (Inside-Outside-Beginning) format is adopted which provides for the affixing of positional prefixes to the various tokens. In order to recognize and classify each address token, the sequence of attributes that can formally compose an address must be provided as input to the model. Using NLP terminology, these attributes are called Part Of Speech (POS) and as in a grammar of any language, an attribute indicates the role the word

[1] 4.4% of BRI units (as at 31/12/2019), about 2.6 million, were non-geocoded.

plays in the sentence / address. The machine learning algorithm, after training the model, was applied to predict and unpack the address keywords, allowing you to remove systematic errors and have a new string to normalize.

4.2 Probabilistic Record linkage algorithm for matching the toponym

The procedure processes the distinct Street considering the specific form of an Italian address, consisting in the fact that a street is divided in Generic Urban Designations named DUG and Official Urban Designations named DUF.

The procedure compares separately set of DUG and DUF of the address (street) not recognized in the basic statistical register of places, using the form obtained by parsing through CFR described above.

The probabilistic matching algorithm compares the variables of the Street of the unrecognized address with the variables of the Street of the addresses recognized in BRP. In particular, the variable DUG is compared by means of a distance of type Cosine with q-grams equal to 1, the variable DUF is compared by means of a Jaccard distance with q-grams equal to 3 (Fortini and Tuoto, 2020).

The result, obtained by processing the individual provinces, and blocking the Street at the level of the municipality, generates a Cartesian product of combinations. The Cartesian product of the combinations is subject to a probabilistic procedure in order to determine the likelihood ratio (w) and the analogous posterior probability (m.d) that a pair of Street is a match. It was chosen to make the probability of concordance on the DUG dependent on the distance between the DUFs (dnc) to favour the choice of couples with concordant DUG among those with a distance dnc lower than the given threshold. The posterior probability m.d for each pair is determined by Bayes' rule and, similarly, the log-likelihood ratio w is given by:

$$w(dnc, dug, s) = ln\left(\frac{P(dnc|M)P(dug = 1|M, dnc, s)}{P(dnc|U)P(dug = 1|U, dnc, s)}\right).$$

The threshold level “s” is a pre-set parameter as a proportion of the number of roads to be combined on the set of pairs of the Cartesian product. The selection of the candidate pairs is carried out by ordering the pairs of the Cartesian product by decreasing w value and then choosing, for each street, the pair with the largest w value.

Subsequently, the associated Street are supervised by revision activities, which consists in identifying for each province the highest value of the probability w where the matching is doubtful. All Streets above this threshold value are considered to be recognized correctly, so we proceed with the reconstruction of the complete address by adding all the civics and exponents of the Street data and we proceed with the reprocessing in the basic statistical register of places.

5. Imputation procedures for non-geocoded units

The following imputation procedures were defined for the treatment of FRAME units not placed in any enumeration area (residual units):

- Family reconstruction: if one of the members is geocoded, we assign to the other non-geocoded members the enumeration area assigned to the geocoded one.
- Spatial approximation (SA): when the coordinates of the address are known, the EA of the nearest geocoded house number is assigned (distance criteria: “≤10” or “>10” house number).
- Address strings from the 2011 Census (AD2011): the retrieval of the EA is done by searching for the address of residence in the municipality among the municipality addresses in the 2011 Census.

- Real Estate Property (REP): EA retrieval occurs through the real estate property unit owned by the individual.
- Real Estate Rentals (RER): EA recovery occurs through the real estate unit of which the individual is a tenant.
- Probabilistic imputation. It is composed of a sequence of donor imputation steps mainly characterised by different imputation cells. The statistical unit is the household; the imputed value is the EA.

  The sequence of imputations steps is:

  1. Donor imputation with imputation cells: Street, EA in the 2011 Census.
  2. Donor imputation with imputation cells: Street.
  3. Random choice of an EAs belonging to the street of non-geocoded household.
  4. Donor imputation with imputation cells: EA in the 2011 Census.
  5. Donor imputation through random choice of an EA attached to an observed household.

The characteristic of those methods is that of reproducing the observed distributions of the EA with respect to the imputation cells (Little and Rubin, 2019). For example, in step 1, for an household that is in a specific street and that was in a specific EA in the 2011 census, the method reconstruct the behavior (the frequency distribution) of the units that are in the same street and that were in the same EA in 2011. A discussion on geo-imputation can be found in Henry et al. (2008), Dilekli et al. (2018) and Curriero et al. (2010).

6. Experimental study of the imputation procedures

The experiment for the assessment of the imputation procedures was divided into 3 phases:

1) The different deterministic procedures are applied independently on the same database, so that a comparative evaluation is possible, which is also useful for choosing a possible sequence of methods;
2) Based on the evidence of phase 1, an integrated imputation procedure is defined;
3) Application of the integrated procedure for assigning EAs to an updated Frame and over a larger set of municipalities. 63 municipalities are selected by different geographical area, population size, level and quality of geocoding; this set includes all major municipalities.

For deterministic imputation methods AD2011, REP, RER, an empirical evaluation is carried out on a subset of data with an observed EA considered highly reliable. The imputed EA is compared with the observed EA. The percentage of concordant EAs is an indicator of the performance of the methods. Similar evaluations are made when considering Administrative Areas (AdminA), each of which consists of the aggregation of neighbouring EAs (Table 1).

Table 1: Percentage of concordant EA/AdminA imputed by AD2011, REP, RER.

Deterministic methods	Frequency and percentage of concordant EA		Frequency and percentage of concordant AdminA	
RER	179,232	55.06%	255,195	78.40%
REP	1,468,139	82.83%	1,612,729	90.98%
AD2011	2,884,880	98.72%	2,899,427	99.21%

For the SA method, the same assessment cannot be followed, but a similar approach is adopted. SA, AD2011, REP, RER are applied independently. Units having at least two methods imputing the same EA are selected (prevalence criterion); the idea is that this EA is enough reliable. The frequency of times the EA imputed by SA is included in the prevalent EA is considered as an indirect evaluation of the performance of SA (Table 2).

Table 2: Percentage of concordant EA/AdminA imputed by SA method (2 distance criteria).

Possible inclusion of the SA method	SA with Distance ≤ 10 house number		SA with Distance > 10 house number	
	EA	AdminA	EA	AdminA
SA included	93.6%	100.0%	77.7%	99.1%
SA not included	6.4%	0.0%	22.3%	0.9%
Total	100.0%	100.0%	100.0%	100.0%

We notice a general good performance, especially referring to Administrative Area level.

7. Assessments of the accuracy of sub-municipal counts

For the probabilistic imputation methods, a replication approach is adopted for evaluating the uncertainty of the EA counts. The probabilistic imputation is repeated 100 times. The results are used to compute the Coefficient of Variation (CV) and Confidence Interval (CI) for each Enumeration Area and for each Administrative Area. In addition to the number of individuals in BRI and the percentage of non-geocoded units (NG), the average CV% of EAs by some municipality and the Width of the 95% confidence interval (CI) are shown below (Table 3).

Table 3: Average CV% of EAs by municipalities and Width of the 95% Confidence Interval.

Municipality	Livorno	Genova	Torino	Milano	Venezia	Roma	Trento	Verona
Pop. in BRI_31/12/19 *(thousands)*	157.3	573.8	871.4	1,394.7	259.3	2,839.4	119.3	259.5
NG%	0.07%	0.10%	0.08%	0.07%	0.43%	0.29%	0.17%	0.15%
CV%_average	0.1%	0.1%	0.1%	0.1%	0.2%	0.2%	0.3%	0.3%
Width_CI_average	0.2	0.4	0.7	0.8	0.2	0.7	0.9	0.9

Municipality	Firenze	Bologna	Catania	Cagliari	Monza	Napoli	Bari	Messina
Pop. in BRI_31/12/19 *(thousands)*	371.9	392.0	311.1	153.2	124.2	962.7	322.4	229.9
NG%	5.64%	0.39%	0.34%	0.53%	0.36%	2.90%	5.17%	38.84%
CV%_average	0.4%	0.5%	0.6%	0.6%	0.8%	0.9%	1.9%	4.1%
Width_CI_average	1.3	1.6	1.2	0.9	2.1	4.5	8.2	13.1

We notice a general high precision of estimator and very narrow confidence intervals. Only two municipalities have an average error above 1%: “Bari” and “Messina”. They are affected by a high level of units with missing EAs: they have 5.17% and 38.84% missing EAs respectively, while the average of missing EAs in all municipalities considered is around 2%.

8. Census data produced at sub-municipal level

After the allocation of the non-geocoded units of the municipalities involved in the experiment, the sub-municipal data referring to the 2019 Census were determined by applying a corrector for under and over coverage errors to the FRAME population. Data for EA and AdminA were obtained as a weighted sum of individuals residing there. The variables or combinations of variables produced at the sub-municipal level are:

- ✓ Population by gender and age group;
- ✓ Employed by gender and age group;
- ✓ Population by educational attainment;
- ✓ Foreign population by age group.

These data are not official but have a provisional character. The data for administrative areas have been sent to the statistical offices of municipalities with more than 100,000 inhabitants that have such areas, and those for enumeration areas only to a few large municipalities that have high quality spatial archives. The municipalities will use these data to carry out spatial analyses and to provide Istat with feedback on the level of accuracy.

9. Future developments

The definition processes of the BRI and BRP registers are continuously evolving and, together with the improvement of the quality of the information entering these registers, a higher accuracy of the geo-coding operation of individuals and a reduction of non-geocoded residual units are expected. Further quality improvement is expected from the spatial integration in BRP of dwellings and buildings with individuals and households.

The whole process of enumeration area code imputation will have to become structural in the process of producing sub-municipal census estimates.

Finally, the approach for the validation of the final data will have to be defined, also with the indications coming from the municipalities to which the data have been sent. In addition, the impact on the dissemination possibilities of the final results[2] will have to be assessed.

References

Comber, S., Arribas-Bel, D. (2019). Machine learning innovations in address matching: A practical comparison of word2vec and CRFs. *Transactions in GIS*, Vol. 23, Issue 2, pp. 334 – 348.

Crescenzi, F., Lipizzi, F. (2020). The integration of geographic and territorial data sources into the base register of territorial and geographical entities. *Statistical Journal of the IAOS*, Vol. 36, no. 1, pp. 143–149.

Curriero, F.C., Kulldorff, M., Boscoe, F.P., Klassen, A.C. (2010) Using imputation to provide location information for nongeocoded addresses. *PLoS ONE,* 5(2).

Dilekli, N., Janitz, A. E., Campbell, J. E., de Beurs, K. M. (2018). Evaluation of geoimputation strategies in a large case study. *International journal of health geographics*, 17(1), pp. 1-13.

Falorsi, S. (2017). Census and Social Surveys Integrated System. Note by the National Institute of Statistics of Italy, presented at the UNECE/Eurostat Group of Experts on Population and Housing Censuses, Nineteenth Meeting, Geneva, Switzerland, 4–6 October 2017.

Fardelli, D., Orsini, E., Pagano, A. (2021). The address component of the Statistical Base Register of Territorial Entities. In *Book of Short Papers SIS 2021*, pp. 1206-1211. Pearson.

Fortini, M., Tuoto, T. (2020). Probabilistic record linkage with less than three matching variables, in *Book of Short Papers – SIS 2020*, pp. 3-8. Pearson.

Henry, K. A., Boscoe, F. P. (2008). Estimating the accuracy of geographical imputation. *International journal of health geographics*, 7(1), pp. 1-10.

Little, R. J., Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons.

[2] It is expected that data per enumeration area referring to the 2021 census wave will be released in spring 2024.