Big data analysis and labour market: an analysis of Italian online job vacancies data

Francesca Giambona, Adham Khalawi, Lucia Buzzigoli, Laura Grassini, Cristina Martelli

1. Introduction

Economists and social scientists are increasingly making use of web data to address socio-economic issues and integrate existing sources of information. The data produced by online platforms and websites could provide a lot of useful and multidimensional information with a variety of potential applications in socio-economic analysis. In this respect, with the internet growth and knowledge, many aspects of job search have transformed thanks to the availability of online tools for job searching, candidate searching and job matching.

In European countries, there is growing interest in designing and implementing evidence-based decision-making tools to analyse Internet labour market data. The analysis of labour market online data could provide useful information, as big data - jointly with official statistics - could help answer the question namely "How to tackle the mismatch between jobs and skills?".

In this regard, the topic of skills gap, how to measure it, and how to bridge it with education and continuous training have been tackled by using the big data collection, as in the Cedefop (European Center for the Development of Vocational Training) initiative (Cedefop, 2018).

This contribution focuses on the issues arising from the use (and the usefulness) of online job vacancies (OJVs) to analyse the most recent Italian data. Data available for the years 2019 and 2020 are analysed to evaluate whether there has been any change in terms of required skills in occupations after the COVID19 pandemic. We use the index proposed by Deming and Noray (2020) that accounts for the change in skills for each occupation (here considered) between 2019 and 2020. Furthermore, some regional information is provided due to the particular importance that the territory has in the Italian labour market.

2. Online job vacancies and data

For some years on, OJVs have received increasing attention as an important source for real-time information on the labour market: thanks to the availability of more and more efficient big data analysis and text mining techniques, an enormous amount of information can be quickly collected and processed to monitor the changes in job demand.

These data provide a detailed and timely description of the jobs: the set of skills and the level of education and experience requested by the companies; the geographic location of the job; the type of contract; the economic sector of the company, etc..

In this sense, even if they cannot be used directly as a support tool for employment policies, they can be considered part of the modern view of Labour Market Information Systems (LMIS, see ETF, 2019), together with more traditional sources, such as statistical and administrative data. Moreover, OJVs also represent an important link between the labour market and the education system, because they provide updated information on the skills required by the market, an essential input to configure effective training offers (OECD, 2020). On the other hand, this type of data also has evident limitations and drawbacks, mainly related to representativeness and, in general, to quality issues (Cedefop, 2019).

Francesca Giambona, University of Florence, Italy, francesca.giambona@unifi.it, 0000-0002-1760-2062
Adham Kahlawi, University of Florence, Italy, adham.kahlawi@unifi.it, 0000-0003-4040-5590
Lucia Buzzigoli, University of Florence, Italy, lucia.buzzigoli@unifi.it, 0000-0003-3297-1023
Laura Grassini, University of Florence, Italy, laura.grassini@unifi.it, 0000-0003-4678-6507
Cristina Martelli, University of Florence, Italy, cristina.martelli@unifi.it

Our study is based on OJVs data produced for Italy by Burning Glass Technologies[1] (BGT), a company that collects millions of online job posting by scanning daily thousands of Internet sources (dedicated portals and company's websites).

The procedure for creating the database is very articulated and complex (ETF, 2019). The data are collected from different sources with various methods (API, scraping, crawling), based on the web portal characteristics, and are pre-processed to eliminate noise, outliers and duplicate entries. Then with the application of text classification algorithms the content of the ads is coded using categories based on reference taxonomies: in short, the taxonomy of variables is standardised according to the official classifications used in the various countries. These data have received increasing attention and have been analysed in numerous research works. Recently, Cammeerat and Squicciarini (2021) have analysed BGT data from a statistical point of view to assess their representativeness. Our data, in particular, refer to the OJVs posted on 239 online job portals in Italy in the period January 2019 - December 2020. The total number of ads is 1,741,621 in 2019 and 1,748,431 in 2020. They contain about 70 variables, most of them referred to official classifications (shown in brackets in the following): opening and closure date of publication, identification and description of occupation and related skills (ESCO classification), job geographic location (LAU and NUTS), economic sector of the company (NACE), educational level (ISCED).

To the aim of this contribution we use the BGT data to explore if between 2019 and 2020 skill changes occur by considering the occupation and regional classification.

3. Methods

Skill change is measured by the index proposed by Deming and Noray (2020) in order to understand if between 2019 and 2020 changes in skills required occurred.

For each year, BGT data collect all skills required for each job vacancy (*JobAds*) and for each occupation. The formulation of the index for the single occupation o is:

$$SCI_o = \sum_{s=1}^{S} \left| \left(\frac{\# JobAds_{os}}{\# JobAds_o} \right)_{2020} - \left(\frac{\# JobAds_{os}}{\# JobAds_o} \right)_{2019} \right|$$

where $\# JobAds_{os}$ is the number of job ads requiring skill s for the occupation o.
This index measures the net skill change in each occupation: the greater the index value the greater the skill change.

Due to the peculiarities of the Italian labour market, it may be useful to report the index value by region instead of occupation, in order to understand if and in which regions there has been the greatest change in required skills. To this aim, the above equation becomes:

$$SCI_r = \sum_{s=1}^{S} \left| \left(\frac{\# JobAds_{rs}}{\# JobAds_r} \right)_{2020} - \left(\frac{\# JobAds_{rs}}{\# JobAds_r} \right)_{2019} \right|$$

where r stands for each Italian region. And, finally, by crossing occupations and regions

$$SCI_{ro} = \sum_{s=1}^{S} \left| \left(\frac{\# JobAds_{ros}}{\# JobAds_{ro}} \right)_{2020} - \left(\frac{\# JobAds_{ros}}{\# JobAds_{ro}} \right)_{2019} \right|$$

[1] Source: Burning Glass Technologies. burning-glass.com. 2021.

4. Empirical findings

The index SCI_o is calculated for each occupation available in the BGT data to assess if changes occurred between 2019 and 2020. Highest values (i.e. the highest skill changes) concern mainly occupations related to the ICT as: s*tatistical and mathematical technicians and similar*, *software and application developers and analysts not classified elsewhere*, *web and multimedia developers*, s*oftware developers*, *specialists in databases and computer networks not classified elsewhere*, *web technicians* and *specialists in the design and administration of databases*. We find also some occupations as *public transport controllers and conductors*, *pawnbrokers and loan officers* and *education specialists not classified elsewhere*. Occupations as *geologists and geophysicists* have the lowest SCI values.

Overall, some skills required in 2019 disappear in 2020 such as: *MySQL* or *searching online information*; on the contrary, new skills appear in 2020 such as: *buy raw materials*, *maintain relations with suppliers*, *be updated on social media*, *interpreting the automatic call distribution data* and *create animation*.

For specific occupations, we find some skills that in 2019 are not required. For example *Android* in the occupation *social networking* or also *sell the services* in occupation *statistical and mathematical technicians and similar*. Overall, skills required in 2020 (respect to the previous year) mainly concern the (advanced) *use of computer and statistical tools*, *the ability to adapt to change* and *work in a team, offer support to customers.*

Due to the territorial characteristics of the Italian labour market, it is interesting to investigate if between 2019 and 2020 there was a change in the skills required at the regional level using the index SCI_r. Results highlight the index is higher for Molise, Calabria and Lazio, whilst the index is lower for Friuli Venezia Giulia, Marche and Emilia Romagna.
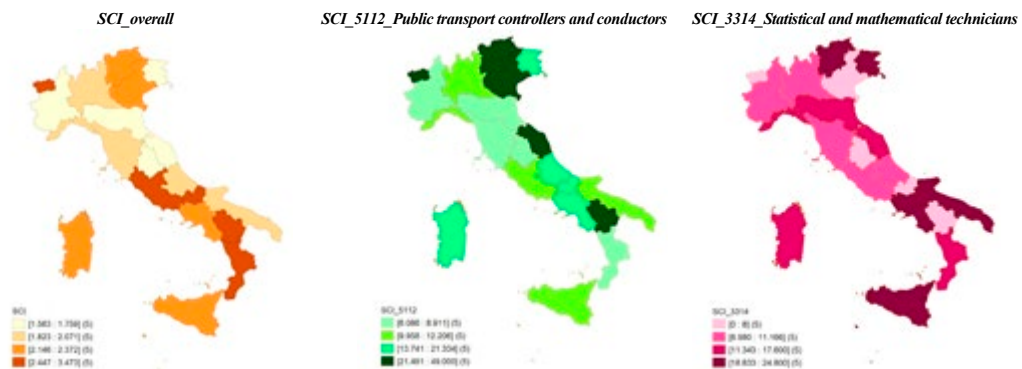
Graph 1: skill change index (SCI) at regional level

If we cross the information about occupation and regions it is possible to analyse, for the occupations with the higher SCI_o, in which region the change was highest and, therefore, whether there are any notable regional differences. Graph 1 displays the SCI_r values quartiles for the overall occupations, and the SCI_{ro} values of some occupations with higher changes.

In this respect, if we consider, for example, those occupations with highest values of SCI_{ro} we can appreciate slight different patterns across regions. In fact, we observe high values of the coefficient of variation (CV) of SCI_{ro} for those occupation with highest skill changes as, for example, CV=0.57 for *mathematicians, actuaries and statisticians*.

5. Some conclusions

The online job vacancies data give us the chance to improve information about labour market with the availability of timely data about the demand of businesses and the skills required for each occupation. In this contribution, by using the BGT data available for the years 2019 and 2020, we apply the skill change index proposed by Deming and Noray (2020) to understand if skills demand changed, for which occupation and if there are Italian regional differences.

Empirical findings suggest that between 2019 and 2020 skill changes occur, especially for some occupations and in some Italian regions. This result proves that the change in the skills required is obviously linked to each occupation (ICT-related occupations are the ones with the greatest dynamism) and to regional business environment.

By crossing occupation with regions, the skills change appears much differentiated between regions proving that for the same occupation the change in skill requirements coming from businesses are not the same, perhaps underlining a different local "perception" with respect to the skills required to carry out the same occupation.

References

Cammeraat, E., Squicciarini M. (2020), Burning Glass Technologies' Data Use in Policy-Relevant Analysis: An Occupation-Level Assessment. OECD Science, *Technology and Industry Working Papers* 2021/05.

Cedefop (2018), Mapping the landscape of online job vacancies. Background report: Italy, https://www.cedefop.europa.eu/en/publications-and-resources/

Cedefop (2019), Online Job Vacancies and Skills Analysis. A Cedefop pan-European Approach, The European Centre for the Development of Vocational Training, Thessaloniki.

Deming, D.J., Noray K. (2020). Earnings Dynamics, Changing Job Skills, and STEM Careers. *Quarterly Journal of Economics*. Forthcoming. ETF (2019), Big Data for labour Market Intelligence", European Training Foundation, pp. 1- 49

ETF (2019), Big Data for labor Market Intelligence, European Training Foundation.

OECD (2020), What skills do employers seek in graduates? Using online job posting data to support policy and practice in higher education, *OECD Education Working Papers,* No. 231.