Experimental OJA based indicators on labour demand changes: opportunities and challenges

Kostadin Georgiev[1], Francesca Inglese[2], Annalisa Lucarelli[2], Renato Magistro[2], Giulio Massacci[2], Giuseppina Ruocco[2], Galya Stateva[1]

[1]BNSI, Bulgaria

[2]Istat, Italy

Abstract

Online Job Advertisements (OJAs) offer great opportunities for analysing labour market trends due to the high level of detail of the information they contain and the high frequency with which they are made available. This work focuses on an attempt to produce new experimental indicators of changes in unmet labour demand based on OJAs broken down by strategic information not currently available in official statistics. In order to fully exploit the information provided by the OJA data, the main quality aspects of the sources – websites/job portals from which OJA data are scraped – have been taken into account, in particular their relevance and stability over time. The attempt to construct experimental indicators on the basis of OJAs has been mainly shared between two countries – Italy and Bulgaria – within the activities of the ESSnet Web Intelligence Network project and on the basis of CEDEFOP data. The main effort concerned the definition of an harmonised methodological framework covering several aspects such as: the choice of the indicators to be produced (levels or changes); the distinction between the concepts of stock and/or flow variables and their definition for the purposes of implementation on the basis of OJA data; the choice of the type of breakdown to be considered (economic activity, occupation, skill, territorial area); the level of detail (to which digit of the International Standard Classifications); and the choice of the reference period. The first experimental indicators based on OJA are the year-on-year changes in the monthly stock and flow of OJA – i.e. the stock and flow in each month compared with the same month of the previous year – broken down by skill (ESCO Skills and Competence Pillar), by major groups of occupations (ISCO-08) and by region (NUTS 2 level) over the years 2019-2022.This first attempt was useful in that the results obtained were to some extent consistent with the economic framework, particularly during the period of the health emergency. It also allowed methodological issues to be addressed and shared between the three countries involved in order to move towards a harmonised framework for the production of these indicators. The production of a new experimental OJA indicator - based on a harmonised European production framework - is particularly relevant as it could enable new requirements of EU Regulation to be met in the future, addressing unsatisfied/emerging information needs.

Keywords: OJA data, quality assessment, experimental indicators, labour demand trends, occupation and skills

1. Introduction

In recent years, National Statistical Institutes (NSIs) have been using web data sources to enrich official statistics reducing costs of data collection and burden on respondents. In order to exploit the potentials of web data, the European Statistical System (ESS) has launched the

WIN (Web Intelligence Network) project. The main objectives of this four-year initiative, which is going to end in March 2025, are:

- To develop a methodological and quality framework for web data at European level
- To implement a set of solutions, to be shared and reused through a common hub (Web Intelligence Hub)
- To foster the cooperation among NSIs for the production of trusted smart statistics (TSS) based on web data.

One of the most mature use case within the WIN project is Online Job Advertisements (OJAs), which aims both to evaluate the quality of the variables extracted from job ads, by the scraping and classification procedures, and to define new labour market indicators complementing official statistics. At the European level, the advertisements published on job portals for the recruitment of workers are centrally collected and processed by the European Center for the Development of Vocational Training – CEDEFOP (Cedefop, 2019). The statistical output is stored in a Data Lab and the databases are released on a quarterly basis, with data available at a daily level. The central repository currently contains datasets from the third quarter of 2018 to the fourth quarter of 2023. The Italian National Institute for Statistics (Istat) is a member of the WIN project and cooperates with the other participants to assess and increase the accuracy of the statistical output and to produce OJA-based experimental statistics.

In this paper, we discuss the potentials and quality issues of OJA data (section 2) and propose some solutions from the perspective of NSIs (section 3). In section 4, we present some examples of experimental OJA-based indicators, focussing on occupation and skills, while making some concluding remarks in section 5.

2. Opportunities and quality issues of OJA data

OJA data is a timely source of information for monitoring labour market trends. They also have great potential for labour market analysis and the production of new experimental indicators due to the granularity of the information. Indeed, the OJA data provide a great deal of information on job characteristics (e.g. occupation, location, type of contract and working hours and pay), employer characteristics (e.g. economic activity), job requirements (e.g. education, skills and experience) and also on the advertisement itself (e.g. job ad posting and expiring date on the website/job portal).

The quality of the derived statistical output must be high in order to make full use of OJA data for statistical purposes. Good data quality is an essential prerequisite for the production of reliable statistics and accurate analyses. The challenges of ensuring OJA accuracy can be grouped into several related areas: data acquisition, data processing (de-duplication and data

validation, classification of variables) and data analysis. Moreover, transparency becomes a fundamental requirement, encompassing all stages of the implemented workflow, starting from the source selection model used for landscaping. OJA data should be treated with caution as they have limitations in terms of coverage and representativeness. OJAs over-represent some occupations and skills while under-representing others: the first is generally related to job offers for workers with high levels of education and posted by large size companies; the second is due to the fact that not all job offers are published online, as employers also use traditional recruitment channels other than advertising on websites (e.g. word-of-mouth, as in the case of Italian small and seasonal enterprises). The representativeness of data sources is an important sub-dimension of data accuracy when assessing the quality of OJA data. It is closely related to their "linkability" and comparability over time. Changes in technology may affect data comparability over time, which can become a problem when producing statistical output to be released regularly. In summary, the main quality issues that need to be taken into account in the construction of OJA-based indicators depends mainly on: the quality of the sources and the criteria adopted for their selection, errors in the data processing stage due to the soundness of the algorithms, language processing, deduplication approach and definition of validation rules (Six et al., 2023). The accuracy of OJA-based indicators also depends on the classification process of the information extracted from each advertisement.

3. Assessing source and classification quality

One of the main aspects further investigated during the project was the quality of the sources, in order to assess the quality of the output used to calculate experimental OJA-based indicators. More in detail, the WIN methodological work package leaders (Six et al., 2023) suggested a set of quality indicators related to two main dimensions: relevance of the selected sources and their stability over time.

In the case of Italy, the selected sources, which cover about 80% of the total amount of OJAs, are very few (from 6 to 8). They have changed over time, with some sources disappearing and others being added. Due to limitations (scraping possibilities, quality of sources and agreements with the portal owners), it is impossible to scrape every existing labour portal (source) for every country. Eurostat has already selected country-specific sources (websites) during the landscaping process. Even if not all sources are selected, it is crucial that the most relevant ones are selected and that the set of sources is stable over time. As far as coverage of the scraped information for each website is concerned, the specificity of the most relevant sources – that is those with more than 5000 ads in at least one year – should be take into account. In particular, the specificity in terms of the main variables of interest – occupation,

skills and territorial area– should be examined in order to identify either the existence of specific sources (mainly covering certain occupational groups, skills and territorial areas) or sources that do not cover certain categories at all.

In addition to the quality of the selected sources, one of the main tasks of the OJAs workflow is the translation of the information expressed in Natural Language (NL) into standard international statistical classifications (e.g. ISCO, ESCO, NUTS and NACE). The classification algorithms, for extracting information from OJAs, are based on ontologies (keyword lists) and a machine learning model. Within the WIN project, to improve the accuracy of classifications, two data annotation exercises were performed. Stratified samples of OJAs containing the raw data and the classification results were selected for several countries to carry out the data annotation process. Some reviewers checked the selected OJAs, enhancing the correspondence between the ad text and the results of the classifiers. Istat participated in both data annotation exercises by analysing a sub-sample of 400 and 300 OJAs respectively. The results of the second annotation exercise show an improvement in the quality of the classification of the occupation variable: from the first to the second exercise, the percentage of cases correctly classified increases from 45% to 62%. Furthermore, for the regional variable the distribution of correctly classified cases reaches a percentage of about 60%.

4. Changes in skill demand in Italy during the pandemic emergency: some signals from OJA-based experimental indicators

Some of the most useful experimental indicators to be produced on the basis of OJA data are those on the dynamics of unmet labour demand, disaggregated by strategic information not currently available in official statistics, mainly skills, employment and territorial area (Catanese et al., 2023). In other words, the use of experimental indicators from the OJA data provides timely, detailed and high-frequency information on how enterprises' labour demand is changing in relation to changes in labour market characteristics. During the WIN project, Italy and Bulgaria have cooperated to construct experimental indicators to monitor the labour demand trends over time. This collaborative and effective approach resulted in a common and standardised code script that can be shared and reused by WIN members to produce the same indicators. The main efforts concerned the definition of a harmonised methodological framework covering several aspects such as:

- The choice of the indicators to be produced (levels or changes)
- The distinction between the stock and flow OJA variables and their definition
- The choice of the type of breakdown to be considered (economic activity, occupation, skill, territorial area, etc.)

- The level of detail (to which digit of the International Standard Classifications)
- The choice of the reference period.

To illustrate some of the key findings and the relevance of the approach shared with Bulgaria, this section reports a descriptive analysis of skill demand changes in Italy during the COVID emergency phase, from the first quarter 2020 to the fourth quarter 2022. The skills demanded by the employers by means of the online channels were measured as a stock at the last day of each quarter. Changes in the demand for skills have been assessed in terms of year-on-year changes (i.e. percentage changes between the stock in each quarter and the stock in the same quarter of the previous year). Skills are classified on the basis of the international standard classification (ESCO), as are the other variables taken into account in this section (ISCO-08 for occupation, NUTS 2 level for territorial area).The following figures show changes in the demand for skills at national level and by macro-regions.

Figure 1: Skill demand from OJA year on year percentage changes – Italy 2020Q1-2022Q4
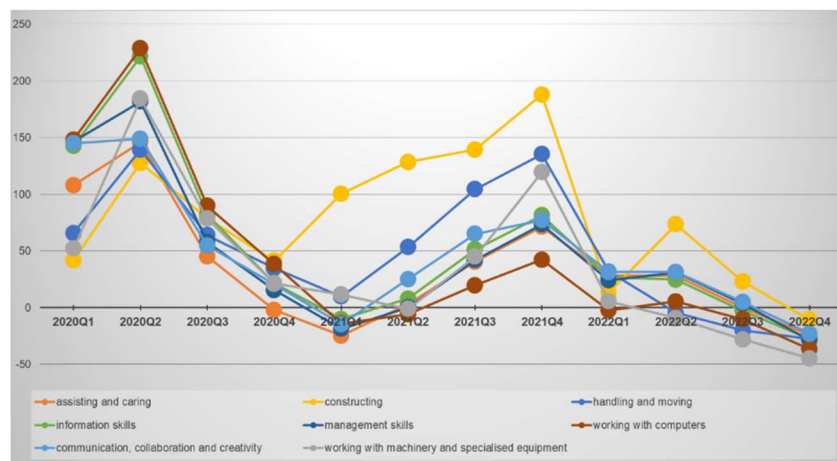

At national level, the analyses (Figure 1) show a first peak in demand for skills "working with computers" and "information skills" in the second quarter 2020, consistently with the period under analysis characterized by the pandemic crisis. Moreover, the upward trend in the demand for "constructing skills" (in the ESCO classification, this skill groups building, repairing, installing and finishing interior and exterior structures), during the year 2021, could be attributed to the bonus effect in the construction NACE economic activity sector. A similar trend is also observed in the demand for "handling and moving skills".

The dynamic of the skill demand shows marked differences at macro-region break down (Figure 2). In the Centre of Italy, the demand for skills in the construction sector shows a higher increase than the national average, approximately four times higher. In Southern Italy, the demand for skills in the construction sector shows a higher increase too – even if to a lesser

extent than in the Centre – with a peak occurring two quarters earlier. A peak is also observed in the second quarter of 2022, with an upward trend. In the North-East, on the other hand, the signal for an increase in constructing skills is much weaker than the national average and concerns the first two quarters of 2021. While in the North-West, the signal for an increase in constructing skills is halfway between that of the South and the North-East with a peak in the fourth quarter 2021 later than that one observed for the South.

Figure 2: Skill demand from OJA year on year percentage changes – Italian macro regions 2020Q1-2022Q4 (skill colours same as Figure 1)
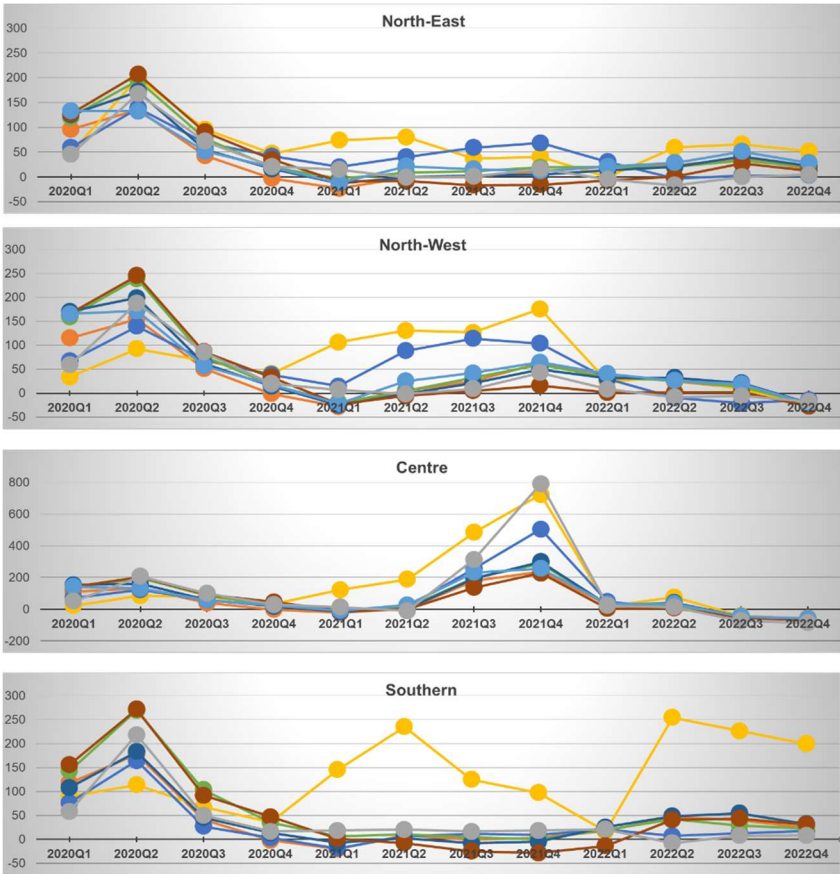

Furthermore, while at national level the increase in demand for constructing skills is also followed by an increase in other kind of skills, this is not the case in the Southern Italy, where an increase in constructing skills is the only one observed. Overall, signals from the OJA dynamics by ESCO skills are in line with the changes in the macroeconomic framework during the period considered, characterised by the pandemic crisis. The analysis took also into account changes in the percentage composition by skill of each major occupational groups, between the third quarter 2019 (prior to the health emergency period) and the correspondent quarter in 2022. For the construction of the next figures, we select two skills in each

occupational group with the higher increase (Figure 3) and higher decrease (Figure 4) in percentage points.

As Figure 3 shows, the skill "communication, collaboration and creativity" was the skill with the largest increase across all occupational groups, except for the Elementary occupation and Professional occupation groups.

Figure 3: Skill percentage composition increase by occupation – 2019Q3 and 2022Q3
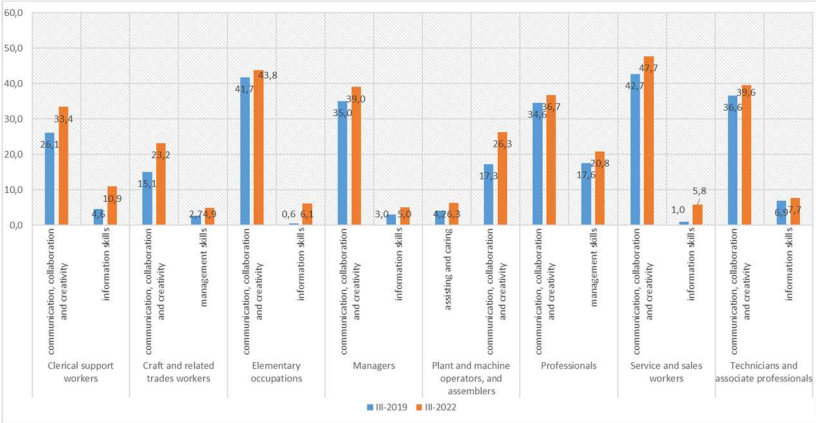
For these two groups, the greatest increase was for the information skill and management skill. These last two skills represent the second largest increase of all other occupational groups. The widespread increase in the weight of the skill "communication, collaboration and creativity" should be also due to the impact of the pandemic period. This skill actually refers to the ability in communicating, collaborating, relating and negotiating with other people, developing solutions to problems, creating plans or specifications for the design of objects and systems and also imparting knowledge to others.

Figure 4: Skill percentage composition decrease by occupation – 2019Q3 and 2022Q3
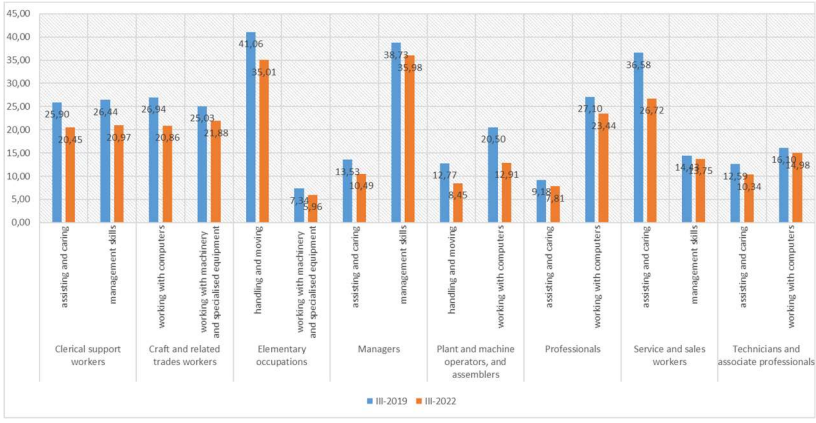
The demand for skills that showed a decrease between the two periods across occupations (Figure 4), were "assisting and caring" and "working with computers". This evidence is also in line with the effects of the Covid crisis, which led to an increased demand for these skills and a subsequent return to normal levels in the years after the health emergency.

5. Conclusions and open issues

This paper describes the potentials of OJA data, a new data source that does not replace traditional surveys but can be used to timely enrich the analysis on the labour market trends. The OJA-based experimental indicators, resulting from the cooperation with Bulgaria, are consistent with the Italian economic changes observed during the health emergency phase. This evidence highlights the usefulness of these indicators in describing the dynamics of labour demand, given their timeliness and granularity. Although the potential of the OJA data source is evident, some methodological challenges need to be addressed. OJA data are a non-probability sample and suffer from coverage issues, which remain open questions. Selectivity is strongly related to the linkability of web data. Linking online job postings with Job Vacancy Survey microdata is not possible, except in very few cases, but would be aimed at understanding the data coverage issues of scraped OJAs. Finding the appropriate reference population is the main obstacle in assessing the representativeness of statistics on the number of job vacancies derived from online job ads (Beręsewicz et al., 2021). Moreover, there is the open question of how to assess representativeness. In order to understand whether OJAs complement or overlap with data derived from vacancy surveys (Branka et al., 2022), the evaluation of the representativeness can be done by comparing information collected from online sources with information from traditional data sources (Labour Force Survey, Job Vacancy Survey).

References

Beręsewicz, M., & Pater, R. (2021). Inferring job vacancies from online job advertisements. *Statistical Working Papers*. Eurostat

Branka, J., & Kvetan, V., & Napierala, J.N. (2022). From the online job advertisements to official statistics – the aspects of quality assurance. *Q2022, Vilnius (2022)*

Cedefop (2019). Online job vacancies and skills analysis: a CEDEFOP pan-european approach. https://www.cedefop.europa.eu/files/4172_en.pdf

Six, M., & Kowarik, A. (2023). Issue 10 - Quality aspects of web scraped data - Focus on landscaping and selection of sources. *Web Intelligence Network Blog*. https://cros.ec.europa.eu/book-page/issue-10-quality-aspects-web-scraped-data-focus-landscaping-and-selection-sources

Catanese, E., & Inglese, F., & Lucarelli, A., & Ruocco, G. (2023). Analysing the labour market through a web data source: the skill dynamic based on Online Job Advertisements. *Rivista Italiana di Economia Demografia e Statistica,* Volume LXXVII n.4,
http://www.sieds.it/wp-content/uploads/2024/02/21_04531RV_Lucarelli.pdf