Napierała, Joanna; Kvetan, Vladimír; Branka, Jiri

Working Paper

Assessing the representativeness of online job advertisements

Cedefop working paper series, No. 17

Provided in Cooperation with:
European Centre for the Development of Vocational Training (Cedefop), Thessaloniki

Suggested Citation: Napierała, Joanna; Kvetan, Vladimír; Branka, Jiri (2022) : Assessing the representativeness of online job advertisements, Cedefop working paper series, No. 17, ISBN 978-92-896-3456-4, Publications Office of the European Union, Luxembourg, https://doi.org/10.2801/807500

This Version is available at:
https://hdl.handle.net/10419/337392

Working paper series
No 17 / December 2022

ASSESSING THE REPRESENTATIVENESS OF ONLINE JOB ADVERTISEMENTS

Joanna Napierala,
Vladimir Kvetan and
Jiri Branka

cedefop

The **European Centre for the Development of Vocational Training** (Cedefop) is the European Union's reference centre for vocational education and training, skills and qualifications. We provide information, research, analyses and evidence on vocational education and training, skills and qualifications for policy-making in the EU Member States.
Cedefop was originally established in 1975 by Council Regulation (EEC) No 337/75. This decision was repealed in 2019 by Regulation (EU) 2019/128 establishing Cedefop as a Union Agency with a renewed mandate.

Europe 123, Thessaloniki (Pylea), GREECE
Postal: Cedefop service post, 570 01 Thermi, GREECE
Tel. +30 2310490111, Fax +30 2310490020
Email: info@cedefop.europa.eu
www.cedefop.europa.eu

Please cite this publication as:
Napierala, J.; Kvetan, V. and Branka, J. (2022). *Assessing the representativeness of online job advertisements*. Luxembourg: Publications Office. Cedefop working paper, No 17. http://data.europa.eu/doi/10.2801/807500

Luxembourg: Publications Office of the European Union, 2022

PDF ISBN 978-92-896-3456-4
ISSN 1831-2403
doi:10.2801/807500
TI-BA-22-010-EN-N

Acknowledgements

This publication was produced by the European Centre for the Development of Vocational Training (Cedefop), Department for VET and Skills (DVS), under the supervision of Antonio Ranieri. Joanna Napierala, Vladimir Kvetan and Jiri Branka Cedefop experts, were responsible for this publication.

This publication is based on results of a feasibility study which was conducted under the project *Towards the European web intelligence hub: European system for collection and analysis of online job advertisement data* (WIH-OJA) (contract number 2020-FWC7-AO-DSL-VKVET-JBRAN-WIH-OJA002/20). Cedefop would like to acknowledge the contribution of representatives of the Interuniversity Research Centre on Public Services (Crisp) of university of Milano Bicocca.

An earlier version of this paper was presented at the 10th European Conference on Quality in Official Statistics, Lithuania, 8 to 10 June 2022.

Contents

Tables and figures

Tables

Figures

Executive summary

This working paper presents the results of an assessment of the representativeness of collected information on online job advertisements (OJA) in indicating the number of vacancies on the labour market in EU Member States. Two external data sources were used, the Labour force survey (LFS) and the Job vacancies survey (JVS), available in most EU countries. The coverage biases in OJAs data, compared to existing data sources, were evaluated at sectoral, occupational and geographic levels. While noticeable variation was observed in terms of coverage biases at sectoral level across Member States, there are also common patterns, such as high-skilled level occupations and more industrialised regions being overrepresented in OJAs.

Comparison of the job titles, present at least once in OJAs in the period analysed, indicates the increasing popularity among employers in using online job advertisements in recruitment for all kinds of roles, not just for high-skilled or managerial ones. The further analysis of occupations that do not appear in OJAs at country level seems to support that the coverage bias is not related to underusage of this approach to talent recruitment but could be related to the size of the labour market; in smaller Member States the number of missing occupations in OJAs was higher.

The conclusions of this comparison suggest that both the sources of information used as a benchmark for carrying out this evaluation also have their weaknesses. The reliability of information collected in surveys is heavily dependent on response rates, which in turn rely on various factors (e.g. willingness to participate in a survey). Moreover, the methodology of JVS is not homogenous across Member States, which also impedes comparisons. Lastly, the analysis indicates that none of the three sources of information allows precise estimates on the number of vacancies to be derived.

CHAPTER 1.

Introduction

Data of good quality are a prerequisite for creation of reliable statistics and analysis; in turn these are strategic elements for society and the economy to allow sound policy decision making in both the private and public sectors. The abundance of existing open data and big data sources is often characterised by the five Vs: volatility, variety, velocity, veracity, and volume ([1]). They cover almost any aspect of our life, creating opportunities for the production of official statistics. Since 2014, the European Centre for Development of Vocational Training (Cedefop) has been developing a system to collect and analyse data from online job advertisements (OJA) to produce skills intelligence.

The success of this endeavour, and seeing the potential of online data sources to produce new indicators to guide policy makers dealing with education and labour market policies, led Cedefop and Eurostat to push for the integration of big data into the production of official statistics. Since 2020 the development of a database based on OJAs became part of 'trusted smart statistics' belonging to the Web Intelligence Hub, referred to as the WIH-OJA database ([2]). Several years of joint project implementation brought more understanding of using OJAs as the data source, as well as of challenges related to data quality assurance.

This paper focuses on one aspect of data quality assurance: assessment of the representativeness of collected information. Two external data sources, the Labour force survey (LFS) and the Job vacancies survey (JVS), were used to evaluate the selectivity of data in terms of several criteria (e.g. comparisons on sectoral, occupational and geographic levels). The comparison of occupations listed in the European skills, competences and occupations classification (ESCO) ([3]) taxonomy was used to identify occupations on the labour market absent from OJAs.

([1]) For the explanation of terms see for example Ishwarappa, J. Anuradha (2015).

([2]) More information about the project can be found on the European Web Intelligence Hub website.

([3]) ESCO levels of taxonomy up to 4th digit level are compatible with ISCO.

CHAPTER 2.

Representativeness as quality assurance

The Big data quality framework (BDQF) presented by UNECE (2014) provides a list of quality dimensions to consider when making assessment of the production of statistical outputs based on big data. Accuracy, defined as 'the degree to which the information correctly describes the phenomena it was designed to measure' is one of them. The analysis of representativeness, which is the converse of looking at selectivity, belongs to the standard quality assurance procedures in survey research.

Yet, although the evaluation of representativeness is a standard procedure, it may be defined and carried out by researchers differently (Kruskal and Mosteller 1979, p. 111). Traditionally in representativeness evaluations, researchers focus on the impact of design and coverage (Lavrakas 2008), mainly because the potential causes of inaccuracies identified in datasets are related to aspects such as coverage, sampling, nonresponse, and response (UNECE 2014). If data availability allows, researchers use statistics derived from the general population as a benchmark for comparison.

As representativeness is dependent on context, there is yet no ultimate judgement whether a data set can be considered representative or not (Ochsner 2021). The assessment of representativeness is carried out to improve data collection or to understand potential biases affecting the estimates and inferences. In the case of the new datasets based on big data, the representativeness analysis brings more understanding about potential biases (e.g. Lin 2017).

To evaluate representativeness, this paper uses an approach proposed by Bethlehem (2009, p. 24) who suggests that 'A survey data set is defined to be representative with respect to variable(s) x if the distribution of x in the data set is equal to the distribution of this variable in the population'. The main goal of the representativeness evaluation presented in this paper is to understand to what extent OJA complements or overlaps the data derived on number of vacancies from the two well established surveys conducted by Eurostat, JVS ([4]) and LFS ([5]). Additionally, the comparison of distribution of occupations with the European multilingual classification of skills competences, qualifications, and occupations (ESCO), on 4-digit level indicates which occupations may still not be advertised online.

([4]) Eurostat. Job vacancies.

([5]) Eurostat. European Union labour force survey.

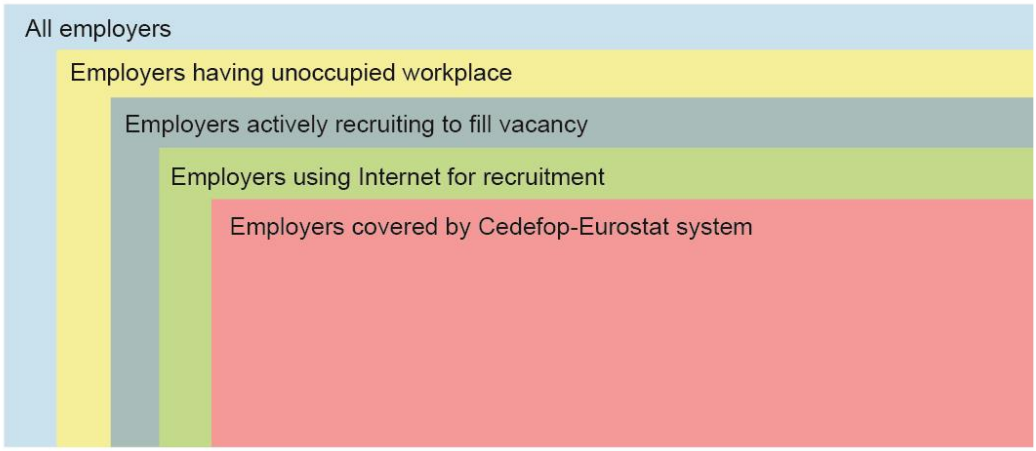
CHAPTER 3.

Evaluating online job advertisement representativeness

It is believed that the use of the term representative is not adequate in situations where no random or probability sampling is involved. Therefore, when basing analyses on big data, in which case random sampling is not the case and, instead, almost all available data is used, representativeness is debatable. For example, Beresewicz (2016) claims that the concept of representativeness is still valid in the context of big data and recommends looking at the self-selection mechanism to measure it.

The size of a dataset is not a guarantee that estimations based on online sources about vacancies and/or skills needed by employers in certain sectors will be unbiased, especially when OJAs will represent only a part of the researched population. Beresewicz and Pater (2021) list reasons for which the two concepts of OJAs and vacancies are distinct (Figure 1).

Figure 1. **Self-selection mechanism underlying relationship between OJAs and vacancies**



Source: Author's elaboration.

The most obvious one comes from the fact that not every vacancy is advertised online. However, the decision to use OJAs as recruitment channels is driven by various factors. Digital literacy in a country (or region) and jobseeker search strategies affect the decision to advertise vacancies online directly. The number of OJAs is also affected by indirect factors. For example, in the event of economic growth or structural changes, the need to increase the reach and pool

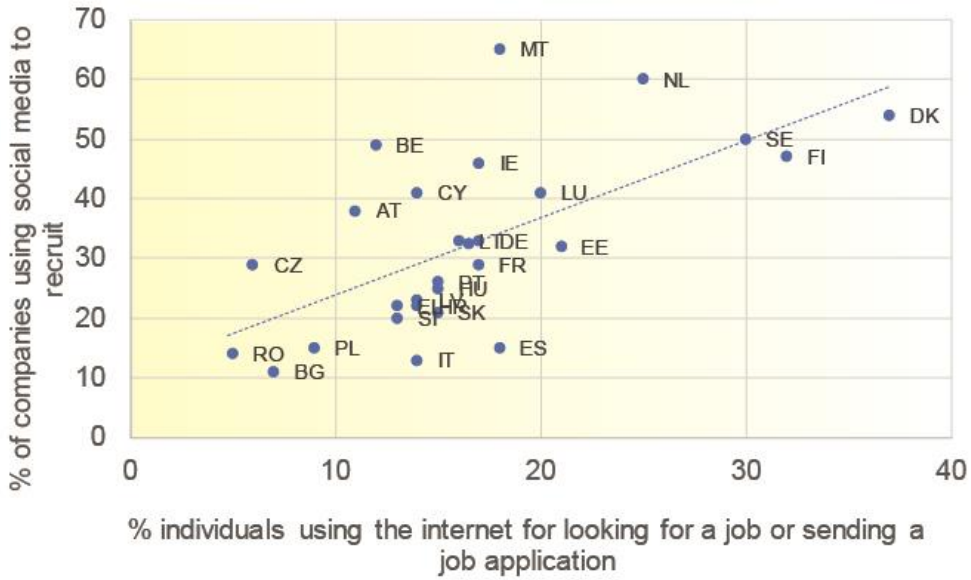
of the potential candidates drives the decision to advertise online. This trend is even more important when local labour supply does not match demand.

Some companies (for example, leading high-tech companies) may even not publish some of their vacancies at all as they receive unsolicited applications from potential candidates. In some cases, alternative channels for searching for workforce are more effective than online advertisements (e.g. using a window of a bar/restaurant to recruit waiters/kitchen helpers); this can lead to underestimation of the total number of vacancies. The preference for other than online channels in recruitment could be also understood in smaller companies due to too higher costs of online recruitment (Swier et al. 2018). Employers may prefer using online sources to search for a worker in the case of hard-to-fill vacancies, to maximise advertisement outreach. Further, the role of OJA as a recruitment channel is subject to various national specificities or driven by the role of PES and concentration of OJA portals (Cedefop 2019b). In many countries, the institutional framework makes it obligatory for employers to report a vacancy to public employment services (PES). As these institutions across the EU increasingly use online tools to create databases of vacancies, more information about openings is eventually published online.

Although both employers and jobseekers are more frequently using online job portals for faster and more effective matching, in 2019 one in three individuals, on average, were using the internet to search for jobs in Europe (Figure 2) and 16% of companies declared using social media ([6]) for recruiting purposes. Usage of social media for recruiting purposes by companies in Europe varied across countries and was much higher in Malta and the Netherlands (above 60% of enterprises), while it was very low in Bulgaria (around 10%). This use of social media was observed more frequently among bigger enterprises than small ones (62% versus 29%). OJAs tend to be biased toward high-skilled professional occupations (e.g. Carnevale et al., 2014; Kureková; Beblavý and Thum-Thysen, 2015).

([6]) This indicator could understate the use of online websites in recruitment of workers because the term used in the question can be misunderstood by employers: ‘social media’ by definition covers websites and applications that focus on ‘communication, community-based input, interaction, content-sharing and collaboration’ and not necessarily on publishing information about vacancies.

Figure 2. **Percentage of enterprises that recruit employees using social media and percentage of individuals who use the internet to search for job, by country (2019)**



Source: TIN00102, ISOC_CISMP, 2019.

Another possible underestimation of the number of vacancies based on online sources may come from employers using one advertisement to fill multiple vacancies; the employer maintaining OJAs that they are not currently seeking to fill could be another issue.

The sample is considered representative when the elements in a population under study have equal chances to be included in the sample. This condition is not always met when collecting information about OJAs, as sometimes providers of services block the possibility to access the information displayed on their website by crawlers. Also, as the web data-gathering algorithms do not operate all the time on all websites, so there is a risk that not all OJAs posted are included in the sample (Beresewicz and Pater, 2021).

Job advertisements gathered from the web may actually refer to apprenticeship or training opportunities, which from a statistical and labour law points of view are not defined as a vacancy (Beresewicz and Pater, 2021). However, this information, if clearly stated in the content of job advertisement, could be used to filter out such advertisements and allow controlling for this overestimation.

CHAPTER 4.

Evaluating the representativeness of WIH OJA data

Finding the appropriate reference population is the main obstacle in assessing the representativeness of statistics on the number of vacancies derived from OJAs (Beresewicz and Pater, 2021). In Europe, apart from administrative data which are made available for wider research community only in a few countries, JVS and LFS may serve as a potential benchmark for such evaluation. However, using either of these sources as the population benchmark brings some challenges.

The main challenge is related to the quality of both surveys, with much depending on response rates. Where response rates are low, the representativeness of the surveys may be impacted and may also complicate the comparison with the OJA data.

The other challenge is related to the unit of measurement and the need to convert information about the flows of OJAs to the stock of vacancies in a given period. This requires an estimate of the duration of the vacancies posted and is derived from websites, which is not always easy (Beresewicz and Pater, 2021). Another challenge is related to data availability, as not all EU countries report the total number of vacancies: France and Italy are examples. There is also a possible time lag between reporting on vacancy and publishing job advertisements. This is particularly important in the last quarter of the year; where a vacancy occurs by the end of the year, there is a tendency to advertise for it only in January. Conversely, some jobs (e.g. temporary, or seasonal jobs) may be advertised by the employer well before these are reported as vacancies.

The advantage of using LFS over JVS as the population benchmark for the representativeness analysis could be related to the level of detail in this source of data; it can allow comparisons not only at sectoral (NACE), but also at occupational (International standard classification of occupations, ISCO) and regional (NUTS) levels. An additional challenge is related to the fact that LFS represents the supply side while OJA represent the demand side of the labour market. To allow for comparisons between these two data sources based on available LFS microdata, the identification of the cases of new hirings or job changes that occurred within the previous 3 months was made as a proxy of the total number of vacancies posted in the quarter.

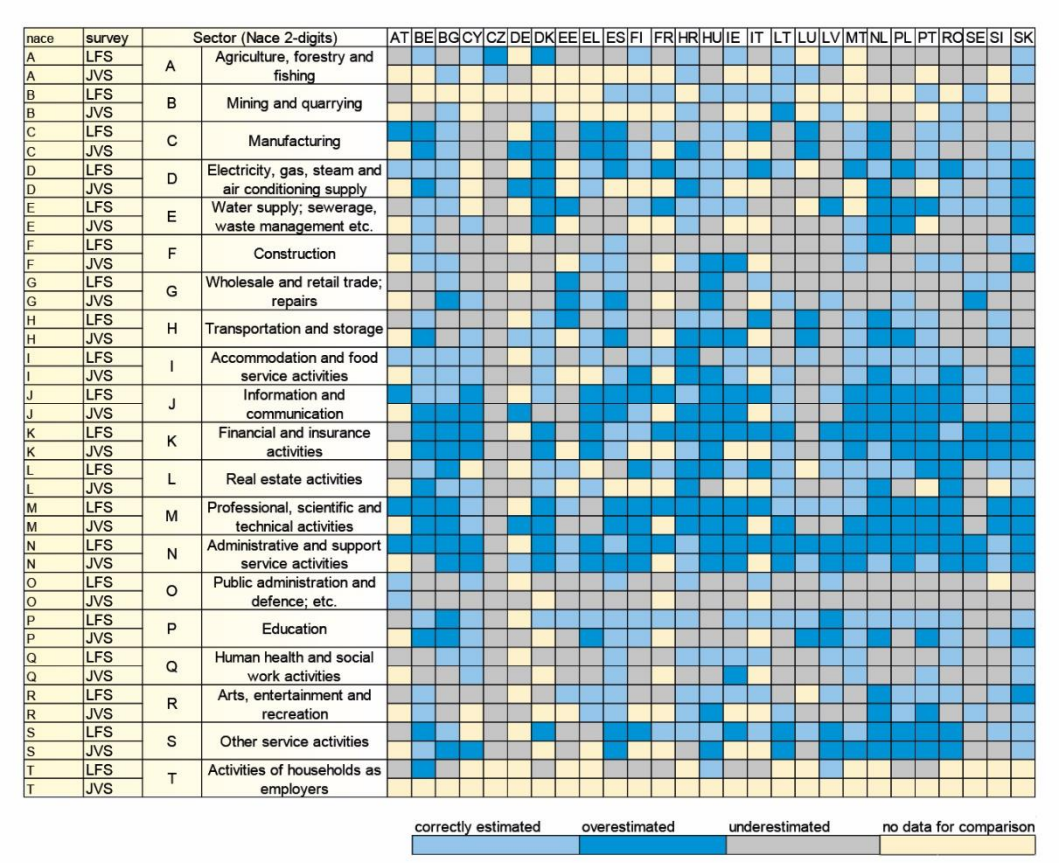
For the fourth quarter of 2020 ([7]) a statistical test on homogeneity of distributions for each country was performed to see how representative OJA were in comparison to JVS or LFS surveys. The values of decomposition of chi-square statistics were also examined. This allowed us to identify the covariate categories of sectors/occupational levels or regions which are responsible for lack of representativeness (non-homogeneity). The representativeness was judged based on the significance of chi-square (and values of normalised chi-square). The significant chi-square value (high value of normalised chi-square) indicates low representativeness of OJA with respect to the reference sample for the selected variable.

In comparing the representativeness of estimates of vacancies derived from OJAs we also calculated shares at the 2-digit NACE, ISCO 1-digit and NUTS 2-digit levels across the country. An observation is considered accurately represented online or correctly estimated if the share of OJAs is within the 95% confidence interval for the corresponding share of vacancies in the JVS/LFS. If the share does not fall within that interval, then it is deemed over- or underestimated (based on whether the OJA share is greater or less than the corresponding job vacancy share). We have calculated confidence intervals, using the coefficient of variation (CV) as the relative standard error (SE).

Overall, the significant values of chi-square statistics for all comparisons of distributions by sector indicated that the estimates derived from OJAs in the fourth quarter of 2020 are not representative in comparison to LFS or JVS distributions. Nevertheless, it is well known that chi-square tests are extremely sensitive to sample size and, in the presence of a very large sample size (as in our case), almost any small difference will appear statistically significant and lead to refusal of the hypothesis of homogeneity. Therefore, the comparison of shares based on confidence intervals was used to indicate which sectors were estimated correctly (Figure 3).

([7]) The latest available data for which all data sources were available at the moment of the analysis.

Figure 3. **Comparison of shares of vacancies by sector and country between OJA with estimates from JVS or LFS surveys (NACE 2-digits, Q4 2020).**



NB: Missing data for EL, LU.

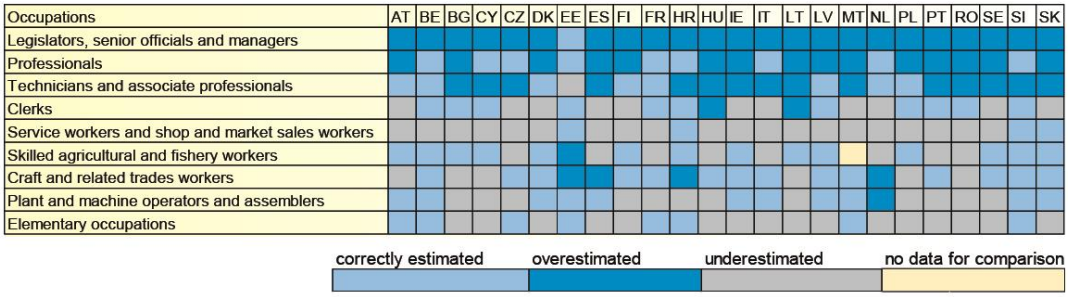
Source: Own calculation based on JVS data, LFS, OVATE.

When looking at the significance of estimates by sector we observe quite high variety across countries. Despite the lack of clear patterns, we observe that OJAs were mostly correctly estimating the vacancies advertised in education (P), accommodation and food service (I) and real estate (L) activities. Also, whenever the data were available for comparison at country level, the estimates were correctly derived for mining and quarrying (B). In most countries the shares derived based on OJAs were not strictly accurate: on the one hand, overestimating the number of vacancies in professional, scientific and technical activities (M), administrative and support service activities (N), and financial and insurance activities (K); and underestimating the shares of vacancies in public administration and defence (O) sectors.

When looking at comparisons on the occupational level across countries, we observe that (with exception of Estonia) the high-skilled occupation groups are

overestimated in deriving estimates based on OJAs, while the middle- and lowest-skilled occupation groups are underestimated (Figure 4).

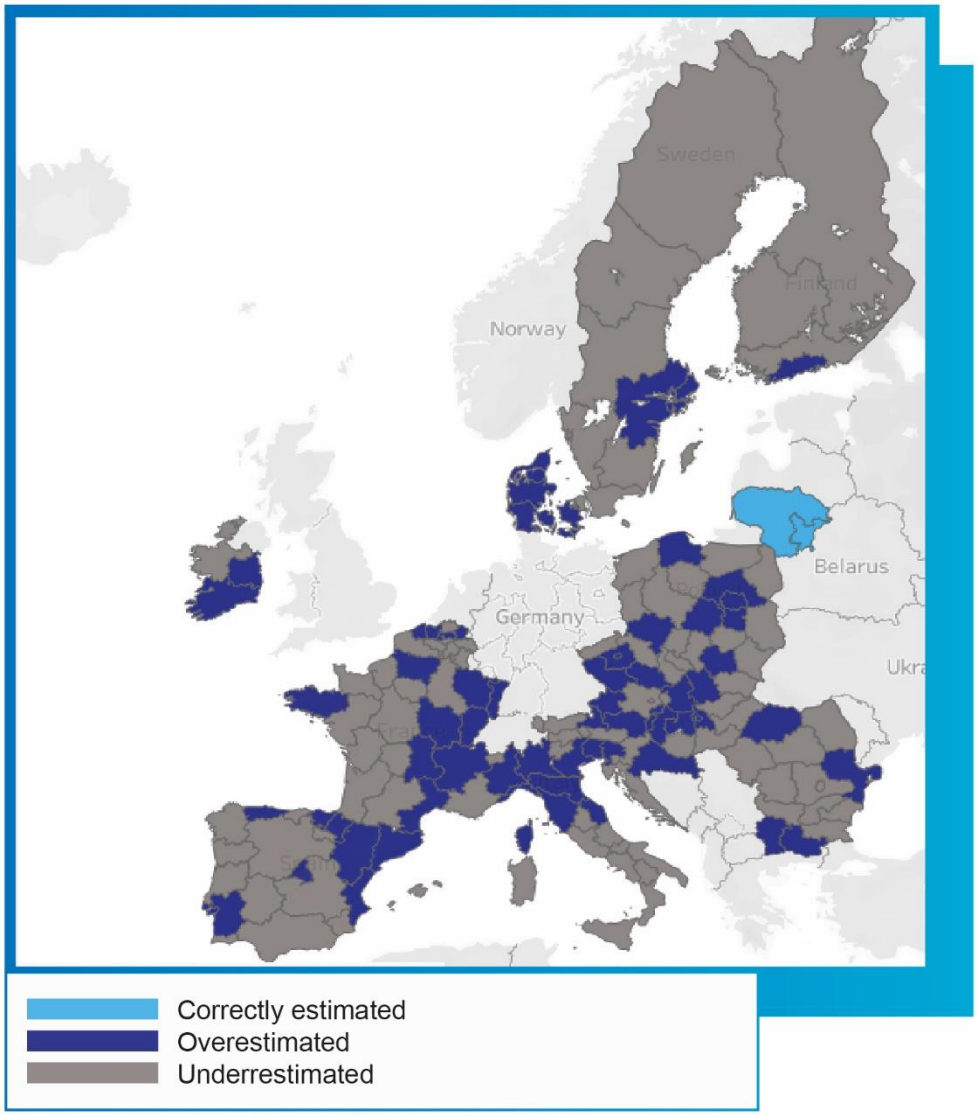
Figure 4. **Comparison of shares of vacancies by occupation (ISCO 1-digit) and country between OJA and estimates from LFS surveys (Q4 2020).**



Source: Author's calculation based on LFS and OVATE.

In comparing vacancies at the regional level, we observe that such assessment is not possible for many countries due to lack of sufficient observations (e.g. small countries like Cyprus, Latvia, Luxembourg and Malta only provide data at country level). We also observe that the share of OJAs in the capital region and in more industrialised regions (against total numbers in the country) tend to be overestimated because of lower coverage of rural regions. (Figure 5). This could be for several reasons: different level of digital skills, different occupational structure, or more direct access to potential candidates in rural areas.

Figure 5. **Comparison of shares of vacancies by region between OJAs and estimates from LFS surveys (Nuts 2-digits, 4th quarter 2020).**



Source: Author's calculation based on LFS and OVATE.

CHAPTER 5.

Representativeness of occupations in WIH-OJA data

The assessment of representativeness on a more granular level than 1-digit ISCO occupations groups could bring a different angle in understanding coverage biases of OJA data. Assuming the ESCO taxonomy includes information about most existing occupations on the labour market ([8]), then such comparison with OJA data should allow us to identify the occupations which were either never advertised online or never picked up by our system ([9]). We have taken information about occupations at 4-digit ISCO ([10]) level which were advertised at least once since the data collection started, and we calculated the share by dividing the number of occupations on 4-digit level we found in OJAs by all existing occupations in ESCO within 1-digit ISCO groups. The results are presented in Table 1 by each language pipeline, which in many cases correspond to the country level data: advertisements in German language are mainly collected from websites in Germany but a few are also collected in other countries such as Belgium and Luxembourg. If we look at language pipelines altogether, we see that, on average, 95% of existing occupations in the ESCO taxonomy were found at least once in OJAs, which is an encouraging result, indicating that employers are moving toward this way of recruiting workers. Coverage is below 90% only in two countries, Estonia and Greece. As expected, the size of the labour market could lower the possibility to observe certain jobs in OJAs. For example, the highest shares, indicating that almost close to 100% occupations were advertised online, were observed in the five biggest countries: France, Germany, Ireland, Italy and United Kingdom. Despite the size of the labour market, the very high share of occupations observed OJAs in Netherlands or Sweden could be explained by the overall higher shares of companies using this way of approaching candidates when recruiting workers (Figure 2).

Comparing the shares of observed occupations with the information about the numbers of extracted occupations at the 1-digit level ISCO occupation groups, presented in the previous chapter, we may conclude that even when OJAs

([8]) The frequency of ESCO taxonomy updates do not allow for capturing adequately the emerging occupations.

([9]) Some websites block the access for our scrapping algorithms.

([10]) ISCO 1-, 2-, 3- and 4-digit levels correspond to respective ESCO groups. Therefore, in this paper both ISCO/ESCO terms are used interchangeably.

potentially overestimate the number vacancies, even in the best-performing countries some occupations are not advertised online. For example, in Estonia, where according to comparison with LFS data the number of vacancies estimated based on OJA was correct for the legislators, senior officials and managers group, as much as 16.7% of occupations listed in the ESCO taxonomy were not observed in OJAs. This, again, may result from the size of the market, which is less diverse in smaller EU countries. ISCO is built under the responsibility of the International Labour Organization (ILO) and contains harmonised information from all labour markets; some occupations may not be present on small labour markets.

Looking at the shares across 1-digit level ESCO occupation groups, we notice that no matter which language pipeline, the skilled agriculture and fishery workers stands out in terms of the lowest shares of occupations observed in OJAs. On average, only 69% of existing ESCO occupations in this group were ever found among OJAs. Even in the best-performing countries this share is 86% and, in some languages like Greek, only 43% of occupations from this group were found online. This is the smallest occupation group containing only 14 occupations, many of which regards very specific positions which may not be observed in certain countries ([11]) or may not be advertised online. Another possible reason for low coverage of these occupations in our data system may be the way the system is designed itself. We may expect that some of these occupations, especially in occupations for which it is difficult to find native workers, may be advertised online by recruitment agencies who will be targeting workers from outside of the EU. This part of the job vacancies market, although it undoubtedly should be included in the total number of vacancies in EU, is unfortunately excluded from the WIH-OJA database in which we only include information from EU-hosted websites and not ones hosted outside of the EU (e.g. Morocco, Turkey, Ukraine).

([11]) The low coverage of this occupation group in Hungary may be related to the fact that, in a country without access to sea, the vacancy for an occupation profile of ‘Deep sea fishery worker’ will simply not occur.

Table 1. **Occupations observed in OJAs as the share of all existing occupations at 1-digit level groups**

Language	Total	Legisltors, senior officials and managers	Profes-sionals	Technicians and associate profes-sionals	Clerks	Service workers and shop and market sales workers	Skilled agricultural and fishery workers	Craft and related trades workers	Plant and machine operators and assemblers	Elementary occupations
fr	99.3	100.0	100.0	100.0	96.4	100.0	85.7	100.0	100.0	100.0
de	99.1	100.0	100.0	100.0	96.4	100.0	85.7	100.0	100.0	97.0
en	98.8	100.0	100.0	100.0	96.4	100.0	78.6	100.0	100.0	97.0
cs	98.1	100.0	100.0	100.0	96.4	100.0	78.6	98.5	100.0	90.9
it	98.1	96.7	100.0	100.0	96.4	100.0	78.6	98.5	100.0	93.9
nl	97.9	100.0	100.0	100.0	92.9	100.0	71.4	100.0	100.0	90.9
sv	97.4	96.7	100.0	98.8	96.4	100.0	71.4	98.5	97.5	93.9
es	96.9	93.3	98.9	100.0	96.4	97.5	71.4	95.4	100.0	97.0
pl	96.9	96.7	98.9	96.4	96.4	100.0	78.6	98.5	100.0	90.9
lt	96.5	100.0	97.8	97.6	92.9	97.5	64.3	98.5	100.0	93.9
pt	96.0	96.7	98.9	92.9	96.4	100.0	71.4	96.9	97.5	97.0
lv	95.1	96.7	97.8	98.8	92.9	97.5	64.3	95.4	97.5	84.8
ro	95.1	100.0	97.8	91.7	92.9	97.5	78.6	95.4	100.0	90.9
hu	94.8	93.3	96.7	97.6	96.4	97.5	57.1	98.5	95.0	87.9
fi	93.9	96.7	98.9	95.2	89.3	100.0	85.7	90.8	92.5	81.8
da	93.4	96.7	98.9	91.7	92.9	97.5	50.0	95.4	92.5	90.9
hr	93.4	93.3	97.8	94.0	89.3	97.5	64.3	93.8	97.5	84.8
bg	92.7	96.7	98.9	94.0	96.4	100.0	64.3	90.8	77.5	90.9
sl	91.8	96.7	95.7	88.1	85.7	97.5	71.4	89.2	95.0	93.9
sk	91.1	93.3	94.6	90.5	92.9	97.5	57.1	93.8	85.0	87.9
el	89.4	93.3	96.7	92.9	92.9	97.5	42.9	73.8	95.0	87.9
et	87.3	83.3	94.6	85.7	85.7	95.0	50.0	92.3	82.5	78.8

Source: Own calculation.

CHAPTER 6.
Discussion

The high coverage of occupations in comparison with the ISCO taxonomy across the majority of language pipelines indicates that only in a few occupations may the potential use of OJAs be less accurate when assessing the number of vacancies. Often the lack of presence of certain occupations in OJAs is related to the size of the market, with small countries having less diverse labour markets and therefore also having lower coverage than larger countries with more complex labour markets. We also foresee that the coverage of OJAs may improve, with an increase in the number of employers who will be using online channels to approach potential candidates. We already see that the changes observed in work organisation introduced by employers in relation to the COVID-19 pandemic, allowing workers more remote working time, has also positively influenced the number of OJAs.

The discussion on the results of the representativeness assessment of OJAs with LFS and JVS surveys data needs to start with few remarks about the compatibility of using both data sources for this purpose, as neither perfectly describes labour market reality. First, JVS data from which reference populations were created are sampling-based surveys, with some exemptions for countries which use administrative data. Although companies are obliged by regulation to report quarterly on the number of vacancies, some may not comply with these rules, making it necessary to impute data in order to correct for non-response bias (as reported by Bulgaria and Denmark). Second, vacancies listed by newly established enterprises may not be reported given that the registry of establishments is not updated regularly; due to delays between selection of the sample and data collection, these companies may not be included in the sampling frame. Third, as the JVS data is collected in the form of declarations it may also be that some reports include errors arising from questionnaire complexity (as reported by Ireland). For example, some countries report problems in correctly defining the number of vacancies in large establishments, especially for an establishment in the public sector (as reported by Finland), where the recruitment of new employees is decentralised to the department level.

In the case of the LFS, as with other European social surveys, declining response rates may also translate into greater risk of errors in estimates. Additionally, LFS data may underestimate the number of vacancies if the estimates of new hires by the established definition require a change of employer: they may

not capture well the change from being an apprentice to an employee if this is within the same enterprise.

Besides the problems with JVS and LFS, discrepancies may also come from the ways OJA data is processed. Although information like job title or employer name is often contained in a separate field in OJAs, and therefore is easy to extract and classify, other information may need to be extracted from the full text of the job description and converted into structured elements using natural language processing (NLP) and classification algorithms. The NLP algorithms applied to annotate the information from the content of an OJA may not catch everything correctly and this may lead to misclassifications. For example, part of the discrepancies between OJAs and JVS data observed at sectoral level may come from the fact that the company NACE code is usually not stated in the OJA. The job advertisement is classified to a corresponding code based on the information included in the content of the OJA, while for the vacancy surveys this information comes from the registry of companies and covers the main economic activity of the enterprise.

Comparisons of OJAs with both surveys would benefit from better identification of the reference population. This could be achieved, for example, by starting data collection from whether websites were used to advertise the vacancy in JVS data by reporting companies, or, for the LFS, if the website was how the newly hired person found his job. The comparison with JVS data would benefit from harmonization of methodologies across countries, as now not all countries report, for example, on all NACE sections. There is also no internationally agreed rule for the time of recording of job vacancy statistics. Depending on countries, the time of recording for quarterly job vacancy statistics may be one specific day in the quarter (e.g. the 15th of the middle month, the last calendar or working day of the quarter) or a 3-month average or flow of the vacancies throughout the quarter.

Acronyms

General acronyms

BDQF	big data quality framework
CEDEFOP	European Centre for Development of Vocational Training
CV	coefficient of variation
ESCO	European skills, competences, and occupations (multilingual classification)
EU	European Union
EUROSTAT	Statistical office of the European Union
ILO	International Labour Organization
ISCO	International standard classification of occupation
LFS	labour force survey
JVS	job vacancy survey
NACE	nomenclature of economic activities
NLP	natural language processing
NUTS	nomenclature of territorial units for statistics
OJA	online job advertisements
OVATE	online vacancy analysis tool for Europe
PES	public employment services
SE	standard error
UNECE	United Nations Economic Commission for Europe
VET	vocational education and training
WIH	web intelligence hub

EU countries

AT	Austria	ES	Spain	LV	Latvia
BE	Belgium	FI	Finland	MT	Malta
BG	Bulgaria	FR	France	NL	Netherlands
CY	Cyprus	HR	Croatia	PL	Poland
CZ	Czechia	HU	Hungary	PT	Portugal
DE	Germany	IE	Ireland	RO	Romania
DK	Denmark	IT	Italia	SE	Sweden
EE	Estonia	LT	Lithuania	SI	Slovenia
EL	Greece	LU	Luxembourg	SK	Slovakia

References

[URLs accessed 9.11.2022]

Beresewicz, M. (2016). *Measuring representativeness of Internet data sources through linking with register data*. Presentation for the European Conference on Quality in Official Statistics, Madrid, May 31 – June 3.

Beręsewicz, M. and Pater, R. (2021). *Inferring job vacancies from online job advertisements*, Luxembourg: Publications Office, 2021. http://data.europa.eu/doi/10.2785/96387

Bethlehem, J. (2009). *Applied survey methods: a statistical perspective*. New York: Wiley.

Carnevale, A.P., Jayasundera, T. and Repnikov, D. (2014). *Understanding online job ads data: a technical report*. Center on Education and the Workforce. https://cew.georgetown.edu/wp-content/uploads/2014/11/OCLM.Tech_.Web_.pdf

Cedefop (2019a). *Online job vacancies and skills analysis: a Cedefop pan-European approach*. Luxembourg: Publications Office. http://data.europa.eu/doi/10.2801/097022

Cedefop (2019b). *The online job vacancy market in the EU: driving forces and emerging trends*. Luxembourg: Publications Office. Cedefop research paper; No 72. http://data.europa.eu/doi/10.2801/16675

Garasto, S. et al. (2021). *Developing experimental estimates of regional skill demand*. ESCoE Discussion Paper 2021-02, March 2021.

Ishwarappa and Anuradha, J. (2015) A brief introduction on big data 5Vs characteristics and Hadoop technology. *Procedia Computer Science*, Vol. 8, pp. 319-324. https://doi.org/10.1016/j.procs.2015.04.188

Kruskal, W. and Mosteller, F. (1979). Representative sampling II: Scientific literature excluding statistics. *International Statistical Review*, Vol. 47, pp. 111-123.

Kureková, L.M.; Beblavý, M. and Thum-Thysen, A. (2015). Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. *IZA Journal of Labor Economics*, Vol. 4, Art. 18. https://doi.org/10.1186/s40172-015-0034-4

Lavrakas, P. J. (2008). Representative sample. In: *Encyclopedia of Survey Research Methods*, 2008. https://dx.doi.org/10.4135/9781412963947.n469

Lin, E. (2017). *Representativeness analysis: how our data reflects the real labor market dynamics.* Data science for social gold blog. https://www.dssgfellowship.org/2017/10/06/representativeness_analysis/

Ochsner, M. (2021). *Representativeness of surveys and its analysis*. Lausanne: Swiss Centre of Expertise in the Social Sciences (FORS). FORS guide, No 15. http://dx.doi.org/10.24449/FG-2021-00015

Swier, N. et al. (2018). *Web scraping / job vacancies*, deliverable 2.2: final technical report. ESSnet big data special grant agreement No 2 (SGA 2). https://ec.europa.eu/eurostat/cros/sites/default/files/SGA2_WP1_Deliverable_2_2_main_report_with_annexes_final.pdf

UNECE – United Nations Economic Commission for Europe (2014). *A suggested framework for the quality of big data.* Geneva: UNECE.

ASSESSING THE REPRESENTATIVENESS OF ONLINE JOB ADVERTISEMENTS

This working paper presents results of an assessment of the representativeness of information collected from online job advertisements (OJA) in establishing the number of labour market vacancies in EU Member States. Two external data sources were used, Labour force survey (LFS) and Job vacancies survey (JVS), available in most EU countries. The coverage biases in OJA data, in comparison to existing data sources, are evaluated at sectoral, occupational and geographic levels.

6217 EN - TI-BA-22-010-EN-N - doi:10.2801/807500

cedefop | European Centre for the Development of Vocational Training

Europe 123, Thessaloniki (Pylea), GREECE
Postal: Cedefop service post, 570 01 Thermi, GREECE
Tel. +30 2310490111, Fax +30 2310490020
Email: info@cedefop.europa.eu

www.cedefop.europa.eu

Publications Office of the European Union

ISBN 978-92-896-3456-4