Inferring job vacancies from online job advertisements

MACIEJ BERĘSEWICZ AND ROBERT PATER

2021 edition

STATISTICAL
WORKING PAPERS

eurostat

Inferring job vacancies from online job advertisements

MACIEJ BERĘSEWICZ AND ROBERT PATER

2021 edition

Manuscript completed in January 2021

The Commission is not liable for any consequence stemming from the reuse of this publication.

Luxembourg: Publications Office of the European Union, 2021

Theme: Population and social conditions
Collection: Statistical working papers

PDF: ISBN 978-92-76-23299-5 ISSN 2315-0807 doi: 10.2785/963837 KS-TC-20-008-EN-N

Abstract

Online data creates possibilities as a supplement to official statistics. With the development and popularity of online job boards, one of such big data are job advertisements. Online job advertisements are an example of data that may support and deepen labour market statistics. At least two applications of such data are especially promising: (i) job offers as a leading indicator of the labour market situation, (ii) a source of additional structural and qualitative information at a detailed level, for example on skills. However, online data come from non-probability samples. Before supplementing official statistics problems of coverage and representativeness of online job offers need to be addressed.

This study analyses data produced by Cedefop in a pan-European approach, based on online job advertisements (OJA) that it has collected and from which it has extracted several statistical variables. Cedefop supplies the largest database of OJA data in Europe and involves representatives of each of the European Union Member-States. Such a rich database may possibly supplement job vacancy statistics, obtained with probability-based surveys conducted at national level.

The main objective of the study was to develop estimator(s) for the number of job vacancies from OJA data on online job advertisements, accounting for the differences in the statistical unit and coverage. We start with defining the relation between a job advertisement and a job vacancy, and review approaches to analysis of online job offers presented in economic literature. Next, we describe potential statistical methods that may be used to infer job vacancies from job ads. We start our empirical analysis with comparison of historical data on both statistics. Then we proceed with the application of presented methods to Cedefop and Eurostat datasets.

We find that although Cedefop's OJA data are promising, at the present stage they are still experimental and could not be used to estimate the total number of vacancies. Both the time trends and the structure of job advertisements according to country, industry and occupation significantly differ from the ones of job vacancies. However, we do find promising results for certain countries and industries. We also explain the considerable limitations of our current analysis that come from: unknown quality of some of the procedures followed to produce the dataset, short time series of available data, and lack of unitary data from the surveys of job vacancies. We recommend taking a subsample of Cedefop's data that will be suitable to predict job vacancies and redesigning the study that will cover certain disadvantages of the current approach. We make a number of detailed suggestions, among which: (i) conduct an audit sample, (ii) reconsider the procedure of choosing sources for online job offers, focusing on obtaining more specified information, (iii) link the final information on the estimated job offer with the source of job advertisement and provide measures of the quality of machine learning methods in order to be able to apply proper inference methods, (iv) collect information about a company advertising a job and link it with the business register, (v) use Cedefop's data as source of additional qualitative information about job vacancies.

Keywords: job vacancies, online job advertisements, big data, non-probability samples.

Authors: Maciej Beręsewicz [1] , Robert Pater [2] .

Acknowledgement: On behalf of Eurostat - Unit B1 the study was coordinated by Mátyás Mészáros and Fernando Reis, with contributions from Raquel Paulino. The authors would like to thank to all contributors from Eurostat and Cedefop as well to Fabio Bacchini (external collaborator) for their helpful comments.

The study was implemented by GOPA, as contractor of Eurostat for the framework contract on Methodological Support (ref. 2018.0086). Project management was ensured by Florabela Carausu.

[1] Poznań University of Economics and Business, maciej.beresewicz@ue.poznan.pl; Statistical Office in Poznań, Poland, m.beresewicz@stat.gov.pl

[2] University of Information Technology and Management in Rzeszów, rpater@wsiz.rzeszow.pl; Educational Research Institute in Warsaw, Poland, r.pater@ibe.edu.pl

Table of contents

Abstract **3**

1 Introduction **5**

2 General quality issues **7**
- **2.1 Online Job Advertisements, Vacancies and Capacity Utilization** 7
- **2.2 Using online job advertisements for economic research** 10

3 Methods **14**
- **3.1 Basic settings** 14
- **3.2 Unit-level methods** 15
  - 3.2.1 Modelling misclassification 15
  - 3.2.2 Non-ignorable selection 16
  - 3.2.3 Single source capture-recapture 17
  - 3.2.4 Multiple source capture-recapture 19
- **3.3 Domain-level methods** 22
  - 3.3.1 Measurement error models 22
  - 3.3.2 Modelling of under-reporting 23
- **3.4 Summary** 26

4 Comparing OJA data with job vacancy statistics **27**
- **4.1 Aligning job advertisements data to Job Vacancies Statistics** 27
- **4.2 Relation between online job offers and job vacancies** 32
  - 4.2.1 Descriptive statistics 32
  - 4.2.2 Cross-sectional regressions 37
  - 4.2.3 Panel data regressions 39

5 Results **45**
- **5.1 Methods and their assumptions** 45
- **5.2 Naïve trimming** 47
- **5.3 Zero-one truncated capture-recapture results** 49
  - 5.3.1 Estimation procedure 49
  - 5.3.2 Application 50
  - 5.3.3 Countries with under 200k advertisements 51
  - 5.3.4 Countries over 200k advertisements 53
- **5.4 Conclusions** 55

6 Findings and recommendations **56**
- **6.1 Main findings** 56
  - 6.1.1 Correlation between OJA and JVS 56
  - 6.1.2 Population size estimation 56
- **6.2 Main limitations** 57
- **6.3 Recommendations** 58
  - 6.3.1 For the OJA dataset 58
  - 6.3.2 Data sources 58
  - 6.3.3 Audit sample 58
  - 6.3.4 Additional scraping 59
- **6.4 Conclusion** 60

7 References **61**

References **61**

Annex **64**

1 Introduction

The literature on the use of online job offers has been thriving in the latest years. Online job offers may provide much detail about behaviour of economic entities. However, the literature and methods are still at the early stage of fast development. They especially need statistical accuracy. So far, researchers paid special attention to obtain similar distributions of online data with data from probability-based surveys for certain variables. The literature still needs to develop an approach to reduce representation bias observed in online job offers.

The largest collection of online job offers for analytical purposes in Europe is conducted by the European Centre for the Development of Vocational Training (Cedefop). This pan-European approach applies methods to collect online job advertisements in all European Union member states. Cedefop's system of collecting online job offers provides experimental data on online job advertisements (OJA), as well as experimental data on skills' demand (Skills Online Vacancy Analysis Tool for Europe (Skills OVATE)([3])). The former is the subject of this study.

Experimental data on online job advertisements provided by Cedefop are the result of the cooperation between various institutions from the European Union Member-States within the project "Real-time labour market information on skill requirements: feasibility study and working prototype". The advantage of the data is not only their European scope, but also their continuous collection. This provides the opportunity to track market trends and possibly support official statistics in monitoring job vacancies.

In order to collect big data on online job offers, Cedefop uses extensive cooperation with national experts, institutions providing data on job advertisements, as well as webscraping algorithms, downloading publicly available data from online job boards. The landscaping of online job advertisements is wide (Cedefop, 2019b). At the stage of writing this report, results from Cedefop's research were available for the period 2018Q2-2020Q1.

Cedefop uses dedicated ontologies created to collect, transform, and classify job advertisements (Cedefop, 2019c). Those methods are based on artificial intelligence techniques (Colombo et al., 2019). Thanks to them Cedefop is able to generate a very detailed data. These data include more variables than probability-based surveys can possibly contain. These include mainly:

- time
- area
- occupation
- skill
- contract
- education

([3]) See https://www.cedefop.europa.eu/en/data-visualisations/skills-online-vacancies

- industry
- source.

Thanks to this approach, the obtained job offers estimates can be controlled for some pitfalls of online data. For example, the ”type of contract” indicates whether an advertisement concerns a work contract or an internship/traineeship. This can be used to properly calculate the number of job positions.

However, it needs to be underlined that at this stage, methods to collect and analyse online big data are still developing and are not unerring. As Cedefop underlines, ontologies developed to organise collected data are still imperfect, and are subject to continuous correction. Also, even though Cedefop gathers data from many sources, OJA represent only part of the vacancy market (Cedefop, 2019a).

The objective of this study is to develop estimator(s) for the number of job vacancies from data on online job advertisements, accounting for the differences in the statistical unit and coverage.

In order to fulfil the general objective of the study, we first review the literature on online job offers, and review methods that can be used to infer job vacancies from job offers. The theoretical part is based on economic and statistical literature. In the empirical part we compare online job offers and vacancies. We suggest a method for inference job vacancies from Cedefop's online job offers, and present limitations of such a study. Finally, we apply the chosen method, calculate its error and formulate recommendations for further studies.

In Chapter 2 we identify the relationship between the populations of the two objects of our study, "online job advertisement" and "job vacancy". We point various sources of potential bias in job advertisements used as proxies of job vacancies. Then we review economic literature on the approach to collecting, transforming, and analysing online job ads.

In Chapter 3, we investigate possible inference methods that can be applied to estimate job vacancies from online job advertisements data. We assess the available auxiliary information. We distinguish those methods that can be applied with the existing auxiliary data.

In Chapter 4, we compare online job advertisement (OJA) data with job vacancies statistics (JVS), by analysing the relations between Cedefop's OJA dataset with Eurostat's JVS. We start by aligning both datasets, discuss data limitations and merging approach. Then we analyse them, using a general to specific procedure. We consider available breakdowns: time in quarters, countries, NACE sections (industries), and ISCO major occupation groups. Finally, we calculate aggregate statistics and regressions, and we disaggregate them to individual cross-sections.

In Chapter 5, we perform the estimation of the number of job vacancies using the method selected in Chapter 3. We also estimate the accuracy of the estimations of the number of job vacancies based on OJA data.

The 6th chapter contains conclusions and recommendations for Cedefop and Eurostat.

2 General quality issues

2.1 Online Job Advertisements, Vacancies and Capacity Utilization

A company may always have non-utilised capacity. Capacity utilisation rate is the share of potential output a company could produce if all installed equipment (capital owned by a company) was fully used. Economic theory points out that it happens when the average production cost function is at its minimum. Most of the time, (physical) capital is not fully used and therefore capacity utilization is not full. This means that there are workplaces that are not filled in the company. Even holding capital constant, i.e. in the short run from the economic point of view, most companies may still hire new workers to increase output.

However, non-full capacity utilisation does not mean that there are job vacancies. A vacancy is an unoccupied workplace that a company wants to have filled. Company representatives must take action to hire a worker. This makes vacancy statistics highly susceptible to economic expectations. It also connects vacancies to the short-run economics. That is why vacancies are generally strongly pro-cyclical, and exhibit a weak trend. A vacancy does not explicitly refer to the long run, because a company does not publish a notice that will hire a worker after certain investments increase the firm's capital. It takes a long time for investments made by a company to be fully reflected in terms of job vacancies.

On one side, we have economic expectations, which are imaginary, even though they are based on real factors. On the other side, we have *employment*, which is well documented for legal and other reasons. Job vacancies fall in between these two categories. Also, in some companies employment decisions are formalized and centralized while in others they are not, which poses a measurement problem (Dunlop, 1966). Figure 1 shows the theoretical relation between non-utilized capacity, job vacancies and online job advertisements (i.e. job offers)([4]) .

Job openings can refer to one of two types of data: 1) newly created jobs across a certain period, for example a month, that is, flow of job openings, or 2) vacancies in the end of a given period, that is, stock of job openings.

Eurostat([5]) defines a job vacancy as a *paid post that is newly created, unoccupied, or about to become vacant under two conditions*:

1. *employer is taking active steps and is prepared to take further steps to find a suitable candidate for a job from outside the enterprise concerned; and*

([4]) A "Job advertisement" is a notice containing a "job offer". Online job advertisement is placed on an Internet website with online job offers (commonly referred to as an "online job board"). Because of this, we use terms "job advertisements" and "job offers" as synonyms. However, for consistency purposes, throughout the report we generally use the term "job advertisements" (JA), and for their Internet portion the term "online job advertisements" (OJA)

([5]) See https://ec.europa.eu/eurostat/cache/metadata/en/jvs_esms.htm

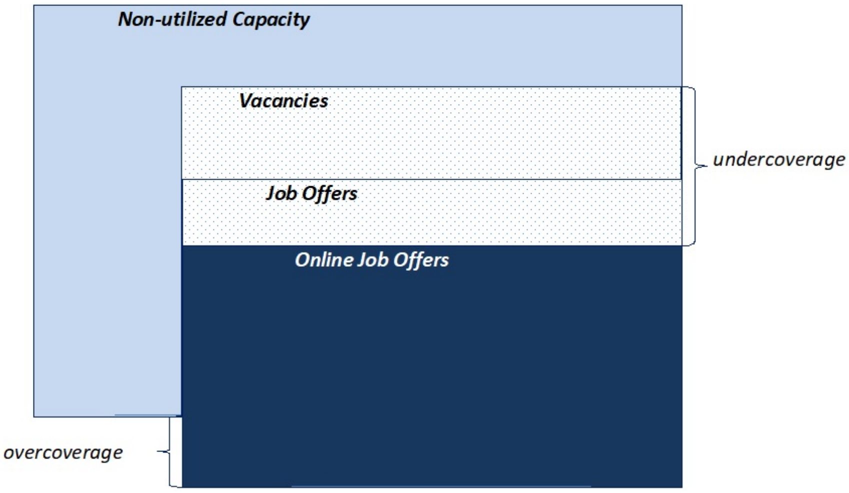

Figure 1: Non-utilised capacity, vacancies and online job offers

2. *employer intends to fill the job position either immediately or within a specific period*.

Eurostat also clarifies what ”Active steps to find a suitable candidate” means by specifying the following activities:

- notifying the job vacancy to the public employment services,
- contacting a private employment agency/headhunters,
- advertising the vacancy in the media (for example internet, newspapers, magazines),
- advertising the vacancy on a public notice board,
- approaching, interviewing or selecting possible candidates/potential recruits directly,
- approaching employees and/or personal contacts,
- using internships.

The first part of the first condition is fulfilled by an online job advertisement, since posting one might be considered an ”active step” to find a candidate for a job. For the online job advertisements we do not know whether the second part of the first condition is fulfilled. Enterprises might or might not be prepared to ”take further steps”, e.g. to make an interview with potential candidates for the job. While in the case of paid advertisements, a company is most likely willing to take further steps to employ a person, it might not always be true for free-of-charge postings. By posting a job advertisement, a company might potentially only try to investigate the market – the number and potential skills of job candidates (this is sometimes referred to as ”ghost vacancies”).

Likewise, we cannot be certain about the first part of the second condition. It is likely that by using paid job advertisements websites, an employer intends to ”fill the job position”, but it might not hold for free-of-charge websites. This means that it is important, when using OJA data, to differentiate between these types of websites. We might consider the second part of the second condition fulfilled by online job advertisements, as Eurostat does not specify the ”specific period”.

To sum up the definition stage, we can see that publishing a *job advertisements* is one of the methods to notify a *job vacancy*. Our doubts on whether vacancies fully encompass job advertisements are connected to the truthfulness of a company willingness to hire a job seeker. However, this heavily depends on the credibility of a considered website with job advertisements. Credible websites stipulate that doubtful job advertisements will be removed and list conditions when this happens. Providing that the initial landscaping and selection of websites ensures that only credible websites are considered, we might assume that in terms of definition, publishing online job advertisements is a valid method of notifying a job vacancy.

The sources of potential bias in online job advertisements as vacancy measures are the following (types of errors indicated in square brackets):

- job advertisements do not include the part of vacancies for which the employer is looking for an employee using methods other than an advertisement on the Internet; for low-paid jobs it might be an informal way or a sign in the window; also some high-paid or scarce occupations may be underrepresented, because companies seek such workers through HR agencies, e.g. directors, data scientists; even online job search may take different forms, not only job postings but also job-task search or crowdwork **[under-coverage error]**,

- online job advertisements are not considered by households and companies that do not have or rarely use the Internet **[under-coverage error]**,

- not all online job advertisements are collected, since not all websites are covered; it is virtually impossible to cover all of them **[under-coverage error due to selection]**,

- since the web data gathering algorithm does not operate all the time in a given spot on all websites, not all advertisements published in each month will be considered, as they are characterized by varied and unknown length of stay on the website **[under-coverage error due to data collection]**,

- it is not perfectly clear whether job advertisements constitute a stock or flow; generally job offers are gathered by a web data gathering algorithm in a specific moment, so they should constitute a stock of vacancies; however, some websites keep job advertisements for a fixed period (it is especially the case for paid job advertisements), and they may impair the differentiation between stock and flow measures (temporal aggregation bias); price policies and market strategies of job ads providers can also affect the number and composition of advertised job postings **[over-coverage error]**,

- online job advertisements over-represent certain occupations and skills and under-represent other; overrepresentation is generally connected to vacancies for workers with higher education and industries from the private sector, which is the opposite to job advertisements from public employment services; both sources of job advertisements are, to a high extent, supplementary to each other **[under-coverage error due to self-selection]**,

- several vacancies may be placed in one job advertisement, which disrupts the estimation of the number of vacancies; it might especially be true for positions not requiring higher education **[over-coverage error, unit error]**,

- it is difficult to check whether job postings are fully updated; filled or withdrawn positions might not mean that a job advertisement is withdrawn immediately and might bias their stock upwards. This bias might last until the job advertisement is automatically removed from the websites (most commonly, websites offer 30 days of publishing a job ad, but it is not the only option available). Turrell et al. (2019) call it an ”aggregate outflow bias” **[over-coverage error]**,

- online job market evolves in the long-run with technological development (logistic function-shaped technological diffusion, see Barnichon (2010) and Pater (2017)), and over the business cycle, through costs of posting vacancies online relative to costs of other recruitment methods (Cajner et al., 2016), labour market tightness, and through changes in the number of required skills (Pater et al., 2019); popularity of Internet will affect coverage of vacancies by online job advertisements, as digital divide does **[under-coverage error due to selection]**,

- even if the structure of job advertisements is corrected for ”representativeness” bias (e.g. by occupation, sector of activity etc.), skills requirements in job advertisements are likely to be accurately specified in advertisements for high-paid jobs and only generally or not-at-all specified for low-paid jobs; also many skills are implicit, or too trivial to be included in a job advertisement, for example basic computer skills, even though they are still necessary for the job **[selection error]**,
- ”job” advertisement placed online might not always refer to ”employment”. For example, internships and traineeships advertisements might refer to a job from employer’s perspective, while they may not be considered ”employment” from a statistical and labour law points of view **[overcoverage error]**.

Online job advertisements form a specific fraction of the job vacancy market. The advantages of online job advertisements as vacancy measures are as follows:

- it is possible to obtain very detailed information on job vacancies, e.g. in terms of skills and qualifications; traditional sources, in particular questionnaire based sample-surveys, do not allow obtaining such detailed information,
- the employers are interested in placing their preferences in the advertisement, including those related to new trends, which is an advantage over questionnaire based surveys, in which the employer does not benefit directly from accurately addressing the questions,
- job advertisements can potentially be examined in many cross-sections,
- job advertisements from previous periods can be re-examined, because they are saved in the database; this allows for re-examination and re-calculation of measures based on these job advertisements should methodology or classifications change; this also ensures comparability of data over time,
- collecting job advertisements is relatively low-costly, compared to questionnaire based surveys,
- Internet is currently the main method used to publish job advertisements, and this market is still developing,
- research can be conducted with high frequency that can be increased in the future,
- sample surveys are likely to under-represent new firms, as they are drawn from business registers (see Davis et al. (2013) and Turrell et al. (2019)); job advertisements might be less susceptible to this bias.

2.2 Using online job advertisements for economic research

There is a growing collection of economic literature on the use of online job offers, with increasing attention paid to selectivity. Most of the literature is based on the US economy, which has the longest tradition in collecting first newspaper, and then online job offers as vacancy measures (see e.g. Abraham and Wachter (1987), Kuhn and Skuterud (2004), Marinescu and Wolthoff (2016), Marinescu and Rathelot (2018), Deming and Kahn (2018) and Hershbein and Kahn (2018)). The European literature that recognizes the selectivity problem includes works of Colombo et al. (2019), Pater et al. (2019) and Turrell et al. (2019).

Carnevale et al. (2014) estimate that in the US economy the share of online job offers in all vacancies in 2014 was between 60% and 70%. The Conference Board discontinued traditional, newspaper-based Help-Wanted Advertising Index in 2008, after having begun a Help-Wanted Online Index (HWOL) in 2005. Ads in the HWOL index are collected in real-time from over 28,000 online job boards including traditional job boards, corporate boards, social media sites, and smaller job sites that serve niche markets and smaller geographic areas. In construction of HWOL, special attention is paid to time series properties, by removing outliers, seasonal adjustment etc. Also, HWOL does not include online job aggregators (The Conference Board, 2018). The Conference Board uses Wanted Analytics data to calculate Help-Wanted Online Index. They show data disaggregated by region, sector of activity and occupation, but not by skill requirements.

Cajner et al. (2016) state that there have been significant discrepancies between the stock of vacancies implied by two US series, the JOLTS (Job Openings and Labor Turnover Survey) and the Conference Board Help Wanted Online, which may be caused by changes in the price charged to employers when posting online job vacancies.

State vacancy surveys is another source of job offers data in the US economy. However, they are conducted by a limited number of states, sometimes with certain skill requirements, but cover only a few geographic areas; such job offers are not comparable between states.

Marinescu and Wolthoff (2016), Marinescu and Rathelot (2018), Deming and Kahn (2018), and Hershbein and Kahn (2018) analyse the US economy. They compare the distributions of online job offers they use, or a subsample of them, to the results of probability-sample surveys and to the distributions of other online job offers measures. For highly correlated results they assume that the non-probability sample job offers are "representative". If the correlation is low, they either take a subsample of gathered job offers or weight their data. Weights are based on the relation between vacancies obtained from probability-sample survey and their online job offers. Turrell et al. (2018) and Turrell et al. (2019) for the UK economy use weights in a few job breakdowns to obtain comparable distributions.

Marinescu and Wolthoff (2016) and Marinescu and Rathelot (2018) gather data on job offers from one US website (CareerBuilder). This website is chosen because it provides many variables, it provides data for job seekers and data about responses of job seekers to job offers as well. All vacancies are used or only a random subsample within. Comparing online job offers they collected to the ones from a probability-sample survey JOLTS it was stated that CareerBuilder represents 35% of the total number of vacancies in the US economy. It was also ascertained that CareerBuilder data overrepresents following industries: information technology, finance and insurance, and real estate, rental, and leasing, and underrepresents state and local government, accommodation and food services, other services, and construction.

They merge three datasets extracted from CareerBuilder's. The first one is a random sample of registered users. Data includes residence location at the ZIP code level. The second data set is a sample of vacancies published on the website. These vacancies are available to the job seekers. This data also contains a ZIP code. The third dataset connects the two previous ones by showing which jobs each job seeker applied to. An application is defined as a click on the "Apply now" button that can be found on the full job listing webpage.

The authors state that CareerBuilder data are "not representative" in terms of the industry breakdown. In terms of occupation (2-digit Standard Occupational Classification, SOC codes), the distribution of unemployed job seekers' occupations in CareerBuilder data are very similar to the one from CPS (Current Population Survey, correlation of 0.71 between the shares of job seekers in each occupation in the two datasets), and the distribution of vacancies' occupations in the CareerBuilder data are very similar to the distribution of vacancies in the general measure of online jobs (correlation of 0.95 with Help Wanted Online data). Regional distribution of vacancies from CareerBuilder is very similar to vacancies from the probability-sample survey (Job Openings and Labor Turnover Survey, correlation of 0.96 between the shares of vacancies in each region in the two datasets). The spatial distribution of job seekers was also analysed by comparison to the unemployed from the Current Population Survey (correlation of 0.88).

Rothwell (2014), Deming and Kahn (2018), Hershbein and Kahn (2018), and Azar et al. (2018) use an extremely broad database of job offers. They possess microdata from nearly 100 million electronic job postings in the United States between 2007 and 2015. Data was collected and assembled by Burning Glass Technologies (BGT), that examined 40,000 online job boards and company websites. However, the procedure of data gathering, cleaning, and classifying is largely unknown. They cross-validate the data (e.g. on skills) with other measures obtained from a probability-sample survey. They weight the data by the size of the metropolitan statistical areas (MSA) labour force. BGT shows that the share of their collected jobs online is 85% of the jobs in JOLTS in 2016.

The broad coverage of the database has a substantial strength over datasets based on a single vacancy source, such as CareerBuilder.com, but it also has drawbacks. The drawbacks come from the fact that we do not know the mechanism of job offer posting (e.g. paid or not paid, whether they are immediately withdrawn if candidate found, whether the site also helps in recruitment process or just provide space for ads etc.). The main merit of BGT data is its granularity level. While JOLTS ask a nationally representative sample of employers about vacancies they wish to fill in the near term, it is typically available only at aggregated levels, and contains relatively little information about the characteristics of vacancies. In contrast, the BGT data contain 70 possible standardized fields for each vacancy. Authors use information on occupation, geography, skill requirements, and firm identifiers. The codified skills include stated education and experience requirements, as well as specific skills standardized from the text in each job posting.

Hershbein and Kahn (2018) provide a description of industry and occupation distributions of vacancies in BGT relative to other sources (JOLTS, the Current Population Survey, and Occupational Employment Statistics), an analysis of how these distributions have changed over the sample period, and correlations between the datasets. Authors state that BGT postings are disproportionately concentrated in occupations and industries that require higher skill, the distributions are relatively stable across time, and the aggregate and industry trends in the quantity of job offers track other sources reasonably closely. In comparison to JOLTS, this data overrepresents health care and social assistance, finance, insurance, and education. It underrepresents accommodation and food services, public administration/government, and construction. Authors show that education requirements in BGT data strongly correlate with average education levels of employed workers at the MSA and occupation levels.

Carnevale et al. (2014) show that the occupation-industry composition of the BGT data are similar to that of the Conference Board's HWOL. Moreover, the authors audited a sample of job postings in the BGT database and compared them to the actual text of the postings, finding that the coding for occupation, education, experience were at least 80% accurate.

Rothwell (2014) compares the occupational distributions of BGT data to those from state vacancy surveys for selected metropolitan areas for which data are available. He finds that computer, management, and business occupations are overrepresented in comparison to the state vacancy surveys, while health care support, transportation, maintenance, sales, and food service workers are underrepresented. Furthermore, it is said that BGT regularly revises and attempts to improve its algorithms (applying them retroactively on the complete historical database of postings).

Turrell et al. (2018) and Turrell et al. (2019) transform the text of job adverts into time series data labelled by official classifications (Standard Occupational Classification codes). They deal with flow data and transform these data into stocks. Their data consist of millions of individual online job ads posted by firms and recruitment agencies on the website Reed.co.uk in the UK. The site facilitates matching between firms and jobseekers. In contrast to webcrawling algorithms, on this website companies post their vacancies directly with recruiters. Recruiters may have access to private information about the job vacancy which an aggregator would not, e.g. the salary offered. This is an advantage of using only one website to analyse online job offers trends, with an obvious drawback of lower amount of job offers. Authors process, clean, re-weight, and use this data as a measure of job vacancies by occupation and by region over time, and according to existing official statistical classifications.

They report many sources of bias in online job offers as vacancy measures. They adjust "aggregate outflow bias" with data on an average duration of job offer in the UK economy (2-4 weeks) instead of 6 weeks as it is on the website they use. They report a "differential outflow bias" as the one connected to varying duration of vacancies according to occupation. In contrary, in the OJA dataset job offer duration is not fixed and potentially unknown because Cedefop uses many websites. Turrell et al. (2019) treat coverage bias by applying weights. They reweight data by occupation and sector of activity over time. They also state that "the ads are run at a cost for the posting party so the concerns about an ever-growing stock of vacancies which have, in reality, been filled or withdrawn do not apply."

Turrell et al. (2019) point that some biases in survey-based research are non-existent in online job offers. These are non-response bias, incomplete-response bias, and overestimation of the vacancies posted by large companies. Probability-based surveys also possess the "frequency mismatch", because vacancy statistics in the EU is based on a quarterly survey, while unemployment statistics, at least to some extent, is available on a monthly basis. Small vs. large firms' bias may take a different form in online job offers than in probability-based surveys. It seems reasonable to assume that both types of companies publish job ads online. However, large firms may use paid websites more often.

The newly created vacancy time series, split by occupation, are compared by Turrell et al. (2019) to existing data on UK job vacancies, namely the ONS' Vacancy Survey and JobCentre Plus data. To demonstrate the utility of their analysis, authors use data they obtain to estimate Beveridge curves by occupation and to calculate the rate of mismatch unemployment (by occupation) for the UK.

3 Methods

3.1 Basic settings

The main goal of the project described in this publication was to develop estimators for the number of job vacancies from data on online job advertisements, taking into account the differences in definition of statistical units, coverage and selection bias.

Therefore, from the formal perspective the goal is to estimate total number of job vacancies given by

$$N_t = \sum_{i=1}^{N} y_{it}, \tag{1}$$

where $y_{it} = 0, 1, 2, ...$ is number of job vacancies in the company i at the end of given quarter t. This is the standard way where instead of sampling target population we sample a reference population by means of indirect sampling. See Deville and Lavallée (2006) andLavallée (2009) for more details.

Currently, National Statistical Institutes provide estimates of quantity (1) on the basis of sample surveys and using Horwitz-Thompson estimator

$$N_t = \sum_{i=1}^{n} d_{it} y_{it}, \tag{2}$$

where d_{it} is the design weight denoting inverse of probabilities of inclusion π_{it}^{-1}. However, these surveys suffer from non-response and coverage errors therefore, d_{it} is adjusted by means of post-stratification or calibration. After this procedure, d_{it} is replaced by w_{it} and equation (2) becomes

$$N_t = \sum_{i=1}^{n} w_{it} y_{it}. \tag{3}$$

An alternative way of estimating the total number of job advertisements might be to use the naïve estimator given by

$$N_t = p_t^{-1} N_t^{Internet}, \tag{4}$$

where p_t is the share of job vacancies published online (this may be obtained from sample surveys) and $N_t^{Internet}$ is the total number of vacancies published online. However, obtaining p_t requires additional costs and $N_t^{Internet}$ should be obtained without errors. Moreover, all the cases above require access to sampling frames/lists that cover the target population. This assumption is rather naïve, as we do not have these sources and thus in this chapter we discuss alternative ways of estimating population size based on non-probability data.

3.2 Unit-level methods

3.2.1 Modelling misclassification

Following Liu and Zhang (2017), let Y to represent the true state of the binary response variable which could be modelled by the generalized linear model (i.e. logistic regression) and $\boldsymbol{X}$ is matrix of predictors. Then let $\{(\tilde{y}_i, \boldsymbol{x}_i)\}$ be a set of data collected from n participants where $\tilde{y}_i$ are realisations of the potentially misclassified variable $\tilde{Y}$. Under the assumption of non-differential misclassification, the chance of misclassification is only related to the true status y_i by the transition probability distribution function

$$\begin{aligned}
&\text{Pr}\left(\tilde{y}_i = 1 | y_i = 0\right) = r_0, \\
&\text{Pr}\left(\tilde{y}_i = 0 | y_i = 0\right) = 1 - r_0, \\
&\text{Pr}\left(\tilde{y}_i = 0 | y_i = 1\right) = r_1, \\
&\text{Pr}\left(\tilde{y}_i = 1 | y_i = 1\right) = 1 - r_1,
\end{aligned} \tag{5}$$

where r_0 and r_i are called *false positive* (FP) and *false negative* (FN) rates, respectively, representing the extent of misclassification.

It follows that the regular logistic regression model can be extended to include both false-positive and false-negative misclassification parameters

$$\begin{cases}
y_i \sim & \text{Bern}\left(\pi_i\right), \\
\pi_i = & r_0 + \left(1 - r_0 - r_1\right) F_i, \\
F_i = & \frac{1}{1+\text{exp}(-\eta_i)}, \\
\eta_i = & \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}.
\end{cases} \tag{6}$$

To estimate the parameters of the logistic models, the maximum likelihood (ML) estimation method is used because it provides consistent and asymptotically unbiased parameter estimates and standard error estimates.

The algorithm is based on the estimated equations from the ML estimation. The probability density function of $\tilde{Y}$, conditional on the covariates $\boldsymbol{x}_i$ is

$$\text{Pr}\left(\tilde{Y}_i = \tilde{y}_i | \boldsymbol{x}_i\right) = \pi_i^{y_i}\left(1 - \pi_i\right)^{1-y_i} = exp\left\{y_i\theta_i - \text{log}\left(1 + \text{exp}\left(\theta_i\right)\right)\right\}, \tag{7}$$

where $\theta_i = \text{log}\,\frac{\pi_i}{1-\pi_i}$, $\pi_i = r_0 + (1 - r_0 - r_1)F_i$, $F_i = \frac{\text{exp}(\eta_i)}{1+\text{exp}(\eta_i)}$ and $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. Given n independent observations the likelihood function is

$$L = \text{exp}\left\{\sum_{i=1}^{n} y_i\theta_i - \sum_{i=1}^{n} \text{log}\left(1 + \text{exp}\left(\theta_i\right)\right)\right\}, \tag{8}$$

and the corresponding log-likelihood

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \left[y_i \theta_i - \log\left(1 + \exp\left(\theta_i\right)\right) \right]. \tag{9}$$

This model requires information about which job advertisements are erroneous, outdated, out-of-scope or not job vacancies and auxiliary variables $\boldsymbol{X}$ are required.

In the literature further developments of these methods could be found. For instance, one can refer to Roy et al. (2005), Daniel et al. (2005), Meyer and Mittag (2017), Pires and Quinino (2019) and Roy et al. (2013) to name few.

3.2.2 Non-ignorable selection

Literature on non-ignorable non-response and selection bias is rich (see e.g. Pfeffermann and Sikov (2011), Riddles et al. (2016), Chang and Kott (2008) and Kott and Chang (2010)). Sikov (2018) and Tang and Ju (2018) provide a recent review of some approaches. In general, distinction between these methods is based on the delimitation of two cases:

1. information about characteristics of non-respondents (not covered units) available,
2. only information about characteristics of selected units is available,

and for both cases it is assumed that known population totals are available.

In this paper we will focus only on the second case as only online data from `Cedefop` is available. In particular, we will focus on the model proposed by Chang and Kott (2008) and Kott and Chang (2010).

Let $c_i = (c_{1i}, ..., c_{Ki})$ denote the values of the calibration variables for unit i that was in the non-probability sample. To estimate the unknown parameters by forming the non-linear regression equations

$$\boldsymbol{C}^{POP} = \sum_{i=1}^{r} w_i \frac{c_i}{\rho(y_i, v_i; \gamma)} + \epsilon^*, \tag{10}$$

where $\boldsymbol{C}^{POP} = \sum_{j=1}^{N} c_j$, ϵ^* is a vector of errors, $\rho(y_i, v_i; \gamma) = P(R_i = 1 | y_i, v_i; \gamma)$, and $w_i = \pi_i^{-1}$ denote the sampling weights. The w_i are assumed to be known for every responding unit in the sample. Chang and Kott (2008) proposed an iterative procedure to obtain γ

$$\hat{\gamma}^{k+1} = \hat{\gamma}^{k} + \{\hat{\boldsymbol{H}}(\hat{\gamma}^k)^T \boldsymbol{V}^{-1}(\hat{\gamma}^k) \hat{\boldsymbol{H}}(\hat{\gamma}^k)\}^{-1} \times \left(\boldsymbol{C}^{POP} - \sum_{i=1}^{r} w_i \frac{c_i}{\rho(y_i, v_i; \hat{\gamma}^k)} \right), \tag{11}$$

where

$$\hat{\boldsymbol{H}}(\hat{\gamma}^k) = \frac{\partial w_i \frac{c_i}{\rho(y_i, v_i; \gamma)}}{\partial \gamma} | \gamma = \hat{\gamma}^k \tag{12}$$

and $\boldsymbol{V}^{-1}(\hat{\gamma}^k)$ is the inverse of an estimator for the quasi-randomisation variance $w_i \frac{c_i}{\rho(y_i, v_i; \gamma)}$ of computed at $\gamma = \hat{\gamma}^k$.

Having estimated the response probabilities, the use of this approach allows estimating the population totals of the target variables of interest, but it does not allow imputation of the missing data, because no model is assumed for the outcome values.

This method may be applied to reduce bias due to duplication of records. For instance, the y_i may be the number of advertisements and $\boldsymbol{x}_i$ may be the characteristics of companies with known population totals $\boldsymbol{C}^{POP}$ of these with free vacancies.

3.2.3 Single source capture-recapture

In case when only one data source is available a common approach is to analyse repeated observations of given persons or objects by applying regression models. In such case, when the goal is to estimate population size, we apply single source capture-recapture approach.

Let us assume that dataset A contains units i that are members of the target population U. In this approach, we assume that dataset A is error free i.e. no erroneous units or over-coverage are present, but it allows to identify multiple recordings of the same units. Thus, the dataset contains multiple actions for the same unit.

Example of data is presented in table 1 where column *Captures* refers to number of occurrences, i.e. 1 – only one registration, 2 – two registrations, and so on. Column *Number of units* refers to the number of units that were observed once, twice, or more times. Note that this table does not contain information about zero captures N_0 and should be estimated (i.e. $N = N_0 + N_{obs}$).

Table 1: Example of single source capture-recapture data

Captures	Number of units (N)
0	– (N_0)
1	1500
2	40
3	5
4	2
6	1
Total	$N = N_0 + 1548$

Source: own elaboration.

To tackle this problem, i.e. estimate missing number of units, a flexible approach based on distributional assumptions about number of captures is applied. In particular, zero-truncated count regression models are applied, for instance zero-truncated Poisson or zero-truncated negative binomial regression. These models often use covariates such as sex, age, or nationality to account for heterogeneity in captures. For recent review of approaches see Bohning et al. (2017) and Zhang and Chambers (2019).

Let us focus on the baseline model, where the number of apprehensions/captures of a member of the target population, denoted by y_i for $i = 1, ..., N$, follows a Poisson distribution with parameter λ. We have

$$P\left(y_i = 0\right) = e^{-\lambda} \text{ and } P\left(y_i | y_i > 0; \lambda\right) = \frac{P\left(y_i; \lambda\right)}{P\left(y_i > 0; \lambda\right)} = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!\left(1 - e^{-\lambda}\right)}. \tag{13}$$

The parameter λ can be estimated based on the observed $N_1, N_2, ...$ denoted by $\hat{\lambda}$ under the truncated Poisson distribution. An estimated Horvitz-Thompson estimator is given by

$$\hat{N}_{HT} = \frac{N_{obs}}{1 - e^{-\hat{\lambda}}} = \sum_{i=1}^{N_{obs}} \frac{1}{P\left(y_i > 0; \hat{\lambda}\right)}. \tag{14}$$

However, this approach has some drawbacks that were discussed by Zhang (2008):

- captures are not independent – positive contagion occur if previous apprehensions increase the probability of subsequent apprehensions; whereas negative contagion occur if the probability decreases.
- unexplained heterogeneity – the homogeneity assumption is violated if there are differences in the individual Poisson parameters that cannot be explained by the observed covariates.
- out-of-sample units are like in-sample units – other people with the same x must be outside the sample.
- *closeness* of the population – population is clearly not closed during the data collection period. The concepts of exposure and hit rate, people with different life duration in the population should have different Poisson parameters.

Recently, to overcome some of these limitations, Godwin and Böhning (2017) proposed zero-truncated (positive) one-inflated Poisson regression and Böhning and van der Heijden (2019) showed the identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities. Godwin and Böhning (2017) justified, one-inflation in two ways:

1. The subject discovers that the observational experience is more unpleasant than expected and decides to put forth avoidance effort towards being observed again.
2. Additionally, the subject may learn how to avoid being observed subsequently, which we call *avoidance ability*.

Let us consider two distributions. Let $f_y(\lambda)$ be the mass function for the Poisson distribution. Then, a one-inflated Poisson distribution is

$$\begin{cases} f_y(\lambda), & y = 0, \\ \omega(1 - f_0(\lambda)) + (1-\omega) f_y(\lambda), & y = 1 \\ (1-\omega) f_y(\lambda), & y > 1, \end{cases} \tag{15}$$

where $f_0(\lambda)$ is the probability that a 0 occurs under a Poisson distribution and ω is the proportion of zero-inflation.

Zero-truncating expression (15) results in the following OIPP distribution:

$$\begin{aligned} f_1^{OIPP} &= \omega + (1-\omega)\frac{\lambda}{\exp(\lambda) - 1} \quad & y = 1 \\ f_y^{OIPP} &= (1-\omega)\frac{\lambda^y}{(\exp(\lambda) - 1)y!} \quad & y = 2, 3, \ldots \end{aligned} \tag{16}$$

and the resulting estimator of population size is given by

$$\hat{N}^{OIPP} = \frac{N_{obs}}{1 - \exp\hat{\lambda}}, \tag{17}$$

where N_{obs} is the number of observed units.

Note that this approach was mainly applied for the hard-to-reach populations and suitability for inferring the total number of job vacancies should be verified. However, one should note that this approach requires error-free data i.e. no erroneous records or out-of-scope units.

3.2.4 Multiple source capture-recapture

3.2.4.1 CAPTURE-RECAPTURE FOR DEPENDENT SOURCES

Kiranmoy Chatterjee and his colleagues in the series of papers on dual and triple estimation (Chatterjee and Bhuyan (2017), Chatterjee and Mukherjee (2018), Chatterjee and Bhuyan (2019a) and Chatterjee and Bhuyan (2019b)) proposed estimators that take into account dependence between sources.

Let us focus on Chatterjee and Bhuyan (2017) with two lists. Let (Y, Z) be a paired variable such that Y_i and Z_i denote the List 1 and List 2 inclusion status of the i^{th} individual belonging to U. This is presented in the Table 2. Note that standard Dual-System Estimation have the following assumptions

(S1) Population is closed until the second sample is taken.

(S2) Individuals are homogeneous with respect to their capture probabilities.

(S3) Inclusion of each individual, belonging to U, in List 2 is *causally independent* to its inclusion in List 1.

Table 2: Dual-record-System (DRS)

List 1	List 2		
	In	Out	Total
In	$x_{11}[p_{11}]$	$x_{10}[p_{10}]$	$x_{1.}[p_{1.}]$
Out	$x_{01}[p_{01}]$	$x_{00}[p_{00}]$	$x_{0.}[p_{0.}]$
Total	$x_{.1}[p_{.1}]$	$x_{.0}[p_{.0}]$	$x_{..} = N[p_{..}]$

Source: Chatterjee and Bhuyan (2017, p. 5).

Having that in mind, Chatterjee and Bhuyan (2017) introduced Bivariate Bernoulli model (BBM). Suppose that (Y, Z) follows bivariate random variables distributed as

$$(Y_i, Z_i) \sim \begin{cases} (X_1, X_2) & \text{with prob. } 1-\alpha, \\ (X_1, X_2) & \text{with prob. } \alpha, \end{cases} \tag{18}$$

where X_1 and X_2 are independently distributed Bernoulli random variables with parameters p_1 and p_2, respectively. Let $p_{yz} = Pr(Y = y, Z = z)$ for $y, z = \{0, 1\}$. Thus, based on the parameters involved in the above model (1), we have the following cell probabilities in the DRS

$$\begin{aligned} p_{11} &= \alpha p_1 + (1-\alpha)p_1 p_2, \\ p_{10} &= (1-\alpha)p_1(1-p_2), \\ p_{01} &= (1-\alpha)(1-p_2)p_2, \\ p_{00} &= \alpha(1-p_1) + (1-\alpha)(1-p_1)(1-p_2). \end{aligned} \tag{19}$$

Consequently, the marginal probabilities are

$$\begin{aligned} p_Y &= p_{1.} = p_1, \\ p_Z &= p_{.1} = \alpha p_1 + (1-\alpha)p_2, \end{aligned} \tag{20}$$

with $Cov(Y, Z) = \alpha p_1(1-p_1)$. This introduces Bivariate Bernoulli model for Dual-record System (BBM-DRS).

Based on that, Chatterjee and Bhuyan (2017) proposed two models. We will focus only on model 2. Let us assume that U of size N can be divided into two mutually exclusive and exhaustive sub-populations U_A and U_B with size N_A and N_B (i.e. stratified populations).

In the model 2, Chatterjee and Bhuyan (2017) relaxed the assumption (S3) and considered BBM-DRS for both U_A and U_B with parameters $p_1 = p_{1k}$, $p_2 = p_{2k}$, $\alpha = \alpha_k$, $N = N_k$ for $k = A, B$. Additionally, $\alpha_A = \alpha_B = \alpha_0$, because we have two lists, so we have only one value of correlation between these data sources.

Chatterjee and Bhuyan (2017) considered two methods to estimate vector of parameters $(N_A, N_B, p_1, p_{2A}, p_{2B}, \alpha_0)$ in this model. Method of moments solution is given by (21)

$$\begin{aligned}
\hat{p}_{2A} &= \frac{x_{01B}\left(x_{1\cdot A}x_{10B} - x_{1:B}x_{10A}\right)}{x_{1\cdot B}\left(x_{01A}x_{10B} - x_{10A}x_{01B}\right)} \\
\hat{p}_{2B} &= \frac{x_{01A}\left(x_{1\cdot A}x_{10B} - x_{10A}x_{10A}\right)}{x_{1\cdot A}\left(x_{1A}x_{10B} - x_{10A}x_{10A}\right)} \\
\hat{\alpha_0} &= 1 - \frac{x_{10A}}{x_{1\cdot A}}\frac{1}{1-\hat{p}_{2A}} \\
\hat{p}_1 &= \frac{x_{1\cdot A}}{1 + \frac{x_{1\cdot A}}{x_{10A}}\left(\frac{1}{\hat{p}_{2A}} - 1\right)} \\
\hat{N}_A &= \frac{x_{1\cdot B}}{\hat{p}_1} \\
\hat{N}_B &= \frac{x_{1.B}}{\hat{p}_1}
\end{aligned} \tag{21}$$

and Maximum Likelihood solution is solution of the following likelihood function (22)

$$\begin{aligned}
L\left(\theta|\underline{x}_A, \underline{x}_B\right) \propto & \frac{N_A! N_B!}{\left(N_A - x_{0A}\right)!\left(N_B - x_{0B}\right)!}\left[\alpha_0 p_1 + \left(1-\alpha_0\right)p_{2A}\right]^{x_{11A}} \\
& \times \left[\alpha_0 p_1 + \left(1-\alpha_0\right)p_1 p_{2B}\right]^{x_{11B}} p_1^{\left(x_{10A}+x_{10B}\right)}\left(1-p_1\right)^{\left(x_{01A}+x_{01B}\right)} \\
& \times p_{2A}^{x_{01A}} p_{2B}^{x_{01B}}\left(1-p_{2A}\right)^{x_{10A}}\left(1-p_{2B}\right)^{x_{10B}}\left(1-\alpha_0\right)^{\left(x_{10A}+x_{01A}+x_{10B}+x_{01B}\right)} \\
& \times \left[\alpha_0\left(1-p_1\right) + \left(1-\alpha_0\right)\left(1-p_1\right)\left(1-p_{2A}\right)\right]^{\left(N_A - x_{0A}\right)} \\
& \times \left[\alpha_0\left(1-p_1\right) + \left(1-\alpha_0\right)\left(1-p_1\right)\left(1-p_{2B}\right)\right]^{\left(N_B - x_{0B}\right)},
\end{aligned} \tag{22}$$

where $\underline{x}_k = (x_{11k}, x_{10k}, x_{01k})$ and $x_{0k} = x_{11k} + x_{10k} + x_{01k}$, for $k = A, B$.

In a recent paper, Chatterjee and Bhuyan (2019a) extended the Bivariate Bernoulli model to a dependent triple-record system. Note that, both approaches assume no duplicates and erroneous records (i.e. no over-coverage).

3.2.4.2 TRIMMED DUAL SYSTEM ESTIMATION

Zhang (2015) considered the case of two data sources where the first suffers from over- and under-coverage and the second (i.e. independent survey) suffers only from under-coverage. Note that, Zhang (2015) states *An additional independent coverage survey with only undercoverage error is always needed for estimation*. Further, Zhang and Dunne (2017) proposed Trimmed Dual system estimation (TDSE) and we will focus on this approach.

Let N be the unknown size of the target population, denoted by U. Let A be the first list enumeration that is of size x. Suppose list A is subject to over-coverage, and the number of erroneous records is r, i.e., the size of set $\{i; i \in A \text{ and } i \notin U\}$. Suppose list A is subject to under-coverage as well, so that

$x - r < N$. Let B be the second list enumeration that is of the size n. Suppose list B is subject to only under-coverage, so that $n \leq N$, but there are no erroneous records in B.

Suppose the records in lists A and B can be linked to each other in an error-free manner, which we refer to simply as the assumption of matching. This is a very common assumption, although it can be difficult to satisfy in practice if the two lists do not share a unique identifier. However, the linkage errors are not easy to adjust. For now, suppose that error-free matching between A and B gives rise to the matched list AB with m records.

Because one does not actually know r, i.e., the number of erroneous records in A. But one can (a) score some records in list A, which are most likely to be erroneous, (b) match them to list B, and then (c) calculate the DSE as if list A would have been free of erroneous enumeration once the scored records had been removed. This yields what we call the trimmed DSE, given by

$$\hat{N}_k = n \frac{x - k}{m - k_1}, \tag{23}$$

where k is the number of scored (verified) records in list A (of size n, and k_1 is the number of records among them that can be matched to list B (of size x). The naïve DSE estimator of N would be in this case

$$\hat{N} = nx/m. \tag{24}$$

To summarise, as long as one is able to score the erroneous records in list A more effectively than random scoring and one does not score more records than the total number of erroneous records in list A, the trimmed DSE (23) can be expected to reduce the bias of the naïve DSE and move it closer to the ideal DSE.

Variance of (23) obtained by linearization is given by

$$\hat{V}(\hat{N}_k) = n(n - m_k)x_k(x_k - m_k)/m_k^3, \tag{25}$$

where $x_k = x - k$ and $m_k = m - k_1$.

Now, consider that source B also suffers from over-coverage. The second estimator considers the case when both sources contain over-coverage

$$\hat{N}_{\boldsymbol{k}} = \frac{(n_1 - r_1)(n_2 - r_2)}{n_{12} - r_{12}}, \tag{26}$$

where r_1, r_2 and r_{12} be the number of erroneous records in list A (of size n_1), B (of size n_2) and AB (of size n_{12}), respectively.

Finally, Zhang and Dunne (2017) consider record linkage errors when identifiers are not available. Consider the case when erroneous enumeration is only present in list A. Let m_L be the number of records in the linked list AB. Given the existence of linkage errors, let u be the number of missed matches, and let e be the number of false links. In other words, the true number of matches between A and B is given by

$$m = m_L - e + u. \tag{27}$$

Let f be the rate of missing (matches) and q is the rate of false links and $\hat{\xi} = (1 - \hat{q})/(1 - \hat{f})$ then a trimmed LDSE can possibly be given by

$$\hat{N}_k = \frac{n(x-k)}{\hat{\xi}(m_L - k_{1L})}, \tag{28}$$

where k is the number of records scored in A (of size n) and k_{1L} that among the linked list AB, and $\hat{\xi} = (1-\hat{q})/(1-\hat{f})$ is based on the estimated linkage error parameters; f is the rate of missing (matches) and q is the rate of false links. But it is impossible to conclude on the properties of the trimmed LDSE without some strong additional assumptions involving the linkage errors.

Zhang and Dunne (2017) applied trimming based on subjectively identifying those records that are most likely to contain erroneous information. In this case, the trimming method removes records for persons in list A in several steps, when a person has only an employment record with lower remuneration than a specified amount in EUR. For instance, they consider the following steps:

1. step 1 requires removing records for persons with only an employment record with earnings lower than 1000 EUR,
2. step 2 removes records for persons with only an employment record with earnings lower than 2000 EUR,
3. and so on.

Note that Zhang and Dunne (2017) mainly focus on situation when over-coverage is a result of errors or out-of-scope units but not duplicated records.

Further advances and readings may include:

- Capture-recapture methods in the presence of linkage errors – Zhang and Chambers (2019, Chapter 3),
- Estimating population size in multiple record system with uncertainty of state identification – Zhang and Chambers (2019, Chapter 8),
- Log-linear models for erroneous list data – Zhang and Chambers (2019, Chapter 9).

3.3 Domain-level methods

3.3.1 Measurement error models

In general, for the measurement error model we assume that

$$\gamma = \theta + b + \epsilon, \tag{29}$$

where γ is the proxy measurement (i.e. number of job vacancies based on job advertisements), θ is the parameter of interest (i.e. true number of job vacancies), b is the bias introduced by the differences between concepts and finally, $\epsilon \sim N(0, \sigma)$ is an error.

These models are rather used to estimate the bias and thus require source to evaluate it. For instance, Lohr and Brick (2012) considered small area (domain) estimation with two data surveys, where one is subject to measurement error that results with additive or multiplicative bias. Consider the first case, where $\bar{y}_d$ is estimated from the first survey and $\bar{x}_d$ is from the second survey that suffers from measurement error. Relation between these surveys can be written as in the equation (30)

$$\begin{pmatrix} \bar{y}_d \\ \bar{x}_d \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \theta_d \\ \theta_d + \eta_d \end{pmatrix}, \sigma^2 \begin{pmatrix} 1/n_{yd} & 0 \\ 0 & 1/n_{xd} \end{pmatrix} \right], \tag{30}$$

where θ_d is the true value, η_d is bias, σ^2 is variance and n_{yd}, n_{xd} are effective sample sizes for the first and the second survey, respectively.

Note that, measurement errors models require existence of the *gold standard* that measures the concept without errors. As the JVS in most countries are based on sample surveys, it has measurement error, so the existing estimates cannot be used to correct bias in the estimates based on online data.

3.3.2 Modelling of under-reporting

3.3.2.1 BAYESIAN ESTIMATION OF UNDER-REPORTED COUNT DATA

Recently, Stoner et al. (2019) proposed a Bayesian hierarchical model for under-counting of Tuberculosis in Brazil. Their model can be summarised as follows.

- Let $y_{i,t,s}$ be the number of events occurring in the space $s \in S$, time $t \in T$ and in domain i (e.g. age group) .
- If $y_{i,t,s}$ is believed to have been perfectly observed, the counts are conventionally modelled by an appropriate conditional distribution $p(y_{i,t,s}|\boldsymbol{\theta})$, usually either Poisson or Negative Binomial.
- $\boldsymbol{\theta}$ represents random effects allowing for various dependency and grouping structures (e.g., space and time), as well as parameters associated with relevant covariates.
- In case of under-counting/-reporting instead of $y_{i,t,s}$ we observe $z_{i,t,s}$ and let $I_{i,t,s}$ be the index of under-reporting treated as a random variable.
- It is assumed that $I_{i,t,s} \sim$ Bernoulli$(\pi_{i,t,s})$ but Stoner et al. (2019) assumed that $I_{i,t,s}$ can be continuous in the range $[0,1]$ to be interpreted as the proportion of true counts that have been reported.
- Finally, the proposed hierarchical model given by (31)

$$\begin{aligned} z_{i,t,s} &\sim \text{Binomial}(\pi_{i,t,s}, y_{i,t,s}) \\ \log\left(\frac{\pi_{i,t,s}}{1-\pi_{i,t,s}}\right) &= \beta_0 + \sum_{j=1}^{J} \beta_j w_{i,t,s}^{(j)} \\ y_{i,t,s} &\sim \text{Poisson}(\lambda_{i,t,s}) \\ \log(\lambda_{i,t,s}) &= \alpha_0 + \sum_{k=1}^{K} \alpha_k x_{i,t,s}^{(k)}, \end{aligned} \tag{31}$$

  where $y_{i,t,s}$ are unknown true counts that follow Poisson distribution, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown vectors of parameters associated with the probability of under-reporting $\pi_{i,t,s}$ and level of true counts $y_{i,t,s}$. Note that we have two sets of independent variables $\boldsymbol{Z}$ associated with $\pi_{i,t,s}$ and $\boldsymbol{X}$ associated with $y_{i,t,s}$. Therefore, this model requires access to powerful variables.
- The final model presented in Stoner et al. (2019) was more complicated than model (31) because of inclusion of random effects (including spatial autocorrelation), non-linear relationship with $\boldsymbol{Z}$ and $\boldsymbol{X}$ and offset $log(P_{i,t,s})$ was used instead of α_0 where $P_{i,t,s}$ is the population size.

Another approach to correcting under-reporting is to base inference on the censored likelihood. This is the product of the evaluation of (32)

$$p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta}) = \prod_{I_{i,t,s}=1} p(y_{i,t,s}|\boldsymbol{\theta}) \prod_{I_{i,t,s}=0} p(y_{i,t,s} \geq z_{i,t,s}|\boldsymbol{\theta}). \tag{32}$$

In this framework, the indicator $I_{i,t,s}$ for which data are under-reported is binary (where $I_{i,t,s} = 1$ when $z_{i,t,s} = y_{i,t,}$) and notation is as in Stoner et al. (2019). Strength of this approach is that all of the observed counts contribute to the inference and, by accounting for the under-reporting in the model design, a more reliable inference on $\boldsymbol{\theta}$ is obtained. However, information on which counts are under-reported is not always readily available, introducing the challenge of having to determine or estimate this classification. Oliveira et al. (2017) presented an alternative to this approach, which treats the binary under-reporting indicator $I_{i,t,s}$ as unobserved and therefore random.

Based on this approach the following requirements about the data should be underlined

- $z_{i,t,s}$ should be free of duplicated records,
- strong set of correlates $\boldsymbol{Z}$ and $\boldsymbol{X}$ are required,
- information about the reference population P may be needed.

Fortunately, the model is easily implemented in R (R Core Team, 2019) using `stan` (Carpenter et al., 2017) or `brms` (Bürkner, 2017).

3.3.2.2 ESTIMATION OF HIDDEN POPULATION BASED ON DOMAIN DATA

Now, we focus on an alternative approach proposed by Zhang (2008) based solely on an aggregated data assuming relationship between the size of registered (observed) and unregistered (unobserved) population. This relationship is further modelled by generalized non-linear regression model, in particular Gamma-Poisson hierarchical model. In the paper, Zhang (2008) focused on estimation of irregular migration but we adapt the description and notation to online data.

For both the target and the reference populations, let $i = 1, ..., t$ be the index of the sub-population classified by the country of citizenship and origin, respectively. Assume that the observed number of irregular residents follows a Poisson distribution, with parameter λ_i, denoted by

$$m_i \sim \text{Poisson}\left(\lambda_i\right). \tag{33}$$

It is intuitively plausible that the parameter λ_i should depend on two other quantities: (a) the total number of irregular residents from country i, denoted by M_i, and (b) the probability of being observed, i.e. the probability for an irregular resident denoted by p_i, i.e. $\lambda_i = M_i p_i$. In addition, let $u_i = M_i p_i / E(M_i p_i | n_i, N_i)$ where $E(M_i p_i | n_i, M_i)$ denotes the conditional expectation of $M_i p_i$ given n_i and N_i. The u_i is a random effect that accounts for heterogeneous variation from one country to another. Together, we have

$$\lambda_i = \mu_i u_i, \text{ where } \mu_i = E\left(M_i p_i | n_i, N_i\right) = E\left(M_i | N_i\right) \cdot E\left(p_i | M_i, n_i, N_i\right). \tag{34}$$

We complete the model specification by assuming that

$$\xi_i = E\left(M_i | N_i\right) = N_i^{\alpha}, \tag{35}$$

$$E\left(p_i | M_i, n_i, N_i\right) = E\left(p_i | n_i, N_i\right) = \left(\frac{n_i}{N_i}\right)^{\beta}, \tag{36}$$

$$u_i \sim \text{Gamma}(1, \phi), \tag{37}$$

The target parameter and its estimator are given as, respectively,

$$\xi = \sum_{i=1}^{t} E\left(M_i | N_i\right) = \sum_{i} N_i^{\alpha}, \tag{38}$$

and

$$\hat{\xi} = \sum_{i} N_i^{\hat{\alpha}}, \tag{39}$$

where $\hat{\alpha}$ is the estimator of α. We shall use the maximum likelihood estimator (MLE). Denote by $L(\eta, \mathbf{m})$ the likelihood of $\eta = (\alpha, \beta, \phi)$ given m_i, for $i = 1, ..., t$. Under the Poisson gamma model, we have

$$f\left(m_i, u_i; \eta\right) = \frac{e^{-\mu_i u_i}\left(\mu_i u_i\right)^{m_i}}{m_i!} \cdot \frac{\phi^{\phi} u_i^{\phi-1} e^{-\phi u_i}}{\Gamma(\phi)} = \frac{\mu_i^{m_i}\phi^{\phi}}{m_i!\Gamma(\phi)} e^{-u_i(\mu_i+\phi)} u_i^{m_i+\phi-1}, \tag{40}$$

where $\Gamma()$ is the gamma function. Thus,

$$\begin{aligned} f\left(m_i; \eta\right) &= \int_0^{\infty} f\left(m_i, u_i; \eta\right) d\left(u_i\right) \\ &= \frac{\mu_i^{m_i}\phi^{\phi}}{m_i!\Gamma(\phi)} \int_0^{\infty} e^{-\left(\sqrt{u_i}\right)^2(\mu_i+\phi)} \left(\sqrt{u_i}\right)^{2(m_i+\phi-1)} 2\sqrt{u_i} d\left(\sqrt{u_i}\right) \\ &= \frac{\mu_i^{m_i}\phi^{\phi}}{m_i!\Gamma(\phi)} \left(\mu_i+\phi\right)^{-(m_i+\phi)} \Gamma\left(m_i+\phi\right), \end{aligned} \tag{41}$$

based on the identity $\int_0^{\infty} e^{-\gamma z^2} z^k dz = \frac{1}{2}\gamma^{-\frac{k+1}{2}}\Gamma\left(\frac{k+1}{2}\right)$, with $z = \sqrt{u_i}$ and $k = 2(m_i + \phi) - 1$. Notice that, conditional on m_i, u_i has the gamma distribution with mean $(m_i + \phi)/(\mu_i + \phi)$ and variance $(m_i + \phi)/(\mu_i + \phi)^2$.

The likelihood is given by

$$L(\eta; \mathbf{m}) = \prod_{i=1}^{t} f\left(m_i; \eta\right), \tag{42}$$

The log-likelihood is thus, disregarding constant terms, given by

$$l(\eta; \mathbf{m}) = \sum_{i=1}^{t} l_i(\eta), \tag{43}$$

where

$$\begin{aligned} l_i(\eta) &= m_i \log \mu_i - (m_i + \phi) \log\left(\mu_i + \phi\right) + \log\Gamma\left(m_i + \phi\right) + \phi \log \phi - \log\Gamma(\phi) \\ &\doteq m_i \log \mu_i - (m_i + \phi) \log\left(\mu_i + \phi\right) + \phi \log \phi \\ &+ \left(m_i + \phi - 0.5\right) \log\left(m_i + \phi\right) - \left(m_i + \phi\right) - (\phi - 0.5) \log(\phi) + \phi \\ &= m_i \log \mu_i - (m_i + \phi) \log\left(\mu_i + \phi\right) + \left(m_i + \phi - 0.5\right) \log\left(m_i + \phi\right) + 0.5 \log \phi, \end{aligned} \tag{44}$$

by the Stirling approximation $\log\Gamma(z) \doteq (z - 0.5)\log(z) + 0.5\log(2\pi) - z$.

The mean parameter μ_i is linear on the log scale, denoted by $log\mu_i = x_i^T\gamma$ with generic vector of covariates x_i and parameters γ. Now that $l_i(\eta)$ depends on γ only through μ_i, we have

$$\frac{\partial l_i(\eta)}{\partial \gamma} = \frac{\partial l_i(\eta)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \log \mu_i}\frac{\partial \log \mu_i}{\partial \gamma} = \frac{\partial l_i(\eta)}{\partial \mu_i}\mu_i x_i = \frac{m_i - \mu_i}{\mu_i + \phi}\phi x_i, \tag{45}$$

where $\partial l_i(\eta)/\partial\mu_i = m_i/\mu_i - (m_i + \phi)\,/\,(\mu_i + \phi)$, and

$$\frac{\partial l_i(\eta)}{\partial \phi} = -\log(\mu_i + \phi) - \frac{m_i + \phi}{\mu_i + \phi} + \log(m_i + \phi) + \frac{m_i + \phi - 0.5}{m_i + \phi} + \frac{1}{2\phi}. \tag{46}$$

Moreover,

$$\frac{\partial^2 l_i(\eta)}{\partial \gamma \partial \gamma^T} = \frac{\partial^2 l_i(\eta)}{\partial \mu_i^2}\mu_i x_i \frac{\partial \mu_i}{\partial \gamma^T} + \frac{\partial l_i(\eta)}{\partial \mu_i} x_i \frac{\partial \mu_i}{\partial \gamma^T} = -\left(\frac{m_i + \phi}{\mu_i + \phi}\phi\right) x_i x_i^T, \tag{47}$$

$$\frac{\partial^2 l_i(\eta)}{\partial \phi^2} = -\frac{2\mu_i + \phi - m_i}{(\mu_i + \phi)^2} + \frac{m_i + \phi + 0.5}{(m_i + \phi)^2} - \frac{1}{2\phi^2}, \tag{48}$$

and

$$\frac{\partial^2 l_i(\eta)}{\partial \gamma \partial \phi} = \left(\partial\left(\frac{\partial l_i(\eta)}{\partial \mu_i}\right)/\partial\phi\right)\mu_i x_i = -\frac{\mu_i - m_i}{(\mu_i + \phi)^2}\mu_i x_i = \left(\frac{\partial^2 l_i(\eta)}{\partial \phi \partial \gamma^T}\right)^T. \tag{49}$$

The MLE of η, denoted by $\hat{\eta}$, is given by the solution to the likelihood equations, i.e.

$$\frac{\partial l(\eta; \mathbf{m})}{\partial \eta} = \sum_{i=1}^{t} \frac{\partial l_i(\eta)}{\partial \eta} = 0. \tag{50}$$

The MLE can be obtained using the Newton-Raphson method. As the starting values we use the ordinary least squares fit of the heuristic log-ratio model. We use the estimated α and β as the starting values for the same parameters of the Poisson-gamma model, and the inverse of the estimated $V(\epsilon_i)$ as the starting value for ϕ. We estimate the model using `maxLik` package in *R* software.

3.4 Summary

This chapter summarised the literature on population size estimation in the presence of non-sampling data, in particular under- and over-coverage. Most of these methods require information about auxiliary variables from the study population, audit sample to assess quality or independent data source that is source of error-free records or dependent variables. Moreover, in case of non-probability data estimation and assessing uncertainty are based on the model-based assumptions and therefore require simulation studies.

4 Comparing OJA data with job vacancy statistics

4.1 Aligning job advertisements data to Job Vacancies Statistics

The final goal of the study was to assess methods to estimate the total number of job vacancies based on online job advertisements (OJA) data. In order to render OJA data comparable with currently published official statistics we tried to approximate OJA data to the definition of job vacancy used in job vacancies statistics, as well as to align the type of variable (from flow to stock) and the reference period. We followed the OJA definitions used by Cedefop as well as their deduplication process.

The relation between job advertisements, or job offers (we will use both terms as having the same meaning), and job vacancies is comparable to the one between family and household. They look similar but the definition is in fact different. Actually, the definition taken from the Cambridge dictionary[6] states that:

Cambridge's job advertisement definition

Job advertisement is an announcement in a newspaper, on the Internet, etc. about a job that people can apply for.

Cedefop uses a similar definition of job advertisement.

Cedefop's job advertisement definitions

- The advertisement, defined as the job offer published by a company to search for a new employee.
- The word 'job vacancy' in this document means the document (typically in HTML) that describes the job offer.

Having that in mind, one should align the definitions of job offer and of job vacancy. Cedefop describes the process as follows:

1. The generation process of vacancies from advertisements is called expansion.

[6] Source: https://dictionary.cambridge.org/pl/dictionary/english/job-advertisement

2. Expansion is not necessary for all analysis dimensions but only for those considered distinguishing for the job offer (e.g. expansion can be performed for the profession but not for the contract).
3. There are two possible ways of expanding advertisements:
   - By selection: choosing one of the possible jobs offers found in the advertisement
   - By repetition: generating as many vacancies as the number of jobs offers found in an advertisement.
4. Cedefop also considers using expiry date to align with job vacancy definition: "As discussed in Section 6 a critical parameter that has to be considered in assessing the data quality is the identification of the expiry date of the posted vacancy. This information allows us to define the average duration of vacancies which is necessary to construct a flow measure comparable to the standard definition adopted by statistical offices. In the data used in this preliminary experiment the average duration of a vacancy is approximately one month (39.25 days)."

Job advertisements (including those online) and job vacancies from a probability-based survey are two different measures of job openings. They cover different populations. Job vacancies and job advertisements might also differ by the definition of a job opening. A job advertisement may not refer to a job from the labour law point of view, and we are not certain whether a company is convinced that they want to hire a worker. But after removal of non-work job advertisements and assuming that a company publishes a valid job advertisement (e.g. credible websites with job advertisements are used), online job advertisements by definition are just one way to present a job vacancy.

In order to align the type of variable (from flow to stock) and the reference period of the OJA data to make it comparable to job vacancies statistics, we identified those advertisements that were active at the end of each quarter:

- 2018Q3 = expire_date >= "2018-09-30" & grab_date <= "2018-09-30",
- 2018Q4 = expire_date >= "2018-12-31" & grab_date <= "2018-12-31",
- 2019Q1 = expire_date >= "2019-03-31" & grab_date <= "2019-03-31",
- 2019Q2 = expire_date >= "2019-06-30" & grab_date <= "2019-06-30",
- 2019Q3 = expire_date >= "2019-09-30" & grab_date <= "2019-09-30",
- 2019Q4 = expire_date >= "2019-12-31" & grab_date <= "2019-12-31",

We also calculated the number of days between the *grab_date* and the end of quarters:

- 2018Q3_days = "2018-09-30" - grab_date,
- 2018Q4_days = "2018-12-31" - grab_date,
- 2019Q1_days = "2019-03-31" - grab_date,
- 2019Q2_days = "2019-06-30" - grab_date,
- 2019Q3_days = "2019-09-30" - grab_date,
- 2019Q4_days = "2019-12-31" - grab_date.

After this transformation, we got the following results.

- number of advertisements not active at the end of any quarter – 8,183,067.

- number of advertisements active at the end of one single quarter – 44,799,353,
- number of advertisements active at the end of two quarters – 14,697,645,

Table 3 presents the number of advertisements in the OJA dataset at the end of each quarter. A significant decrease is observed between 2019Q1 and 2019Q2. For instance, in Germany the number of job offers decreased by 2 million during this period. This type of decrease is not visible in job vacancy statistics compiled by National Statistical Institutes (NSIs) in Member-States. As the data for the same period for 2018 is not available, we are not sure what is the reason for this decrease. Such unexpected changes will have significant impact on estimates of job vacancies.

Table 3: Number of advertisements active at the end of each quarter

Country	2018Q3	2018Q4	2019Q1	2019Q2	2019Q3	2019Q4
Austria	281,546	298,612	364,221	214,306	223,323	138,868
Belgium	305,807	367,834	599,394	515,096	452,800	333,538
Bulgaria	11,919	34,977	83,908	60,684	49,132	33,602
Croatia	2,086	7,158	22,813	23,398	18,406	19,502
Cyprus	1,434	2,946	8,175	6,404	7,263	4,298
Czechia	93,145	138,361	264,765	239,250	225,002	138,011
Germany	2,356,289	3,286,338	4,607,939	2,618,403	2,771,006	1,767,932
Denmark	10,337	26,902	87,021	58,743	61,989	59,240
Estonia	1,049	4,972	11,314	10,430	8,879	7,707
Greece	365	3,870	14,326	13,831	10,380	8,957
Spain	226,842	314,062	532,433	511,229	521,692	411,553
Finland	13,975	27,810	81,143	55,687	60,678	42,939
France	1,999,566	2,660,478	3,293,666	2,075,295	2,821,807	3,123,865
Hungary	18,474	29,377	83,125	60,492	73,386	44,816
Ireland	105,512	135,307	180,270	161,850	149,908	111,068
Italy	423,393	576,520	817,059	501,388	730,925	797,023
Lithuania	1,476	9,832	24,365	24,817	24,912	20,973
Luxembourg	10,992	17,810	23,070	23,355	21,214	18,625
Latvia	708	6,446	25,751	20,126	16,144	14,929
Malta	2,648	6,222	2,981	3,218	2,689	1,875
Netherlands	574,150	692,293	864,496	462,977	599,563	752,627
Poland	309,962	320,509	400,852	265,979	402,984	376,891
Portugal	40,806	71,498	79,903	68,078	74,404	45,940
Romania	17,074	43,203	134,329	80,192	91,859	67,940
Sweden	104,367	196,539	355,811	240,500	158,650	126,353
Slovenia	2,833	1,499	12,555	15,262	14,709	11,000
Slovakia	22,710	22,605	63,673	49,356	53,989	44,726

Table 4 presents the number of unique sources of advertisements (websites) in the OJA dataset. We see that from quarter-to-quarter in 2018 the number of sources increases but after 2019Q1 we see a decrease. This may be the reason for the decrease of advertisements in 2019Q2.

Furthermore, table 4 further investigates how many data sources were present over the whole period. The share of unique sources varies between countries. The lowest number of sources observed in all 6 quarters is in Latvia (10.8%), while the highest is in France (48.8%). This indicates that stability of sources varies and may have further influence on estimates of job vacancies based on job advertisements.

Table 4: Number of unique sources of advertisements in Cedefop's OJA dataset, by quarter

Country	Number of sources								
	2018Q3	2018Q4	2019Q1	2019Q2	2019Q3	2019Q4	Whole period	Total	Share
Austria	91	112	135	105	107	90	48	177	27.12
Belgium	132	159	170	160	142	138	84	210	40.00
Bulgaria	111	137	167	139	126	123	66	201	32.84
Cyprus	40	76	94	73	62	55	17	125	13.60
Czechia	63	83	106	87	71	52	30	144	20.83
Germany	147	171	186	175	156	141	102	218	46.79
Denmark	31	54	59	51	36	48	16	95	16.84
Estonia	32	61	91	65	50	44	15	116	12.93
Greece	17	35	49	45	39	57	14	84	16.67
Spain	115	149	170	159	138	125	72	204	35.29
Finland	45	68	86	66	65	61	20	124	16.13
France	152	186	205	192	192	175	118	242	48.76
Croatia	74	94	104	98	83	78	34	155	21.94
Hungary	99	120	132	120	101	91	49	174	28.16
Ireland	119	132	158	129	126	110	71	189	37.57
Italy	133	174	183	177	165	157	96	226	42.48
Lithuania	34	63	83	67	53	49	21	117	17.95
Luxembourg	53	68	63	55	62	55	29	103	28.16
Latvia	16	32	57	34	30	24	8	74	10.81
Malta	50	59	61	50	53	45	20	102	19.61
Netherlands	134	155	187	164	155	142	84	222	37.84
Poland	60	87	103	83	71	66	26	144	18.06
Portugal	112	135	144	132	123	113	63	183	34.43
Romania	117	139	176	151	141	129	80	206	38.83
Sweden	87	98	124	101	91	76	42	163	25.77
Slovenia	42	51	63	49	45	45	20	95	21.05
Slovakia	76	99	117	91	83	70	40	154	25.97

Note: Column *Whole period* contains information about the number of sources observed in each quarter from 2018Q3 to 2019Q4. Column *Total* shows the total number of sources observed in the period and *Share* is calculated as values of *Whole period* divided by *Total*.

Finally, Table 5 presents co-occurrence of advertisements in several web sources by number of sources. We report statistics for one, two, three, and four or more data sources. For instance, in Poland 312,218 advertisements were present only in one source, 14,833 in two, 938 in three and 331 in four or more sources. Thus, we observe one-inflation in the distribution of re-captures of advertisements. We see that for certain countries such as Croatia, Cyprus, Denmark, Estonia, Greece, Latvia, Lithuania, Malta and Slovenia the number of re-captures is very small, in most cases lower than 100.

Table 5: Average number of advertisements occurring on one, two, three and four and more sources advertisements between 2018Q3 and 2019Q4

Country	Number of sources			
	1	2	3	4 and more
Austria	229,314	9,831	1,213	200
Belgium	401,652	12,952	449	41
Bulgaria	43,830	852	51	5
Croatia	15,339	100	6	2
Cyprus	5,009	36	6	1
Czechia	154,305	13,129	718	80
Denmark	50,236	228	6	1
Estonia	7,324	32	1	1
Finland	45,157	867	45	8
France	2,371,491	114,525	15,640	3,448
Germany	2,611,564	116,131	15,134	2,795
Greece	7,827	367	22	3
Hungary	45,909	2,617	108	33
Ireland	122,732	8,505	280	20
Italy	544,532	38,835	4,532	1,205
Latvia	13,961	26	2	0
Lithuania	17,458	119	9	3
Luxembourg	14,869	2,042	68	6
Malta	3,184	42	2	0
Netherlands	549,507	42,588	5,533	1,507
Poland	312,218	14,833	938	331
Portugal	51,508	4,792	677	71
Romania	65,755	2,760	246	118
Slovakia	39,339	1,438	118	94
Slovenia	9,562	39	1	1
Spain	381,104	16,055	1,618	364
Sweden	189,730	3,200	248	58

There is also variability in the number of advertisements observed in two sources in a given quarter. These sources may vary between quarters and it does not mean that these two websites are the same over the whole period. What we observe is high variance and varying pattern in number of ads.

One should note that these results depend on the effectiveness of the deduplication procedure applied.

4.2 Relation between online job offers and job vacancies

4.2.1 Descriptive statistics

Data available at Eurostat was limited to quarterly estimates from job vacancies statistics (JVS). The time span of our analysis was limited to 2018Q2-2020Q1. However, not all countries supplied data on job vacancies to Eurostat for the last quarter, which limited our analysis even more. OJA data in this quarter showed a significant decline (to 20% job ads in 2019Q4 and to 10% of the ones in 2019Q1). One of the possible reasons for this decline may have been the initial impact of Covid-19 in the labour market. The OJA dataset also contained few job advertisements for the first quarter. It constituted only 4% of job ads identified in the consecutive period. This period contained in fact only preliminary results of gathering job ads. For these reasons, and comparability purposes, we decided to exclude the first and last quarters from the analysis. Figure 2 compares both datasets for the period finally taken for the analysis, 2018Q3-2019Q4. It includes aggregated data for the 16 countries for which data from JVS were available. OJA data are less stable over time. We do observe a lower number of vacancies in 2019Q2 (but not as low as in the excluded 2018Q2), and then, their significant increase. In 2019Q1 we can see the largest number of job ads (in contrast to a significant decline in the excluded 2020Q1). Job ads are 2.7 times higher than job vacancies in 2019Q1. During the rest of the period OJA identifies 40%-91% job offers more than JVS does. The Pearson linear correlation between the two datasets is 0.61. OJA data show a lot more variation than JVS with coefficient of variation 0.28 in comparison with 0.03 for JVS and identify more job offers, with mean 4.6 mln job ads in comparison to 2.5 vacancies per quarter.

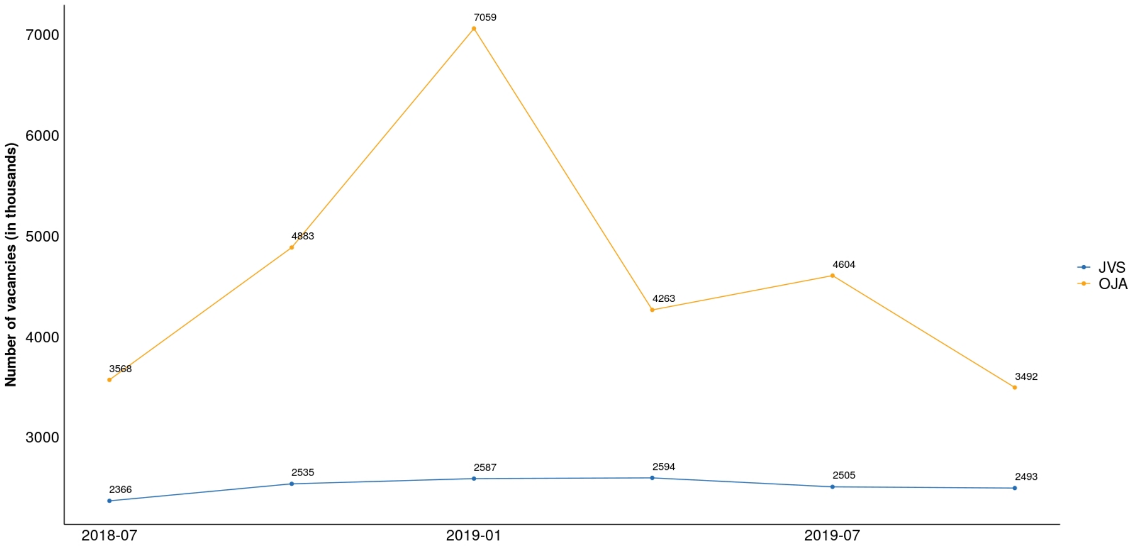

Figure 2: Number of job ads in the OJA dataset and number of vacancies from JVS at the end of quarter for 16 countries

Note: Countries included – Bulgaria, Croatia, Czechia, Estonia, Germany, Hungary, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Sweden

Large variance is typical for indices that are often used for predicting economic activity, that is leading indices. Internet data may be more sensitive to economic activity than probability-based surveys. For this reason seasonal patterns of both statistics may differ as well. Having short time series we cannot seasonally decompose them, let alone to other unobservable components. This led us to the conclusion that we cannot analyse time trends (long-term trends and business cycles) of the time series, but only their general time properties.

Job vacancy statistics were available only for 16 countries, while OJA data contains information on all countries. In terms of variance OJA data is more comparable to JVS data across countries than over time. In this case, the coefficient of variation equals 2.47 for OJA data and 2.14 for JVS data. The average number of job offers across countries differs significantly. Mean OJA number of job offers across all countries equals 185% of the one from JVS. In the case of the median it is 163%.

For nine countries the number of advertisements identified is larger than the number of vacancies estimated by JVS (Figure 3). Large differences between the number of jobs offers in comparison to job vacancies occurred for Luxembourg, Poland, Netherlands, Bulgaria, and Germany. For these countries the number of job advertisements were over two times higher than the number of vacancies. Almost two times lower numbers were estimated for Latvia and Slovenia. The highest similarity was found for Lithuania (2% dissimilarity).

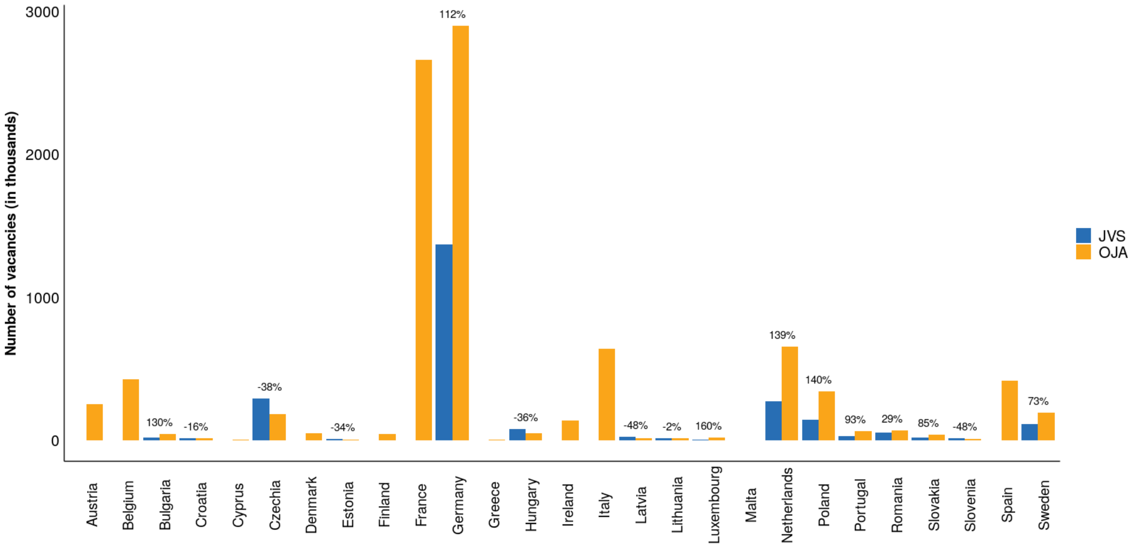

Figure 3: Mean quarterly number of job ads in OJA data and vacancies in JVS at the end of quarter

Note: Percentages over bars mean by how much OJA data differ from JVS data.

Time trends for each country with both available datasets are presented in Figure 4. Country statistics confirm the general conclusion of aggregate data on large variance of OJA data. Most countries share similar pattern of job ads with a peak in the middle of the analysed period. Notable exceptions are Netherlands and Poland. Descriptive statistics of the 96 values for the individual time periods and countries show that the mean is higher for OJA data by 85% and median is higher by 53%, with more deviation around them in OJA data (Table 6). Both datasets are leptokurtic, and OJA data is more leptokurtic, meaning that it may contain outliers. Both distributions are similarly positively skewed (right skewed). They concentrate in lower numbers of vacancies and advertisements.

The definitions of economic activity in OJA data and in JVS differ, even if both are classified using NACE. While the economic activity in OJA data is based on the vacancy description in the advertisement and may refer to any of the activities performed by the enterprise, the establishment or even to a smaller economic unit (e.g. a branch constituted by a separate legal unit), JVS are broken down by the main economic activity of the enterprise. This may produce differences in results. Despite this, data for NACE sections are more comparable between OJA data and JVS than data for the total of all economic activities. However, due to missing data, we were able to include only 13 countries. OJA mean across NACE sections accounted for 186% of JVS mean, while median accounted for 175% of the one for JVS data. Although the range was over two times lower for JVS than OJA, the coefficient of variation was comparable and two times lower than for countries (1.16 for OJA and 0.94 for JVS). The largest differences between the two datasets were for sections D, M, K, and J, for which OJA data showed three to almost seven times

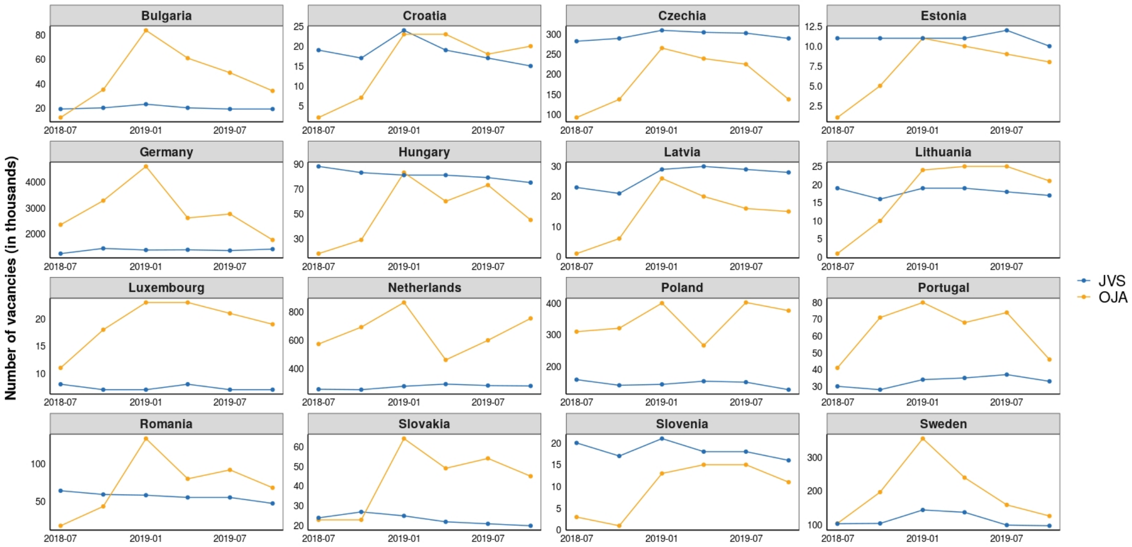

Figure 4: Number of job ads from OJA data and vacancies from JVS in countries over time

Table 6: Descriptive statistics for OJA data and JVS - countries and time

	OJA	JVS
nbr.val	96	96
min	708	6899
max	4607939	1458393
range	4607231	1451494
sum	27869238	15092711
median	45378	29604
mean	290305	157216
var	539026367233	107755914226
std.dev	734184	328262
coef.var	2.53	2.09
skewness	3.86	3.17
kurtosis	15.60	9.00

Note: Countries included – Bulgaria, Croatia, Czechia, Estonia, Germany, Hungary, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Sweden.

more vacancies (Figure 5). The opposite situation was found for sections F, A, and B, for which OJA data was lower by 38%-79% than JVS data. Highest similarity was found for sections I, E, R, and G (+/-20% dissimilarity).

NACE section breakdown (industries) shows a similar time pattern as the country did. JVS data are more stable, while OJA values rise in the beginning of the analysed period most often reaching a peak in 2019Q1, and then decline (Figure 6). For some sections, OJA mimics changes in JVS, e.g. sections A and R. Descriptive statistics show us that the median number of job offers in an average NACE section was 15% lower in OJA data, while the mean number is higher by 86% (Table 7). Again, variance is larger for OJA data, and even larger than in the country breakdown. The distribution is right-skewed, more for OJA than JVS, and also more than for the countries. OJA data distribution exhibits a very high kurtosis, a lot higher than for countries.

For most countries there are no estimates of JVS across occupations nor are uncertainty measures for recent years being reported. This prevents the comparison between OJA and JVS for ISCO occupations,

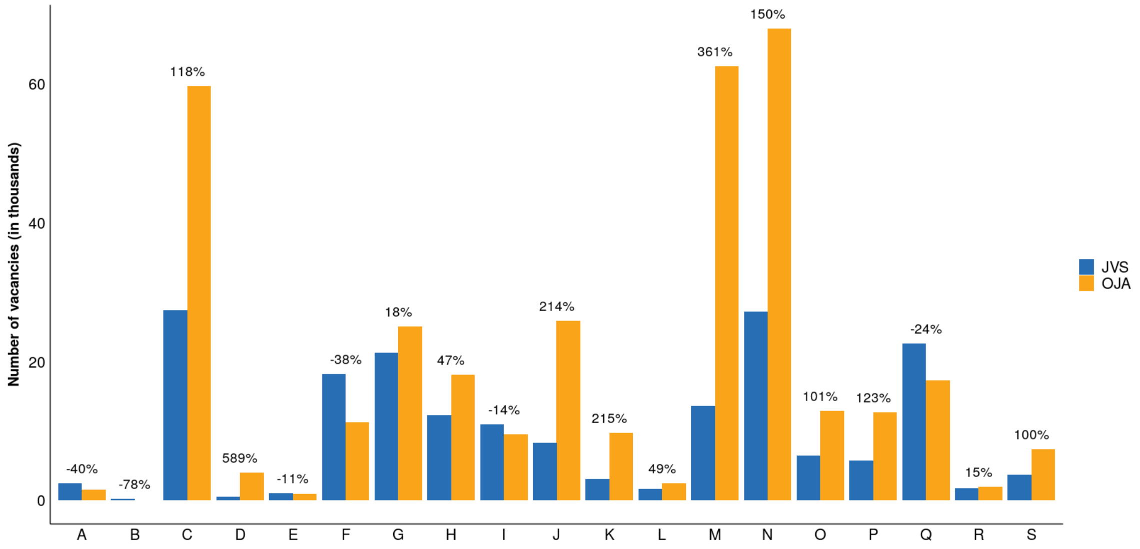

Figure 5: **Mean quarterly number of job ads from OJA and vacancies from JVS at the end of quarter across NACE sections**

Note: Percentages over bars mean by how much OJA data differs from JVS data.

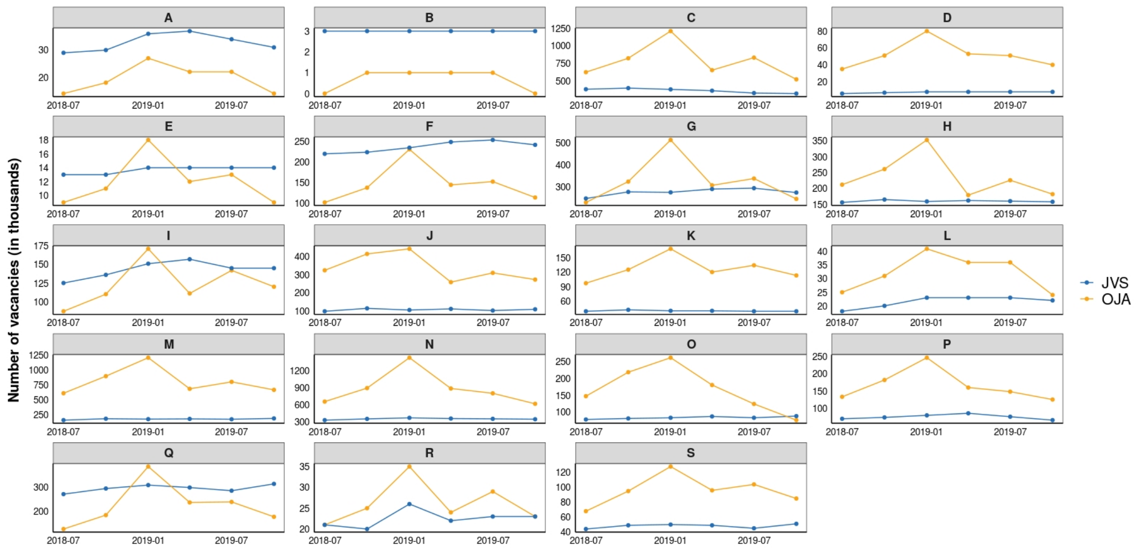

Figure 6: **Number of job ads from OJA and vacancies from JVS in NACE sections over time**

Note: Countries included – Bulgaria, Croatia, Czechia, Germany, Hungary, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Romania, Slovakia, Sweden.

even though OJA data includes detailed data. JVS provides data only for Hungary. At the time of analysis there are no publicly available data for other countries during 2019.

The occupation identified in OJA has misclassification errors that should be corrected through audit samples or linking with administrative data (such as jobs at employment offices). Having this in mind, we provide the comparison of OJA and JVS across ISCO major groups of occupations for Hungary.

Table 7: **Descriptive statistics for OJA data and JVS - NACE section and time**

	OJA	JVS
nbr.val	1482	1482
min	0	0
max	991253	273314
range	991253	273314
sum	27386400	14718277
median	1291	1514
mean	18479	9931
var	5014878703	812404965
std.dev	70816	28503
coef.var	3.83	2.87
skewness	7.95	5.42
kurtosis	77.00	34.08

Note: Countries included – Bulgaria, Croatia, Czechia, Germany, Hungary, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Romania, Slovakia, Sweden.

For almost all ISCO major groups of occupations OJA data provide lower estimates of job offers in Hungary than JVS do (Figure 7). A notable exception is group 1 ('Managers'), for which OJA data presented values 3.3-times higher than those in JVS. In the remaining groups the highest discrepancy in the number of identified jobs offers was for group 6 ('Skilled agricultural, forestry and fishery workers'), where OJA constituted only 19% of JVS, and group 8 ('Plant and machine operators, and assemblers'), where OJA constituted 23% of JVS. The lowest differences were identified for group 5 ('Service and sales workers'), with 3% less ads than vacancies, and group 2 ('Professionals'), with 5% less ads than vacancies. Descriptive statistics show generally larger quantiles of JVS than OJA, but similar coefficient of variation (0.70 for OJA and 0.67 for JVS).

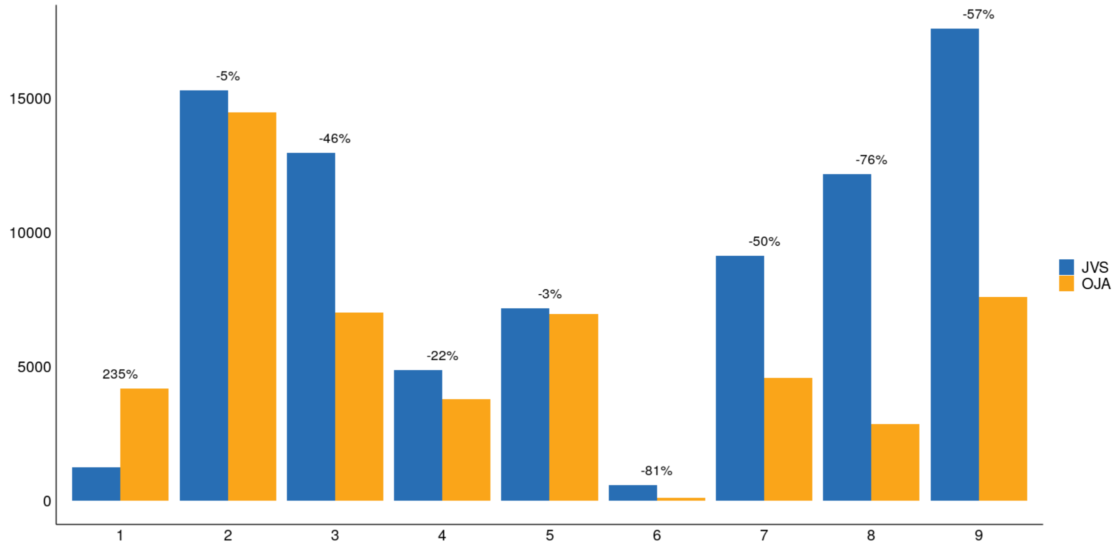

Figure 7: **Mean quarterly number of job ads from OJA and vacancies from JVS for Hungary across ISCO major groups of occupations at the end of quarter**

Note: Percentages over bars mean by how much OJA data differs from JVS.

The comparison of OJA and JVS across occupations in time confirms previous results for countries and industries on larger coefficient of variation of OJA than JVS (Figure 8). The time pattern was also similar, with growth in the beginning, peak in the middle of the analysed period and decline in the end, for each occupational group. In some sections this resembles JVS time trends well (groups 5 and 6), but in other sections changes of JVS and OJA over time are different.

The distributions of both datasets are quite different (Table 8). The median and mean numbers of job offers in an average ISCO group are respectively 42% and 36% lower in OJA data than in JVS. The distribution is right-skewed for OJA data, and slightly left-skewed for Eurostat. Kurtosis of OJA distribution is lower than the one for the normal distribution, and a lot lower than for country and NACE sections. JVS data distribution is visibly platykurtic.

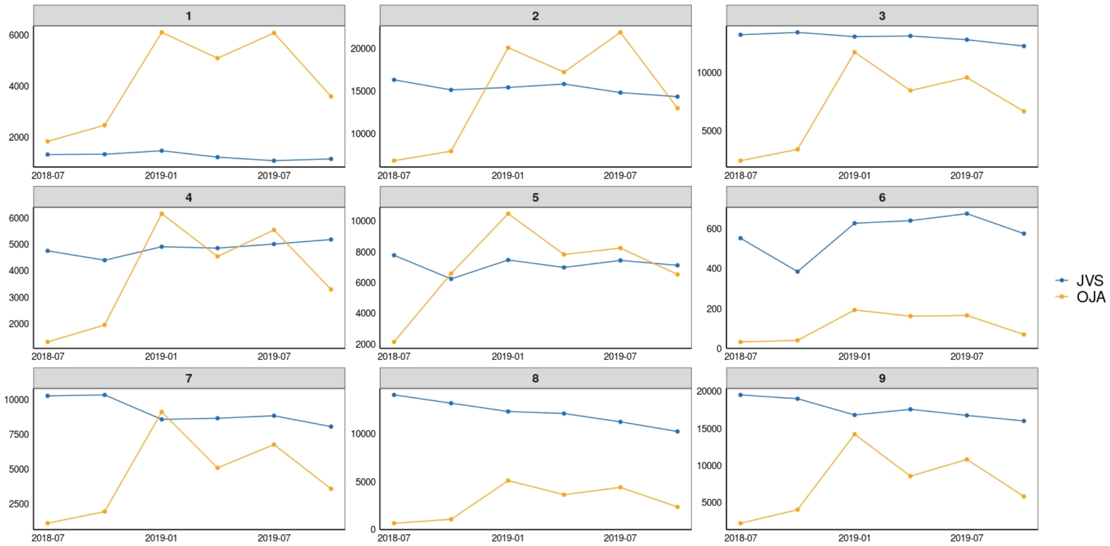

Figure 8: Number of job ads from OJA and vacancies from JVS for ISCO major groups of occupations in Hungary over time

Table 8: Descriptive statistics for OJA and JVS data for Hungary - ISCO major groups of occupation and time

	OJA	JVS
nbr.val	54	54
min	32	384
max	21847	19494
range	21815	19110
sum	309670	485825
median	5083	8749
mean	5735	8997
var	24140390	33419973
std.dev	4913	5781
coef.var	0.86	0.64
skewness	1.30	-0.06
kurtosis	1.67	-1.28

4.2.2 Cross-sectional regressions

Figure 9 shows basic features of the relation between OJA and JVS over time. Regressing JVS on OJA gives a very low, but statistically significant coefficient 0.041 (t-statistic=2.96). Such a low slope parameter shows incomparability of the fluctuations in the OJA dataset used for this study and that present in JVS. The residuals are obviously heteroscedastic. To account for this, we need to decompose the dataset into countries, NACE section (industry), and ISCO occupation. These are the only variables through which we can currently link OJA and JVS.

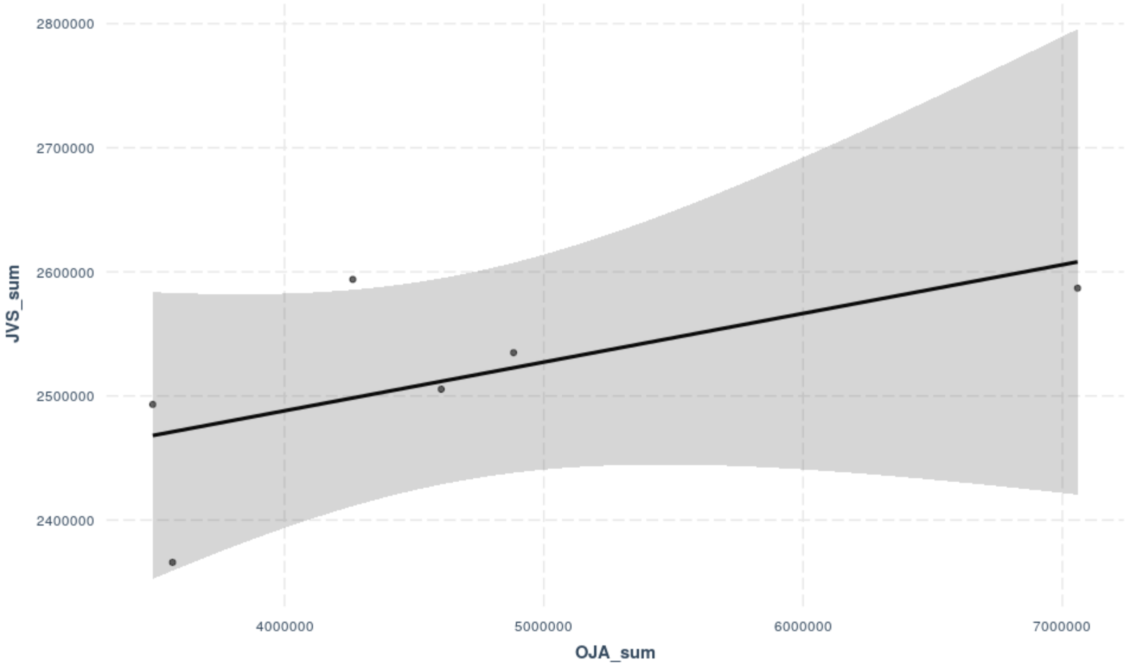

Figure 9: Regression fit between JVS and OJA

Note: Countries included – Bulgaria, Croatia, Czechia, Estonia, Germany, Hungary, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Sweden.

We can observe large differences in OJA-JVS relations between countries. Therefore we calculated separate regressions for each country (Figure 10). Table 9 shows additional statistics. Significant relations were found for four countries, namely Czechia, Latvia, Bulgaria, and Sweden. The results for these countries were satisfactory, with $0.66 < R^2 < 0.99$. In other countries the slope coefficients were not statistical significant. In six countries the estimated coefficient of slope was negative, even if non-significant.

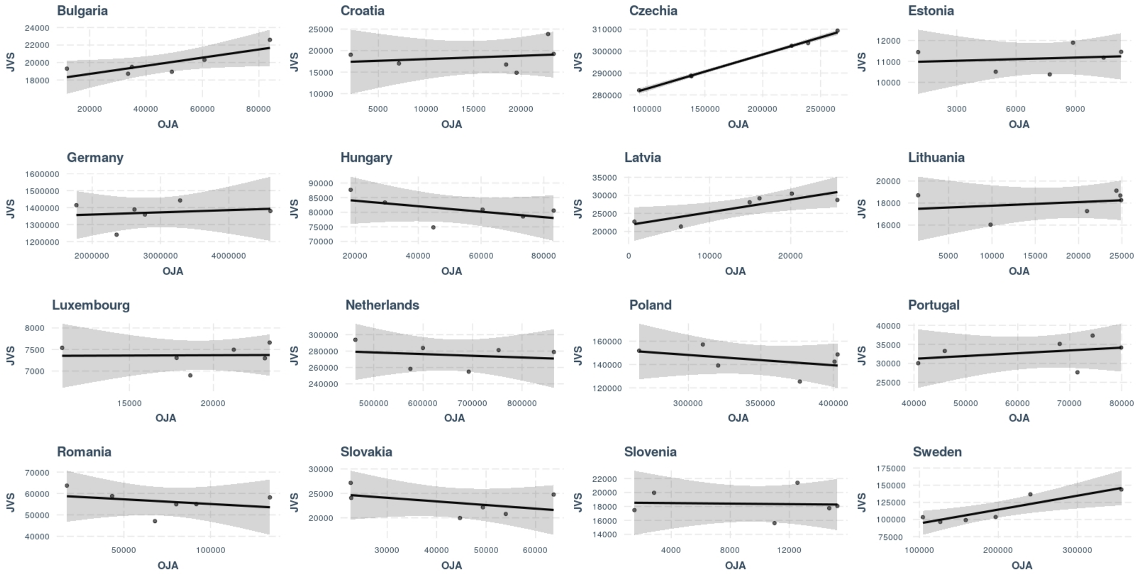

Figure 10: Regression results between OJA and JVS for individual countries

Table 9: Summary of regression results for individual countries

	Intercept	OJA	R^2	DW test
Bulgaria	17765*** (20.67)	0.047* (2.77)	0.66	1.1
Croatia	17238** (5.75)	0.079 (0.46)	0.05	1.5
Czechia	267277*** (325.47)	0.156*** (36.73)	0.99	3.3
Estonia	10954*** (17.56)	0.025 (0.33)	0.03	2.7
Germany	1330457*** (11.65)	0.015 (0.39)	0.04	2.2
Hungary	85817*** (21.07)	-0.094 (-1.3)	0.30	0.9*
Latvia	21783*** (12.54)	0.354** (3.32)	0.73	1.4
Lithuania	17433*** (15.53)	0.033 (0.58)	0.08	2.7
Luxembourg	7338*** (12.82)	0.001 (0.05)	0.00	1.6
Netherlands	288620** (8.16)	-0.02 (-0.39)	0.04	0.8
Poland	174627** (5.54)	-0.088 (-0.98)	0.19	2.2
Portugal	28243* (4.21)	0.074 (0.72)	0.11	1.2
Romania	59478*** (11.33)	-0.044 (-0.68)	0.10	1.0
Slovakia	26318** (8.16)	-0.074 (-1.04)	0.21	0.9*
Slovenia	18549*** (9.91)	-0.018 (-0.11)	0.00	1.9
Sweden	74157** (6.86)	0.202** (4.00)	0.80	2.6

Similarly to country breakdown, the differences between regression parameters calculated for individual NACE sections were large. Parameters representing OJA-JVS relation for all but one (O section is the exception, but the parameter is not statistically significant) are positive (Figure 11). However, only three out of 19 were statistically significant at p=0.05 (Table 10). Significant relations were found for sections A, N([7]) , and R. For these sections, the predictive power of OJA is promising. In these cases OJA explained 68%-76% of JVS variance. For other countries the slope coefficient was not statistically significant.

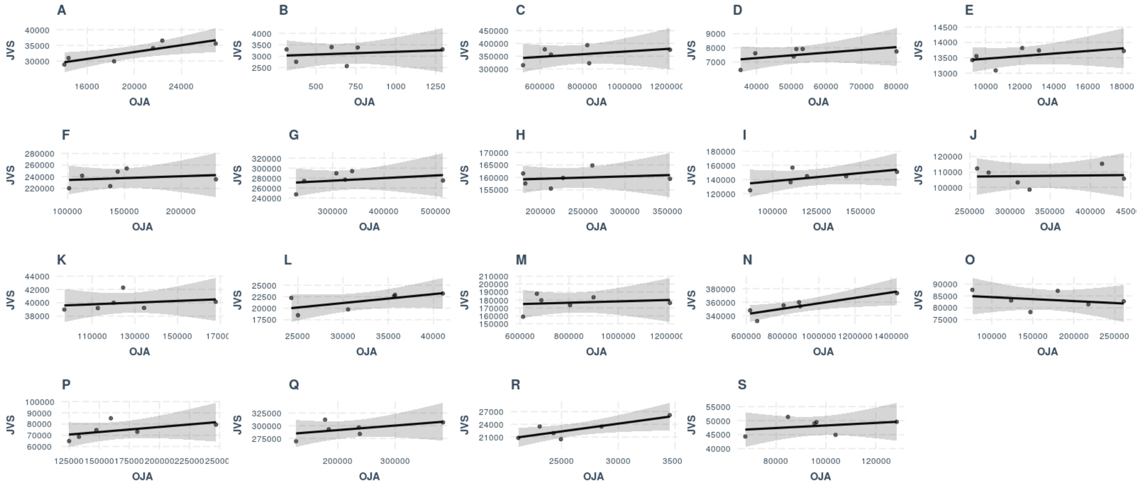

Figure 11: Regression results between OJA and JVS for individual economic activities

4.2.3 Panel data regressions

We apply a panel data regression for JVS as a function of OJA across countries. Such an approach gives us a holistic overview of the analysed relations. It shows us whether OJA data, as a whole, can be a good predictor of JVS data, providing that we account only for simple country differences. Starting with fixed effects model (the within transformation), we follow with testing the random effects (Table 11). The regressions are based on untransformed data. Alternatively, we used the logarithmic transformation of

[7] OJA data shows strong results for section N – Administrative and support service activities. One must take into account though that significant part of job ads within this section is published by sub-section of Employment activities by recruitment agencies that do it on behalf of employers from other sectors. In Skills OVATE, Cedefop treats this NACE sub-section as a separate sector.

Table 10: Summary of regression results for individual economic activities

	Intercept	OJA	R^2	DW test
A	21681** (6.87)	0.56* (3.57)	0.76	1.5
B	2966** (8.49)	0.24 (0.51)	0.06	0.7*
C	316778** (6.96)	0.05 (0.89)	0.16	0.8*
D	6502** (8.02)	0.019 (1.28)	0.29	0.9
E	13062*** (30.29)	0.041 (1.19)	0.26	1.5
F	226570*** (9.92)	0.071 (0.47)	0.05	0.6*
G	260122*** (10.38)	0.05 (0.67)	0.10	1.3
H	157442*** (22.72)	0.012 (0.41)	0.04	3
I	113380** (5.62)	0.239 (1.5)	0.36	2.1
J	105999** (7.56)	0.004 (0.1)	0.00	2.9
K	30606** (4.89)	0.067 (1.37)	0.32	2.9
L	15351* (4.29)	0.193 (1.76)	0.44	0.5
M	168996** (7.51)	0.012 (0.44)	0.05	2.7
N	317115*** (23.65)	0.042* (2.89)	0.68	1.7
O	86157*** (19.64)	-0.016 (-0.66)	0.10	1.3
P	59974** (5.22)	0.088 (1.3)	0.30	1.8
Q	276337*** (16.02)	0.081 (1.12)	0.24	2
R	13223* (4.11)	0.365* (3.00)	0.69	1.7
S	44514** (7.16)	0.038 (0.6)	0.08	2.6

the data. In this case the goodness-of-fit was even lower, with negligible slope coefficient. We proceed with untransformed data.

Table 11: Panel data regression between OJA and JVS - countries

	Within	Random	First differences
Intercept	-	130680*** (3.98)	1637 (0.51)
OJA	0.013 (1.35)	0.091*** (4.78)	0.004 (0.44)
Time dummies	YES	YES	NO
Country dummies	YES	YES	YES
R^2	0.02	0.20	0.002
DW test	2.7	1.1***	2.7
F / Chi^2 statistic	1.8	22.8***	0.2
Hausman	-	22.5***	-

Note: t-statistic in (). *** p<0.001, ** p<0.01, * p<0.05. Balanced Panel: n = 16, T = 6, N = 96.

Hausman test results show that differences between countries are consistent, and the fixed effects rather than random effects can be used to describe the relation between OJA and JVS. However, the relation between OJA and JVS is weak and statistically insignificant. It is far from one-to-one. As the model shows, 100 online job ads are associated with 1 vacancy. Time effects (Figure 12) show stable relation between OJA and JVS in the period 2018Q4-2019Q4. However, 2018Q3 significantly differs from the rest.

The relation between first differences of OJA and JVS might be potentially promising, given that online job ads might not represent the number of vacancies, but may be a good predictor of their changes. First differences transformation of the data also emphasizes the short-run relations between our variables. Despite this, within our limited time span the model shows statistical insignificance of such predictions.

We apply regression with increasing window of quarters (starting with two quarters, ending with 6 quarters) to check how the coefficients change when new information is added. Figure 13 presents the changes in the slope coefficient (marginal effect of OJA). For 4-5 quarters the relation between OJA and JVS is stable. Adding the sixth quarter decreases the coefficient approximately three times. Thus, we cannot say that online job ads is a stable predictor of job vacancies.

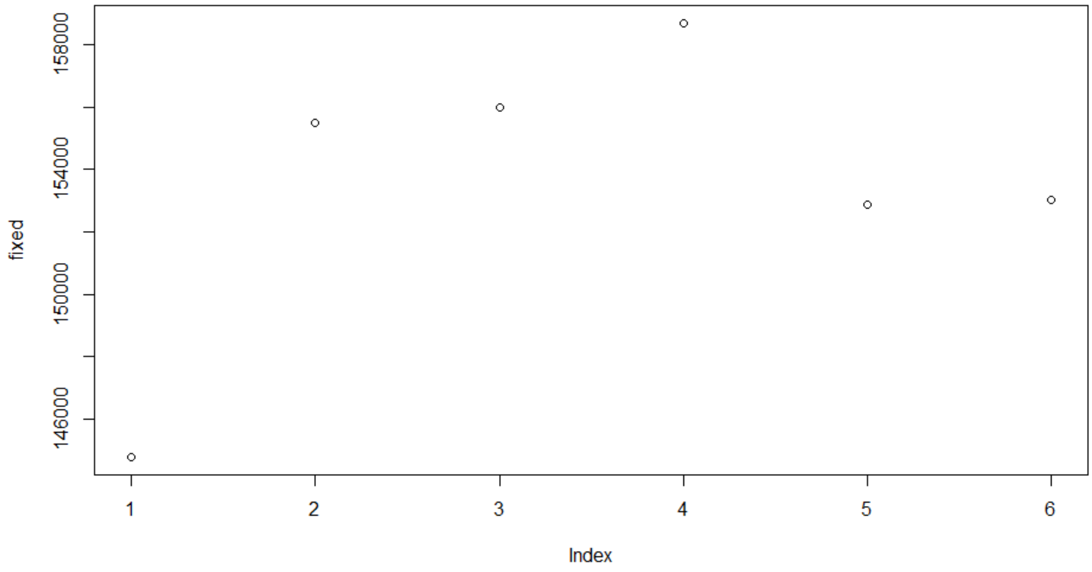

Figure 12: Time effects for the fixed effects regression between OJA and JVS

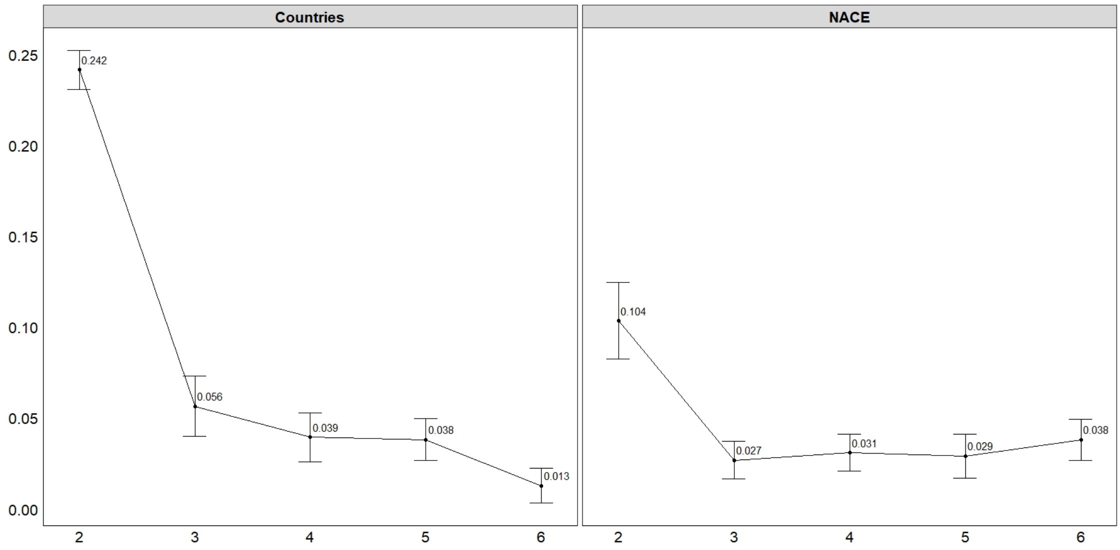

Figure 13: Fixed effects regression between OJA and JVS with increasing time span

Note: Results based on balanced panel data regression: n = 16, T = 6, N = 96. Regression contains individual effects, but not time effects. Horizontal axis represents number of quarters included in the regression, with 2 meaning 2018Q3-2018Q4 and 7 meaning 2018Q3-2019Q4.

Individual country effects, similarly to the slope coefficient, were not stable when we increased time span of the analysis (Figure 14). There are countries for which they remained quite stable after the initial adjustments. Those countries included mainly Estonia, Poland, and Sweden. For Latvia, the effects tended to a stable value, and consecutive increments were getting smaller. The most unstable values were estimated for Hungary, Romania, Slovakia, and Slovenia. For other countries fixed effects were moderately stable.

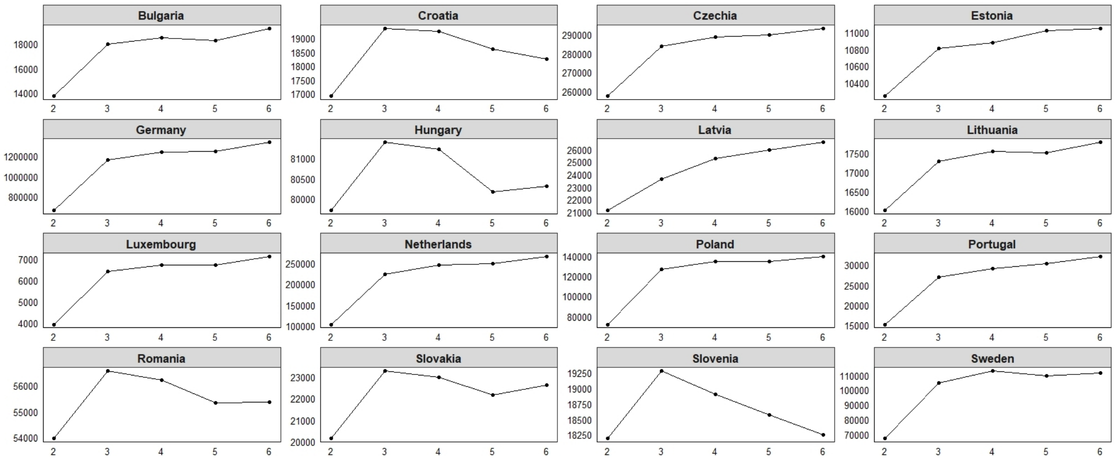

Figure 14: Individual country effects in fixed effects regression between OJA and JVS with increasing time span

Note: Results based on balanced panel data regression: n = 16, T = 6, N = 96. Regression contains individual effects, but not time effects. Horizontal axis represents number of quarters included in the regression, with 2 meaning 2018Q3-2018Q4 and 7 meaning 2018Q3-2019Q4.

Next, we look at the relations between OJA and JVS with panel data regression across industries. Starting with fixed effects model (the within transformation), we follow with testing the random effects (Table 12). Likewise the countries, the regressions are based on untransformed data, which give a better fit to the data.

Table 12: Panel data regression between OJA and JVS - NACE sections

	Within	Random	First differences
Intercept	-	118270*** (5.98)	142 (1.34)
OJA	0.038** (3.39)	0.045*** (4.11)	0.015* (2.22)
Time dummies	YES	YES	NO
Country dummies	YES	YES	YES
R^2	0.11	0.13	0.05
DW test	1.6*	1.3***	2.3
F / Chi^2 statistic	11.5**	19.2***	5.0*
Hausman	-	7.9**	-

Note: t-statistic in (). *** p<0.001, ** p<0.01, * p<0.05. Balanced Panel: n = 19, T = 7, N = 133.

The fixed effects estimator is consistent and better describes the relation between OJA and JVS than the random effects estimator. The relation is stronger than for countries and statistically significant. 100 online job ads are associated with 4 vacancies. The $R^2 = 11\%$. Time effects (Figure 15) for this model are similarly unstable as the ones in the model with countries.

The first differences model shows the short-run relation between OJA and JVS. For this reason, the OJA coefficient is even lower than for the rest of the models. Differencing eliminated autocorrelation of residuals, which was present in the previous models. The relation between OJA and JVS is statistically significant.

Figure 13 presents how the OJA slope coefficient changes if we increase the window of quarters taken to regression, also for the NACE section. As for the full-time span, the slope coefficient is higher than the one

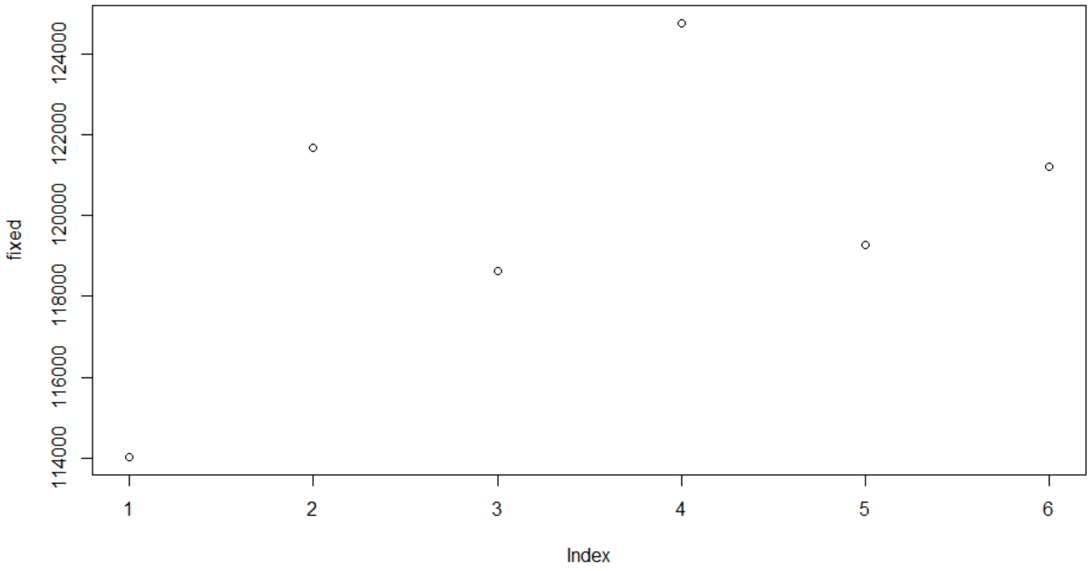

Figure 15: Time effects for the fixed effects regression between OJA and JVS

for countries. It is stable for 3-5 quarters analysis, but adding the last quarter increases the coefficient. Changes in this coefficient are smaller than the one in the country breakdown.

Generally, individual NACE effects show similar pattern if we increase time span for the regression (Figure 16). They tend to be stable after a few quarters. The exception is section C, for which the effect was the most unstable.

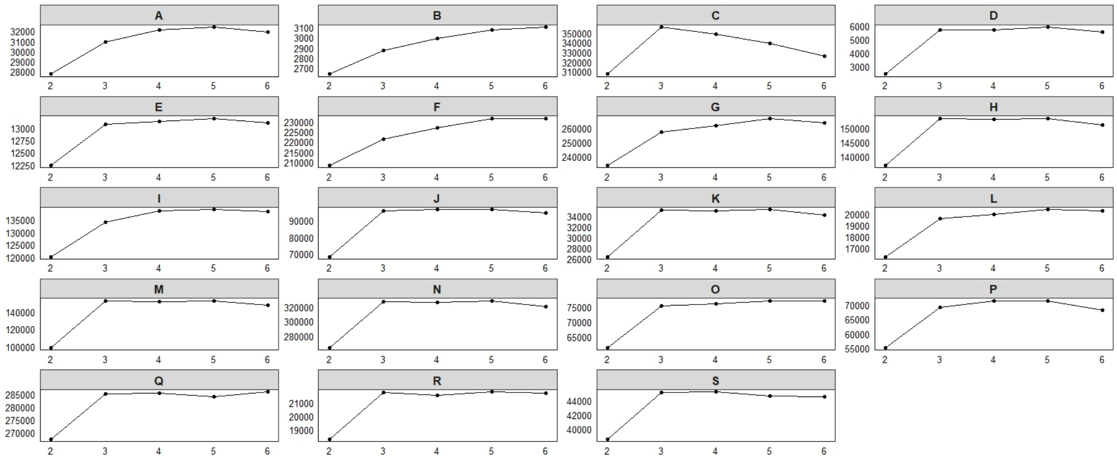

Figure 16: Individual industry effects in fixed effects regression between OJA and JVS with increasing time span

Note: Regression contains individual effects, but not time effects. Horizontal axis represents number of quarters included in the regression, with 2 meaning 2018Q2-2018Q3 and 7 meaning 2018Q2-2019Q4.

The relations between OJA and JVS, estimated with panel data regression do not show promising results (Table 13). The results are not statistically significant, showing no statistical relation between both measures of job vacancies. Hausman test results show that differences between ISCO groups are inconsistent. Despite this, the slope coefficient in random effects model, as well as in the first differences model is not significant.

Table 13: Panel data regression between OJA and JVS - ISCO major groups in Hungary

	Within	Random	First differences
Intercept	-	9426** (5.61)	-260** (2.71)
OJA	-0.083 (1.51)	-0.075 (1.64)	-0.042 (1.74)
Time dummies	YES	YES	NO
Country dummies	YES	YES	YES
R^2	0.05	0.05	0.07
DW test	1.6	1.4*	2.4
F / Chi^2 statistic	2.3	0.8	3.0
Hausman	-	0.07	-

Note: t-statistic in (). *** p<0.001, ** p<0.01, * p<0.05. Balanced Panel: n = 9, T = 6, N = 54.

5 Results

5.1 Methods and their assumptions

Due to the limitations found in the OJA data available to this study, discussed in previous chapter, we decided to use capture-recapture methods that are based on distributional assumptions. We especially focused on:

- naïve trimming, as a way for dealing with over-coverage, based on number of days from grab_date to the end of given quarter,
- zero-one-truncated capture-recapture (ZOT CR), which is the equivalent to zero-truncated one-inflated distributions as proved by Böhning and van der Heijden (2019). We used Poisson, Geometric and Negative-Binomial distribution.

We decided to use capture-recapture methods because these methods are suitable for estimating total number of job vacancies based on limited number of data sources. In order to prepare data for ZOT CR we took the following steps:

1. we treat the whole OJA database as **one data source**,
2. for each quarter we calculated the **number of job vacancies** (based on variable `general_id`) present in one, two, three etc. input data sources (variable `source`). Description of variables can be found in the Appendix in Table 16.
3. based on results from Table 5 we decided to discard the following countries due to low number of records in the OJA dataset: **Cyprus**, **Croatia**, **Estonia**, **Greece**, **Latvia**, **Lithuania**, **Malta**, and **Slovenia**.

Furthermore, due to the high differences in the number of job advertisements we decided to divide the OJA data into two groups in order to apply different adjustments:

1. Group 1 – Bulgaria, Czechia, Ireland, Portugal and Romania – countries with number of advertisements lower than 200 thousand (200k).
2. Group 2 – Belgium, Spain, Germany, the Netherlands and Poland – countries with number of advertisements higher than 200k.

One of the main challenges that occur in OJA data is the lack of knowledge about the over-coverage error due to false or outdated advertisements. Figure 17 presents the distribution of the number of days to the end of the quarter calculated using variable grab_date. Seasonal patterns resemble data collection process. There are also increases in the number of advertisements scraped at specific dates in 2019Q1, 2019Q3 and 2019Q4.

Certainly, some (unknown) over-coverage is present, as there are advertisements that had grab_date over 60 or even 100 days. It is reasonable to specify some cut-off threshold to remove outdated advertisements from the study population.

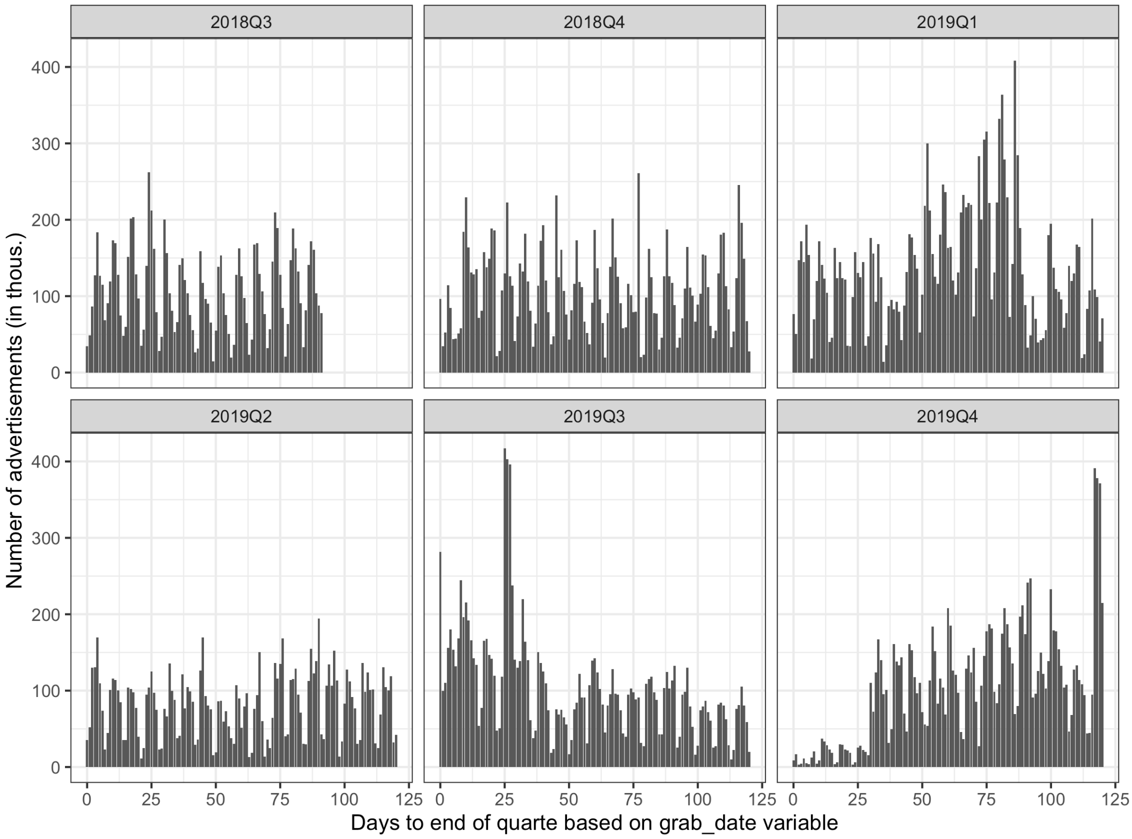

Figure 17: Distribution of number of advertisements by quarters and days to the end of given quarter based on grab_date variable

Based in the information on the number of days we decided to apply different trimming for the two groups of countries (under 200k and over 200k) defined above.

5.2 Naïve trimming

The first method that we applied is naïve trimming, meaning that we just specified a cut-off threshold for number of days from the grab_date to the end of given quarter. For the group of countries with larger number of job advertisements we used 7 to 30 days (by one day) and for the other group of countries we used the range of 30 to 60 by 5 days.

Figure 18 presents the result of the trimming procedure and its impact on the number of job advertisements in the OJA data. The red line indicates the number of vacancies according to job vacancy statistics (seasonally unadjusted), the solid black line shows the number of job vacancies in the OJA data without any trimming and the gray lines show the transformed data for different alpha-transparency scale.

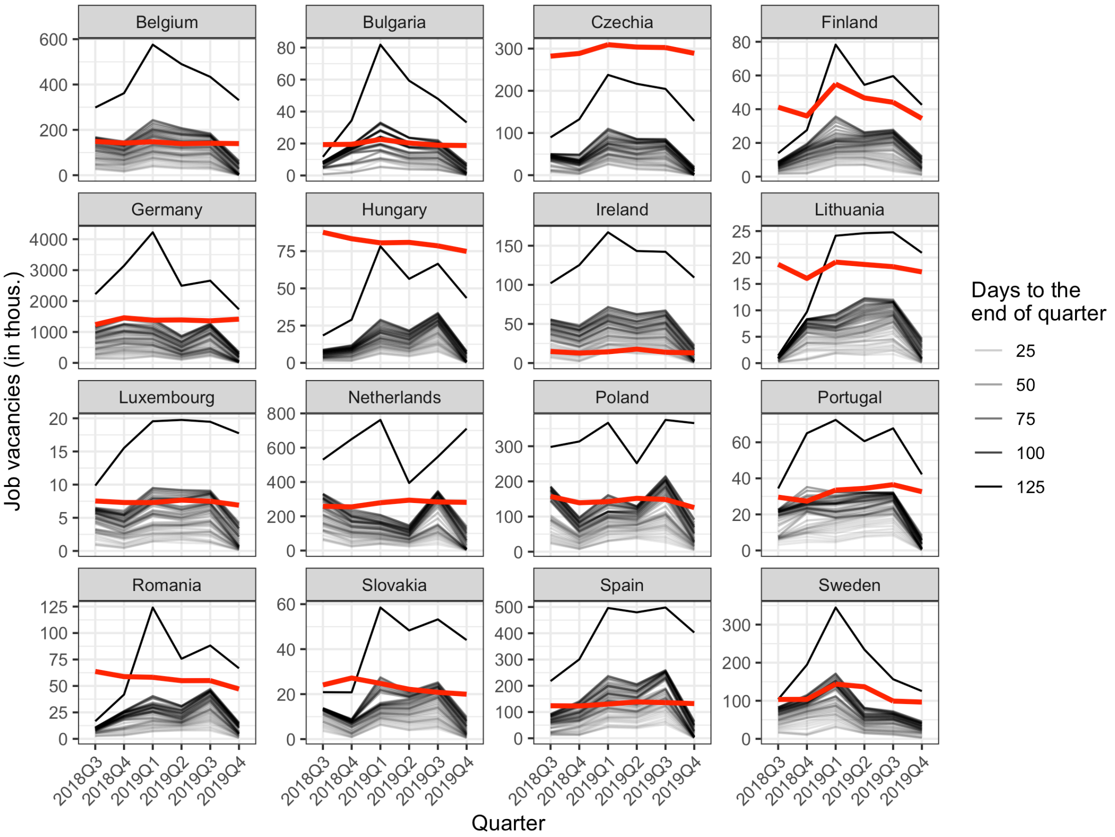

Figure 18: Comparison of estimates based on JVS (solid red line; seasonally unadjusted data), OJA data without trimming (solid black line) and trimmed data (solid gray lines)

The main conclusions from this plot are: 1) for Hungary and Czechia the number of vacancies according to the OJA data are underestimated. In Czechia, JVS are compiled based on administrative data registers of the Labour Office of Czechia and in Hungary they are based on the Quarterly Establishment Survey; 2) Ireland is a special case where JVS is significantly smaller than in other countries; 3) JVS are less variable in comparison to OJA data. A lack of trend is visible in particular for Belgium, Germany or Netherlands; 4) Trimming makes the OJA data less variable but we cannot say anything about seasonal aspect as the time period is short; 5) in most cases trimming makes the data more similar to JVS but there is no clear pattern and 6) results of trimming are connected to the number of job vacancies collected in a given quarter.

The last point from the above discussion is presented in Figure 19 where we calculated absolute relative difference between trimmed OJA data and estimates from job vacancy statistics. In particular, the figure presents the following measure

$$\delta^{t}_{c,q} = \frac{\mid \theta^{\text{OJA},t}_{c,q} - \theta^{\text{JVS}}_{c,q} \mid}{\theta^{\text{JVS}}_{c,q}} \times 100\%, \tag{51}$$

where c represents the country, q represents the quarter, t represents the trimming threshold (i.e. 10 days to the end of given quarter), $\theta^{\text{OJA},t}_{c,q}$ is the number of job vacancies for a given country, in a given quarter, for given trimming threshold based on OJA data, and finally $\theta^{\text{JVS}}_{c,q}$ is an estimate based on JVS.

Figure 19 presents $\delta^{t}_{c,q}$ for each quarter separately to verify if the same threshold may be applied for all quarters. In this plot we set the threshold for grab_date to be between 7 and 50 days before. Countries significantly vary between countries and quarters. There are no clear patterns as regards to which threshold should be applied. For instance, Spain and Poland have different threshold for each quarter, while for some quarters in Germany, Luxembourg, Lithuania, or Finland the same threshold gives similar results.

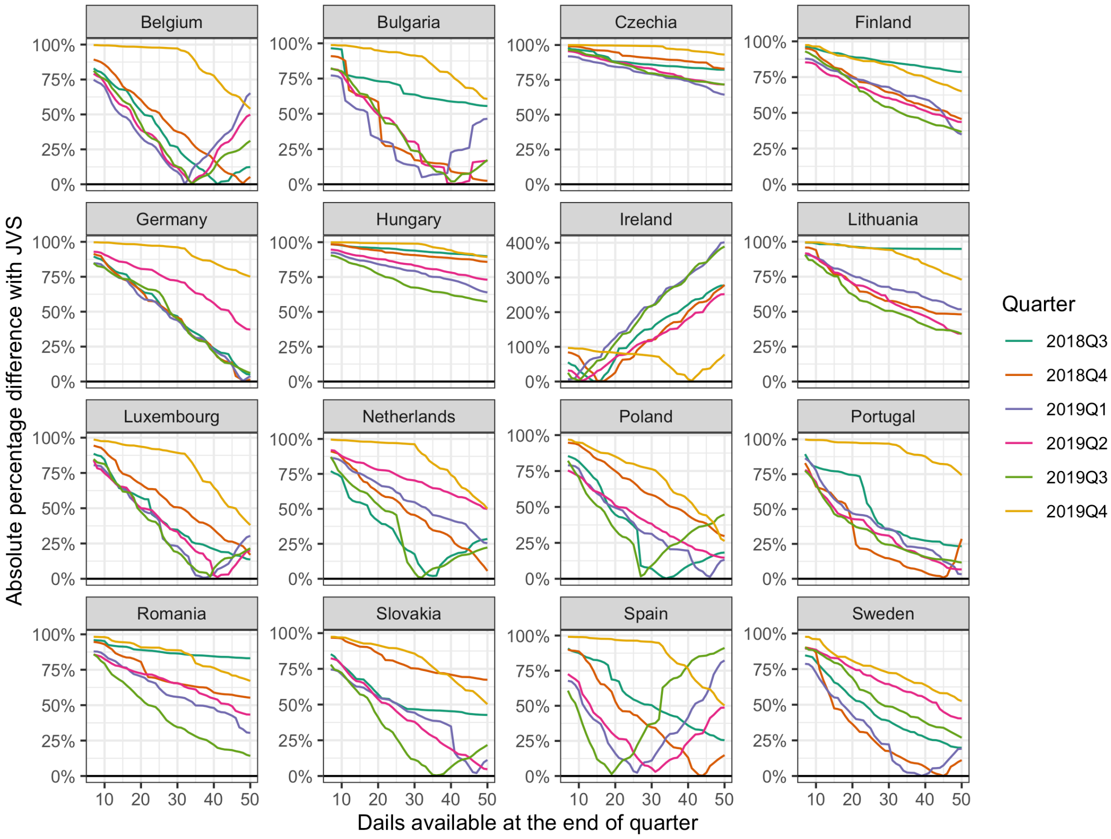

Figure 19: Comparison of absolute difference of estimates between OJA trimmed estimates with JVS for each quarter

To sum up the findings regarding naïve trimming:

- countries significantly vary in patterns of OJA vacancies and JVS,
- it is difficult to set one threshold for all countries,
- setting the same threshold for each quarter in one country is also inconclusive.

5.3 Zero-one truncated capture-recapture results

5.3.1 Estimation procedure

This section covers the estimation of the size of the job vacancies population in selected countries based on zero-truncated one-inflated (ZTOI) or zero-one truncated (ZOT) capture-recapture approach. In this chapter, we followed Böhning and van der Heijden (2019) findings regarding equivalence of ZTOI and ZOT distributions. Under this assumption a Horvitz-Thompson estimator for the target population with no extra-singletons (i.e. one-inflation) is given by the equation (52)

$$\hat{N}_{\text{nes}} = \frac{n_1}{1 - p\left(x_0, \theta\right) - p\left(x_1, \theta\right)}, \tag{52}$$

where $p(x_0, \theta), p(x_1, \theta)$ are probability of occurrence of 0 and 1 respectively. $\hat{N}_{\text{nes}}$ is an unbiased estimator of the population size with no extra singletons. Hence, Böhning and van der Heijden (2019) construct an estimator of the hidden units f_0 as

$$\hat{f}_0 = p\left(x_0, \theta\right) \frac{n_1}{1 - p\left(x_0, \theta\right) - p\left(x_1, \theta\right)}. \tag{53}$$

Finally, to achieve the Horvitz-Thompson estimator of the target population of interest, i.e. under ZTOI/ZOT distribution, as

$$\hat{N} = \hat{f}_0 + f_1 + n_1 = p\left(x_0, \theta\right) \frac{n_1}{1 - p\left(x_0, \theta\right) - p\left(x_1, \theta\right)} + f_1 + n_1, \tag{54}$$

where $\hat{f}_0$ is the estimated hidden population, f_1 is the number of units observed one time and n_1 is the rest. $\hat{N}$ is unbiased estimator of the population size and as θ is unknown and need to be estimated; we can replace θ with maximum likelihood estimator $\hat{\theta}$ under $p_{+1}(x, \theta)$ that leads to:

$$\hat{N} = \hat{f}_0 + f_1 + n_1 = p\left(x_0, \hat{\theta}\right) \frac{n_1}{1 - p\left(x_0, \hat{\theta}\right) - p\left(x_1, \hat{\theta}\right)} + f_1 + n_1. \tag{55}$$

To estimate θ using maximum likelihood method we define likelihood function for zero-one truncated distribution as

$$L_{++} = \prod_{i=2}^{m} p_{++}\left(x_i, \theta\right)^{f_i}, \tag{56}$$

where p_{++} is defined as

$$p_{++}(x, \theta) = p_{+}(x, \theta) / \left[1 - p_{+}\left(x_1, \theta\right)\right] = p(x, \theta) / \left[1 - p\left(x_0, \theta\right) - p\left(x_1, \theta\right)\right], \tag{57}$$

and f_i denotes frequency counts for i-th number. In practice, instead of using the likelihood function defined in (56) we used the log-likelihood given by

$$\log L_{++} = \sum_{i=2}^{m} f_i \log p_{++}\left(x_i, \theta\right), \tag{58}$$

where notation is the same as previously.

In order to provide a measure of uncertainty we use non-parametric bootstrap to estimate standard errors and follow recommendation provided by Böhning and van der Heijden (2019). The procedure to estimate estimated variance given by

- Draw a sample of size $||\hat{N}||$ from the observed distribution defined by the probabilities $\frac{\hat{f}_0}{N}, \frac{\hat{f}_1}{N}, \frac{\hat{f}_2}{N}, ..., \frac{\hat{f}_m}{N}$.
- Derive $\hat{\theta}$ and $\hat{N}$ for the boostrap sample in 1).
- Repeat step 1) and 2) B times, leading to a sample of estimates $N^{(1)}, ..., N^{(B)}$.
- Calculate the bootstrap standard errors as

  $$SE^* = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(N^{(b)} - \bar{N}^*\right)^2}, \tag{59}$$

  where $\bar{N}^* = \frac{1}{B}\sum_{b=1}^{B} N^{(b)}$ and ,
- calculate relative standard error (i.e. coefficient of variation; CV) by

  $$CV^* = \frac{SE^*}{\bar{N}^*} \times 100\%. \tag{60}$$

5.3.2 Application

We calculated multiple models, in particular:

- We assumed Poisson

  $$p(x;\theta) = p(x;\lambda) = \text{Pr}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{61}$$

  Geometric

  $$p(x;\theta) = p(x;p) = \text{Pr}(X = k) = (1-p)^{k-1}p. \tag{62}$$

- We calculated models separately for each quarter (6 quarters),
- We calculated models separately for two groups of countries:
  - Under 200k vacancies in quarter – Czechia, Ireland, Bulgaria, Portugal, and Romania (5 countries) with trimming thresholds (30, 35, 40, 45, 50, 55 and 60 days; 7 thresholds),
  - Over 200k vacancies in quarter – Belgium, Spain, Germany, the Netherlands, and Poland (5 countries) with trimming thresholds (from 7 to 30 days; 24 thresholds).

This gave $2 \times 6 \times (5 \times 7 + 5 \times 24) = 1860$ models.

Unfortunately, for the following countries the optimization procedures were not able to find a solution: Bulgaria, Slovakia, Lithuania, Hungary, Sweden, Luxembourg, and Finland. The main reason is not enough vacancies observed two or more times.

Within each group we selected the most appropriate distribution (Poisson, Geometric or Negative-Binomial) based on AIC criterion and estimated the size of the population. The AIC and other characteristics of the estimated models are reported in Appendix 7. In this exercise, we used trimming for accounting for over-coverage and as a sensitivity analysis to different thresholds.

5.3.3 Countries with under 200k advertisements

Figure 20 presents the comparison of estimates based on JVS (red dashed line), OJA data without any trimming (blue dashed-line) and results from the capture-recapture procedure using zero-one truncated Poisson distribution. On X axis we have the number of days to the end of given quarter (e.g. 30 days means that we used vacancies that are max 30 days old). Each panel is defined by country and quarter.

Based on the results we see that in general, untrimmed OJA data for some countries (ie.g. Bulgaria, Portugal, Romania) provide similar results. For Czechia, the difference between JVS and OJA data decreases as time goes on.

If we look at the results of the estimation procedure, we can see that the most problematic quarters are the first (2018Q3) and the last (2019Q3). We see variability in estimates as trimming threshold increases. For Ireland and Portugal trimming made the estimates closer to the JVS (for selected quarters). In general, there is no clear pattern of estimates and we see that results are sensitive to different trimming.

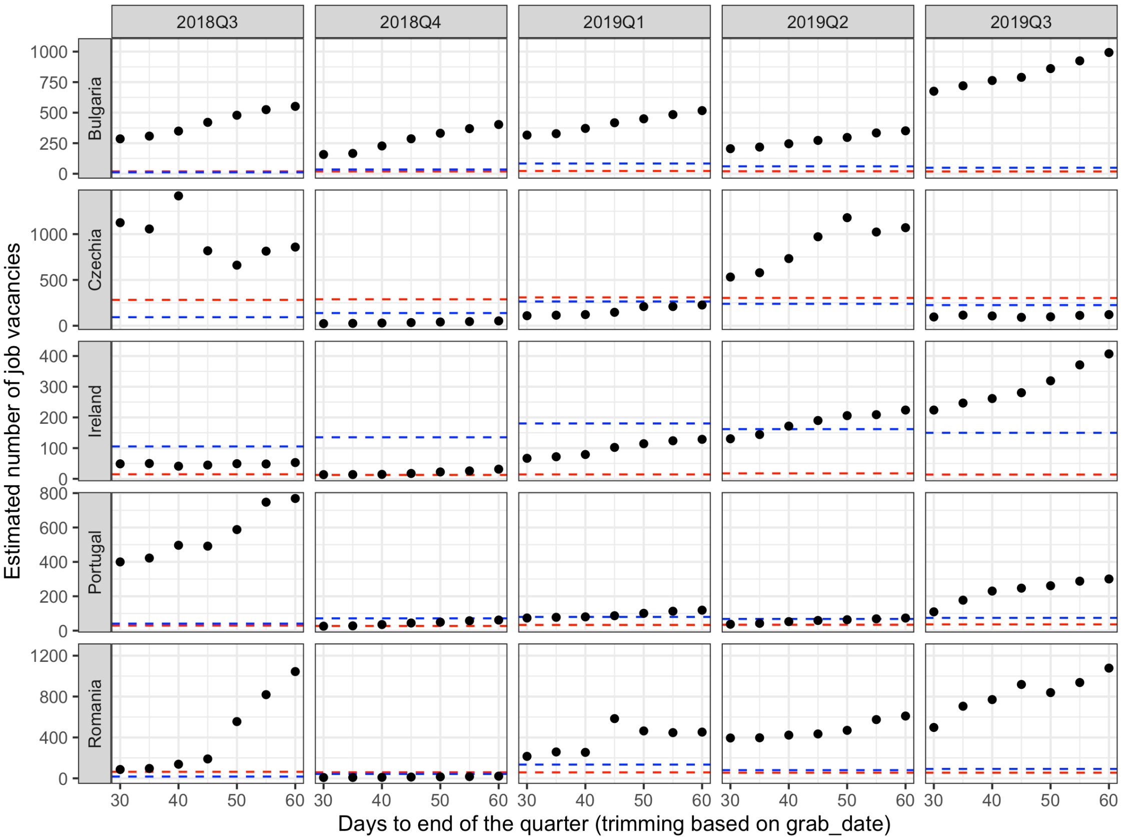

Figure 20: Comparison of estimates based on JVS (red dashed line), OJA data without trimming (blue dashed line) and zero-one-truncated capture-recapture size estimator under different trimming dates (30 to 50 by 5)

To compare numerically, we report results for threshold of 40 days in Table 14. The table contains three variables – JVS, OJA data without trimming and the proposed population size based on zero-one truncated (ZOT) Poisson distribution. This estimate is based on the mean from the bootstrap procedure.

Table 14: **Comparison of estimates (in thous.) based on JVS, OJA data without trimming and zero-one-truncated Poisson capture-recapture size estimator under different trimming to 40 days**

country	quarter	JVS	OJA	Trimmed	ZOT Poisson	CV
Czechia	2018Q3	282.1	93.1	45.7	1417.6	92.9
	2018Q4	288.5	138.4	33.9	29.6	10.7
	2019Q1	309.3	264.8	79.9	121.8	14.3
	2019Q2	303.6	239.2	74.3	733.1	32.5
	2019Q3	302.4	225.0	74.6	106.8	44.3
Ireland	2018Q3	14.8	105.5	47.6	41.2	17.0
	2018Q4	12.8	135.3	38.7	14.5	0.9
	2019Q1	14.3	180.3	58.0	79.2	3.9
	2019Q2	17.8	161.8	50.1	171.7	10.4
	2019Q3	13.8	149.9	55.9	261.5	10.4
Bulgaria	2018Q3	19.3	11.9	8.0	349.3	3.4
	2018Q4	19.5	35.0	17.7	227.4	0.4
	2019Q1	22.6	83.9	27.7	371.2	1.5
	2019Q2	20.3	60.7	20.2	245.7	4.5
	2019Q3	18.9	49.1	18.5	763.3	7.0
Portugal	2018Q3	29.6	40.8	21.5	496.7	11.5
	2018Q4	27.4	71.5	26.1	35.6	0.7
	2019Q1	33.5	79.9	26.2	80.3	3.1
	2019Q2	34.4	68.1	29.2	52.5	2.1
	2019Q3	36.5	74.4	30.8	230.2	7.2
Romania	2018Q3	63.7	17.1	9.9	137.8	128.4
	2018Q4	58.8	43.2	24.2	9.6	0.2
	2019Q1	58.2	134.3	30.2	253.9	27.9
	2019Q2	55.0	80.2	24.9	422.7	27.4
	2019Q3	55.0	91.9	42.9	770.2	22.5

5.3.4 Countries over 200k advertisements

Figure 21 presents the comparison of estimates based on JVS (red dashed line), OJA data without any trimming (blue dashed-line) and results from the capture-recapture procedure using zero-one truncated Poisson distribution for countries with high number of vacancies. On the X axis we have the number of days to the end of a given quarter (e.g. 10 days means that we used vacancies that are max 10 days old). Each panel is defined by country and quarter.

Contrary to the previous results for the group of below 200k, the countries in this group are more stable. Exceptions are the first quarter (2018Q3) and the last quarter (2019Q3). We also note that for Belgium and Poland, the number of job vacancies are about two times higher than JVS. For other countries, over-coverage error is significant as the number of reported vacancies based on OJA data is three or more times higher. We note that in comparison to countries with a lower number of advertisements, trimming and usage of ZOT Poisson capture-recapture, is more stable and closer to estimated JVS. This pattern is clearly visible for Germany, where trimming and ZOT Poisson significantly decreased the number of vacancies initially estimated from OJA data.

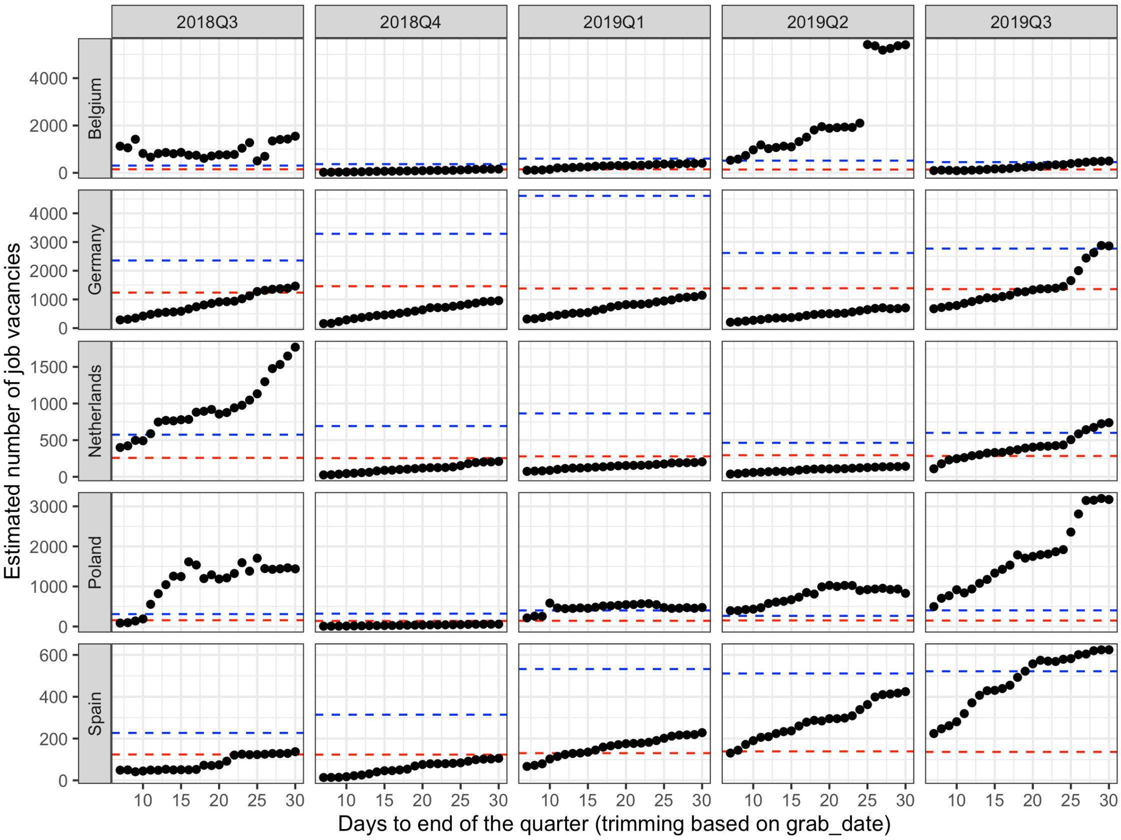

Figure 21: Comparison of estimates based on JVS (red dashed line), OJA data without trimming (blue dashed line) and zero-one-truncated capture-recapture size estimator under different trimming dates (7 to 30)

To compare numerically, we reported results for threshold of 40 days in Table 15. The table contains three variables – JVS, OJA data without trimming and the proposed population size based on zero-one truncated (ZOT) Poisson distribution. This estimate is based on the mean from the bootstrap procedure.

Table 15: Comparison of estimates (in thous.) based on JVS, OJA data without trimming and zero-one-truncated Poisson capture-recapture size estimator under different trimming to 20 days

country	quarter	JVS	OJA	Trimmed	Proposed	CV
Belgium	2018Q3	149.2	305.8	70.9	678.9	42.4
	2018Q4	141.7	367.8	56.2	89.7	10.2
	2019Q1	147.7	599.4	97.8	304.0	9.5
	2019Q2	139.0	515.1	85.1	1827.9	19.7
	2019Q3	140.9	452.8	80.3	248.7	22.9
Spain	2018Q3	123.8	226.8	39.2	73.3	10.1
	2018Q4	123.1	314.1	57.7	75.3	4.0
	2019Q1	130.3	532.4	109.3	175.0	2.7
	2019Q2	138.5	511.2	99.2	293.9	6.2
	2019Q3	135.9	521.7	142.0	554.2	8.4
Germany	2018Q3	1237.4	2356.3	427.5	911.2	1.9
	2018Q4	1458.4	3286.3	513.7	637.0	0.5
	2019Q1	1380.3	4607.9	546.0	817.9	1.1
	2019Q2	1389.2	2618.4	269.2	500.6	2.9
	2019Q3	1359.4	2771.0	428.1	1330.4	5.4
Netherlands	2018Q3	258.3	574.1	151.5	856.1	6.3
	2018Q4	254.9	692.3	97.1	118.9	1.6
	2019Q1	278.9	864.5	89.3	153.3	1.8
	2019Q2	294.0	463.0	65.0	108.0	2.1
	2019Q3	283.7	599.6	133.6	404.3	4.7
Poland	2018Q3	157.2	310.0	87.0	1084.0	36.0
	2018Q4	139.2	320.5	34.8	37.9	0.1
	2019Q1	142.5	400.9	72.0	535.0	16.2
	2019Q2	151.8	266.0	74.5	1003.5	18.8
	2019Q3	148.6	403.0	98.1	1715.0	16.8

5.4 Conclusions

Based on the results we conclude that:

- Results are inconclusive as one method provides different results for different study countries,
- With unknown quality of OJA data it is not possible to derive unbiased estimates of total number of job vacancies,
- Audit sample is needed to assess OJA data,
- As the OJA data used in this study significantly differ between countries, application of the same method leads to different results,
- Results for bigger countries measured by number of vacancies tend to provide less variable estimates, thus application of these methods may be more suitable for such countries,
- Sensitivity analysis of the trimming procedure suggests that there is additional variability due to over-coverage in OJA data,
- Over-coverage is still a serious problem in OJA data, but it is not possible to verify to what extent from the current study,
- We consider the OJA data as available from Cedefop at the time of the study more appropriate for analysing skills rather than estimating the number of job vacancies.

The OJA dataset available from Cedefop that was the subject of the analysis certainly is an interesting source of information but at the time of the study has a limited usability for estimating the total number of vacancies. The main limitation is the unknown (statistical) quality of the OJA data. We suggest re-designing the study that will cover certain disadvantages of the current approach. Another option, but in our opinion only temporary, is to take a subsample of Cedefop's data that will be suitable to predict job vacancies. More detailed conclusions are in the next chapter.

6 Findings and recommendations

6.1 Main findings

6.1.1 Correlation between OJA and JVS

Analysing the relation between the Online Job Advertisements (OJA) dataset available from Cedefop at the time of this study and Job Vacancies Statistics (JVS) published by Eurostat, we reached the following findings:

- The OJA dataset identified 40%-91% job offers more than JVS did in the analysed period.
- The OJA dataset show a lot more variation than JVS.
- The distributions found in the OJA dataset are more leptokurtic and right skewed than analogous distributions in JVS.
- The relation between both measures of vacancies across time and countries was not statistically significant, whilst across time and industries it was found to be statistically significant. The latter relation was relatively stable over time.
- The relations between OJA and JVS differ significantly between countries, industries, and occupations.
- Satisfactory predictions of JVS based on OJA were made for four countries: Czechia, Latvia, Bulgaria, and Sweden, and for three NACE sections: A, N, and R.
- For ISCO major occupational groups, but only based on one country (Hungary), we found large differences between OJA and JVS.

6.1.2 Population size estimation

Based on the estimation analysis we conclude that:

- Results are inconclusive and further studies are needed.
- Unknown quality of the OJA dataset severely limited the possibility to derive unbiased estimates of total number of job vacancies.
- Audit sample is needed to assess the OJA data.
- As the OJA data significantly differs between countries, the application of the same method to all the countries may lead to different results.
- Results for bigger countries, measured by number of vacancies, tend to provide less variable estimates, thus application of these methods may be more suitable for such countries.

- Sensitivity analysis of trimming suggest that there is additional variability due to over-coverage in the OJA data.
- Over-coverage is still a serious problem in the OJA data, but it is not possible to verify to what extend from the current study.

6.2 Main limitations

Based on our analysis we conclude that in its current form the OJA data available from Cedefop is not suitable for deriving the total number of job vacancies. At this stage, it should be rather considered as a source for statistics on skills and further improvements need to be made in order to attempt to make estimations for the total number of job vacancies.

We identified the following limitations of using the available OJA data from Cedefop for the estimation of the total number of job vacancies:

1. Date of collection and date of expiration – they are based on web-scraping not actual dates from the advertisements;
2. Overcoverage – unknown outdated or erroneous ads;
3. Duplicates – there still may be duplicates despite the procedures applied;
4. Overlap – in most cases overlap between data sources is very small;
5. Variability in data collection – data are collected from various websites and with various frequencies. There are high discrepancies between quarters;

We also identified limitations regarding Job Vacancies Statistics conducted by Member-States and the availability of data at Eurostat's website:

1. Member-States differ in the methodology of calculating job vacancy statistics – they refer to different populations, use different sources and different estimation methods;
2. Data available at Eurostat's website is not complete – some countries report all NACE sections, some only totals, or specific sub-populations. Only Hungary reports information about occupations;
3. Information regarding the quality of JVS estimates is outdated (e.g. 2017) and limited to specific subpopulations – we cannot use these data to properly compare JVS with OJA data;

Further, we faced considerable limitations during the study.

Time series dimension. Having the data for 2018.7-2019.12 (not a complete period of two years) severely limits the application of time series models and analysis. Such data does not allow for a proper seasonality analysis. Without this, we cannot properly extract and compare other time series components of JVS and OJA, and most importantly trends and cycles. At least three years are required to perform such an analysis. Seasonality patterns in online data might differ from the ones in survey data, which is what should be analysed when having longer time series. Those differences might be one of this data specificities and should not lower the usefulness of online data. Longer time series will enable the analysis of OJA as high frequency data in comparison to quarterly JVS.

NACE section dimension. We have serious doubts about the accuracy of the NACE variable extracted from the job advertisements and available in the OJA dataset. Moreover, this is the only variable reported in Eurostat's JVS and due to low quality in Cedefop, significant differences are encountered. According to Cedefop's documentation (Cedefop, 2019c) Annex I: Occupations, skills, and other variables classification), NACE section is classified either based on a structured field or searched for in a workplace description. The latter suggests that the NACE section is not based on the NACE of a company that published a job ad (as it is in official statistics, e.g. JVS), but rather based on a workplace that might potentially be various. For example, an IT company may want to hire an accountant. This leads to a different treatment of the NACE section by Cedefop than by Eurostat.

ISCO occupation dimension. While ISCO occupation is present in the OJA dataset, it is not reported by almost any of the EU countries in Eurostat's JVS statistics. This also limits the use of Cedefop's data on skills. Skills data is potentially a large value added of Cedefop's OJA data in comparison to official statistics. However, before using data on skills, it would be important to compare the occupational distributions of OJA and JVS to analyse representation of occupations in online data. The next step would be to analyse demand for skills.

Longitudinal data. There are no job offers with the same `general_id` between different `grab_date` and `expire_date` periods. This limits tracking of the same advertisement in time and performance of longitudinal analysis (see Summary regarding the data section).

6.3 Recommendations

6.3.1 For the OJA dataset

- Include additional table with links between advertisements and source websites;
- Include information about the quality of the automatic classifiers used to predict the statistical variables from the content of the job advertisement;

6.3.2 Data sources

We recommend focusing on data sources from Public Employment Offices (PEO) which may be of higher quality than commercial websites or job ads aggregators. For instance, in the Polish Ministry of the Ministry of Family, Labour and Social Policy that supervises PEOs provides online service – the Central Job Advertisements Database – with all job ads that are registered in PEOs. Each job ad is verified by a clerk at local PEOs, classified to occupation and all information about the employer is provided. Moreover, outdated, erroneous or if employee(s) was/were found, ads are removed by PEOs which minimize potential over-coverage error. These data may be further used as a training data for advertisements from commercial websites both in terms of occupation as in terms of delays in reporting.

Cedefop uses machine learning algorithms to collect, deduplicate, classify, and present job advertisements. However, to obtain measures of estimation errors at the stage of OJA-JVS inference, measures of accuracy of Cedefop's algorithms are needed. Another problem is that we do not know the share of possible obsolete job advertisements, that is job advertisements for job positions that have been already filled. Companies might not inform timely the job board about filling the job vacancy. Since most websites introduce a job advertisement publication period, after which this job advertisement is automatically removed from the website (most often 30 days), this bias will last for this period.

6.3.3 Audit sample

Furthermore, to assess the quality of the OJA data, an audit sample should be considered. This approach is recommended when one is interested in estimating coverage or classification error. Examples regarding this approach can be found in:

- *the prevalence of cybercrime in the Netherlands* – See The accuracy of estimators based on a binary classifier,
- *correcting misclassification in enterprises* – Delden et al. (2016),

- *correcting labour force status* – Zhang (2005),
- coverage error of register – Zhang (2015)

Having that in mind, the audit sample may be designed for:

1. verification of over-coverage error due to false, outdated or out-of-scope units;
2. verification of over-coverage error due to insufficient deduplication;
3. verification of classification error due to automatic classification error;

We recommend preparing two independent audit samples to tackle both over-coverage errors separately (out-of-scope units and duplicated records) and the process should be conducted separately on each country. Moreover, we suggest that the samples may be drawn from the final OJA dataset. In particular, we propose the following procedure:

- Let the final OJA dataset be the reference data used to draw a sample. The dataset of size n_t may contain deduplicated job advertisements that (according to the OJA data) are open at day t,
- Draw two independent audit samples – for over-coverage (s_1) and for deduplication (s_2) according to the following schema:
    1. sample s_1
        - simple random sampling or Poisson sampling proportional to number of days between placing advertisement and time t,
        - size of the sample should be manageable for team of people who conduct clerical review – this requires contacting the company or recruiting agency to verify if the vacancy is in fact open. For instance, one may consider 1% sample, but it depends on the available funds,
        - based on the result of clerical review, estimate the error due to over-coverage that might be followed by modelling procedure i.e. binary classification procedure where 1 = out-of-scope and 0 = open vacancy,
    2. sample s_2
        - we recommend applying some probabilistic procedure to duplicate the records. Currently, Cedefop duplicates rows based on comparison of specific fields,
        - compute distances / probabilities of duplication between pairs
        - sample pairs that have the lowest distance / highest probability and have the highest distance / lowest probability,
        - verify which pairs are in fact duplicates,
        - conduct supervised record linkage to deduplicate records.
    3. Combine results from sample s_1 and s_2 to verify overlap between out-of-scope units and duplicated records. Based on that result, estimate the over-coverage error, and use it to downsize n_t records.

This approach requires that additionally to web-scraping, a survey is conducted in parallel. However, this may exceed the goal of this study, that is to estimate the number of vacancies solely based on the existing online data. Nevertheless, it would increase the quality of obtained data.

6.3.4 Additional scraping

We also recommend that Eurostat and Cedefop additionally scrapes, where it is possible, information about the company that publishes an online job advertisement. It could be classified or linked with the business register. We are aware that it will not be always possible or the company might differ from the one that in fact offers a job position (i.e. recruiting agency).

6.4 Conclusion

OJA data are certainly very useful and promising. Cedefop's OJA dataset is large, it includes many countries, and many breakdowns of job ads. Among them, there are estimates of individual skills, that can be connected to occupations. However, for now, the quantitative results of inferring official job vacancies statistics from Cedefop's OJA data are poor. The number of job offers differ from vacancies by means, variation and other central moments of both distributions. This might be due to fast-developing structures of online job boards, as well as variable interest of job seekers and companies in existing online job boards, and formation of new job boards. Such changes happen randomly, so they are not easily predictable. A continuous monitoring of job boards might be supplemented with a validation of job advertisements with companies' websites. It might also lead to elimination of possible outdated job offers. Furthermore, we recommend development of a selection method for data sources on online job ads to account for various types of job offers, e.g. small and large companies, local and country-wide job ads, low-skilled and high-skilled occupations etc. Currently the composition of job boards considered in different countries varies. It may be one of the key factors behind varying properties of data on OJA.

The distributions of OJA and JVS are different. The more disaggregate data we use, the bigger are those differences. We find some promising results for forecasting JVS with aggregate data on OJA for four countries and three NACE sections. However, these relations need to be analysed when longer time series are available.

The number of online job offers gathered by Cedefop is 85% higher than the stock of Eurostat's JVS job vacancies. Having so many job advertisements creates the opportunity of using only some of them as predictors for job vacancies. One of the typical solutions, most often used in the literature, is to draw a subsample of online job offers to ensure comparability of distributions according to specific variables, out of which country and industry are the most important.

Classification algorithms are based on dictionaries, for example dictionary of occupations and their synonyms. However, some languages contain a lot richer variety of words than English language does, for example Poland, Slovakia, Czechia, Hungary. In such countries the dictionary of basic forms of words might not be enough. We recommend lemmatizing words from dictionaries and job advertisements by transforming these words to their basic forms, for example verbs to infinitives, nouns to first person, first case, singular forms. It should improve recognition of phrases from dictionaries in job advertisements.

We also suggest considering OJA data as a source of information about skills, and showing qualitative data, for example, skills rankings, classifications of skills across occupations, and other characteristics of job offers, like type of contract etc. Such detailed data are obvious advantages of OJA data over probability-based surveys. With these data, structural changes in the labour market may be analysed more carefully than traditional official statistics allow it. Also, despite large variance, OJA data may potentially show leading properties in comparison to other labour market aggregates. They may possibly lead the business cycle turning points, especially as OJA data are high frequency, high granularity, and may be gathered, analysed, and published with a small lag in comparison to job vacancy statistics. These functions should be further explored with longer OJA data time series, and more accurate data from particular EU Member-States' statistical offices, especially for the NUTS2 regions and ISCO occupational breakdowns.

7 References

References

Abraham, K. G. and Wachter, M. (1987). Help-wanted advertising, job vacancies, and unemployment. *Brookings papers on economic activity*, 1987(1):207–248.

Azar, J. A., Marinescu, I., Steinbaum, M. I., and Taska, B. (2018). Concentration in us labor markets: Evidence from online vacancy data. Technical report, National Bureau of Economic Research.

Barnichon, R. (2010). Building a composite help-wanted index. *Economics Letters*, 109(3):175–178.

Böhning, D. and van der Heijden, P. G. (2019). The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in britain. *The Annals of Applied Statistics*, 13(2):1198–1211.

Bohning, D., Van der Heijden, P. G., and Bunge, J. (2017). *Capture-recapture methods for the social and medical sciences*. CRC Press.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Cajner, T., Ratner, D., et al. (2016). A cautionary note on the help wanted online data. *FEDS Notes, Board of Governors of the Federal Reserve System https://www. federalreserve. gov/econresdata/notes/feds-notes/2016/acautionary-note-on-the-help-wanted-online-data-20160623. html*.

Carnevale, A. P., Jayasundera, T., and Repnikov, D. (2014). Understanding online job ads data. *Georgetown University, Center on Education and the Workforce, Technical Report (April)*.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Cedefop (2019a). Online job vacancies and skills analysis: a Cedefop pan-european approach.

Cedefop (2019b). The online job vacancy market in the eu: driving forces and emerging trends. *Cedefop research paper*, 72.

Cedefop (2019c). Project "real-time labour market information on skill requirements: feasibility study and working prototype". final report.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95(3):555–571.

Chatterjee, K. and Bhuyan, P. (2017). A New Capture-Recapture Model in Dual-record System A New Capture-Recapture Model in Dual-record System.

Chatterjee, K. and Bhuyan, P. (2019a). On the estimation of population size from a dependent triple-record system. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 182(4):1487–1501.

Chatterjee, K. and Bhuyan, P. (2019b). On the estimation of population size from a post- stratified two-sample capture – recapture data under dependence. *Journal of Statistical Computation and Simulation ISSN:*.

Chatterjee, K. and Mukherjee, D. (2018). A new integrated likelihood for estimating population size in dependent dual-record system. *Canadian Journal of Statistics*, 46(4):577–592.

Colombo, E., Mercorio, F., and Mezzanzanica, M. (2019). Ai meets labor market: exploring the link between automation and skills. *Information Economics and Policy*.

Daniel, C., Silva, G., and Alberto, J. (2005). Bayesian analysis of correlated misclassified binary data. *Computational Statistics & Data Analysis*, 49:1120–1131.

Davis, S. J., Faberman, R. J., and Haltiwanger, J. C. (2013). The establishment-level behavior of vacancies and hiring. *The Quarterly Journal of Economics*, 128(2):581–622.

Delden, A., Scholtus, S., and Burger, J. (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics*, 32:619–642.

Deming, D. and Kahn, L. B. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, 36(S1):S337–S369.

Deville, J. and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, 32(2):165.

Dunlop, J. T. (1966). Job vacancy measures and economic analysis. In *The measurement and interpretation of job vacancies*, pages 27–47. NBER.

Fader, P. S., Hardie, B. G., and Lee, K. L. (2005). “counting your customers” the easy way: An alternative to the pareto/nbd model. *Marketing science*, 24(2):275–284.

Godwin, R. T. and Böhning, D. (2017). Estimation of the population size by using the one-inflated positive poisson model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(2):425–448.

Hershbein, B. and Kahn, L. B. (2018). Do recessions accelerate routine-biased technological change? evidence from vacancy postings. *American Economic Review*, 108(7):1737–72.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105(491):1265–1275.

Kuhn, P. and Skuterud, M. (2004). Internet job search and unemployment durations. *American Economic Review*, 94(1):218–232.

Lavallée, P. (2009). *Indirect sampling*. Springer Science & Business Media.

Liu, H. and Zhang, Z. (2017). Logistic regression with misclassification in binary outcome variables : a method and software. *Behaviormetrika*, 44(2):447–476.

Lohr, S. L. and Brick, J. M. (2012). Blending domain estimates from two victimization surveys with possible bias. *Canadian Journal of Statistics*, 40(4):679–696.

Marinescu, I. and Rathelot, R. (2018). Mismatch unemployment and the geography of job search. *American Economic Journal: Macroeconomics*, 10(3):42–70.

Marinescu, I. and Wolthoff, R. (2016). Opening the black box of the matching function: The power of words. Technical report, National Bureau of Economic Research.

Meyer, B. D. and Mittag, N. (2017). Misclassification in binary choice models. *Journal of Econometrics*, 200(2):295–311.

Oliveira, G. L. D., Loschi, R. H., and Assunção, R. M. (2017). A random-censoring Poisson model for underreported data. (June):4873–4892.

Pater, R. (2017). Is there a beveridge curve in the short and the long run? *Journal of applied economics*, 20(2):283–303.

Pater, R., Szkola, J., and Kozak, M. (2019). A method for measuring detailed demand for workers' competences. *Economics: The Open-Access, Open-Assessment E-Journal*, 13(2019-27):1–30.

Pfeffermann, D. and Sikov, A. (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, 27(2):181–209.

Pires, M. C. and Quinino, R. d. C. (2019). Repeated responses in misclassification binary regression: A bayesian approach. *Statistical Modelling*, 19(4):412–443.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Riddles, M. K., Kim, J. K., and Im, J. (2016). A propensity-score-adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, 4(2):215–245.

Rothwell, J. (2014). Still searching: Job vacancies and stem skills. *Report. Washington: Brookings Institution*.

Roy, S., Banerjee, T., and Maiti, T. (2005). Measurement error model for misclassified binary responses. *Statistics in Medicine*, 24:269–283.

Roy, S., Das, K., and Sarkar, A. (2013). Analysis of binary data with the possibility of wrong ascertainment. *Statistica Neerlandica*, 67(3):293–310.

Schmittlein, D. C., Morrison, D. G., and Colombo, R. (1987). Counting your customers: Who-are they and what will they do next? *Management science*, 33(1):1–24.

Sikov, A. (2018). A brief reivew of approaches to non-ignorable non-response. *International Statistical Review*, 86(3):415–441.

Stoner, O., Economou, T., Drummond, G., Stoner, O., Economou, T., Drummond, G., Stoner, O., Economou, T., and Drummond, G. (2019). A Hierarchical Framework for Correcting Under- Reporting in Count Data A Hierarchical Framework for Correcting Under-Reporting in Count Data. *Journal of the American Statistical Association*, 0(0):1–17.

Tang, N. and Ju, Y. (2018). Statistical inference for nonignorable missing-data problems: a selective review. *Statistical Theory and Related Fields*, 2(2):105–133.

The Conference Board (2018). The conference board help wanted online® technical note. Technical report.

Turrell, A., Speigner, B., Djumalieva, J., Copple, D., and Thurgood, J. (2018). Using job vacancies to understand the effects of labour market mismatch on uk output and productivity.

Turrell, A., Speigner, B. J., Djumalieva, J., Copple, D., and Thurgood, J. (2019). Transforming naturally occurring text data into economic statistics: The case of online job vacancy postings. Technical report, National Bureau of Economic Research.

Zhang, L.-C. (2005). On the bias in gross labour flow estimates due to nonresponse and misclassification. *Journal of Official Statistics*, 21.

Zhang, L.-C. (2008). Developing methods for determining the number of unauthorized foreigners in norway. *Statistisk Sentralbyrå/Utlendingsdirektoratet, Oslo Garcia, Jose Miguel Morales*.

Zhang, L.-c. (2015). On Modelling Register Coverage Errors. 31(3):381–396.

Zhang, L.-C. and Chambers, R. L. (2019). *Analysis of Integrated Data*. CRC Press.

Zhang, L.-C. and Dunne, J. (2017). Trimmed dual system estimation. In *Capture-Recapture Methods for the Social and Medical Sciences*, pages 237–257. Chapman and Hall/CRC.

Annex

Description of Cedefop's OJA dataset

We got access to the database (`cedefop_datalab.ft_advertisement_en` and `cedefop_datalab.ft-_skill_analysis_en`) with the following variables

Table 16: Description of variables in Cedefop's OJA dataset (ft_advertisement_en and ft_skill_analysis_en)

Variable	Description
general_id	Ad identifier
grab_date	First date of web-scrapping
expire_date	Last date of web-scrapping or expire date
lang	Language
idesco_level_4,3,2,1	ID of ESCO level 4,3,2,1
esco_level_4,3,2,1	Name of ECSCO level 4,3,2,1
idescoskill_level_3,2,1 (ft_skill_analysis_en)	ID of ESCO skills level 3,2,1
escoskill_level_3,2,1 (ft_skill_analysis_en)	Name of ESCO skills 3,2,1
idcity	City ID
city	City name
idprovince	Province ID
province	Province name
idregion	Region ID
region	Region name
idmacro_region	Macroregion ID
macro_region	Macroregion name
idcountry	Country ID
country	Country name
idcontract	Type of contract ID
contract	Type of contract name
ideducational_level	Educational level ID
educational_level	Educational level name
idsector	Sector ID
sector	Sector name
idmacro_sector	Macrosector ID
macro_sector	Macrosector Name
idcategory_sector	Category sector ID
category_sector	Category sector name
idsalary	Salary ID
salary	Salary name
idworking_hours	Working hours ID
working_hours	Working hours name
idexperience	Experience years ID
experience	Experience years name
source_category	Data source (website)
sourcecountry	Source country
source	Type of source
site	Exact name of the website
companyname	Company name

Detailed results of models

ZERO-ONE TRUNCATED CAPTURE-RECAPTURE METHODS

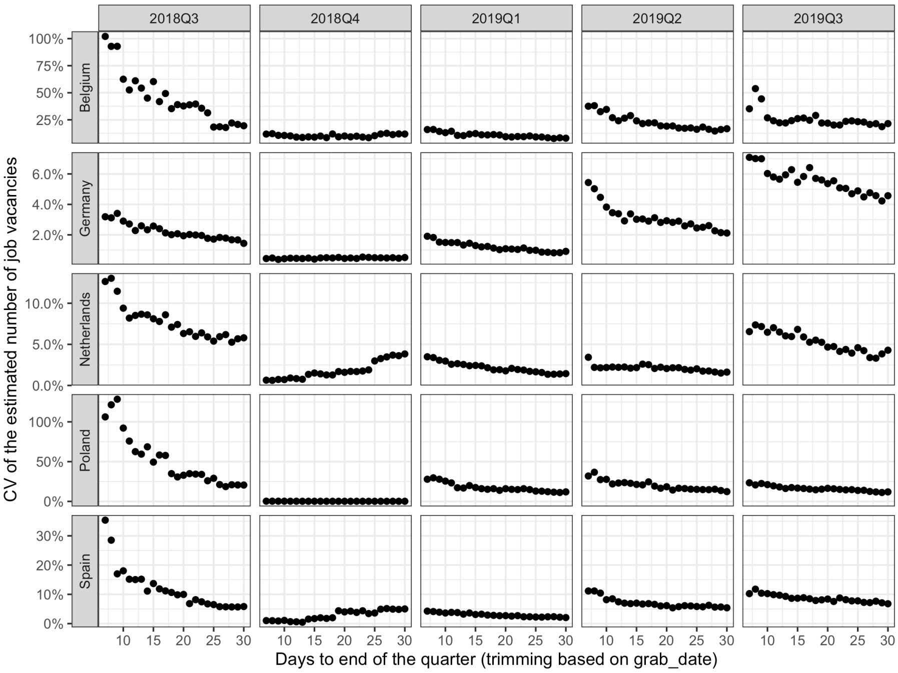

Figure 22: **Comparison of CV of estimates based on JVS (red dashed line), OJA data without trimming (blue dashed line) and zero-one-truncated capture-recapture size estimator under different trimming dates (7 to 30)**

Table 17: **Summary statistics for information criterion for big countries**

Country	# Poisson better	# Geometric better	Avg AIC Poisson	Avg AIC Geometric
Belgium	74	46	720.62	721.73
Spain	108	12	2588.86	2600.42
Germany	70	50	14279.85	14322.83
Netherlands	117	3	6564.95	6601.33
Poland	98	22	1230.90	1255.11

Note: *# Poisson better* denotes how many times the Poisson model was better, *# Geometric better* denotes how many times the Geometric distribution was better, *Avg AIC Poisson* and *Avg AIC Geometric* denotes average AIC criterion for Poisson and Geometric distribution, respectively.

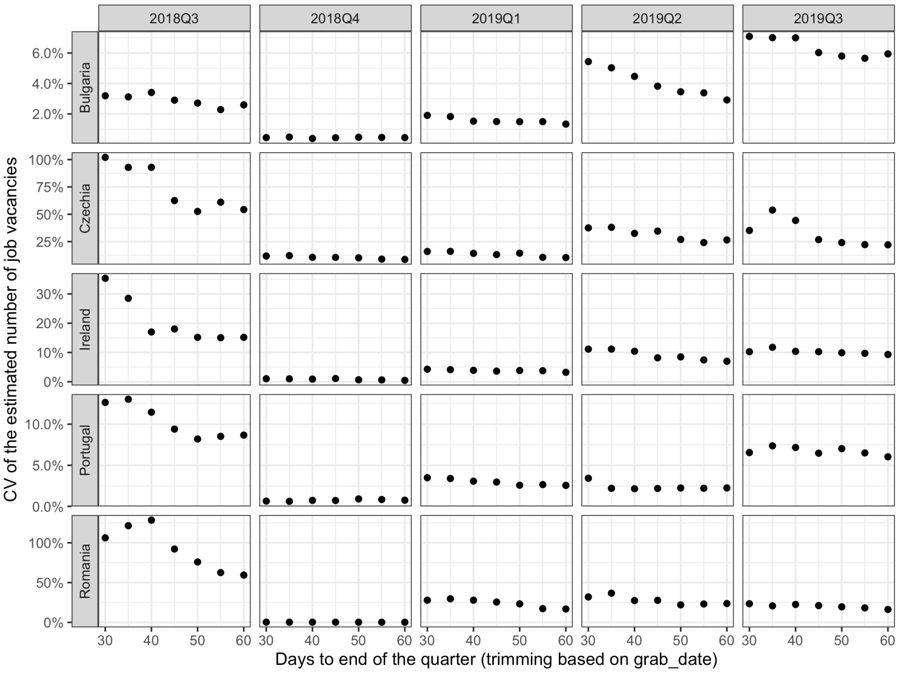

Figure 23: Comparison of CV of estimates based on JVS (red dashed line), OJA data without trimming (blue dashed line) and zero-one-truncated capture-recapture size estimator under different trimming dates (30 to 60 by 5)

Table 18: Summary statistics for information criterion for small countries

Country	# Poisson better	# Geometric better	Avg AIC Poisson	Avg AIC Geometric
Czechia	20	20	429.05	429.55
Ireland	37	3	1588.00	1594.15
Bulgaria	24	16	9202.68	9225.69
Portugal	37	3	4097.21	4122.33
Romania	32	8	851.73	871.23

Note: *# Poisson better* denotes how many times the Poisson model was better, *# Geometric better* denotes how many times the Geometric distribution was better, *Avg AIC Poisson* and *Avg AIC Geometric* denotes average AIC criterion for Poisson and Geometric distribution, respectively.

R codes for estimation of total number of job vacancies

ZERO-ONE TRUNCATED CAPTURE-RECAPTURE MODELS

1. Log-likelihood of zero-one truncated Poisson distribution

```
ll_trun_pois <- function(x, par) {
    vec <- as.vector(table(x))
    lev <- as.numeric(names(table(x)))
    den <- 1 - dpois(x = 0, lambda = par) - dpois(x = 1, lambda = par)
    probs <- dpois(x = lev, lambda = par)
    ll <- vec * log(probs / den)
    sum(ll)
  }
```

2. Log-likelihood of zero-one truncated Geometric distribution

```
ll_trun_geom <- function(x, par) {
    vec <- as.vector(table(x))
    lev <- as.numeric(names(table(x)))
    den <- 1 - dgeom(x = 0, prob = par) - dgeom(x = 1, prob = par)
    probs <- dgeom(x = lev, prob = par)
    ll <- vec * log(probs / den)
    sum(ll)
  }
```

4. Estimation using *maxLik* package

```
est_pois <- maxLik(logLik = ll_trun_pois,
    start = 0.1, method = "NR", x = data)
est_geom <- maxLik(logLik = ll_trun_geom,
    start = 0.1, method = "NR", x = data)
```

5. Estimation of population sizes

```
popsizes <- c(

poisson = dpois(0, coef(est_pois))*sum(res$times[-1])/
(1 - dpois(0, coef(est_pois)) - dpois(1, coef(est_pois))) +
sum(res$times),

geom = dgeom(0, coef(est_geom))*sum(res$times[-1])/
(1 - dgeom(0, coef(est_geom)) - dgeom(1, coef(est_geom))) +
sum(res$times))

)
```

6. Bootstrap for Poisson distribution

```
pop_size <- popsizes[1]
```

```
probs <- c(pop_size- sum(res$times), res$times)/pop_size
boots <- 200
sampled_freqs <- rmultinom(n = boots, size = pop_size, prob = probs)
sampled_freqs <- t(sampled_freqs)
pop_boot_est <- numeric(boots)

for (b in 1:boots) {
      yyy_boot <- rep(res$N[-1], sampled_freqs[b,-c(1,2)])
      est_boot <- maxLik(logLik = ll_trun_pois,
      start = coef(est_pois), method = "NR", x = yyy_boot)
      pop_boot_est[b] <- dpois(0, coef(est_boot))*sum(res$times[-1])/
      (1 - dpois(0, coef(est_boot)) - dpois(1, coef(est_boot))) + sum(res$times)
}
```

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:
– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
– at the following standard number: +32 22999696 or
– by email via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications at: https://op.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

PDF: KS-TC-20-008-EN-N

Inferring job vacancies from online job advertisements

This study analyses online job advertisements collected by Cedefop in a pan-European approach. The main objective of the study is to develop estimator(s) for the number of job vacancies from data on online job advertisements taking into account the differences in the statistical unit and coverage. Although there are promising results for certain countries and industries, at the present stage the data are still experimental and could not be used to estimate the total number of vacancies. Based on the results, recommendations are provided to overcome the current limitations for future analysis..

For more information
https://ec.europa.eu/eurostat/

Publications Office
of the European Union

PDF: ISBN 978-92-76-23299-5