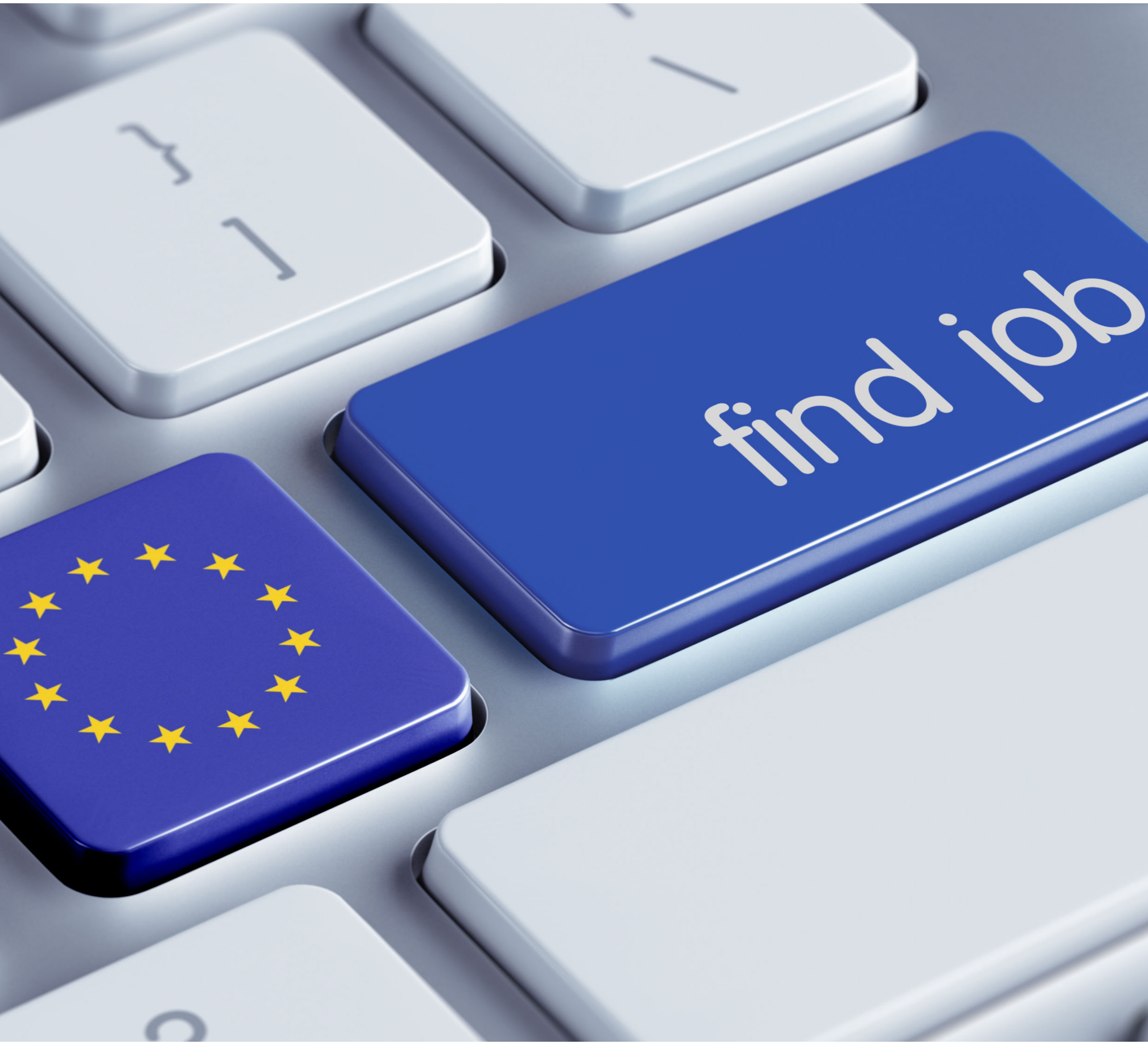
Competition in urban hiring markets: evidence from online job advertisements

ASCHERI ANDREA, KISS NAGY ANCA-MARIA, MARCONI GABRIELE, MESZAROS MATYAS, PAULINO RAQUEL, REIS FERNANDO

2021 edition



STATISTICAL WORKING PAPERS

eurostat

Competition in urban hiring markets: evidence from online job advertisements

2021 edition

Manuscript completed in November 2021

This document should not be considered as representative of the European Commission's official position.

Luxembourg: Publications Office of the European Union, 2021

The information and views set out in this publication are those of the authors and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Theme: Population and social statistics
Collection: Statistical working papers

PDF: ISBN 978-92-76-43554-9 ISSN 2315-0807 doi: 10.2785/667004 KS-01-21-430-EN-N

Executive summary

Innovation in official statistics has been further stimulated by the recent health crisis, which has generated new opportunities and requests to provide new indicators to monitor the economic, social and environmental impacts of the pandemic. Online Job Advertisements (OJAs), which provide a real chance to support and deepen labour market statistics, offer one of these opportunities, created through collaboration with other governmental and private institutes. The main aim of this paper is to demonstrate one of the possible uses of this data source to show its potential and possibly inspire future use cases. Online job ads are available at Eurostat and the European Statistical Systems thanks to an administrative agreement reached in 2020 with the European Centre for Development of Vocational Training (Cedefop). The job advertisements for all 27 EU Member States are collected from hundreds of job portals through web scraping and agreements with the portals. The information contained in the ads are then processed to classify the advertisements according to variables such as occupation type, employer, type of contract and location.

This paper presents an urban labour market concentration index calculated using online job ads. There are currently no statistics available on labour market concentration at European level. Nevertheless, there is a growing interest in measuring concentration of labour markets since it implies limited competition among firms, which in turn can drive down workers' bargaining power and deteriorate job conditions. The study provides evidence of the level of labour market concentration across nearly all occupations and for every functional urban area (FUA) of the 27 EU Member States, using OJA data for 2019 and 2020. More specifically, the Herfindahl-Hirschman Index (HHI) is used to calculate the concentration for labour markets at the occupation (ISCO level 4), functional urban area and quarterly level.

The results indicate that the largest urban areas in Europe tend to have lower level of concentration of the hiring market, indicating more competition among employers and more job opportunities for workers across all occupations. This is also confirmed by migration trends that show how these urban areas attract more people in search for better job conditions. Some occupation types appear to be more concentrated than others on average, but this may be partly due to the fact that some occupations are more frequently advertised online than others. With respect to time series, an average increase in labour market concentration can be seen in the second quarter of 2020, when the pandemic crisis hit Europe stronger.

While providing new analytical opportunities, online job advertisements do not cover all job offers. Data on OJAs cannot replace other sources of labour market information, but they can complement them by providing comprehensive, detailed, and timely insights into labour market trends and allowing the early identification of new emerging jobs and skills.

In conclusion, the contribution of this paper is twofold: first, it describes a new source of data, addressing some of the challenges related to its application to policy-relevant work; second, it provides the first Europe-wide evidence on demand-side labour market competition in urban labour markets. The paper also provides directions for future work to improve the accuracy and reliability of both the data sources and the statistical outputs.

Keywords: Online Job Advertisements; Labour Markets**;** Functional Urban Areas, Herfindahl-Hirschman Index; Employment; Competition; New data sources; Big data

Authors: Ascheri Andrea, Kiss Nagy Anca-Maria, Marconi Gabriele, Meszaros Matyas, Paulino Raquel, Reis Fernando.

Acknowledgements: The authors would like to thank the colleagues participating in the ESSnet Big Data project on Online Job Vacancies, and in particular Jakob de Lazzer (Destatis), who provided the idea behind this experimental work, and colleagues from the Slovenian National Statistical Institute who contributed to the coding of company names. Jiri Branka, Joanna Napierala and Vladimir Kvetan (Cedefop) provided very useful comments. Mauro Pelucchi and Alessandro Vaccarino (Burning Glass Technologies) answered all the questions on data engineering.

Table of contents

List of figures

List of tables

List of abbreviations

CEDEFOP:	European Centre for the Development of Vocational Training
EFTA:	European Free Trade Association
ESCO:	European Skills, Competences, Qualifications and Occupations
ESS:	European Statistical System
ESSC:	European Statistical System Committee
FUA:	Functional Urban Area
HHI:	Herfindahl-Hirschman Index
ISCED:	International Standard Classification of Education
ISCO:	International Standard Classification of Occupations
LAU:	Local Administrative Unit
NACE:	Statistical classification of economic activities in the European Community
NSI:	National Statistical Institute
NUTS:	Nomenclature of territorial units for statistics
OECD:	Organisation for Economic Co-operation and Development
OJA:	Online Job Advertisement
OJV:	Online Job Vacancies
TSS:	Trusted Smart Statistics
UNECE:	United Nations Economic Commission for Europe
WIH:	Web Intelligence Hub
WIN:	Web Intelligence Network

1 Introduction

This paper presents the first experimental results on the use of Eurostat’s Online Job Advertisements (OJAs) data for estimating an urban labour market concentration index, together with the underlying methodology. It provides evidence of the level of labour market concentration across nearly all occupations and for every functional urban area (FUA) of the 27 EU Member States, using OJA data for over two years. More specifically, the Herfindahl-Hirschman Index (HHI) is calculated for labour markets at the occupation (ISCO level 4), functional urban area and quarterly level. Therefore, the contribution of this paper is twofold: first, it describes a new source of data, addressing some of the challenges related to its application to policy-relevant work; second, it provides the first Europe-wide evidence on demand-side labour market competition in urban labour markets.

New data sources available in today's *datafied* world (Ricciato, 2019), or "big data" (UN Statistical Commission, 2013), presents important opportunities for new applications in official statistics, but also serious challenges (Yongdai, 2013; Hackl, 2016; Tam, 2015). Besides the large amount of data that they produce, new data sources are also qualitatively different from traditional ones, because they come from different data ecosystem generated by different actors and dynamics (Ricciato, 2019). New computing and processing technologies that became available in the current century help dealing with the complex challenges that this novelty poses. However, these methods are not yet mature or applicable to official statistics, so that more research is still needed to overcome the methodological difficulties implicit in using new data sources" (UN Statistical Commission, 2013). By describing the main processes leading from the ingestion of data from the internet to the production of an experimental statistical indicator, this paper contributes to this growing stream of research. The fact that the work on online job ads is still in the research and development phase also implies that its results are to be considered as experimental. Although they are produced in a robust statistical quality context, their data and methodology display a lower level of maturity as compared to official European statistics and do not meet the same quality criteria.

This paper presents new evidence on the state of urban labour markets in Europe. Cities play a very important role in modern economies, as sources of jobs and innovation (OECD & European Commission, 2020). In the European Union, 72% of employed people is found in predominantly urban regions (authors' calculations based on Eurostat (2021) - data are missing for France). Urban labour markets, particularly in large urban conglomerates, offer opportunities for labour specialisation, knowledge spill over and good matching between skills and job tasks (Gordon & Turok, 2005). However, this happens only when the labour market is "thick", meaning that workers have sufficient job options available for their occupations within a reasonable distance (Brown & Scott, 2012). Thick labour markets give workers more security to specialise in a domain, and the actual ability to move across companies, bringing their knowledge and network along. In contrast, unavailability of alternative job options within commuting distance often makes labour markets "thin", meaning that firms have more market power, with a negative effects on wages (Manning, 2003). Demand-side competition in labour markets (i.e., firms competing to hire workers) also reduces the risk of abuses of monopsony power (e.g. abusive adoption of non-compete provisions, preventing workers from disclosing information about salary, incorrectly treating employees as self-employed, preventing workers from bringing legal action), which tend to be difficult to prevent by legal means (OECD, 2020).

Measuring labour market thickness (i.e., labour market competition for a given occupation within reaching distance from home) has been broadly unfeasible for statisticians and researchers due to the demanding information requirements on job offers (location, occupation, and employer) and commuting options. However, a new empirical approach to address this challenge became possible due to two concurrent statistical developments: the availability of granular and detailed web-scraped data on OJAs (see Azar et al. (2020) who first applied this approach to the US); and the development by international organisations of a new definition of geographic areas based on commuting distance.

This new definition (OECD, 2012) makes it possible to aggregate OJAs at the level of the functional urban area based on their location identifiers. By using additional information on occupation and company name for each OJA, it is possible to calculate measures of demand-side competition in the labour market. Following Azar et al. (2020), the HHI concentration index is computed and used as a measure of thinness of the urban labour market: the highest the HHI, the more concentrated is the demand side of the labour market, implying that fewer companies compete for workers or, in other words, that the labour market is thinner.

While the benefits of measuring a phenomenon of great relevance to labour markets are clear, the disadvantage of using OJAs as a data source is that they do not cover all job offers (Cedefop, 2019). For example, many jobs could be advertised only outside the web, either informally or through other means (e.g. in newspapers). Therefore, our HHI measures are likely to overestimate concentration across occupations because they capture only jobs that are advertised online. However, the results presented in this paper remain relevant for online recruitment activities. These play a large and increasing role in modern labour markets (De Hoyos, 2013; Cedefop, 2019), to the point that some authors argue that recruitment is moving towards "digital by default" (Green, 2017). Online job posting allows employers to advertise quickly and cheaply. In addition, "web 2.0" tools like social media and videoconference apps add further possibilities to provide more detailed information or to follow up job applications with requests or interviews. All this made online posting ubiquitous (Green, 2017). At the same time, job seekers can take advantage of online job searches in order to scan many potential jobs in a short time and make use of the additional functionalities made available by the web (De Hoyos, 2013). Possibly, because of this, a large majority of job seekers looks for jobs online. For example, 57% of unemployed people in Europe reported to have looked for jobs online in the three months before being surveyed in 2019 (Eurostat, 2021).

The analysis presented in this paper shows a core of large urban areas with thick labour markets in Central and Western Europe including Amsterdam, Brussels, Hamburg, Milan, Munich, Paris, the Ruhr, Stockholm and others. In these urban areas, many firms compete to hire employees within each occupation, resulting in a low HHI level. In contrast, urban labour markets are thinner all along the southern and eastern periphery of the European Union (in Greece, Lithuania, Romania, Portugal and other countries and regions), particularly in smaller towns. Labour market concentration (or thinness, implying that workers have less choice of employers resulting in a high HHI level) across urban areas in the European Union is associated with lower employment rates, lower satisfaction with personal job situation and outward mobility of workers towards areas with thicker labour markets. Comparing different methods for calculating the HHI shows that the results allow for a robust comparison across urban areas. These comparisons also show that, despite the level of the HHI is substantially affected by methodological assumptions, the level of demand-side competition can be considered problematic in at least a quarter (and at most, the large majority) of European urban areas.

After the introduction in Section 1, Section 2 presents the institutional setting behind the collection of online job ads and its use for statistics. Section 3 gives a comprehensive overview of the data source used for calculating the indicator, including references to the process of extracting information from online job ads as well as its quality aspects. Section 4 presents in detail the methodology developed and implemented for this study. Section 5 are showing the results obtained. Section 6 discusses the results and draws some conclusions, including ideas for further work on this topic. Finally, the Annex I – Company names provides more information on the methodology to deal with company names for the purpose of the index.

2 OJA data: institutional setting

Following the increasing internet penetration and information and communication technology literacy, the use of the internet for publishing job advertisements has increased over the past years. While job advertisements published online were initially targeting predominantly highly skilled workers, today it contains job advertisements for all almost all occupations and skills. In addition to simplifying the process leading to a match between employers and jobseekers, the increasing use of OJA portals also has great potential for labour market and skills analysis (Cedefop; European Commission; ETF; ILO; OECD and UNESCO, 2021). OJAs have a great potential to complement official statistic thanks to their higher timeliness, relevance and granularity (location data and occupation-skills information). The real-time nature of job ads data also allows for the early detection of labour demand trends, which gives job seekers, employers, and policymakers a forward-looking analytical tool. Compared to survey-based labour market data (which have the advantage of relying on random sampling), OJA data are cost-effective and offer the potential to improve the accuracy of labour market forecasts while producing supplemental estimates of demand at a detailed level of occupations, industries, locations and skills.

OJAs are a powerful source of information on job requirements. However, they are not usually available in a structured format to researchers, and they are difficult to gather through the conventional means of official statistics. OJAs cannot replace other types of labour market information like job vacancy statistics and labour forces survey, but they have the potential to provide complementary, granular, and timely insights into labour market trends, for example on emerging jobs and skills and on the geography of labour markets.

The European Centre for the Development of Vocational Training (Cedefop) and several National Statistical Institutes (NSIs) of the European Statistical System (ESS), grouped into networks called *ESSnet,* have engaged in parallel projects to assess the feasibility of using OJAs for labour market analysis and job vacancy statistics. After an initial feasibility study finalised in 2016, Cedefop developed a Pan-European system providing information on skills demand in OJAs (Cedefop, 2019). Around the same time, the ESS launched two projects, the ESSnet Big Data I (2016-2018, focused on exploring potential statistics that can be derived from OJAs) and the ESSnet Big Data II (2018-2020, concentrated on creating the conditions for a larger scale implementation of the project). Altogether, this Pan-European approach has developed methods to collect OJAs in all European Union Member States, making the data available for analysis and interpretation. The advantage of the data is not only their European scope, but also their collection on a continuous basis. This provides the opportunity to track market trends and support official statistics.

Eurostat and Cedefop have been discussing for many years ways and conditions to augment the Cedefop system to produce official statistics at European and national level. In 2018, the ESSnet Big Data I recommended considering close collaboration between Cedefop and the ESS to set up a system that would be beneficial to both sides as well as to the broader scientific community (Descy, Kvetan, Wirthmann, & Reis, 2019). Following these developments, Eurostat -representing the ESS- and Cedefop are currently working towards a joint data production system based on OJAs. Based on a formal agreement establishing the basis for this cooperation (Cedefop, 2020), Eurostat is building a new system that will replace the current system for OJA data collection. The new infrastructure will be modular to enable inclusion of national and European processes as well as using intermediate and final data for different purposes at national and European level.

Substantial effort is being put in improving the data collection so that it will be fit for the production of official statistics in the future. This goes beyond the initial aim of the project launched by Cedefop, and will require substantial inter-institutional cooperation between Eurostat, Cedefop itself, the NSIs, and other private or public partners contracted to support the data collection. At its meeting on 16 May 2019 in Luxembourg, the European Statistical System Committee (ESSC) discussed the principles of Trusted Smart Statistics (TSS) and priority areas for producing European statistics from new data sources (Eurostat, 2019). This includes the creation of a Web Intelligence Hub (WIH) that collects various data from the web to enhance statistical information in various domains.

The purpose of the WIH is to provide to Eurostat and subsequently to the ESS, the necessary building blocks for harvesting information from the web and produce statistics out of it. In order to do so, the WIH will set up those building blocks required for the collection and processing of data for the specific use-cases as defined in the TSS portfolio. In addition to work on IT infrastructure and business architecture, priority areas include ensuring stable access to sources of primary data, unifying classifiers for jobs and skills developed in Cedefop's project, assessing and improving data quality, aligning OJA data with official statistics standards and conventions, and developing comprehensive documentation.

This experimental study is based on the work done within the Work Package B of the ESSnet project on Big Data II (ESS, 2020) co-financed by Eurostat on online job advertisements. In particular, this study is built on one of the prototypes developed within this project, which uses German OJA data to calculate a labour market concentration index (ESS, 2020). The methodology used in this use case has been adapted and replicated on all the 27 EU Member States.

3 Data

The data used in this study consist of 116 851 363 distinct online jobs ads collected from 316 distinct sources in all 27 EU countries. Data for the UK are also available (though they have not been used for the analysis presented here), and an extension of the dataset is expanded to cover the member states of EFTA: Norway, Iceland, Switzerland and Liechtenstein. Statisticians and labour market experts from all the countries involved in the data collection, through the procedure described in Cedefop (2019), have identified the main data sources. These sources are job search engines (e.g., Indeed, Monster) and websites with job advertisements managed by public employment services.

OJAs refer to advertisements published on the World Wide Web revealing an employer's interest in recruiting workers with certain characteristics for performing certain work. Employers can publish job ads for various reasons, for example to fill a current vacancy or to explore potential recruiting opportunities.

OJAs usually include data on the characteristics of the job (e.g. occupation and location), characteristics of the employer (e.g. economic activity) and requirements (e.g. education/skills). Part of this information is only available as unstructured data (natural language text). ”Big” data from novel sources like those analysed here offer much more granular information compared to traditional data sources, but also require specific methodologies for processing and analysis, which are described below.

The OJA data is released quarterly. The timeliness of the data release has been improving over time, starting with a lag of 7 months between collection and data release, which has been reduced to the current lag of 2 months. The results presented in this document are based on the most recent version of the dataset (i.e. *v9*) released during the first quarter of 2021. This dataset contains data from the third quarter of 2018 to the end of 2020. However, only data from 2019 and 2020 has been used in this paper, because of its better coverage of important sources.

3.1. OJA data collection and processing

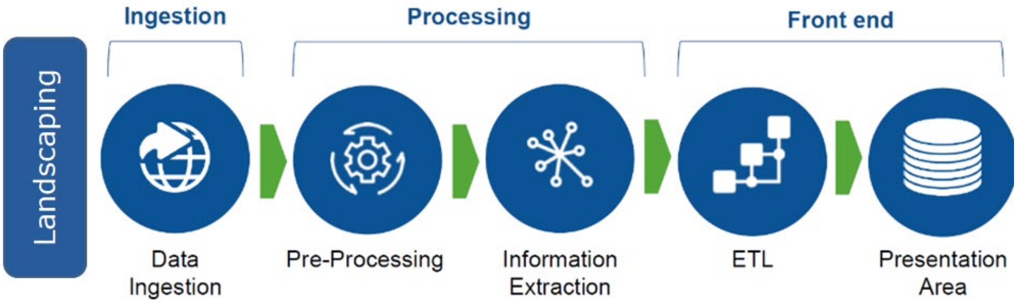
In order to set up the OJA data collection system, a first landscaping exercise provided the state of the art on the use of online job advertisements for labour market analysis in different countries. This exercise achieved the following goals:

- landscape of the data space,
- define strategy for source selection,
- assess websites data quality,
- define a representative list of portals,
- establish agreements to obtain stable data supply, test the selected list of sources.

This first landscaping exercise led to the selection of the sources to be included in the data release used for this paper. Every two or three years, the landscaping exercise is repeated and the list of data sources is updated.

The current system data flow can be summarized in five main steps shown in Figure 1 and described thereafter.

Figure 1: Online Job Advertisement data pipeline



Source: adapted from CEDEFOP (2019a)

Data ingestion includes the activities related to the data collection, for example crawling, fetching, scraping and storing data. Data ingestion is the process of obtaining and importing data from web portals and storing it in a database. It mainly focuses on volume rather than quality, in the sense that the priority is to collect as much of the available information as possible so that it can be processed and refined at a later stage. The key element of data ingestion is to ensure a stable data flow preventing potential loss of data due to harvesting issues. For this reason, direct agreements with the most important data sources were established in order to obtain a stable data supply, agree on a data format and minimize the impact of the data ingestion activity for the portal. In order to maximise the coverage, which might suffer because of website unavailability, blocking or changes in data structure, data from the most important sites are ingested from two or more sources. The data ingestion phase deals with both structured sources, which store information in structured fields, and unstructured sources, where information is extracted from large chunks of text.

Pre-processing includes all the activities related to data preparation for further analysis; data preparation, translation, data cleaning and text processing tasks the main activities of this phase. Pre-processing is a critical step of the data processing pipeline that can be divided into three steps: merging, cleaning and text processing / summarizing. After these steps, data is (i) put into a single complete dataset, (ii) cleaned from noise, (iii) summarised and (iv) prepared for information extraction step. The pre-processing starts with the language detection and redirection to a language specific pipeline. Pages in the websites which do not refer to job ads (e.g. ads of training courses and blog posts) are identified via a combination of simple heuristics and machine learning with an estimated precision of 99% and are eliminated. The process of summarisation deals with job ads that are found (almost) identical in several sources (duplicates). Duplicates are estimated to account for some 20% of the collected job ads. These cases are identified so that duplicate ads can be eliminated (de-duplication). During pre-processing metadata from structured fields, for example the release date of the job ad and the job title, are extracted in addition to the full job description.

Information extraction is the process of extracting structured data from unstructured text to classify it into standard statistical classifications. This process is defined through a set of processing pipelines. A pipeline can be defined as a portion of information extraction dealing with a specific variable and with a particular language. Pipelines can be combined to define the whole information extraction process. During the processing of each pipeline, jobs advertised are analysed to classify the contents of the pipeline (contract, occupation…) according to the specified language. In practice, the information retrieval process is composed by: one pipeline for each language considered and one pipeline for each attribute to be classified (occupation, skill, contract, educational level, experience, economic activity, location, salary, working hours). Each job is processed once for each attribute (variable) by selecting the pipeline related to the language detected. The total number of pipelines supported can be calculated as the product of the number of attributes (9) and the number of supported languages (29). During the information extraction step, the unstructured data is classified into standard statistical classifications.

Extraction, Transformation and Loading (ETL) concerns data preparation for the front-end tool. Apache Spark® is used as the main component for ETL and processing activities. It is a fast and general-purpose engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing.

Data presentation targets on one hand data scientists and analysts, and on the other hand decision makers and business users. In the case of data scientists and analysts it will cover use cases such as data discoverability, machine learning integration and embedded advanced analytics. In the case of decision makers and business users, it will cover self-service analytics, visual-based investigation and storytelling.

3.2. Data quality assessment

While useful for a variety of statistical purposes, online job ads data come with limitations. OJAs do not represent the entire job ads population. Some occupations and economic activities are less well represented in web advertisements than others are. In some regions, digital tools may not be widespread enough to encourage employers to publish job advertisement online. The quality of statistical data is effectively assessed with reference to quality frameworks, which systematically address quality by referring to quality dimensions. The United Nations Economic Commission for Europe (UNECE, 2014) quality framework provides a useful model for assessing the quality of new (big) data sources, such as web data. The remainder of this section will assess some of these dimensions in the context of OJA data.

3.2.1. Data accuracy

UNECE (2014) stresses the importance of selectivity as a quality dimension of web data. The selectivity of the data sources is its degree of representativeness, and it is a sub-dimension of data accuracy. The problem of selectivity must be kept in mind when working with OJAs, because of the following reasons:

1. OJAs represent only a part of job demand, as not all job vacancies are advertised online.
2. The penetration of OJA markets (i.e. the extent to which job demand is captured by OJAs) varies in and across countries and may change over time.
3. The volume, variety and quality of the data depend on the selection of portals in each country. The OJA landscape is different across countries, and it comprises multiple actors with different business models (Cedefop, 2019).

Beresewicz and Pater (2021) and Carnevale et al. (2014) both assess selectivity in job ad data, by comparing with data sources of a similar (though different) nature like job vacancy and labour force surveys. As could be expected, they observe that the distribution of ads across industries and occupation significantly differ from those derived from these surveys. Beresewicz and Pater (2021) attempt to correct for selectivity in OJA data through a post-stratification based on the distributions observed from job vacancy and labour force surveys, taking into account auxiliary information and using model-based estimates and Bayesian inference. Similar estimators could potentially lead to indicators that reflect better the distribution of occupations and sectors calculated from representative data sources. The results of that study are promising, but still experimental as the accuracy of the resulting indicators depend on the country and other factors. Therefore, the present paper takes a simpler approach, by aggregating the main indicator through an arithmetic mean across all occupations. This avoids giving excessive weight in the estimator to occupations that are over-represented among OJAs (a comparison with the weighted average is provided in the section Methodology).

3.2.2. Comparability over time

Comparability over time relates to the UNECE quality dimension of "time related factors", such as "Timeliness", "Periodicity", and "Changes through time". The algorithms for data scraping and processing are still being refined and documented in the context of the OJA data collection, which implies that the data may not be directly comparable across time and data releases. Therefore, the current algorithms and data sources are tuned to the current and most recent data acquisitions, meaning that currently the OJA data is best suited to cross-sectional comparisons for recent periods. In addition, the most accurate time measure in the online job ad dataset is the grab date, i.e. the day in which an ad was scraped. This tend to happen after the ad was originally posted, with a lag that can range from a few days to over a month (e.g. if a hosting website changes structure and the scraping algorithm has to be redefined). The quality aspect of comparability over time will be improved with the introduction of more standardised statistical processes for regularly produced statistical products. For example, an undergoing data improvement project focuses on the identification of OJA sources that are stable over time, so that they can be used to produce more reliable time series.

In addition, online job portal landscape is shaped by broad trends reflecting technological, market and social changes, such as the spreading of online activity, portal consolidation and the introduction of new tools (Cedefop, 2019). For example, more jobs are expected to appear online in the near future, as more companies will only hire online. These changes may affect comparability over time between countries, because they do not happen at the same pace across countries (Cedefop, 2019).

3.2.3. Punctuality

With respect to the quality of results (*output*), the latter quality criteria refer to two additional criteria:

- punctuality which refers to the delay between the date of the release of the data and the target date;
- comparability referring to the measurement of the impact of differences in applied statistical concepts, measurement tools and procedures where statistics are compared between geographical areas, sectoral domains or over time.

This paper does not fully exploit the potential advantages of OJAs in terms of punctuality, as currently the application of quality assurance procedures for data production and publication of results still implies a delay of several months between the data reference period and the publication date. In contrast, it makes full use from the comparability of statistical concepts included in official classification like the ISCO occupation classification (European Commission, 2021) and the determination of functional urban areas (Dijkstra, Poelman, & Veneri, 2019; OECD, 2012).

3.2.4. Missing data

The OJA data may also encounter some data accuracy issues, especially due to the complexity and the amount of data that has to be classified in the data processing phase. Some degree of miss-classification may be encountered for all the categorical variables in the OJA dataset (Table 1). This working paper reports an accuracy measure for occupation (Section 4), measured during the development of the classification model. Future work on the OJA dataset includes a formal procedure to measure the accuracy for all the categorical variables reported in Table 1.

Table 1: Online Job Advertisement database: Missing values by variable (including variables not used in this paper)

Variable	Missing values	Variable type
economic activity of the employer	2%	Categorical (NACE at 2. level)
type of contract	29%	Categorical ("permanent", "self-employed" and "temporary")
working hours	38%	Categorical ("full-time", "part-time")
education level required	1%	Categorical (ISCED 2011)
salary	74%	Categorical (13 levels)
experience	51%	Categorical (8 levels)
place of employment (region)	36%	Categorical (NUTS3)
place of employment (city)	47%	Categorical (LAU)
occupation	0%	Categorical (ISCO level 4)
skills	1%	Categorical (ESCO level 3)
time (grab and expired dates)	0%	Date
company names	20%	String

Finally, due to a variety of other reasons, such as: scarce information in the ads, inadequate classifiers, interruptions in the data collection pipeline for some of the selected sources (due to spams, problems with portal/site access, ...), some variables can be affected by a substantial amount of missing values in the OJA dataset (Table 1). In this experimental study, the issue of missing data is handled using imputation methods, as described more in detail in Section 4 and Annex I – Company names.

Another important potential source of bias, besides missing data, is selectivity. The fact that only online job ads are observed in the OJA database is likely to impose an upward bias on the estimates, due to the fact that a number of companies and their (offline) job ads are not observed (a market with fewer companies is likely to be more concentrated). This problem is less serious for occupations for which a large fraction of ads are likely to be posted online. This is the case, for example, for jobs related to IT, which are found to be more represented than other occupations in web-collected job ad data and are represented by very large numbers of ads in the OJA database.

The described limitations of the data, combined with the novelty of the applied methodology imply that the results currently obtained with these data are of an experimental nature. In the coming years, a further improvement of the quality of the OJA data is foreseen. This will be combined with more detailed work on quality assessment and reporting in the use of web data for statistical production (including a set of quality indicators), with the intention of aligning the OJA database with the standards of official statistics.

4 Methodology

4.1. Labour market definition

Calculating concentration in labour markets requires a market definition. In our study, based on the approach used by (Azar et al., 2020), a labour market is defined as the intersection of three dimensions: time (quarter), occupation (ISCO level four) and geography (FUA). This definition of labour market it is one of many alternative market definitions, suited for our methodological approach. The HHI is calculated for each identified labour market, by computing ads shares by employer using the company name as the employer unique identifier.

4.1.1. Time

The quarter was used as time unit for the calculation of firms' ad shares. This means that the market share is calculated for each quarter within a year, based on all the ads posted in that given quarter (with the exception of ads classified as internships or traineeships, which are excluded from the calculations). This three-month period analysis suits the pace of the quarterly releases of the OJA dataset and the typical measurement of transitions in labour markets (i.e. see dedicated Eurostat's experimental statistics website (Eurostat, s.d.)). In addition, the trimester coincides with the period of aggregation used by Azar et al. (2020).

The usefulness of a concentration index based on a three-month period probably depends on the type of job search. The median duration of unemployment was about 10 weeks in 2016, suggesting that a period between two and three months provides a good approximation of the duration of the job search of an unemployed person. However, an employed person looking for a change for employer may have a different time horizon, browsing the web for six months or one year, or even just waiting for the right opportunity for an undefined time until it comes. This paper does not explore the impact of alternative time aggregations on the HHI estimates. However, an approximate idea of this impact can be based on Azar et al. (2020), who found that the average HHI level in the United States in 2016 was 26% lower when using semesters as the time unit than when using trimesters.

To compute the reference quarter of an online job advertisement the variable '*grab_date'* is used. For example if a job advertisement has a value '*grab_date*' = 15/02/2020 (i.e. the advertisement was fetched from the web on 15/02/2020), then the advertisement will be counted for the index of Q1 2020. Grab dates indicate the day in which a job ad was scraped, which is generally different from the one in which the ad was posted online.

As already mentioned, the OJAs dataset contains data from the third quarter of 2018 to the last quarter of 2020. The web scraping process described in Section 3.1 section has improved iteratively over time since its start. Due to this, the data for 2018 is affected by a less mature scraping method (e.g. less sources scraped) and it is not included in the analysis of this paper. The index is calculated for eight quarters, covering the two-year period 2019-2020.

4.1.2. Geography

Functional Urban Areas (FUAs) are used to define geographic labour markets. A **functional urban area** consists of a city and its commuting zone (European Commission; FAO; UN-Habitat; OECD and The World Bank (2021); OECD, 2012). Functional urban areas therefore consist of a densely inhabited city and a less densely populated commuting zone whose labour market is highly integrated with the city. A **city** is a local administrative unit (LAU) where a majority of the population lives in an urban centre of at least 50 000 inhabitants. A **commuting zone** contains the surrounding travel-to-work areas of a city where at least 15% of employed residents are working in the city. In cases where cities are connected by commuting, the functional urban area may consist of multiple cities and their single commuting zone. There are a few cases where cities do not have a commuting zone: for these, the city is equal to the functional urban area. The definition of functional urban areas betters captures local economies, labour markets and commuting networks making this unit of analysis more economically meaningful than the traditionally used regional or administrative boundaries.

More information on the definition of cities, commuting zones and functional urban areas is available in the Eurostat territorial typologies manual dedicated section (Eurostat, 2018). The OECD identifies FUAs with a population of 250 000 or more as “Metropolitan Areas”. Metropolitan regions are NUTS3 regions or a combination of NUTS3 regions which represent all agglomerations of at least 250 000 inhabitants (Eurostat, s.d.).

The location of the job advertised in the ad was determined from the ad text through ontology matching. Ontology matching means that some keywords (or text patterns) are found in the job ad text. The keywords used to find matches in the ad texts were those contained in the database (Geonames, 2021). Not all the job ads in the dataset are compiled with geoinformation (see Table 1).

After finding location classifiers and mapping them (when possible) on the LAU and NUTS classification, this information must be matched to functional urban areas. This task is performed in a two-step approach described below. The correspondence between Local Administrative Units and NUTS region is obtained from Eurostat data. 2018 data are used for the EU local administrative units.

The most granular information available are the variables ‘*city_id’* and *city* (i.e. city name). The variable ‘*city_id’* generally refers to the LAU code of the city, but this is not always the case. When possible, city id codes that do not match the LAU classification have been aligned with Eurostat classifications. By using the city id, each ad is matched to a FUA when the city is part of one.

Where city information is not available, the dataset variable ‘*id_province’* has been used, which contains the NUTS level 3 code of the area where the advertised job is located. NUTS level 3 codes have been used to infer the FUA of an ad in those cases in which a FUA coincides exactly with a NUTS3 area. To do this, the codes of the variable ‘*id_province’* have been changed from NUTS2013 (the classification used in the OJA dataset) to NUTS2016 (the first classification for which Eurostat published the correspondence between NUTS, LAU and FUA areas), according to Eurostat data.

Of the ads for which geoinformation is available, some cannot be matched to a FUA. This could happen for two reasons:

- The jobs are located outside a functional urban area.
- The *city_id* does not match a valid LAU code.

Table 2 shows the percentage of ads which ‘*city_id’* or ‘*id_province’* could be matched to a valid functional urban area, by country. While this percentage exceeds 60 % in most countries, it is at most 10 % in Ireland, Lithuania and Slovakia. Therefore, the data for these three countries should be interpreted with caution.

Table 2: Percentage of ads with a match between the extracted location and a functional urban area, by country

	Percentage of ads
Austria	72
Belgium	69
Bulgaria	76
Cyprus	49
Czechia	58
Germany	85
Denmark	59
Estonia	41
Greece	49
Spain	78
Finland	63
France	56
Croatia	40
Hungary	70
Ireland	4
Italy	77
Lithuania	7
Luxembourg	100
Latvia	75
Malta	99
Netherlands	82
Poland	69
Portugal	50
Romania	28
Sweden	64
Slovenia	48
Slovakia	10

4.1.3. Occupations

For the purpose of our analysis the ISCO level 4 code (e.g. *2313*: computer programmers) is considered to be a reasonable baseline to define a labour market. According to (Eurostat, 2021), this choice can be deemed conservative in that the ISCO level 4 code is likely too broad, and therefore labour market concentration will tend to be underestimated. However, it can be counter-argued that a typical worker could apply for jobs in different occupation groups. An example could be that of a research-trained economist that could look for job as an economist (*OC2631*) but also as a university and higher education teacher (*OC2310*), research and development manager (*OC1223*), statistical, mathematical and related associate professional (*OC3314*), or policy administration professionals (*OC2422*). In view of the different arguments related to the definition of the labour market, it was decided to keep the ISCO level 4 code as the relevant labour market for this study, but excluding that, this would lead to a serious underestimation of demand concentration in the labour market. The occupation code of each advertisement is stored in the variable idesco_level_4 of the OJA dataset.

The occupation has been determined based on the text of the job ad based on ontology matching and a machine-learning model. The list of keywords for ontology matching comprises all the keywords used by the ESCO classification (European Commission, 2021) to describe occupations, augmented with keywords suggested by human reviewers or validated by human reviewers after being identified through automatic search processes – Word2vec and Latent Dirichlet Allocation (Mikolov, Chen, & Corrado, 2013; Blei & Ng, 2003). Job ads that could not be classified through the ontology matching were classified through a supervised machine learning model based on word frequencies within the ad (naive Bayes - see e.g. Galindo (2008)), trained on a manually coded gold set of 70 000 job ads (of which 60% were used for training, 20% for testing and 20% for evaluation). The weighted precision of the whole classification model (ontology matching and machine learning) was 80%, meaning that 80% of the observations were accurately classified in the evaluation dataset, after applying weights proportional to the estimated share of ads in each occupation.

4.1.4. Companies

Once the labour market is defined, a crucial factor in computing the indicator is the calculation of the market share for each company in the market. For calculating market shares, it necessary to know the company that is posting the job ad, i.e. the prospective employer looking for an employee. Out of all the ads for a given labour market, the ones that are coming from the same employer are grouped together to build up the market share of that given company.

The variable *'companyname'* in the dataset is the one that allows identifying the company posting a vacancy. However, due to the residual noise present in webscraped data, the variable companyname contains ambiguous string that required a thorough dataset cleaning. In practice what is recorded in the dataset is a self-reported denomination provided by the entity that posted (or re-posted, in some cases) the job ad, presenting the following challenges:

- The advertiser's name can correspond to an actual employer but also to a job agency recruiting on behalf of the employer, a job portal or a generic string (e.g. "confidential")
- Different names can be used for the same employer, for example if abbreviations are used or if a division of a company posts an ad under its own name.

Due to the great importance of the process of cleaning company names and to the many challenges involved, a more detailed description of the methods used is presented in Annex I – Company names.

COMPANY NAMES SPELLING AND VARIANTS

It is frequent to see in the '*companyname'* variable multiple ways to refer to the same company. For example, it is easy to find strings such as "ABC" or "ABC s.a.r.", "ABC company", referring to the same company but with different spelling (depending on how the name of the company was written in the ad fetched from the web). It is also common to find different symbols, characters or cases in strings referring to the same company.

In addition, different branches, local units or franchisees of a same company can post ads under their own names. The name of the prospective employer that is seeking to recruit the worker(s) through the job advertising could correspond to a company, a branch or division of a company, or a holding group, depending on what is the level at which the post is advertised, and possibly representing one (e.g. company branch) or distinct (e.g. franchisee) legal units. For example, different franchisees of the same company could be present in a given country (e.g. "XYZ Madrid 1" and "XYZ Madrid 2"), or a division could advertise under a different name than its owner company (e.g. Z being the chemical division of a mining conglomerate Y). In this case, it is assumed that branches of the same company are not "competing" between them and therefore are consolidated under the same company name (i.e. XYZ). In the future, efforts will be spent investigating the possibility of linking company names to business registers, which would substantially improve the quality of the information collected.

To deal with all these issues, basic cleaning operations are applied to the strings in the '*companyname*' field of the dataset (e.g. convert to lower case, delete punctuations, symbols and white spaces). In addition, a dictionary of companies has been developed (starting with international companies with very large number of ads) where several variants of the same company name are listed so that they can be identified in the dataset and attributed to the original company name. The original/master name is chosen based on the most frequent and simpler version of the name of the company. This dictionary is evolvable and can be improved with new company names for future versions of the indicator.

INTERMEDIARY AGENCIES

Many OJAs, usually posted by agencies or intermediary companies, do not include the name of the company that has the actual job openings a.k.a. paying company. Sometimes it is the paying company itself that do not want its name to be disclosed. The concept of "intermediary agency" used by our methodology is quite wide as it includes all sorts of company names that can appear on ads scraped from the internet but that clearly do not identify the company having a job vacancy. This definition includes job portals (e.g. monster, expatjobs, etc.), recruiting agencies (e.g. Adecco, Manpower, etc) and online job boards. Correctly identifying intermediary agencies can be complex. In general, agencies are entities that hire workers on behalf on a company where the worker will actually work. However, there are some exceptions. Many consulting companies hire staff (possibly on a permanent basis) to be leased to customers while maintaining a relation with the management in the hiring company – these were not flagged as agencies. In addition, some companies provide some HR services, including recruitment, as a side activity, while their main activity is something else (e.g. generic consulting) – these were also not flagged as agencies.

Our methodology aims at identifying these companies and flagging them as intermediary agencies, so that then their job ads can be trated for analysis (imputed or dropped). A two-step method has been implemented to identify intermediary agencies based on [i] ontology matching (i.e. keywords list) and [ii] a classification tree machine-learning model using regression-based rules. Once identified, the '*companyname*' variable of the ads coming from intermediary is coded as missing, and these ads are treated for the analysis in the same way as other ads with missing '*companyname*'.

MISSING COMPANY NAMES IMPUTATION

The initial OJA dataset presents missing values in the '*companyname'* field (i.e. cases where the advertisements downloaded from the web did not contain any indication on the company advertising the job). In addition, the missing values are complemented with the company names that have been recoded to 'missing' because classified as posted by intermediary agencies.

The HHI is calculated using two different approaches to deal with missing data, which yield a lower and upper bound for the estimates. First, the missing data are imputed under the assumption that every missing companyname value corresponds to a unique company name (from a company that posted one single ad). This method is likely to introduce a downward bias to the HHI calculation (i.e. more company names therefore lower values of the HHI). Second, the missing data is dropped from the dataset – which is equivalent, in the calculation of the index, to impute it by re-assigning it to the companies in the dataset proportionally to the companies' number of ads. This second method is likely to introduce an upward bias in the calculation of the indicator, because if all the missing data were available, it would be probable to observe that part of it belong to different companies.

4.2. Herfindahl-Hirschman Index (HHI)

The Herfindahl–Hirschman Index (HHI) is a commonly accepted measure of market concentration calculated as the sum of the squares of the market shares of each firm competing in a market. The HHI ranges from close to 0 under perfect competition to 10 000 in monopoly/monopsony (i.e., 100% market share). The lower the index, the more competitive (or less monopolistic) the market is. For example:

- For a job market with five companies advertising jobs, each one with a 20% share of the total job ads, the HHI will be equal to (20x20 + 20x20 + 20x20 + 20x20 + 20x20) = 2 000
- For a job market with five companies advertising jobs, one representing 80% of the total ads and the other four accounting for 5% each, the HHI will be equal to (80x80 + 5x5 + 5x5 + 5x5 + 5x5) = 6 500

In a situation of equally shared markets, the HHI values is equal to 10 000 divided by the number of companies competing in the market. In the first example given above, a market that is equally shared by five (5) competing firms has an HHI of 10 000 / 5 = 2 000. Similarly if the number of companies increases to 10 (while the market remains equally shared among them), the HHI value will halve to 1 000. In contrast, in case of different market shares with the same number of companies, the HHI values changes. Table 3 provides some examples of markets with the same HHI level but different share distributions.

Table 3: Examples of markets with the same HHI level but different share distributions

Herfindahl-Hirschman Index value	Market shares' distribution in the special case of equal shares	Other example of market shares' distribution
0	Perfect competition between a large number of firms posting very small shares of the ads in the same labour market	(There is no other valid example)
1000	10 companies with equal shares (10%) of the ads in the labour market	2 companies with a share of 20% each, 18 smaller competitors sharing the rest of the market
2500	4 companies with equal shares (25%) of the ads in the labour market	1 dominant firm with a share of 40%, 4 competitors with a share of 15% each
5000	2 firms posting 50% of the ads each	1 dominant firm with a share of 70%, 9 smaller competitors sharing the rest of the market
10000	Monopsony of one firm posting 100% of the ads in the labour market	(There is no other valid example)

The HHI thresholds used to classify market concentration are not fixed and can vary across applications. The empirical literature defines HHI < 1 000 as the threshold for low levels of concentration and HHI > 1 800 as highly concentrated markets. In the US, the agencies generally consider markets in which the HHI is between 1 500 and 2 500 points to be moderately concentrated, and consider markets in which the HHI is in excess of 2 500 points to be highly concentrated (US Department of Justice, 2018).

In this working paper, the HHI is calculated based on the share of vacancies of all the firms that post vacancies in that market, with a market defined as a triplet of time, occupation and urban area. The market share of a firm in a given market is defined as the number of vacancies posted by the given firm in that market divided by the total number of vacancies posted in that market. The HHI is then aggregated through arithmetic averages at levels suitable for the analysis, for example:

- The HHI for urban area A and year T is the arithmetic average of the HHI in area A, calculated across all occupations and across all quarters of year T
- The HHI for the European Union at quarter Q is the arithmetic average of the HHI in quarter Q, calculated across all occupations and across all urban areas

Some of this paper's tables also report the “equivalent” number of firms in the market. This is the hypothetical number of firms found in a market with a given HHI level and in which each firm has the same share, and it is calculated by multiplying the inverse of the HHI by 10 000.

4.3. Sensitivity to methodological assumptions

Working with new sources of data (“big data”) often calls for greater attention to the underlying assumptions used for statistical productions than to traditional measures of statistical uncertainty like standard errors. One reason is that standard errors are typically very small when using large number of observations (Whitaker, 2018), potentially leading to an unjustified sense of confidence in the results (Cox, Kartsonaki, & Keoghc, 2008). A second reason is that new sources of data are not a representative sample, so that there is no population to which the inference derived through standard errors could apply (Cox, Kartsonaki, & Keoghc, 2008). The online job ad dataset contains all the ads from the set of job ad sources that is deemed sufficient to cover the job market, and therefore it is more correctly conceptualised as conveying information on a population than on a sample. For example, some online job ads may be excluded from small specialised platforms may be excluded from OJA data, but this is a consequence of the design of the data collection, and not of sampling.

In contrast, it is very important to assess the impact of at least some of the assumptions underlying the statistical methodology. This subsection assesses the impact of two key assumptions for the calculation of the HHI:

- The imputation of missing company names as unique names, which is compared in Table 4 to excluding missing data
- The aggregation of the HHI through arithmetic means, which is compared in Table 4 to using weighted means where the weights are proportional the total number of ads found in each occupation.

The comparison between arithmetic and weighted means is interesting because the two estimators may be more affected by different sources of bias. Using an arithmetic average gives an equal weight to all occupations, which is a way to reduce the bias induced by the different distribution of occupations among OJAs, compared to the general labour market. Figure 10 illustrate this clearly, by showing for example that “software developers” is the most common occupation found in OJAs. On the other side within each occupation, estimates of concentration suffer from a bias due to the fact that a fraction of vacancies are not posted online (as mentioned in Section 3). Since this fraction is likely to be smaller for the most represented occupations in the OJA dataset (e.g. software developers), the weighted average of the HHI is likely to be less affected by this source of bias.

Table 4: **EU Average HHI calculated through different methodologies for year 2020.**

	Aggregation: Through arithmetic averages Missing data:		**Aggregation: Through weighted averages Missing data:**	
	1a. Imputed as unique companies (baseline methodology)	**1b. Dropped from analysis**	**2a. Imputed as unique companies**	**2b. Dropped from analysis**
Correlation with baseline	1.00	0.94	0.94	0.94
Average	5748.08	7067.55	3189.30	4717.15
Median	4187.00	5714.75	1460.65	2825.07
Lower quartile	3235.25	4417.50	800.18	1704.59
Upper quartile	5824.25	7117.25	2944.99	4771.14

Table 4 shows that using different assumptions for calculating the HHI can lead to substantially different results (a conclusion already reached by Azar et al. (2020)). In reading Table 4, it is useful to keep in mind that dropping missing data makes the market seem more concentrated, imposing an upward bias on the estimate; and using occupation weights means that the estimate reflects occupations where there are more ads, lowering the HHI estimate. Therefore, it is particularly interesting to compare the average results in the first column, reporting the baseline methodology (1a), with the second and third columns, representing the upper bound (1b) and the lower bound (2a) to the index, respectively. This exercise shows that the lower bound of the average level of the HHI is 45% lower than the baseline estimate, while the upper bound is 23% larger. The methodological choice particularly affects the lower part of the distribution, with the bottom quartile of the lower bound equal to just about one fourth of baseline estimate bottom quartile.

Despite the substantial differences in terms of the absolute HHI levels, there is a good level of correlation between the results obtained with the baseline methodology and those obtained with each of the other three methodologies (the three correlation coefficients are all equal to 0.94). This gives some confidence that the results allow for a robust comparative analysis of the HHI across European urban areas.

This methodological comparison also allows drawing one substantial conclusion: whatever the threshold used to characterise a market as "concentrated”, i.e. from 1 800 to 2 500 (see the beginning of this subsection), this is exceeded by the lower bound (2a) of the average and the upper quartile of the HHI across all urban areas. While this paper will take a comparative perspective in the analysis of the results, it is fair to say that there is evidence of limited competition among firms in a substantial fraction of European urban labour markets, consistently with similar evidence provided for the US (Azar et al., 2020).

4.4. Implementation of the methodology in R

The code underlying the methodology is written in R and it is available for scrutiny in a dedicated GitHub repository. The code is based on the work within the ESSnet Big Data II project on Online Job Vacancies (ESS, 2020).

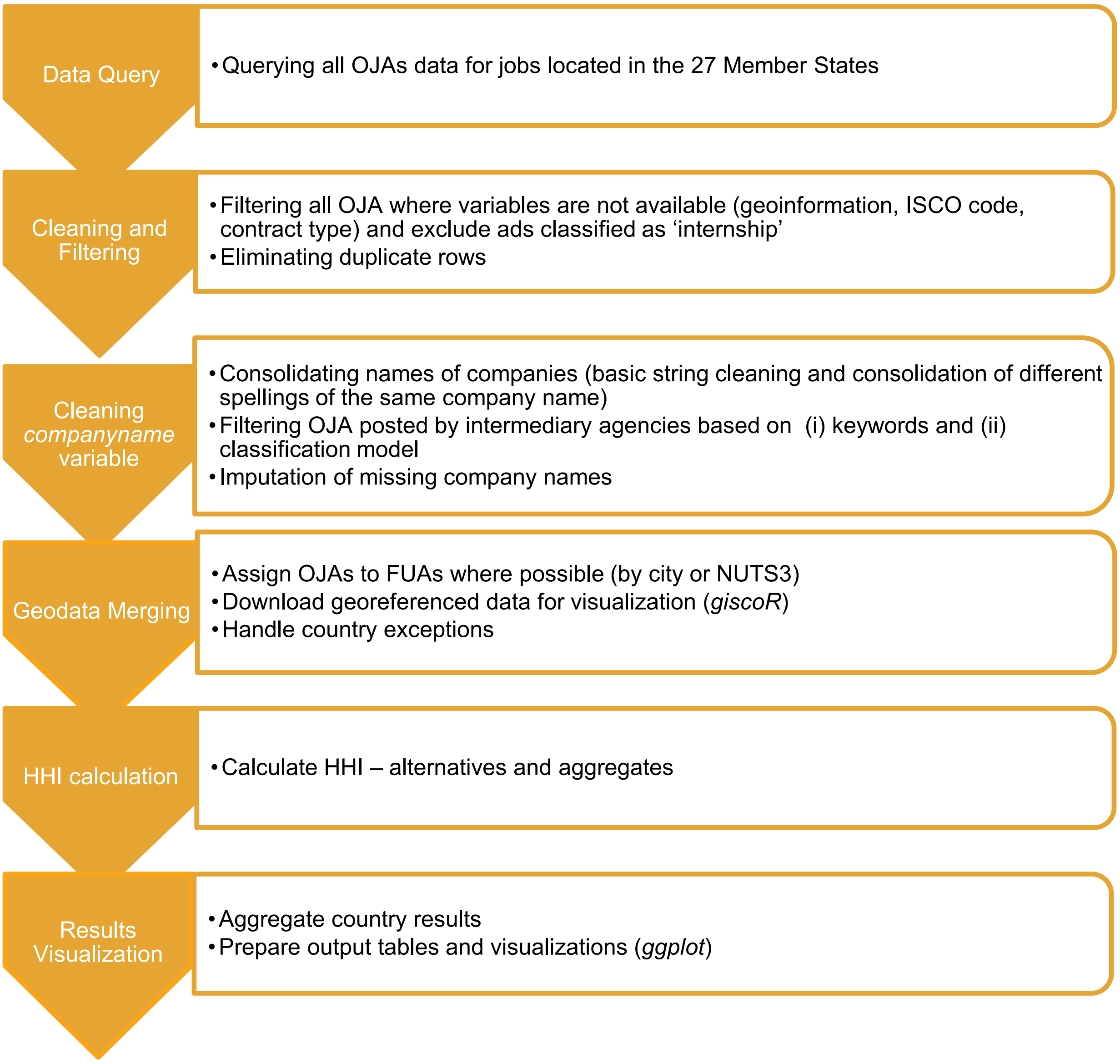
As explained in Section 2, OJAs are availabe at Eurostat thanks to a cooperation with Cedefop. Access to these data can be granted to interested users on a case-by-case basis, following a formal request. Enquiries can be directed to ESTAT-WIH@ec.europa.eu.

The execution of the code is parallelized on three cores to reduce the time needed to process the data for all the 27 countries. The parallel computation is used in several phases: data query, calculation of the index and the production of the maps.

The visualization of the results performed by the R code (using the *ggplot* library) are not used for the purpose of this paper. The graphs and maps in this paper are reformatted respecting the graphical standards of Eurostat's publications.

Future work will focus on streamlining the code, with appropriate documentation, making it fit a regular production of the indicator. The overall workflow can be summarised by the simplified block diagram below.

Figure 2: Simplified view of the methodological workflow



Data Query

- Querying all OJAs data for jobs located in the 27 Member States

Cleaning and Filtering

- Filtering all OJA where variables are not available (geoinformation, ISCO code, contract type) and exclude ads classified as ‘internship’
- Eliminating duplicate rows

Cleaning *companyname* variable

- Consolidating names of companies (basic string cleaning and consolidation of different spellings of the same company name)
- Filtering OJA posted by intermediary agencies based on (i) keywords and (ii) classification model
- Imputation of missing company names

Geodata Merging

- Assign OJAs to FUAs where possible (by city or NUTS3)
- Download georeferenced data for visualization (*giscoR*)
- Handle country exceptions

HHI calculation

- Calculate HHI – alternatives and aggregates

Results Visualization

- Aggregate country results
- Prepare output tables and visualizations (*ggplot*)

5 Results

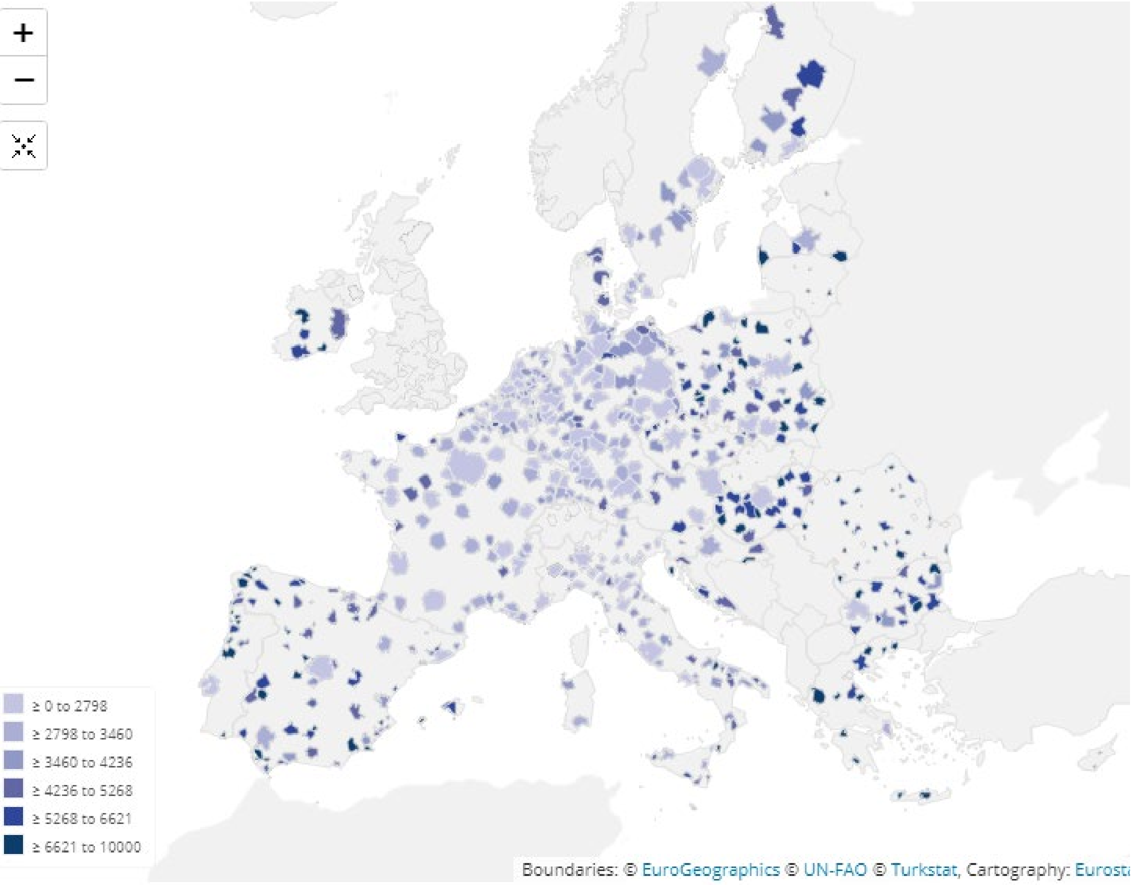
In this section, the HHI values are calculated for each functional urban area, and are compared across urban areas and -in an aggregate form- across countries. The concentration indices are correlated with other variables collected by Eurostat at the urban-area and country level, namely the change in population / migration, employment rates and survey data satisfaction with personal job situation in urban areas. Finally, the evolution of the aggregate HHI over the considered timeframe (2019-2020) is analysed, and a glimpse of the recurring occupations (at 4-digit ISCO code level) is provided, together with their corresponding average concentration index.

The map shown below (Figure 3) explores the HHI, a measure of market concentration ranging from 0 (extremely low) to 10 000 (extremely high), for all urban areas of the European Union (EU) for which data are available for 2020. The index, averaged across all occupations, shows very low levels of concentration in large urban areas with thriving labour markets like Berlin, Milan and Paris (indicating that job-seekers in these urban areas tend to have more online ads to consider). In contrast, there tends to be less choice of employers (as indicated by a higher level of concentration) all along the southern and eastern periphery of the European Union (Greece, Lithuania, Romania, Portugal and other countries and regions), particularly in smaller towns. These results are discussed in more detail in the next sub-section.

Given the novelty of the data source and the methodological assumptions discussed above, the concentration values of the EU urban labour markets needs to be considered with care. Changing some of the methodological assumptions can affect the absolute values of the HHIs, but has a limited effect of the relative position between FUAs. Therefore, the greatest value of the results presented in this paper lies in the opportunity to compare different urban labour markets in terms of their labour market concentration.

The dataset with the results discussed in this paper has been published as an Online Annex at the following link: https://github.com/eurostat/oja_hhi/tree/main/Results. This file includes the data contained in the tables and charts of this paper, as well as the full list of HHI values for each urban area, broken down by quarter. In addition, an interactive map for a better visualization of the results by urban area and quarter for the period 2019-2020 is available at this link: https://ec.europa.eu/eurostat/cache/RCI/rcit/lmci.html.

Figure 3: The HHI across functional urban areas in the European Union. Interactive Choropleth map with data for 2019 and 2020.



Source: Eurostat, Web Intelligence Hub (full data accessible at: https://ec.europa.eu/eurostat/cache/RCI/rcit/lmci.html)

Note: the HHI for a given urban area is the arithmetic average across all occupation categories

5.1. Labour market concentration in European urban areas

Berlin has the lowest HHI point estimate among European functional urban areas for 2020: 1 022, equivalent to a situation of 9.8 firms sharing the market equally in each occupation. Berlin is followed by other urban areas of Central and Western Europe with an HHI around or below 1 500: Amsterdam, Brussels, Hamburg, Milan, Munich, Paris, the Ruhr and Stockholm (Table 5). The low value of the HHI in these cities suggests that they are thick markets with fierce competition for human resources among firms, even within specific occupations. In these urban areas, job-seekers during a three-month period of browsing the web (i.e. a quarter of a year) can expect to find ads from around 8 or 9 different companies for the average occupation. As discussed above, this could potentially drive up salaries and improve working conditions, while giving workers more choice over other characteristics of their jobs (e.g. exact location and job tasks).

While large countries in Western Europe (France, Germany, Italy and Spain) have at least two urban areas in the European top 20 (Table 5), other Member States, particularly smaller ones in Southern and Eastern Europe, have no urban areas with a low level of concentration (Table 6). For example, the urban areas with the lowest labour market concentration in Lithuania (Kaunas), Romania (Timisoara) and Slovakia (Nitra) all have an HHI above 5 000 (a level equivalent to having a duopoly splitting the market in equal shares). In these urban areas, job seekers can find online job ads from at most two different companies within one quarter, on average across all occupations.

Table 5: Top 20 functional urban areas in the European Union in terms of average HHI, 2020

Functional Urban Area	Herfindahl-Hirschman average Index, 2020 (equivalent number of companies)
Berlin	1 022 (9.8)
München	1 218 (8.2)
Hamburg	1 238 (8.1)
Paris	1 238 (8.1)
Milano	1 291 (7.7)
Ruhrgebiet	1 367 (7.3)
Stockholm	1 416 (7.1)
Bruxelles	1 489 (6.7)
Amsterdam	1 525 (6.6)
Düsseldorf	1 689 (5.9)
Lisbon	1 692 (5.9)
Köln	1 696 (5.9)
Leipzig	1 728 (5.8)
Lyon	1 739 (5.8)
Utrecht	1 743 (5.7)
Madrid	1 748 (5.7)
Dresden	1 759 (5.7)
Barcelona	1 772 (5.6)
Roma	1 794 (5.6)
Stuttgart	1 796 (5.6)

* in the hypothesis of an equally shared market (for each occupation / quarter)

Source: Eurostat, Web Intelligence Hub (full data accessible at: https://github.com/eurostat/oja_hhi/tree/main/Results)

Note: the HHI for a given urban area is the arithmetic average across all occupation categories and all quarters of 2020

Table 6: Functional urban areas with the lowest HHI in all European Union countries, using online job advertisement data, 2020

Country	Functional Urban Area	Herfindahl-Hirschman Index, 2020	(equivalent number of companies)
Belgium	Brussels	1 489	(6.7)
Bulgaria	Sofia	2 099	(4.8)
Czechia	Prague	2 439	(4.1)
Denmark	Copenhagen	3 198	(3.1)
Germany	Berlin	1 022	(9.8)
Estonia	Harju	4 016	(2.5)
Ireland	Dublin	4 618	(2.2)
Greece	Athens	3 426	(2.9)
Spain	Madrid	1 748	(5.7)
France	Paris	1 238	(8.1)
Croatia	Zagreb	3 509	(2.8)
Italy	Milan	1 291	(7.7)
Cyprus	Lemesos	4 790	(2.1)
Latvia	Rīga	3 214	(3.1)
Lithuania	Kaunas	6 042	(1.7)
Luxembourg	Luxembourg	3 054	(3.3)
Hungary	Budapest	2 443	(4.1)
Malta	Malta	4 894	(2.0)
Netherlands	Amsterdam	1 525	(6.6)
Austria	Vienna	1 808	(5.5)
Poland	Warsaw	2 101	(4.8)
Portugal	Lisbon	1 692	(5.9)
Romania	Timisoara	5 456	(1.8)
Slovenia	Ljubljana	3 701	(2.7)
Slovakia	Nitra	5 133	(1.9)
Finland	Helsinki	2 647	(3.8)
Sweden	Stockholm	1 416	(7.1)

* in the hypothesis of an equally shared market (for each occupation / quarter)

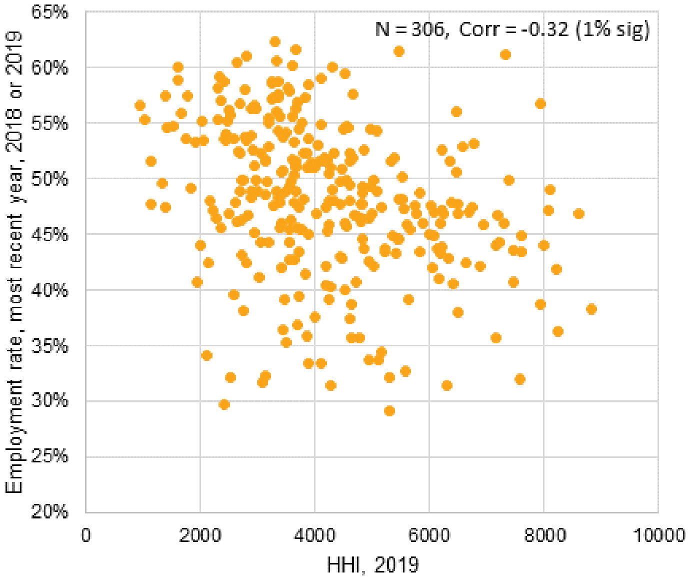
Source: Eurostat, Web Intelligence Hub (full data accessible at: https://github.com/eurostat/oja_hhi/tree/main/Results)

Note: the HHI for a given urban area is the arithmetic average across all occupation categories and all quarters of 2020

5.2. Relationship with other labour market and population indicators

The level of competition / concentration on the demand side of the labour market, as measured by the HHI, is expected to be correlated to a variety of outcomes, such as un/employment rates and wages (Azar et al., 2020), working conditions (OECD, 2020), and people's mobility (Brown and Scott, 2012). It would be naive to expect very high levels of associations between labour market concentration and these outcomes, as social and economic phenomena are determined by a large variety of factors that this analysis will not be able to control for. Yet, observing statistically significant associations between the HHI and some related outcomes serves as a useful validation exercise, increasing confidence in the underlying measure.

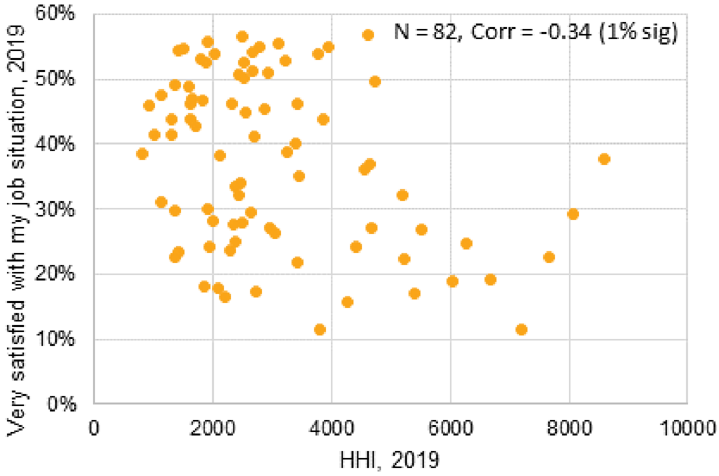
Figure 4: **Relationship between the Herfindahl-Hirschman Index (2019) and the employment rate in European functional urban areas**



Source: Eurostat, Web Intelligence Hub and dataset (urb_llma). Full data accessible at: https://github.com/eurostat/oja_hhi/tree/main/Results

Note: Data on the employment rate refer to the most recent year with available data between 2018 and 2019. The HHI for a given urban area is the arithmetic average across all occupation categories and all quarters of 2019

Figure 5: **Relationship between the Herfindahl-Hirschman Index (2019) and the satisfaction with job situation in European functional urban areas**



Source: Eurostat, Web Intelligence Hub and dataset (urb_percep). Full data accessible at: https://github.com/eurostat/oja_hhi/tree/main/Results

Notes: The HHI for a given urban area is the arithmetic average across all occupation categories and all quarters of 2019

To start with, urban areas with thinner labour market (marked by higher HHI levels) tend to have lower employment rates and lower average levels of satisfaction with the employment situation. Figure 4 plots the HHI level (horizontal axis) for 2019 against the employment rate (vertical axis, source: Eurostat (2021)) for 306 urban areas with data on both variables (out of 587 urban areas for which the HHI was calculated). The correlation coefficient (-0.32) is negative and significant at the 1% confidence level, even though it is not particularly strong (as could be expected, given the large number of other factors influencing employment rates).

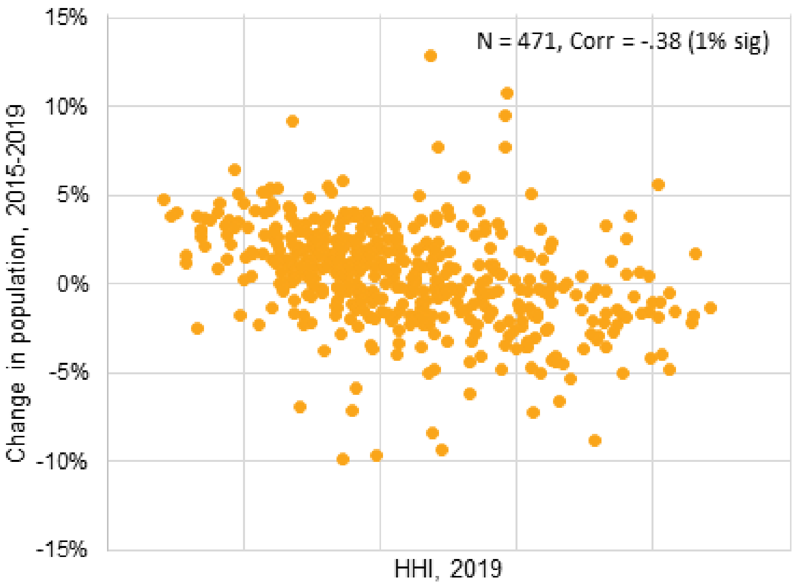
Data on the HHI can be matched to urban-area-level data from the Urban Perception Survey, a survey measuring the local perceptions of quality of life in selected European cities. Figure 5 shows the relationship between the HHI and the proportion of respondents reporting to be "very satisfied" with their personal job situation. The correlation coefficient (-0.34) is again negative and significant at the 1% confidence level, even though it is based on only 82 urban areas with data from the perception survey. This shows that on average, inhabitants of urban areas with thinner labour markets are less satisfied with their job situation.

Employment opportunities are a major pull factor in relocation decisions, attracting workers to regions and cities where they expect to find better jobs (Eurostat and the Netherlands Interdisciplinary Demographic Institute, 2000). Therefore, it can be expected that urban areas with thicker labour markets tend to have higher population growth than other areas, due to higher rates of migration.

It is important to note that a correlation between labour market concentration and population growth or migration patterns falls short of showing a causal relationship. For example, the general level of income (e.g. GDP per capita), are likely to be correlated with both the afore-mentioned variables. This and other factors are likely to induce a correlation between indicators of demand-side labour market competition and population change, independently on the possible existence of a causal relationship between the two. Nonetheless, the fact that the correlation of the HHI with other economic or social outcomes conforms with expectations increases confidence in this experimental indicator.

Figure 6 shows the relationship between the HHI level in 2019 and the change in population between 2015 and 2019 for 471 functional urban areas with available data (see notes under the chart for more details – source: Eurostat (2021)). The correlation coefficient (-0.38) is negative and significant at the 1% level, indicating that urban areas with thinner labour markets (i.e., with high levels of HHI indicating less competition among firms to hire workers) present a relatively low rate of population growth.

Figure 6: Relationship between the Herfindahl-Hirschman Index (2019) and the change in population (2015-2019) in European functional urban areas



Source: Eurostat, Web Intelligence Hub and dataset (urb_lpop1). Full data accessible at: https://github.com/eurostat/oja_hhi/tree/main/Results

Notes: Eight urban areas with values above or below three standard deviations from the mean (i.e., outside the range –16% / +16%) have been excluded from the chart. One of these urban areas is in Germany (Goettingen), all the others in the Czech Republic (Ostrava, Ustí nad Labem, Liberec, Pardubice, Zlín, Karlovy Vary, Chomutov-Jirkov), with values up to -52%. The HHI for a given urban area is the arithmetic average across all occupation categories and all quarters of 2019.

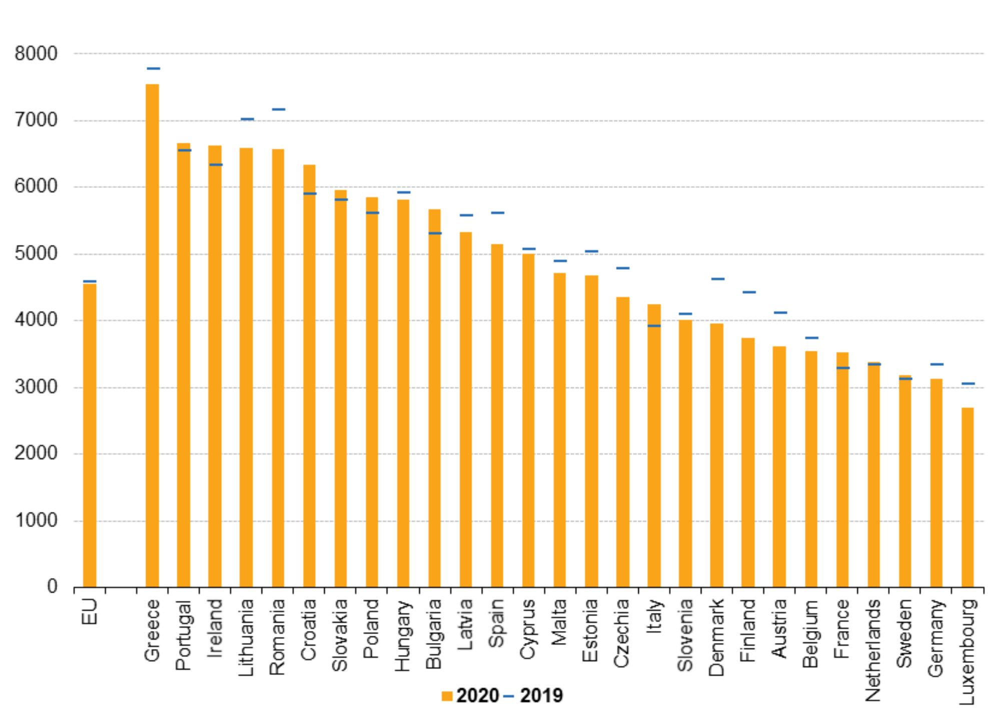
A limitation to the analysis in Figure 6 is that, by looking at overall demographic changes, it puts together migration and other factors, like birth and mortality rates.

To exclude these other factors, Figure 7 and Figure 8 show, for each European Union country, the average of the HHI across all urban areas. This country-level measure is linked to the proportion of citizens living in another European country (Eurostat, 2021), which can be interpreted as an emigration rate.

Luxembourg, Germany and Sweden are the EU countries with the lowest average labour market concentration (around 3 000 or less, on average across all urban areas, in the country in 2020, please see Figure 7). These countries also have strong economies with GDP levels and growth rates well above the EU average in the period 2009-2019 (Eurostat, 2021), which probably helps explain the high level of competition among firms for recruiting.

The high level of the HHI shown in Figure 7 for Ireland, Lithuania and Slovakia could reflect difficulties in the categorization of these countries' data, due to the novelty of the data source. In these three countries, only around 10% or less of the data could be matched with a valid functional urban area code, compared with 65% on average across the other countries. The highest average HHI is in Greece, where it exceeds 7 000, as compared to 4 551 on average across the EU (also reflecting the relatively low prevalence of OJAs as a recruiting channel in this country – see Cedefop (2019)). Romania and Portugal follow Greece with the next highest values of the labour market concentration index. This exemplifies a general pattern of higher labour market concentration in urban areas on the eastern and southern edges of the EU.

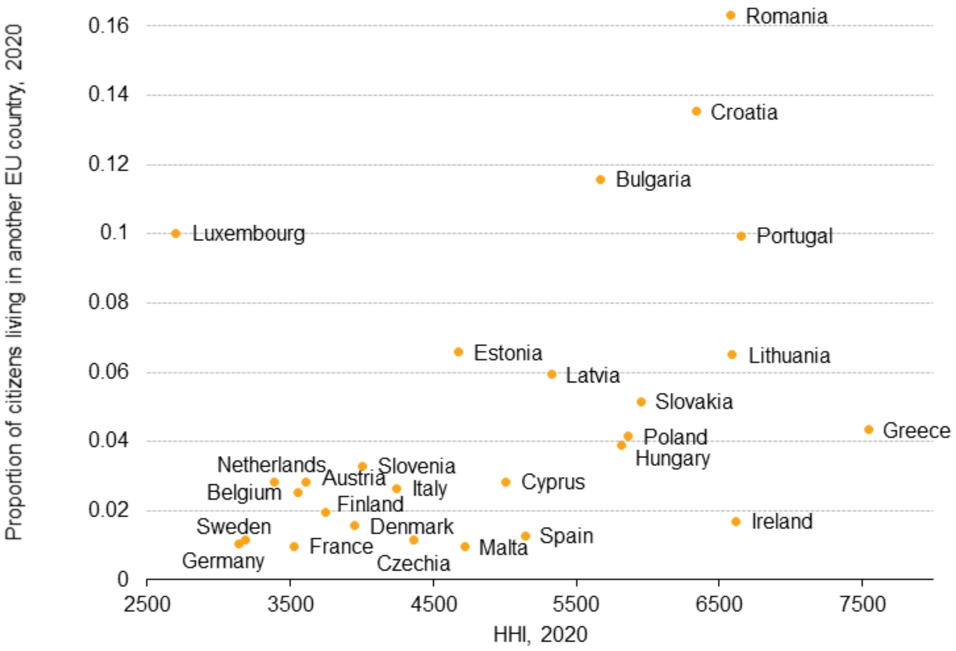
Figure 7: The Herfindahl-Hirschman Index: country averages 2019-2020



Source: Eurostat, Web Intelligence Hub (full data accessible at: https://github.com/eurostat/oja_hhi/tree/main/Results)

Note: The HHI for a given country is the arithmetic average across all functional urban areas, occupation categories and quarters of the same year.

Figure 8: Relationship between the average national Herfindahl-Hirschman Index and the proportion of citizens living in another EU country (emigration rate)



Source: Eurostat, Web Intelligence Hub and dataset (migr_pop1ctz). Full data accessible at: https://github.com/eurostat/oja_hhi/tree/main/Results

Note: The HHI for a given country is the arithmetic average across all functional urban areas, occupation categories and quarters of 2020

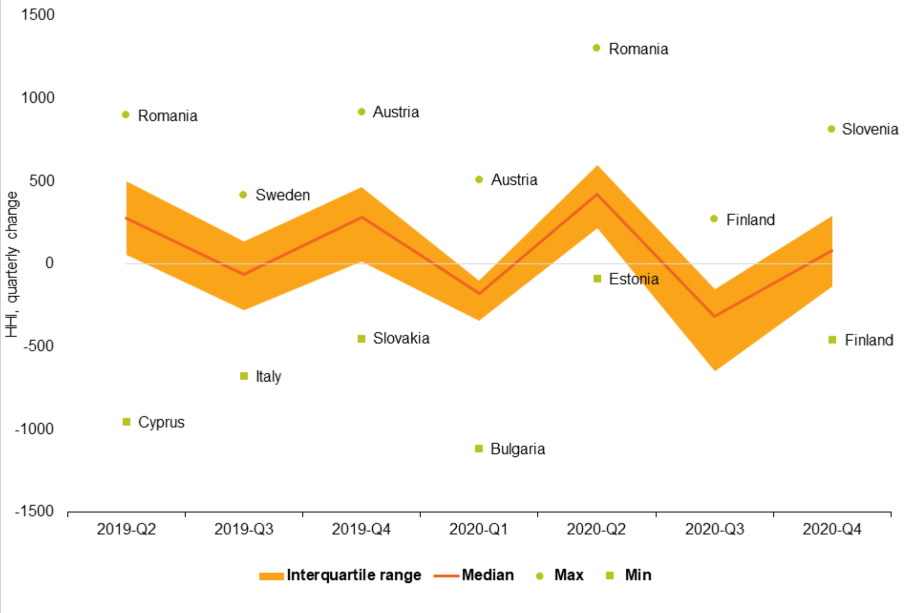
The national HHI averages mirror migration patterns, as shown in Figure 8. Some southern and eastern countries like Bulgaria, Croatia, Portugal and Romania have very large proportions of their citizens living in other EU Member States (emigration rate), while at the same time displaying very thin labour markets, on average across their urban areas. In contrast, countries like Germany and Sweden combine very low emigration rates with thick labour markets (i.e. low HHI levels) across their urban areas. This link between availability of diverse job opportunities (as measured by the HHI) and emigration aligns with expectations, as explained above.

5.3. Evolution over time

There is not a clear, Europe-wide trend in the average level of the HHI across occupations and functional urban areas (Figure 9). The cross-country average of the HHI increased during the second and fourth quarters of both 2019 and 2020, while it decreased during the first and third quarters. In addition, there is not a single quarter in which the HHI increased in all EU Member States or decreased in all EU Member States, as exemplified by the fact that the maximum change is always larger than 0 and the minimum is always smaller than 0.

The largest increase in the cross-country average (+422 HHI points) was observed during the second quarter of 2020. The largest increase for a single country (+1 303 HHI points in Romania) and for the first and third quartiles of the country distribution have been observed in the same quarter. This is possibly related to the job market disruption experienced in many countries and industries at the beginning of the COVID-19 pandemic. A reduction in hiring due to the uncertainty associated with the COVID-19 pandemic would translate into thinner labour markets with a higher concentration of demand. These results are consistent with other analyses on the decline of job advertisements in the first outbreak of pandemics (Cedefop, 2021). However, the short span of the time-series and the methodological limitations (see Section 4) do not yet allow for solid conclusions (De Lazzer & Rengers, 2021).

Figure 9: Evolution over time of the Herfindahl-Hirschman Index across EU countries, changes over the previous quarter.



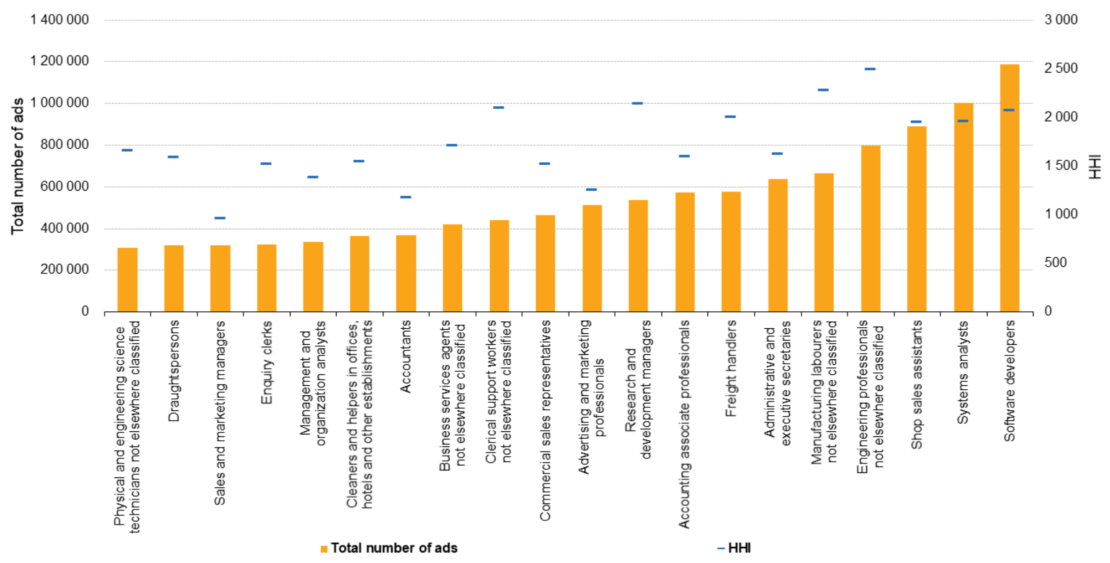
Source: Eurostat, Web Intelligence Hub (full data accessible at: https://github.com/eurostat/oja_hhi/tree/main/Results)

Note: The minimum quarterly change for 2020-Q3 (Austria, -1 579) is not show in the chart. The quarterly change for each country is the difference between the HHI in two consecutive quarters. The HHI for a given country is the arithmetic average across all functional urban areas and occupation categories of the same quarter.

5.4. Labour Market concentration across occupations

While the focus of this paper is on the geographic analysis of data at the urban-area level, the data also lends itself to analysis of occupation categories. Figure 10 illustrates this point by showing, for the 20 most widely advertised occupations in the dataset, the average HHI level across all European FUAs. The chart shows that widely advertised occupations tend to have relatively low levels of HHI. On average across 2019 and 2020, some of the 20 most advertised occupations (with the largest number of total OJAs at EU level) have a very low average concentration, with the lowest levels (below 1 200) observed for sales and marketing managers and accountants. Some of the 20 most advertised occupations have a higher concentration of HHI values (e.g. engineering professionals not elsewhere classified), but never exceeding 2 500 (as compared to an EU average of 4 551, see Figure 10).

Figure 10: The Herfindahl-Hirschman Index for the 20 most advertised occupations in the EU, ranked by total number of ads (2019-2020).



Source: Eurostat, Web Intelligence Hub (full data accessible at: https://github.com/eurostat/oja_hhi/tree/main/Results)

Note: The HHI for a given occupation is the arithmetic average across all functional urban areas and quarters of 2019 and 2020.

6 Conclusions

This study presents a concentration index for online job ads in European urban labour markets in 2019 and 2020, based on over 100 million ads gathered online. The Herfindahl-Hirschman concentration index was calculated under several methodological assumptions, and estimates of the sensitivity of the results to some of these assumptions have been provided. This confirmed that changes in the underlying assumptions could result in substantially different values for the index, but also that these different values are highly correlated across functional urban areas. Therefore, the indicator presented in this paper provides a robust picture of the comparative level of concentration across European urban labour markets, even though it would be premature to define one labour market as "concentrated" (or not) solely based on its score. Given the novel nature of the data source and the methodological improvements still under way, the results of this study must be considered as experimental.

This work achieved two distinct objectives:

- Producing new evidence on firm (i.e. demand-side) competition in European urban labour market. This evidence is based on a new empirical approach (first applied to the United States by Azar et al. (2020)) enabled by two recent developments: the availability of large, granular databases on online job ads (Cedefop, 2019) and the development of new definitions of functional urban areas based on commuting patterns (OECD, 2012), allowing to identify urban commuting areas.
- Advancing the statistical debate on the new methodologies required to deal with new sources of ("big") data (Ricciato, 2019; UN Statistical Commission, 2013), by showcasing an application to a new data source of statistical tools that are not yet commonly used in official statistics (e.g. automatic classification models). Together with this application, the most important methodological challenges and data limitations have been discussed.

This study demonstrates that new data sources have a potential to complement official statistics providing new, timely and detailed information on crucial aspects of our society. Moreover, it is the result of the collaboration between several national statistical offices of the EU showing the importance of sharing experiences and working together to build competences in using non-traditional data sources. Using EU-level data collection from new sources, such as online job ads, would facilitate the creation of harmonised and comparable statistics at European level, possibly reducing in the long term, the burden on EU Member States.

In terms of the evidence on hiring concentration, the results show large differences across urban labour markets. Some urban areas like Amsterdam, Brussels, Hamburg, Milan, Munich, Paris, the Ruhr and Stockholm offer multiple job options to job-seekers even within a detailed occupational field - in other words, they are "thick" labour markets (Gordon & Turok, 2005; Brown & Scott, 2012). However, many other areas, particularly small towns in Southern and Eastern Europe, show worrying signs of "thin" labour markets where there is not much choice of employer for people looking for jobs within their occupational domains.

Limited competition among employers could translate in lower salaries and worse employment conditions (Manning, 2003; OECD, 2020). This is consistent with correlational evidence reported in this paper, showing that people in urban areas with thicker labour markets tend to be more satisfied with their employment situation and more likely to work at all. The possibility to choose between many jobs could also be a migratory pull factor. This again is consistent with correlational evidence presented in this paper that emigration rates are higher in areas with thinner labour markets, while population growth is stronger in areas with thicker labour markets. This cross-sectional, univariate correlation analysis does not provide evidence of causal relationships, but provides some external content validation for the index developed in this study (i.e. it confirms that it is associated to other indicators according to expectations).

With regards to the challenges, this paper discusses a number of limitations including representativeness of the data, comparability over time and the accuracy of the automatic classification models. The discussion on the cleaning, coding and classifying of the variable recording the name of the advertisers (i.e., *companyname*), which is crucial for the determination of the market shares used to calculate the HHI, provides an example of the extensive work needed to prepare and clean the data for a specific application. Future work will focus on developing a quantitative assessment of the data source quality, including a validation of the data coming out of the processing pipeline.

The novelty of the data source and of the methodology described above imply that this study can be improved in different ways. For example:

- Improve the cleaning and classification of employer names. Currently the keywords list of intermediary agencies is refined only for a limited number of countries/languages (i.e. German, Italian, Portuguese, Romanian and Slovenian). The intermediary agencies in the remaining countries are identified by using some EU generic keywords in English language, and then by using the agencies so identified to train a decision tree machine learning model (see Annex 1). In a similar way, different name spelling referring to the same employers are consolidated through an algorithm fed with a sample of alternative spelling variants of the names of large international employers. Including more languages and employer names in the data processing would improve the quality of this crucial variable.
- Calculate the index using a broader definition of labour market. This can be achieved by changing one or more of the three variable that define the labour market: (1) Extend the period of analysis to the entire year (instead of quarter), (2) use the three-digit ISCO code for identifying occupations (instead of the narrower four-digit code), (3) use entire NUTS3 regions instead of FUAs.
- Analysis of the results. Matching the labour market concentration with other variables collected by Eurostat. While using FUA as the unit of analysis remains the most appropriate methodological choice, calculating the HHI also at the level of NUTS3 or NUTS2 regions would increase the choice of available indicators for correlation or multi-variate analysis. This will include finding correlation with other labour market aspects such as wages. For example, due to the specific nature of the data source used (i.e. online job ads), it would be interesting to investigate more in detail the relationship between the HHI and the uptake of IT technologies across regions.
- Extend the scope of the calculation of the index to more countries, such as the United Kingdom (for which data are already available in the dataset) and EFTA countries (data under development).
- Broaden the scope of the analysis by focusing on skills. The high level of detail available from online job ads would allow an in-depth analysis of EU-wide skills requirements, including new and emerging labour market skills, such as "digital" and "green" skills, needed to support the transition towards the ambitious goals of the EU Green Deal.

In addition to improving the methodology, the quality and coverage of the data will be strengthened in several ways. For example:

- A structured, standardised quantitative assessment of the data source quality, including a validation of the data coming out of the processing pipeline and evaluation of the classification models for each variable, is foreseen
- More dates relevant to the appearance online of the job ads could potentially be made available, allowing for more analytical options when dealing with time series
- The coverage will be extended to more countries - data for the UK have already been included in the dataset, and work is in progress to include the other EFTA countries (Iceland, Liechtenstein, Norway and Switzerland).

7 References

Azar, J. M. (2020). Concentration in US labor markets: Evidence from online vacancy data. *Labour Economics, 66*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0927537120300907

Beresewicz, M., & Pater, R. (2021). *Inferring job vacancies from online job advertisements.* Retrieved from https://ec.europa.eu/eurostat/documents/3888793/12287170/KS-TC-20-008-EN-N.pdf/6a86d53e-d0b8-d608-988d-d91f0cef6c21?t=1611673495829

Blei, D., & Ng, A. a. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022. Retrieved from https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?TB_iframe=true&width=370.8&height=658.8

Brown, W., & Scott, M. (2012). Human Capital Location Choice: Accounting for Amenities and Thick Labor Markets. *Journal of Regional Science, 52*, 787-808. Retrieved from . https://doi.org/10.1111/j.1467-9787.2012.00772.x

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

Cedefop. (2019). *Online job vacancies and skills analysis: a Cedefop pan-European approach.* Luxembourg: Publications Office. Retrieved from https://www.cedefop.europa.eu/files/4172_en.pdf

Cedefop. (2019). The online job vacancy market in the EU: driving forces and emerging trends. *Cedefop Research Paper*(72). Retrieved from http://data.europa.eu/doi/10.2801/16675

Cedefop. (2020, May 28). News. *Cedefop and Eurostat formalise joint approach to online job advertisement data*. Retrieved from https://www.cedefop.europa.eu/en/news-and-press/news/cedefop-and-eurostat-formalise-joint-approach-online-job-advertisement-data

Cedefop. (2021). Briefing note - Trends transitions and transformation. *Briefing Note*. Retrieved from https://www.cedefop.europa.eu/files/9157_en.pdf

Cedefop; European Commission; ETF; ILO; OECD and UNESCO. (2021). *Perspectives on policy and practice: tapping into the potential of big data for skills policy.* Luxembourg: Luxembourg: Publications Office. Retrieved from http://data.europa.eu/doi/10.2801/25160

Cox, D., Kartsonaki, C., & Keoghc, R. (2008). Big data: Some statistical issues. *Statistics & Probability Letters, 136*, 111-115. Retrieved from https://doi.org/10.1016/j.spl.2018.02.015

De Hoyos, M. G. (2013). *Literature Review on Employability, Inclusion and ICT, Report 2: ICT and Employability.* Publications Ofice of the European Union, Luxembourg. Retrieved from https://publications.jrc.ec.europa.eu/repository/handle/JRC78601

De Lazzer, J., & Rengers, M. (2021). Auswirkungen der Coronakrise auf den Arbeitsmarkt: experimentelle Statistiken aus Daten von Online-Jobportalen. WISTA - Wirtschaft und Statistik, Vol. 3, pp. 71-89. Retrieved from https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2021/03/auswirkungen-coronakrise-arbeitsmarkt-032021.pdf?__blob=publicationFile

Descy, P., Kvetan, V., Wirthmann, A., & Reis, F. (2019). Towards a shared infrastructure for online job advertisement data. *Statistical Journal of the IAOS, 35*(4), 669-675. Retrieved from https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji190547

Dijkstra, L., Poelman, H., & Veneri, P. (2019). The EU-OECD definition of a functional urban area. In *OECD Regional Development Working Papers* (2019/11 ed.). Paris: OECD Publishing.

ESS. (2020). *ESSNet big data II.* Retrieved from WPB Online job vacancies: https://ec.europa.eu/eurostat/cros/content/WPB_Online_job_vacancies_en

ESS. (2020). *R codes for Labour-market-concentration-index-from-CEDEFOP-data.* Retrieved from https://github.com/OnlineJobVacanciesESSnetBigData/Labour-market-concentration-index-from-CEDEFOP-data

European Commission. (2021). *ESCO Occupations*. Retrieved October 20, 2020, from ec.europa.eu: : https://ec.europa.eu/esco/portal/occupation

European Commission; FAO; UN-Habitat; OECD and The World Bank (2021). (n.d.). Applying the degree of urbanisation: A methodological manual to define cities, towns and rural areas for international comparisons. Retrieved from https://doi.org/10.2785/706535

Eurostat. (2018). Methodological manual on territorial typologies. Statistics Explained. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Territorial_typologies_manual

Eurostat. (2019, May 16). Trusted Smart Statistics Strategy and Roadmap, implementation of the Bucharest Memorandum on "Official Statistics in a datafied society (Trusted Smart Statistics)". *40th meeting of the European Statistical System Committee*. Retrieved from https://ec.europa.eu/eurostat/cros/system/files/item_02_-_background_document_-_essc_2019_40_07_tsssr.pdf

Eurostat. (2021). *EUROBASE.* Retrieved October 14, 2021, from Employment by sex, age and other typologies: https://ec.europa.eu/eurostat/databrowser/view/URT_LFE3EMP__custom_1416050/default/table?lang=en

Eurostat. (2021). *Individuals - Internet Activities.* Retrieved October 14, 2021, from https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_ac_i/default/table?lang=en

Eurostat. (2021). *Labour Market* . Retrieved from ec.europa.eu: https://ec.europa.eu/eurostat/web/euro-indicators/labour-market

Eurostat. (2021). *Labour Market, including Labour Force Survey (LFS) — Overview*. Retrieved from ec.europa.eu: https://ec.europa.eu/eurostat/web/labour-market/overview

Eurostat. (2021, March). Migration and migrant population statistics. *Statistics Explained*. Retrieved July 15, 2021d, from ec.europa.eu: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Migration_and_migrant_population_statistics

Eurostat and the Netherlands Interdisciplinary for Demographic Institut. (2000). Push and pull factors of international migration: A comparative report. Luxembourg. Retrieved from https://www.nidi.nl/shared/content/output/2000/eurostat-2000-theme1-pushpull.pdf

Eurostat. (n.d.). *Eurostat experimental statistics*. Retrieved from ec.europa.eu: https://ec.europa.eu/eurostat/web/experimental-statistics

Eurostat. (n.d.). *History of NUTS*. Retrieved from ec.europa.eu: https://ec.europa.eu/eurostat/web/nuts/history

Galindo, J. (2008). Handbook of Research on Fuzzy Information Processing in Databases. *Hershey: IGI Global,*. https://doi:10.4018/978-1-59904-853-6

Geonames. (2021). *geonames*. Retrieved October 20, 2021, from Geographical database: http://www.geonames.org/

Gordon, I., & Turok, I. (2005). How Urban Labour Markets Matter. In I. H. Gordon, *Changing Cities - Buck I* (pp. 242-264). London: Palgrave.

Green, A. (2017). Implications of technological change and austerity for employability in urban labour markets. *Urban Studies, 54*, 1638-1654. https://doi:10.1177/0042098016631906

Hackl, P. (2016). Big Data: What can official statistics expect? *Statistical Journal of the IAOS, 32*, 43–52. https://doi:10.3233/SJI-160965

Manning, A. (2003). The real thin theory: monopsony in modern labour markets. *Labour Economics, 10*, 105-131. Retrieved from https://doi.org/10.1016/S0927-5371(03)00018-6

Mikolov, T., Chen, K., & Corrado, G. a. (2013). Efficient estimation of word representations in vector space. *arXiv preprint, arXiv:1301.3781*. Retrieved from https://arxiv.org/abs/1301.3781

OECD. (2012). *Redefining "Urban": A New Way to Measure Metropolitan Areas.* Paris: OECD Publishing. Retrieved from https://doi.org/10.1787/9789264174108-en

OECD. (2020). *Competition in Labour Markets.* Retrieved from http://www.oecd.org/daf/competition/competition-concerns-in-labour-markets.htm

OECD. (2020). *Competition issues in labour markets.* Directorate for Financial and Enterprise Affairs. Retrieved from https://www.oecd.org/daf/competition/competition-concerns-in-labour-markets.htm

OECD, & Comission, E. E. (2020). *Cities in the World: A New Perspective on Urbanisation.* (P. OECD Publishing, Ed.) OECD Urban Studies. Retrieved from https://doi.org/10.1787/d0efcbda-en

Ricciato, F. W. (2019). Trusted smart statistics: Motivations and principles. *Statistical Journal of the IAOS, 35*, 589–603. https://doi:10.3233/SJI-190584

Tam, S.-M. C. (2015). Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics. *International Statistical Review, 83*(3), 436-448. Retrieved from https://doi.org/10.1111/insr.12105

UN Statistical Commission. (2013). *Big data and modernization of statistical systems.* Report of the Secretary-General , Economic and Social Council. Retrieved from http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf

UNECE. (2014). *A Suggested Framework for the Quality of Big Data.* Retrieved from https://statswiki.unece.org/download/attachments/108102944/Big Data Quality Framework - final-Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2

US Department of Justice. (2018, 07 31). Herfindahl-Hirschman Index. Retrieved from https://www.justice.gov/atr/herfindahl-hirschman-index

Whitaker, S. (2018). Big Data versus a survey. *The Quarterly Review of Economics and Finance, 67*, 285-296. https://doi:10.1016/j.qref.2017.07.011

Yongdai, K. K. (2013). Big data and statistics. *Journal of the Korean Data and Information Science Society, 24*, 959-974.

Annex I – Company names

The labour market concentration index (Herfindahl-Hirschman index or HHI) is calculated with the purpose of improving our understanding of demand competition in urban labour markets. Ideally, information would be available on the prospective employer for each job advertisement ("ad") appearing in the dataset. In the dataset, the variable most closely associated to this informational need is 'companyname'.

The variable 'companyname' contains raw, unprocessed data. It corresponds to whatever has been filled in by the entity/person posting the ad in the field "employer" (or similar, e.g. "company" or "hiring organisation") of the online form. This can correspond to:

- The name of the prospective employer that is seeking to recruit the worker(s) through the job advertising. The employer name could correspond to a company, a branch or division of a company, or a holding group, depending on what is the level at which the post is advertised.
- A recruiting agency that is looking for candidates on behalf of the actual employer. An example in many countries is "adecco".
- A smaller job portal. The word "smaller" is used here to indicate the fact that the job portal is not among the job advertisement sources included in the landscaping of the data collection, but its advertisements are re-posted by one of the landscaping sources. An example in many countries is "superprof", a platform where teachers and people looking for private tutoring can meet and agree on the delivery of tutoring services.
- A generic text sequence that is not a company name. Example from various countries are: "confidential", probably implying that the entity posting the job ad did not want to disclose its name; generic words that mean "company"; phone numbers or other numeric and non-alphabetic codes.

Therefore, to calculate the HHI, a strategy is needed to edit the variable 'companyname' so that it contains only words that can be assumed to be names of prospective employers. This implies identifying names that do not belong to prospective employers (in particular, staff recruiting agencies, which represent the bulk of advertising in many countries).

The main challenge to overcome, in choosing a procedure for an automatic identification of non-employer companynames, has been the lack of a proper training set. Only limited resources were available for human coding, which were employed to code manually companynames with a relatively large number of ads for three pilot countries (Italy, Portugal and Romania, chosen because of the language skills available in the team), to which Slovenia was added later (thanks to the collaboration with the Slovenian National Statistical Institute). For smaller companynames and all other countries covered in the dataset (i.e. all other EU countries and the United Kingdom), no human-coded training data set was available.

The overall approach to overcome the lack of a large training dataset consisted of:

1. Manually classifying (based on desk research) a set of entries of the variable 'companyname' (hereafter, "company names") as employer or non-employer. This was done for company names with at least 100 ads in a sample of 1 mn ads extracted for each pilot country (a total of 690 company names, referred to as "training set" hereinafter). A set of text strings (hereafter, "keywords") used to filter out these companynames from the dataset was developed during this phase.
2. Using the keywords to filter out (i.e., classify as non-employer) company names in other portions of the data (i.e., in other countries and among company names with fewer ads in the pilot countries). Since there are recurrent patterns in the names of recruiting agencies

and job portals, the keywords extracted from the limited portion of human-coded data worked well when applied to smaller companynames and other countries. For example, in a sample of 200 company names randomly extracted for the evaluation of the model (see below), the keyword search correctly classified almost half of non-employer companynames without any false positive case.

3. Identifying data functions allowing to discriminate between employers and non-employers that are based solely on non-employer data. This made it possible to apply these rules also to other countries, because a substantial number of non-employer company names are identified in each country based on the keyword list (step 2). A typical function would be an algorithm that flags a companyname as non-employer if it is very similar to other non-employers in terms of some observed relationships.
4. Re-parameterise these functions for each country and use them to automatically classify companynames.

This approach led to a two-stage model for the classification of companynames, composed of: (1) an ontology model that classifies companies based on a set of keywords; and (2) a decision tree machine learning model that automatically classifies companynames that the keyword search has not already identified as non-employers.

For the decision tree machine learning model, three empirical rules have been chosen based on the Gini impurity index out of a set of potential rules. Each of these three rules is based on the estimation for each country of an empirical relationship for non-employer companynames in the training data set by linear regression. The regressions have been run in two stages, with the half of the observations with the worse fit excluded in the second stage. The following relationships have been estimated, and the following decision-tree rules applied:

- Rule 1: companynames outside the training data set have been flagged as non-employer if they do not lie significantly (at least 1.96 standard deviations) below the curve generated by regressing the log number of distinct occupation codes for a company name on a quartic polynomial function on the log number of de-duplicated ads
- Rule 2: companynames outside the training data set have been flagged as non-employer if they do not lie significantly (at least 1.96 standard deviations) above the curve generated by regressing the total log number of ads for a company name on a quadratic function on the log number of de-duplicated ads
- Rule 3: companynames outside the training data set have been flagged as non-employer if they do not lie significantly (at least 1.96 standard deviations) above the curve generated by regressing the log number of distinct NUTS3 codes for a company name on the log number of 2-digits NACE code, the log number of de-duplicated ads and the interaction between the two independent variables.

Company names outside the training data set (i.e. company names that have not been coded manually or through the keyword search) have been classified as non-employer if they are flagged as non-employer by all three rules, implying that they lie close to all the three curves that have been estimated for non-employer company names. Therefore, the automatic classification model can be described as a three-nodes decision tree (more specifically, a decision list) in which every rule forms a node. If a companyname is not flagged as non-employer at a node, then it is classified as an employer company name by the model. If it is flagged, then the company name goes on to the next node. If the companyname is flagged at all three nodes, then it is classified as non-employer.

The performance of the companyname classification model has been evaluated on a random sample of 200 manually-coded sample of companynames from all countries in the dataset, with sampling probabilities proportional to the number of ads of each companyname. The accuracy rate (i.e. the proportion of cases correctly classified as either employers or non-employers) is 72%. The recall rate (i.e. the proportion of non-employers that have been correctly identified) is 58%.

KS-01-21-430-EN-N

Competition in urban hiring markets: evidence from online job advertisements

This paper provides the first Europe-wide evidence on competition among firms in urban hiring markets. It calculates a labour market concentration indicator (Herfindahl-Hirschman Index) by occupation for every functional urban area (FUA) of the 27 EU Member States, using over 100 million Online Job Advertisements (OJAs) collected from hundreds of job portals in 2019-2020. The results show that across urban areas, hiring market concentration is associated with migration patterns and employment prospects. It tends to be low in large urban areas in Europe (e.g. Berlin, Milan, Paris), indicating a robust degree of competition among employers and more choice for job-seekers across all occupations. In contrast, urban labour markets are thinner all along the southern and eastern periphery of the European Union, particularly in smaller towns. An increase in hiring market concentration across European countries is observed in the second quarter of 2020, when the pandemic crisis hit Europe stronger. These are the first experimental results using OJAs available at Eurostat thanks to the collaboration with the European Centre for Development of Vocational Training (CEDEFOP) and with European National Statistical Institutes. The data and methodology used in this paper are still in an experimental phase, and some potential improvements are discussed in the paper.

For more information

https://ec.europa.eu/eurostat/

Publications Office
of the European Union

978-92-76-43554-9