EUROPEAN COMMISSION
EUROSTAT

Directorate A: Resources
Unit A-5: Methodology; Innovation in official statistics

Experimental Statistics: Labour Market Concentration Index using OJA data

Methodological Note

Luxembourg, August 2021

Contents

1. Introduction

This document presents the methodology underlying the experimental statistics on labour market concentration based on Online Job Advertisements (OJAs) data.

These statistics represent the first experimental results of the Web Intelligence Hub run by Eurostat as part of the Trusted Smart Statistics initiative. These results are to be considered experimental since they are based on a new methodology and new data sources. Like the other Eurostat's experimental statistics [1], they are still in the research and development phase. 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. The main purpose of an experimental statistics is to experiment, seek user feedback, learn from it and use it to improve the statistics themselves.

The purpose of this experimental study is to showcase one of the possible uses of OJAs. We calculate 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. Using OJA data, we calculate Herfindahl Hirschman Indices (HHIs) for labour markets at the occupation (ISCO level 4), functional urban area and quarterly level. The concentration of job advertisements by particular firms within a labor market is taken as a measure of the concentration of labor demand in a market. The concentration of job advertisements in turn can affect negatively the competition in labour markets and thus it may reduce input prices (labour). According to recent studies [2], this might drive down the workers' bargaining power and ultimately wages.

1.1. The Web Intelligence Hub and OJAs

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 predominantly targeting highly skilled workers, today it contains job advertisements for all occupations and skills. In addition to simplifying the process leading to a match between employers and jobseekers, the increasing use of online job advertisements (OJA) portals also has great potential for labour market and skills analysis.

OJAs are a powerful source of information on job requirements, which is not available and is difficult to gather by the current official statistical sources and methods. OJAs do not replace other types of labour market information, on the opposite, provide comprehensive, detailed, and timely insights into labour market trends and allows to identify early new emerging jobs and skills.

The European Centre for the Development of Vocational Training (Cedefop) and the ESSnet Big Data have engaged in parallel projects to assess the feasibility of using online job advertisements (OJA) for labour market analysis and job vacancy statistics. After an initial feasibility study finalised in 2016, Cedefop developed a Pan-EU system providing information on skills demand in OJAs [3]. The ESSnet has focussed on statistics that can be derived from OJAs and in a second phase on creating the conditions for a larger scale implementation of the project.

This pan-EU approach applies methods to collect online job advertisements in all European Union member states and provides experimental data on online job advertisements, as well as experimental data on skills' demand.

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

Eurostat and Cedefop discussed ways and conditions to augment the Cedefop system for the purpose of producing official statistics at European and national level. In its conclusions the ESSnet Big Data team recommended to reflect on the best use of NSIs' time and resources, not underestimating the efforts needed for making OJA data fit for analysis. They recommend considering close collaboration between Cedefop and the ESS to set up a system that would be beneficial to both sides.[4]

Eurostat, representing the European Statistical System, and Cedefop are working towards a joint data production system based on OJAs. Based on a formal agreement establishing the basis for this cooperation [5], Eurostat is building a parallel system that will be able to replace the current system. The new system should be highly 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 discussed the principles of Trusted Smart Statistics and priority areas for producing European statistics from new data sources [6]. This included 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.

Along the production of the relevant methodologies, recommendations, specifications and statistical software, the production of experimental statistics demonstrating the capabilities to produce statistics was one of the objectives of ESS Big Data II.

At least two applications of OJA data that may support and deepen labour market statistics are especially promising: (i) using job offers as a leading indicator of the labour market situation, (ii) providing structural and qualitative information at a highly granular level, for example on skills.

1.2. Previous Work

This experimental study is based on the work done within the Work Package B of the ESSnet project on Big Data II [7] co-financed by Eurostat on online job advertisements. In particular, this study is built on one of the prototypes developed by Destatis [8] as one of the partners of the work package, which in turn is inspired by the academic study by Azar et al [9] on calculating concentration in the US market using OJA data collected by Burning Glass. The methodology used by Destatis for the calculation of the index based on Online Job Advertisement (OJA) data has been adapted and replicated on all the 27 EU Member States.

In a recent publication [10], the OECD provides pertinent contextual and policy-relevant information on competition and the risks of concentration in labour markets.

2. Eurostat Labour market Statistics

Labour market statistics measure the involvement of individuals, households and businesses in the labour market. They cover short-term and structural aspects of the labour market, both from the supply and the demand side, in monetary and non-monetary terms.

The following aspects are covered:

- Labour market situation providing data on employment, unemployment, inactivity, working time, temporary employment, labour market transitions, etc. from the EU Labour Force Survey (LFS);
- Job vacancies;
- Labour costs such as the quarterly labour cost index, labour cost levels, Labour Cost Survey;
- Earnings such as gross and net earnings, gender pay gap, minimum wages;
- Quality of employment.

Infra-annual statistics such as monthly unemployment rates, the Labour Cost Index (LCI) and the quarterly job vacancy statistics (JVS), which belong to the Principal European Economic Indicators (PEEIs) (Labour market data), provide important information for business cycle analysis and policy decisions. Both the unemployment and labour cost indices play an essential role in the compilation of key indicators for the analysis of long-term economic equilibria and the related cyclical adjustments, e.g. through the Beveridge curve (relationship between the unemployment rate and the job vacancy rate) and the Phillips curve (the relationship between inflation and unemployment).

The Eurostat statistics currently are not broken down geographically beyond the national level and by occupation type. Geographical granularity and split by occupation are two potential advantages of Online Job Advertisement data.

For job vacancies, quarterly data on the number of job vacancies and occupied posts collected from businesses do not have a full coverage in some countries (excluding e.g. small businesses, the public or non-business sector, or the education and human health industries) and do not allow producing breakdowns by ISCO and NUTS in most countries.

The purpose of this study is to showcase possible uses of OJAs, without producing an official European labour market indicator and without replacing or substituting in any way current labour statistics.

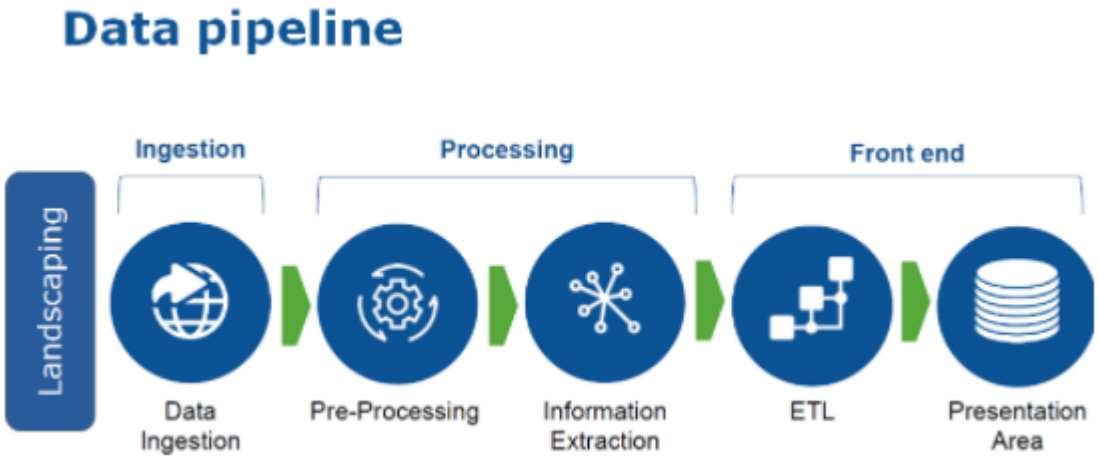
3. Data

3.1. Data source

Online job advertisements (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. This could be motivated by the employer's need to fill a current vacancy, by an exploration of potential opportunities, or other reasons. 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 available only as natural language textual data. This type of big data requires specific methodologies for processing and analysis but also provides much more detailed information (compared to alternative data sources) and avoids pre-conceived classifications (e.g. important to identify emerging skills).

In order to set up the system, the development team has run a **landscaping** phase, where the state of the art on the use of online job advertisements for labour market analysis in different countries was described. Additionally, the OJA data market was assessed and the relevant sources were identified. This is a phase that in a running system will need to be re-run regularly as expectably the data landscape will change over time.

The current system data flow can be summarized in 5 main steps shown in the figure below and described thereafter.



1. **Data ingestion:** includes all the activities related to data collecting. Crawling, fetching, scraping and storing activities are the main tasks of this phase. **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. 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 relevant 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, leading to OJV ads redundancy. The data ingestion phase deals with both structured sources, which store information in structured fields, and unstructured sources, where information is largely extracted from large chunks of text. The data ingestion activity includes the collection of data through scraping, crawling and direct access (e.g. API).

2. **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 3 steps: merging, cleaning and text processing & summarizing in which data is put into a single complete dataset, cleaned from noise, summarised and prepared for information extraction step. The pre-processing starts with the language detection and redirection to a language specific pipeline. Noise originates from pages in the websites which do not refer to job ads, such as ads of training courses and blog posts. These are identified via a combination of simple heuristics and machine learning with a precision of 99% and are eliminated. Summarisation deals with job ads duplicated in several sources. Duplicates can reach 20% of the job ads. These cases are identified and their information is combined (de-duplication). During pre-processing some metadata, such as the release date of the job ad, is extracted.
3. **Information extraction**: related to 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. The classified as well as the extracted variables are listed in the table below, together with the variables that are "extracted" without any classification.
4. **Extraction, Transformation and Loading (ETL**): concerns data preparation for the front-end tool
5. The **data presentation** component of the system will target 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 and BI, visual-based investigation and story telling.

The current system domain include 27 EU countries + UK and will be expanded to the member states of EFTA: Norway, Iceland, Switzerland and Liechtenstein. For this study, 116.851.363

distinct online jobs advertised collect from 316 distinct sources were considered, mainly coming from job search engines and public employment services.

OJA data are released **quarterly**. The timeliness of the data release has been increasing over time, starting with 7 months, in the last release it was already possible to decrease it to 2 months. The production of the experimental statistics presented in this document is based on the most recent version of the dataset (i.e. v9) released during the first quarter of 2021. This dataset contains data from Q32018 to Q42020, however data from 2018 are not used for the calculation of this index.

Table 1: Online Job Advertisement database: Missing data by variable.

Variable	**Missing data (%)**	**Notes**
economic activity of the employer	2%	(NACE at 2. level)
type of contract	29%	divided into "permanent", "self-employed" and "temporary"
working hours	38%	divided between "full-time" and "part-time"
education level required	1%	(ISCED 2011)
salary, classified into 13 levels	74%	divided into 13 levels
experience, classified into 8 levels	51%	divided into 8 levels
place of employment (region)	36%	(NUTS3)
place of employment (city)	47%	(LAU)
occupation	0%	(ESCO 4th level)
skills	1%	classified to skills from the ESCO classification, level 3
Time (grab and expired dates)	0%	extracted variable
Company names	20%	extracted variable

3.2. Quality of the data

OJAs have a huge potential to complement official statistic thanks to their higher timeliness and relevance including the higher level of granularity (location data and occupation-skills information). The real-time nature of job ads data also allows for the early detection of labor

demand trends, which gives job seekers, employers, and policymakers a forward-looking analytical tool. Compared to point-in-time snapshots provided by survey-based labor market data, which rely on random sampling, these data are cost-effective and provide the ability to improve the accuracy of labor market forecasts while producing supplemental estimates of demand within detailed occupations, industries, and geographies. [13]

Eurostat statistics have to meet and serve the users' needs (European institutions, governments, research institutions, business concerns and the public generally) and comply with the European quality standards. Quality of statistical data is effectively assessed with reference to quality frameworks. These quality frameworks systematically address quality by referring to "quality dimensions". The UNECE (2014) quality framework addresses quality issues following an "*input -throughput -output*" model for statistical production and adopts a hierarchical structure (as already suggested for administrative data) where quality dimensions are nested in three hyper-dimensions: *source*, *metadata* and *data*.

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.** Selectivity is strongly related to the **linkability** of web data that refers to linking or combining web data with other data sources. A study on inferring job vacancies (JVS) from online job advertisements (OJA) [14] has been done, attempting to derive proper estimators for the number of job vacancies from OJA data. The differences in the statistical unit and coverage have been taken into account, as well as possible existing auxiliary information and the possibility of using model based estimates and Bayesian inference. For instance, the results of this study are promising for certain countries and industries, but still experimental as the both the time trends and the structure of job advertisements according to country, industry and occupation significantly differ from the ones of job vacancies.

Comparability over time is beyond 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 data may not be directly comparable across time and data releases. 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 of time. This quality aspect will be improved with the introduction of more standardised statistical processes for regularly produced statistical products.

In addition, technological changes, as well as spreading of technology, could affect comparability over time. In addition, as a result of "technological development", more job ads will appear online (for example on social media), because more companies will only hire online. Moreover, technological changes may affect **comparability over time between countries**, because technological evolution is not uniformly distributed across countries.

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 and "**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.

The OJA data may also encounter some **data accuracy** issues, especially due to the use of different versions of standard classifications used, but also due to the complexity and the amount of data that has to be classified in the data processing phase. More precisely, miss-classification may be encountered regarding the occupation (ESCO), skills (ESCO), geo location (NUTS), economic activities (NACE), as well consistency issues between classification levels or data releases.

The OJA data may also be affected by some **consistency issues** that are to be taking into account in the data quality management. These inconsistencies can result from the coverage of sources and/or sites in data collection phase, that may lead to missing information. Additionally, due to a variety of other reasons, such as: scarce information in the ads (i.e. misspelling of company names), 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 data** in the OJA dataset. Thus, for the future use of the time-series, this will be considered in the future releases, for the improvement of the OJA data.

In this experimental study, the issue of missing data is handled using imputation methods, as described more in detail in Section 4. *Methodology.*

While useful in measuring labour demand and honing in on previously inaccessible variables, 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. In some regions digital tools may not be widespread enough to encourage employers to publish job advertisement online. Although this problem is likely to decrease over time as the use of web for job advertising becomes widespread, it is nevertheless a issue that has to be addressed when using OJA.

Online job advertisements data has limitations in terms of representativeness and must be treated with appropriate caution. Key considerations regarding the reliability of OJA data include:

1. The OJA represent a part of the job demand, as not all job vacancies are advertised online.
2. The penetration of OJA markets 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, as the OJA market comprises, in most countries, multiple actors with different business models, resulting in multiple sources of OJAs.
4. The tools used for the processing of the data may be subject to errors even though they use the most up to date technologies.
5. A constant improvement is to be considered of the ontologies developed and used to sort and organise the diverse and complex universe on OJA data.

Some important limitations of the data, which are particularly relevant for the calculation of a labour market concentration index, concern the variable identifying the name of the entity posting the job ad. Ideally, the complete name of the prospective employer initiating the recruitment process would be available for each ad in the dataset. This information would then be used to

compute the market shares on which the labour market concentration index is based. However, 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:

- 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.

Section 4. *Methodology* and *6. Annex* 1 explain how these problems have been addressed in the production of the experimental statistics on labour market concentration. The described limitations of the data source combined with the methodology applied 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

4.1.1. Time

The time unit used for this study is the year's quarter. This means that the market share is calculated for each quarter, 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).

One could argue that variations between quarters are limited in terms of labour market concentration. As a matter of fact is unlikely that the changes in labour market offerings are significant over a 3-month period. However, we decided to keep this three-month period analysis to align with the pace of the quarterly releases of the OJA datasets versions. In addition, the median duration of unemployment is about 10 weeks in 2016 [10] and transitions in labour markets are measured on a quarterly basis (i.e. see dedicated Eurostat's experimental statistics [15]).

To compute the reference quarter of an online job advertisement the variable *grab_date* is taken into considerations. 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.

As already mentioned, the OJAs dataset used contains data from the third quarter of 2018 (Q3-2018) to the last quarter of 2020. The webscraping process described in the "Data source" chapter has improved iteratively over time since its start. Due to this, the data for 2018 are affected by a less mature scraping method (e.g. less sources scraped), therefore we decided not to use them for our analysis. We calculate the index and present results for eight quarters in total covering the two-year period 2019-2020.

4.1.2. Geography

We use Functional Urban Areas (FUAs) to define geographic labor markets. A **functional urban area** consists of a city and its commuting zone. 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 [16]. 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 [17]. FUAs with a population of 250 000 or more are identified as "Metropolitan Areas" by the OECD. Metropolitan regions are NUTS 3 regions or a combination of NUTS 3 regions which represent all agglomerations of at least 250 000 inhabitants.

Not all the job ads in the dataset are compiled with geoinformation. The matching of the OJAs to a FUAs is performed in a two-step approach described below. The correspondence between Local Administrative Units and NUTS region is obtained from Eurostat data. We use 2018 data for the EU local administrative units.

1. The most granular information available are the variables *city_id* and *city* (i.e. city name). The variable *city_id* is based on the LAU code of the city but not for all countries. For some countries, when the id code of the city was not matching the LAU classification, it has been corrected to align it with Eurostat sources. By using the city id, each ad is matched to a FUA when the city is part of one.
2. Where city information is not available, the dataset variable *id_province* has been used, which contains the NUTS3-level code of the area where the advertised job is located. NUTS3-level 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 [18].

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 areas
- The city_id does not match a valid LAU code and it is not possible to correct the problem.

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, data for these three countries should be interpreted with caution.

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

AT	BE	BG	CY	CZ	DE	DK	EE	EL	ES	FI	FR	HR	HU	IE	IT	LT	LU	LV	MT	NL	PL	PT	RO	SE	SI	SK
72	69	76	49	58	85	59	41	49	78	63	56	40	70	4	77	7	100	75	99	82	69	50	28	64	48	10

4.1.3. Occupations

For the purpose of our analysis we consider the ISCO level 4 code (e.g. 2313: computer programmers) to be a reasonable baseline to define a labour market. According to Azar et al. (2020) [10], this choice can be deemed conservative in that the ISCO level 4 code is likely too broad, and

therefore labor 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.

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 firm in the market. For calculating the market share of each company we need to look at the company which is posting the job vacancy, 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 will be grouped together to build up the market share of that given company.

The variable *'companyname'* in the dataset is the one that allow us to identify the company posting a vacancy. However, due to the residual noise present in webscraped data, usually the variable *companyname* is filled with ambiguous string that required a thorough dataset cleaning (see also Section 3.2). Due to the great importance of the process of cleaning company names and due to the many challenges faced, we present a more detailed description of the methods used in *6. Annex* 1.

In particular, the main issues can be summarised by:

4.1.4.1. 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", etc. clearly 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 symbol, characters or different 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 we assume 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* column of the dataset (e.g. convert to lower case, delete punctuations, symbols and white spaces). On top of this, we have developed a dictionary of companies (starting with the biggest and most well-known in three EU countries) 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.

4.1.4.2. 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 an advertisment scraped from the internet but that clearly does 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, as exemplified by the following criteria devised to deal with some particular cases. 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. 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 (thus the company name is not the real name of the company having the vacancy). We implemented a two-step method to identify intermediary agencies based on [i] an 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 set to missing (NA).

4.1.4.3. 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). On top of these, we add to the missing values also the company names that have been recoded to 'missing' because classified as posted by intermediary agencies.

Given the large number of total missing values for *companyname*, we decided to calculate the HH index with two different approaches to deal with missing data, which yield a lower and upper bound for the estimates. First, we impute the missing data 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 HHI index). Second, we drop the missing data from the dataset - which is equivalent, in the calculation of the index, to impute them by re-assigning them 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, we would probably observe that some of them belong to different companies.

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 also represented by very large numbers of ads in the OJA database. Following this reasoning, the average (for each FUA) weighted by the number of ad in each occupation has been calculated, as it is likely less affected by the afore-mentioned upward bias. However, weighting occupations by their number of ads potentially makes the sample less representative, because a relatively large weight is given to occupations that are more likely to have ads online. Therefore, an arithmetic average across occupations at the FUA level is also provided.

4.2. Herfindhal-Hirschmann Index (HHI)

The Herfindahl–Hirschman Index (HHI) is a commonly accepted measure of market concentration calculated by squaring the market share of each firm competing in the market and then summing the resulting numbers. The HHI index 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.

One advantage of the HHI is that it does not only take into account the equality of market shares across firms but also the number of firms in the market. In a situation of equally shared markets, the HHI values gives an indication of the number of companies competing in the market. For example, a market that is equally shared by 10 competing firms has an HHI of 1000. Similarly if the number of companies decreases to 5 (while the market remains equally shared among them), the HHI value will double to 2000.

However, in cases of different market shares markets with the same number of companies can have very different HHI values.

A job market with **five companies** advertising jobs, **each one with a 25% share** of the total job ads, the **HHI index** will be equal to (20x20 + 20x20 + 20x20 + 20x20 + 20x20) **= 2000**

A job market with **five companies** advertising jobs, **with one company representing 80% of the total ads** and the other four accounting for 5% each, the **HHI index** will be equal to (80x80 + 5x5 + 5x5 + 5x5 + 5x5) **= 6500**

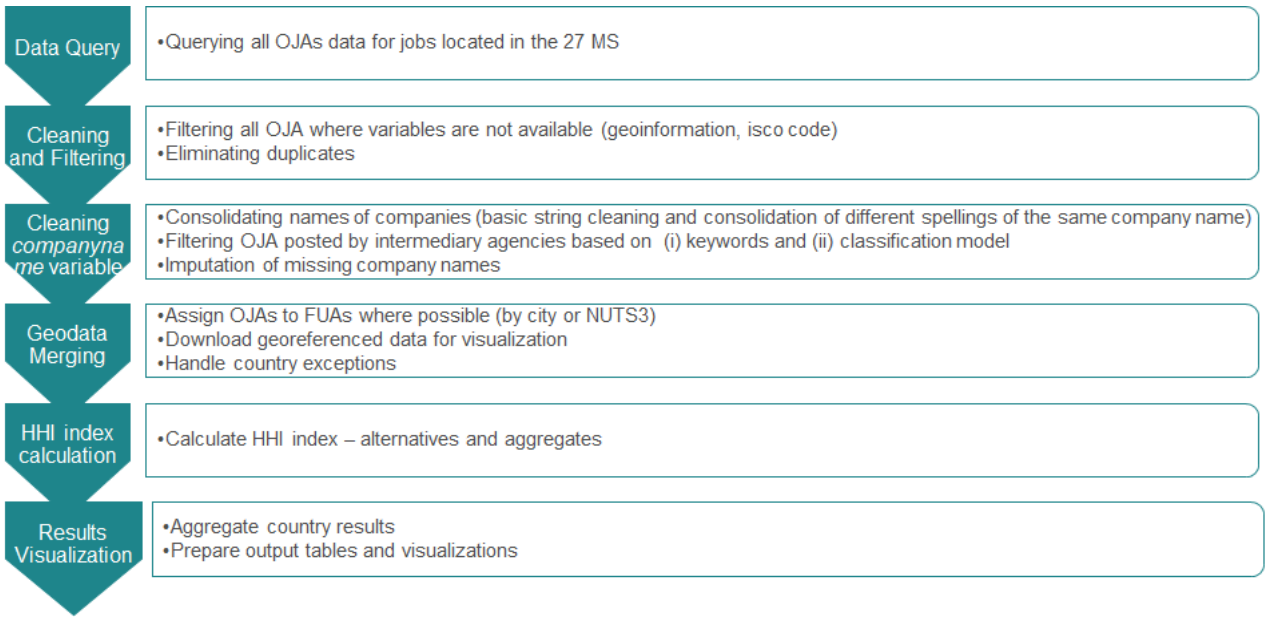
The HHI threshold for defining market concentration are not fixed and can vary. The empirical literature defines HHI < 1000 as the threshold for low levels of concentration and HHI > 1800 as highly concentrated markets [19]. In the US, the agencies generally consider markets in which the HHI is between 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 (https://www.justice.gov/atr/herfindahl-hirschman-index).

In this work the HHI is calculated based on the share of vacancies of all the firms that post vacancies in that market. The market share of a firm in a given market and time is defined as the number of vacancies posted by a given firm in a given market and time divided by total vacancies posted in that market and time. The inverse of the HHI multiplied by 10 000, 10 000/HHI, gives the "equivalent" number of firms in the market, or the number of firms that would result in such an HHI if each had the same share of the market.

4.3. Workflow

The code to explore and process data from the OJA dataset is written in R and it is shared on GitHub: https://github.com/eurostat/oja_HHI

The overall workflow can be summarised by the block diagram below.



4.4. Results

The results of this experimental study are shared via static files and not disseminated through Eurostat's database. Given the nature of the data source and the assumption of the methodology the concentration values of the EU urban labour markets needs to be considered with care. Changing some of the methodological assumptions can impact the absolute values of the HHIs but has a limited effect of the relative position between FUAs. Therefore, the value of these results lies in the opportunity to compare different EU urban labour markets in terms of market concentration.

The baseline for calculating the results is the arithmetic mean of the hhi values across FUAs and occupations, starting from the full dataset with imputed missing company names as described by the previous sections (mean). This value is shown also with the help of an interactive EU choropleth map. Nevertheless we also present hhi values based on different assumptions:

1. hhi mean weighted by the number of distinct job ads starting from the same data set with imputed values (weighted_mean)
2. arithmetic mean calculated from a smaller dataset that exclude missing advertisements with missing company name (mean_upper)
3. hhi mean weighted by the number of distinct job ads starting from a smaller dataset that exclude missing advertisements with missing company name (weighted_mean_upper)

With the obtained hhi values we provide and compare country averages and the evolution over the timeframe analysed (2019-2020).

Additionally we correlate the concentration indices with other variables collected by Eurostat at FUA level, namely size of the FUA (population), un/employment rates and survey data on perception indicators (ease to find a job, job satisfaction).

Finally, we provide a glimpse of the most recurring occupations (at 4-digit ISCO code level) in OJA data and their corresponding average concentration index.

5. Future Work

Given the explorative nature of the data source and the still-uncertain impact of the methodological assumptions on the results, the outputs of this study have an experimental value.

During the course of this study a concentration index of labour markets was calculated under several assumptions. The output tables that we provide try to provide an overview of the different index values that can be obtained. Changes assumptions can result in relevant changes in the output values. Therefore, this indicator provides a picture of the comparative level of concentration of various European urban labour markets. However, it would be premature to define one labour market as "concentrated" (or not) solely based on the score of the indicator.

The nature of the data source implies a bias in the index, since we are considering only online advertisements. One issue for example is the limited coverage of jobs in the public sector that are usually less advertised on the main online job portals. In this regard, it is useful to compare the list of the largest employers in Luxembourg compiled by STATEC and the list of the largest employers in Luxembourg obtained from OJAs data.

There are several possible improvements to the present study:

- Improve the cleaning and classification of the company names. Currently the keywords list of intermediary agencies is refined only for a limited number of countries/languages (i.e. DE, IT, PT and RO). The intermediary agencies in the remaining countries are filtered out by using some EU generic keywords in English language. This include also further work on the classification model. At the same time cleaning the company names includes the consolidation of company names that have different spelling referring to the same company. The current list of company names is based on a sample of big companies from a few countries.
- 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. This would be easier with a definition of labour extended to NUTS3 regions instead of FUAs where fewer Eurostat's dataset are available. This will include finding correlation with other labour market aspects such as wages. Furthermore it is part of future plans to understand the relation between the uptake of IT technologies across countries and the estimated level of labour market concentration, due to the specific nature of the data source used (i.e. online job ads).
- Extend the scope of the calculation of the index to extra EU countries, such as UK (data are already present in the dataset) and EFTA countries (currently sources landscaping in development).

6. Annex 1

The labour market concentration (Herfindahl–Hirschman, or hh) index 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 what the data extraction algorithms employed in data production identify as the entity posting the job ad. 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 text sequence identified by the data extraction algorithm as the company name, even though it 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 hh index, a strategy is needed to edit the variable "companyname" so that it contains only words that are (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).

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 (because their impact on the estimation is larger than for companynames with smaller number of ads) for the three pilot countries (Italy, Portugal and Romania, chosen because of the language skills available in the team). 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 classify (based on desk research) a set of entries of the variable "companyname" (hereafter, "companynames") as employer or non-employer. This was done for companynames with at least 100 ads in a sample of 1 mn ads extracted for three pilot countries. A set of text strings (hereafter, "keywords") used to filter out these companynames from the dataset was developed during this phase.
2. Use the keywords to filter out (i.e., classify as non-employer) companynames in other portions of the data (i.e., other countries and smaller companynames in the three pilot countries). Since there are many 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 companynames 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. Identify data functions allowing to discriminate between employers and non-employers that are based only on non-employer data. This made it possible to apply these rules also to other countries, because a substantial number of non-employer companynames are identified in each country based on the previous step. 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 an ontology model that classifies companies based on a set of keywords; and of a decision tree machine learning model that automatically classify companynames that the keyword search has not already identified as non-employers.

For the automatic classification of companynames, three empirical rules have been chosen out of a set of potential rules based on the Gini impurity index. Each rule 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:

- 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 companyname 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 companyname 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 companyname on the log number of 2-digits NACE code, the log number of de-duplicated ads and the interaction between the two independent variables.

Companynames outside the training data set (i.e. companynames 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 the lie close to all the three curves that have been estimated for non-employer companynames. Therefore, the automatic classification model can be described as a three-nodes decision tree (more specifically, 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 companyname by the model. If it is flagged, then the companyname 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%.

7. References

1. Eurostat experimental statistic website, https://ec.europa.eu/eurostat/web/experimental-statistics
2. OECD, Competition issues in labour markets, https://www.oecd.org/daf/competition/competition-concerns-in-labour-markets.htm
3. Cedefop (2019). Online job vacancies and skills analysis: a Cedefop pan-European approach. Luxembourg: Publications Office. https://www.cedefop.europa.eu/files/4172_en.pdf
4. 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. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji190547
5. Cedefop and Eurostat formalise joint approach to online job advertisement data, https://www.cedefop.europa.eu/en/news-and-press/news/cedefop-and-eurostat-formalise-joint-approach-online-job-advertisement-data
6. 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, Luxembourg, 16 May 2019
7. https://ec.europa.eu/eurostat/cros/content/WPB_Online_job_vacancies_en
8. https://github.com/OnlineJobVacanciesESSnetBigData/Labour-market-concentration-index-from-CEDEFOP-data
9. Azar, J., Marinescu, I., Steinbaum, M., & Taska, B. (2020). Concentration in US labor markets: Evidence from online vacancy data. *Labour Economics*, *66*, 101886. https://www.sciencedirect.com/science/article/pii/S0927537120300907
10. OECD (2020), Competition in Labour Markets, https://www.oecd.org/daf/competition/competition-in-labour-markets-2020.pdf
11. Eurostat, Labour Market Statistics Overview, https://ec.europa.eu/eurostat/web/labour-market/overview
12. Eurostat, Labour Marked Data, https://ec.europa.eu/eurostat/web/euro-indicators/labour-market
13. Carnevale, A. P., 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*. https://cew.georgetown.edu/wp-content/uploads/2014/11/OCLM.Tech_.Web_.pdf
14. Beresewicz M., Pater R. (2021), *Inferring job vacancies from online job advertisements*, https://ec.europa.eu/eurostat/documents/3888793/12287170/KS-TC-20-008-EN-N.pdf/6a86d53e-d0b8-d608-988d-d91f0cef6c21?t=1611673495829
15. Eurostat, experimental statistics on Labour Market Transitions, https://ec.europa.eu/eurostat/web/experimental-statistics/labour-market-transitions
16. **Dijkstra, L., H. Poelman and P. Veneri (2019), "The EU-OECD definition of a functional urban area", *OECD Regional Development Working Papers*, No. 2019/11, OECD Publishing, Paris,**
17. Eurostat, *Methodological manual on territorial typologies*, (2018) https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Territorial_typologies_manual

18. Eurostat, History of NUTS: **https://ec.europa.eu/eurostat/web/nuts/history**
19. European Central Bank, Concentration, market power and dynamism in the euro area, 2019 ECB Working Paper, **https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2253~cf7b9d7539.en.pdf**
20. UNECE (2014), A Suggested Framework for the Quality of Big Data,

    https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2

21. Marinescu, I. and Rathelot, R. (2018). Mismatch unemployment and the geography of job search. American Economic Journal: Macroeconomics, 10(3): 42–70, https://www.aeaweb.org/articles?id=10.1257/mac.20160312