Fabo, Brian; Kureková, Lucia M´ytna

**Working Paper**

# Methodological issues related to the use of online labour market data

ILO Working Paper, No. 68

**Provided in Cooperation with:**
International Labour Organization (ILO), Geneva

*Suggested Citation:* Fabo, Brian; Kureková, Lucia M´ytna (2022) : Methodological issues related to the use of online labour market data, ILO Working Paper, No. 68, ISBN 978-92-2-037282-1, International Labour Organization (ILO), Geneva, https://doi.org/10.54394/ZZBC8484

This Version is available at:
https://hdl.handle.net/10419/263129

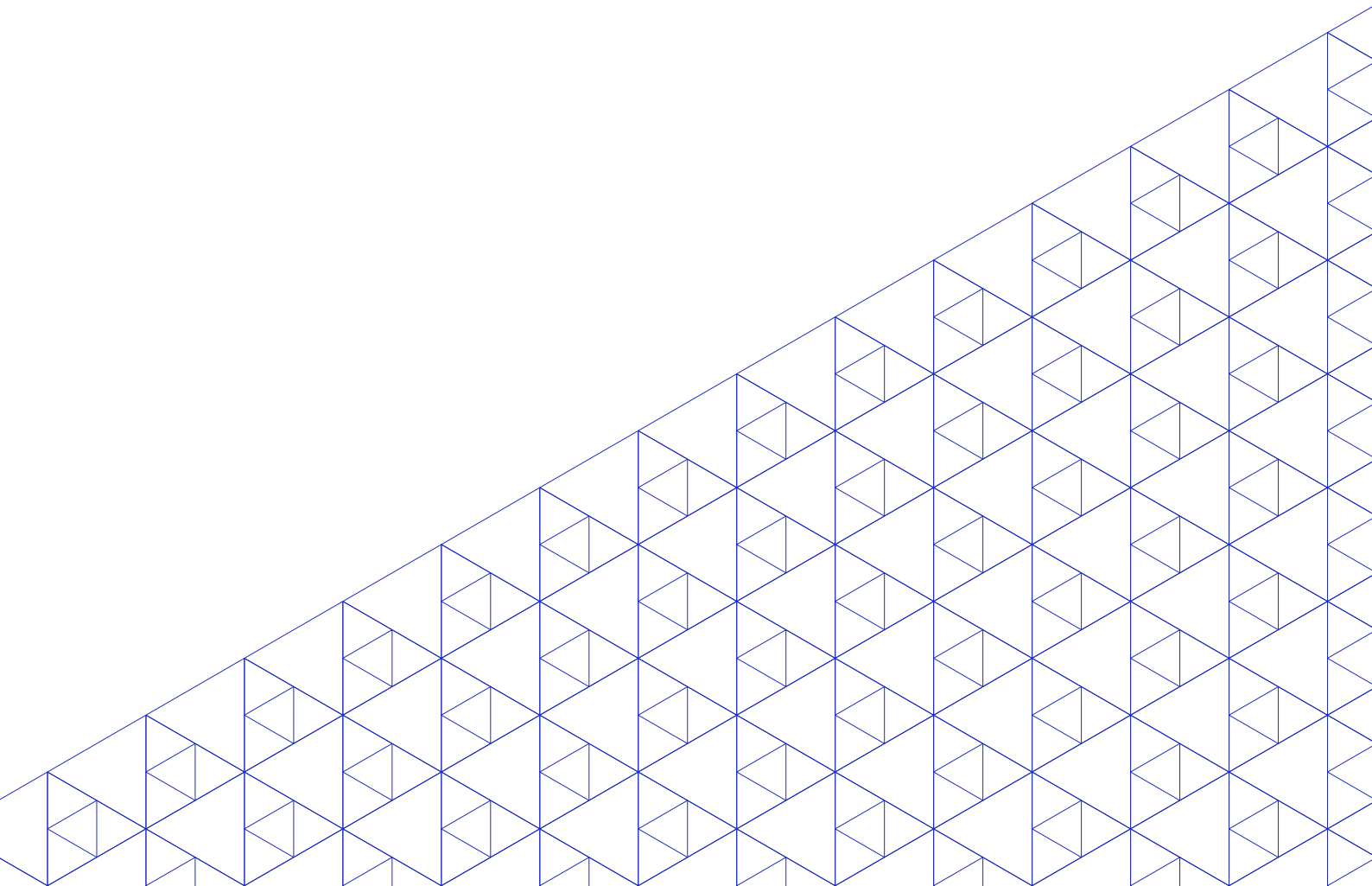
International Labour Organization

ILO Working Paper 68

June / 2022

# Methodological issues related to the use of online labour market data

Authors / Brian Fabo, Lucia Mýtna Kureková



---

ISBN: 9789220372814 (print)
ISBN: 9789220372821 (web-pdf)
ISBN: 9789220372838 (epub)
ISBN: 9789220372845 (mobi)
ISBN: 9789220372852 (html)
ISSN: 2708-3446

https://doi.org/10.54394/ZZBC8484

---



ILO Working Papers summarize the results of ILO research in progress, and seek to stimulate discussion of a range of issues related to the world of work. Comments on this ILO Working Paper are welcome and can be sent to RESEARCH@ilo.org, liepmann@ilo.org.

Authorization for publication: Richard Samans, Director RESEARCH

ILO Working Papers can be found at: www.ilo.org/global/publications/working-papers

**Suggested citation:**
Fabo, B., Mýtna Kureková , L. 2022. *Methodological issues related to the use of online labour market data*, ILO Working Paper 68 (Geneva, ILO).

# Abstract

This report provides a mapping of existing research that employs online labour market data, covering both online job vacancies (demand side) and online applicant data (CVs) (supply side). We discuss and assess a variety of tools and empirical methods that have been used to address specific disadvantages of this data, such as non-representativeness or fluctuations in data quantity and structure; these may be due to external shocks, such as the COVID-19 pandemic. We find that while this research field has expanded rapidly, including with respect to geographical coverage, many empirical studies do not engage with the methodological aspects and weaknesses of online labour market data and take them at face value. We highlight that there are legitimate research approaches, which are inductive in nature, focused on discovering patterns and trends in underlying data. These are by definition less concerned with generalizability of findings, as they have different objectives. For this body of research, online labour market data open new avenues for understanding developments in labour markets. We also argue that biases in online labour market data emerge due to multiple factors. With respect to the order of discrepancies between online labour market data and representative data sources, these are typically not paramount. Different techniques have been adopted to deal with the non-representativeness problem, such as statistical techniques; adapting the research questions and research focus to the quality of data; and use of mixed methods, including qualitative methods, to increase the robustness of results.

# About the authors

**Brian Fabo** (PhD) is a lecturer in Economics at the Department of Public Policy, Comenius University in Bratislava, Slovakia, and a Senior Economist at the National Bank of Slovakia. His research focuses on the application of novel data sources in social science research, digitalization, and bias in research.

**Lucia Mýtna Kureková** (PhD) works as a senior researcher at the Slovak Academy of Sciences, Centre for Social and Psychological Sciences. She is a labour market researcher focusing on skills demand and skill changes, big data in labour market research, labour migration and migrant integration, and labour market inequalities and social inclusion.

# Table of contents

# List of Tables

# List of Boxes

# Introduction

Mismatches between workforce skills and employers' demands represent a key obstacle to economic growth; they are closely associated with factors such as productivity, unemployment, labour force participation and informality (Acemoglu and Autor 2010; Beblavý, Maselli, and Veselková 2014; 2015; CEDEFOP 2014; Ernesto and Francesco 2016). Skill and task demand is changing due to rapid technological advancement, automation and digitalization – processes that are relevant not only to developed economies, but also to emerging and developing ones (Comyn and Strietska-Ilina 2019). The ability to respond to the changing skills demand is considered key to successful economic transitions that are inevitably sought by economies and individuals throughout the world, and for developmental catch-up between higher- and lower-income countries. The improved understanding of how required skills and work tasks are changing, and the quest to better align skill supply to employer demand, have inspired efforts to create more demand-driven labour market policies. Timely and reliable data are a key prerequisite for these efforts.

The online data on labour markets have in the past years become an important source of information for better understanding how labour markets function. This process has been affected by the spread of the Internet and emergence of online labour market intermediary platforms (e.g. Babajobs in India, Glassdoor in the United States (US), Profesia in Slovakia), online vacancy aggregators (e.g. Burning Glass Technologies), and professional websites and social media (e.g. LinkedIn, Twitter, Facebook). These types of labour market data currently provide a source of timely, granular and often comprehensive data that has been increasingly used by academics and policy makers to analyse labour markets around the world.

The use of such data has been driven throughout the world by the fact that traditional representative surveys might not be available, or do not cover more specific aspects of labour markets in sufficient detail and frequency. The absence of high-frequency, detailed survey data has led researchers to revert to a second-best solution of using online data to study diverse questions. More traditional micro-economic questions (focusing on the behaviour of firms and individuals, skill supply, skill demand, matching, and skill-biased technological change) or macro-economic questions (such as predictions of the unemployment rate, aggregate demand, broader phenomena, and changes at the national, regional or local levels) have been analysed with the use of online data (Boselli, Cesarini, Marrara, et al. 2018). Furthermore, new questions or approaches to a structured understanding of labour market characteristics or changes have also emerged in relation to online data availability (for instance, building skill or task taxonomies, building curricula based on identified demand, a deeper understanding of job changes, etc.).

In general, research using online data to study labour market issues has been organized around five related aspects of research: labour market monitoring and analysis; studying demand for workforce skills; observing job search behaviour and improving skill matching; predictive analysis of skill demand; and experimental studies (Nomura et al. 2017). Due to the granularity of the data, research in these areas has been conducted also at sub-national levels, examining regional or local labour markets (e.g. Azar et al. 2019; 2020; Marinescu and Rathelot 2018).

The use of online labour market data, however, is not without disadvantages. The key concern is the non-representativeness of online data, and the implications this has for various aspects of research and policy making. Other data quality issues relate to data validity, reliability, scalability, generalizability, integrity or privacy, and legal issues (Blazquez and Domenech 2018). This paper situates itself within the debate on the methodological appropriateness of using online labour market data for academic and policy purpose, and provides a systematic review and discussion of: (1) *the types and forms of biases* present in online labour market data, and ways in which these are understood, discussed and addressed by research; and (2) *measures and tools – statistical and other* that have been used in past academic and policy research to remedy biases of online labour market data, with a particular focus on two dimensions: *representativeness*, and *fluctuations* in online labour market data.

We build on previous studies that have discussed methodological aspects of big data more generally, including implications for the development of new analytical approaches and tools (Blazquez and Domenech 2018; Einav and Levin 2014; Mezzanzanica and Mercorio 2019; Varian 2014). We differ from these studies by providing a narrower focus on online labour market data, such as job vacancies and applicant data. Nevertheless, in particular parts of this paper we refer to broader conceptual issues of different motivations for research: for instance, deductive versus inductive approaches to gathering and analysing information.

# ▶ 1 Methodology

This analysis builds on an earlier article co-authored by one of the authors (Kureková, Beblavý, and Thum Thysen 2015), which discussed the methodological aspects of using online vacancy data and voluntary Web-based labour market data (i.e. WageIndicator). In this analysis we concentrated on empirical works published after 2015, meaning new empirical papers focusing on or using online labour market data, and a set of methodological and conceptual papers about online labour market data and big data more generally, also prior to 2015. Our approach to a systematic analysis in this paper comprised two related consecutive steps.

In the first step, we conducted a Google Scholar search to identify empirical (applied) studies that have used online labour market data to study the skills demand or supply in the labour market. We focused on the abstracts of the retrieved documents in order to identify papers of interest to us. We paid particular attention to the fast-growing body of studies in developing countries, where representative survey data might be less readily available, and where online data has a specific potential for further expanding the analysis of these labour markets (Table 2). Importantly, the first-step mapping showed that many empirical studies do not engage with the methodological aspects and weaknesses of online labour market data, and take them at face value. The global expansion of this research field, thus, has mainly proceeded through empirical applications, without sufficiently understanding the methodological problems of using such data.

In the next step, we therefore restricted our search by imposing a number of criteria, in order to gather a varied and robust sample of studies which have taken a methodologically engaged approach to applying online data. The criteria for selecting studies for methodological discussion were the following: (1) studies were published in peer-reviewed journals or by recognized international research institutions, such as the Organisation for Economic Co-operation and Development (OECD), the European Centre for the Development of Vocational Training (CEDEFOP), and World Bank; (2) studies were varied geographically, to cover developed and developing countries; and (3) they focused primarily on skills analysis at the micro-level. This search yielded nearly 40 empirical papers, which we systematically reviewed with the aim of understanding how biases were identified and accounted for (see the Annex). In addition, our discussion and analysis is also informed by more conceptual papers, especially those explicitly concerned with and addressing methodological aspects of online data (e.g. Blazquez and Domenech 2018; Einav and Levin 2014; Gandomi and Haider 2015; Kotu and Deshpande 2019; Mezzanzanica and Mercorio 2019; Varian 2014).

We mapped a range of studies using online labour market data, covering both online job vacancies (demand side) and online applicant data (CVs) (supply side). Accordingly, we will discuss and assess a variety of tools and empirical methods that have been used to address specific disadvantages of this data, such as non-representativeness or fluctuations in data quantity and structure, including those caused by external shocks, such as the COVID-19 pandemic. Whereas most research using online labour market data has drawn on sources in advanced economies, we conduct our analysis with an enhanced focus on the specific needs and varied contexts of emerging and developing economies. We are particularly interested in research related to different aspects of skills, such as skill demand, skill supply, matching, changes in skills; and in the relationship between skills and other changes in society and economy, such as automation, digitalization and technological change.

The remainder of the paper is organized as follows. Chapter 2 presents the characteristics and trends in research that relies on labour market data, Chapter 3 discusses the advantages, while Chapter 4 addresses the main disadvantages of online labour market data. Chapter 5 offers an overview of methodological approaches to address non-representativeness and fluctuations in data, and the final chapter concludes.

# ▶ 2 Online data in labour market research: Trends and characteristics

Efforts to understand labour market matching date back to at least the 1960s. Nonetheless, until relatively recently, the empirical work on job vacancies has been quite sparse. Even today, basic questions remain difficult to answer: for instance, how workers are assigned to jobs; what share of jobs is filled through a formal application process; how many people apply to a typical advertised vacancy; how many applications a typical job seeker submits; and how job applicants decide where to apply. This situation reflects the fact that these aspects of matching are rather time- and context-specific (Kuhn 2014).

One likely reason for the slow progress in the field is that before the Internet started taking on the role of "labour market matchmaker", obtaining reliable data about job vacancies and applicants was relatively difficult. Regardless of whether the data were obtained through surveys (Abraham 1983; Barron and Bishop 1985; van Ours and Ridder 1992), public employment office databases (van Ours 1989), or by painstakingly collecting individual advertisements published in newspapers (Jackson 2007; Dörfler and van de Werfhorst 2009; Álvarez and Hofstetter 2014), the data suffered from selection bias and other assorted representativeness issues.

Over the course of the 2000s and 2010s, labour market matching shifted increasingly towards the Internet; this has benefited job seekers by improving access to information about vacancies, and allowed employers to benefit from a larger pool of job candidates (Autor 2001; Kuhn and Mansour 2014; Piróg 2016). This shift to an online labour market has also opened new opportunities for research and policy applications. At the same time, newspapers stopped being relevant as a data source. In the US, since the 1950s the Conference Board organization had systematically surveyed the number of vacancies posted in the "Help Wanted" section of the newspapers, but it stopped publishing its findings in 2008, after failing to see any increase since the 1990s boom. This failure was attributed to the migration of vacancies to the Web (Anastasopoulos et al. 2021).

The Internet allows researchers and policy practitioners to access a large volume of information about jobs and job candidates. Most commonly, the data used originate from individual job sites (Beblavý, Kureková, and Haita 2016; Drahokoupil and Fabo 2022; Marinescu and Wolthoff 2020), as well as commercial websites where people post their personal profiles or CVs, such as LinkedIn or Indeed (Kureková and Žilinčíková 2018; Mamertino and Sinclair 2019; Apaza, Vidal, and Chire 2021; Pejic-Bach et al. 2020). In some cases, however, data are instead taken from an aggregator, such as the European Public Employment Services network, EURES (Kureková et al. 2016), or companies such as Emsi Burning Glass, which collect and extract information from large volumes of job vacancies from many websites (Hershbein 2016; Deming and Kahn 2018; Fabo and Kahanec 2020; Acemoglu et al. 2020). A major advantage of this data source is that the data collection is regular rather than a one-off event. Thus, the data can be analysed as time series and not just cross-sectionally (e.g. Acemoglu et al. 2020; Leitner and Reiter 2020).

Social media is a comparatively less developed alternative. Nonetheless, several notable studies have used social media data to analyse the labour market (see the more detailed overview of alternative online data sources in Box 1). Specifically, Twitter has been used to predict labour market flows by counting the incidence of searches such as "lost my job" over time (Antenucci et al. 2014). Additionally, there has been growing research conducted with LinkedIn data (Mamertino and Sinclair 2019; Tambe 2014; Tambe et al. 2020). Overall, social media remains a powerful but underutilized tool for studying labour markets; this is due to difficulty in obtaining data, as well as "making sense" of often rather opaque signals present in the social media content (Lenaerts, Beblavý, and Fabo 2016).

Importantly, the Internet has a dual role in labour market research: it serves as an vital source of data, but at the same time it is transforming the labour market (Horton 2011). A large number of studies have tried to estimate and/or forecast important macroeconomic variables such as unemployment using Google Trends data; mostly in developed countries (Askitas and Zimmermann 2009; Fondeur and Karamé 2013; McLaren and Shanbhogue 2011), but also in China (Su 2014, using Baidoo). Empirical analyses have shown that online job vacancies are suitable data sources for measuring aggregate economic activity in the labour market (Hershbein and Kahn 2016; de Pedraza et al. 2019).

As a consequence, using online job vacancies as a data source for calculating common labour market indicators (such as number of vacancies, labour market tightness, degree of skill mismatch) has become a relatively common practice for scholars and practitioners (Japec and Lyberg 2020; Štefánik, Lyócsa, and Bilka [2022]; Turrell et al. 2019). Interesting micro-level applications of online labour market data, beyond skills analysis, include those focusing on the value of the migration experience in employers' demands (Kureková and Žilinčíková, 2018); the role of occupational mismatch in explaining the productivity puzzle (Turrell et al. 2021); the relationship between firm credit crunch and employee job search behaviour (Gortmaker, Jeffers, and Lee 2021); discrimination against women in the labour market (Kuhn and Shen 2013); or links between the introduction of unemployment benefits, job searches and job postings during the Great Recession in the US (Marinescu 2017).

Internet labour market data typically cover a specific labour market, such as a country; but in some cases they include multiple countries (or sub-country units, such as the US states), in a comparative design (Azar et al. 2020; Fabo, Beblavý, and Lenaerts 2017; Modestino, Shoag, and Ballance 2020). Most studies focus on the entire labour market, but several investigate specific occupations or industries, such as logistics (Kotzab et al. 2018), nursing (Kobayashi et al. 2016), or data science (Debortoli, Müller, and vom Brocke 2014; Ecleo and Galido 2017).

A clear majority of published studies of online labour market data focus on the US, UK or EU labour markets, which is in line with the general over-representation of research focusing on developed countries in academia (Das and Do 2014). Nevertheless, the Western-centric focus in the literature should not overshadow the growing number of important publications that cover developing and emerging nations' economies. In particular, they deal with big labour markets such as China (Fang et al. 2020; Kuhn and Shen 2013; Maurer-Fazio and Lei 2015; Xu et al. 2017; Zhu et al. 2016), India (Chowdhury et al. 2018; Nomura et al. 2017), Russia (Pitukhin, Astafyeva, and Astafyeva 2020; Skhvediani et al. 2021), and Pakistan (Bilal et al. 2017; Matsuda, Ahmed, and Nomura 2019).

The large size of these markets generates a huge amount of labour market data: for instance, a recent study focusing on the Chinese market identified 20 million job adverts, offering 105 million job vacancies, posted on just one online platform in four months (Fang et al. 2020). Other emerging and developing countries that have been studied include countries as diverse as the Philippines (Ecleo and Galido 2017), Ukraine (Muller and Safir 2019), Belarus (Vankevich and Kalinouskaya 2020), Kosovo (Brancatelli, Marguerie, and Brodmann 2020), Peru (Apaza, Vidal, and Chire 2021), and Mexico (Campos-Vazquez, Esquivel, and Badillo 2021).

Lastly, research using online vacancy data appears more frequently than research relying on job applicant/CV data; these two aspects of labour markets are rarely studied simultaneously, although some examples exist (Fabo and Kahanec 2020; Matsuda, Ahmed, and Nomura 2019). This imbalance might be due to the generally easier access to information about vacant jobs, rather than applicant information, which some portals offer on a paid basis (e.g. Profesia.sk). However, it could also reflect the greater interest in labour demand research, for which there are fewer alternative data sources of vacancy data.

### ▶ Box 1: Alternative online data sources relevant for labour market analysis

Online vacancy postings and CVs are the most widely used Web-based data sources, but they are not the only ones. Some of the widely used alternative Web-based or Web-collected data sources are:

- **Web-based surveys**. Probably the most widely used is the WageIndicator survey, which covers 130 countries, including many emerging economies, for which other data are often unavailable. The respondents of the survey are self-selected on a voluntary basis in response to the online campaign, which results in a biased dataset. There is a lively discussion among scholars as to what extent this bias can be sufficiently addressed using methods such as weighting (Fabo and Kahanec 2018; Smyk, Tyrowicz, and van der Velde 2018; Tijdens and Steinmetz 2016). In developed countries, online panels with probabilistic sampling have recently emerged as a promising means of collecting high-quality, reliable data (Das, Ester, and Kaczmirek 2018). Nonetheless, they are costly and rely on a high level of Internet access, and thus are less practical for most emerging countries.
- There has been an immense body of literature produced recently on the role of **labour market platforms**, such as Uber, Amazon Mechanical Turk, or Upwork. These platforms are relevant for the "standard" labour market, because they often disrupt traditional dynamics present in these markets. This can happen because the platforms allow outsourcing of some tasks that were previously performed by employed workers to people working through platforms, who are classified as self-employed (e.g. some institutions have replaced in-house drivers with an Uber service); or alternatively, because they allow enterprises to perform offshore work overseas (Drahokoupil and Fabo 2016).

# ▶ 3 Advantages of online labour market data

Notwithstanding the well-known limitations, the online labour market data represent a powerful tool that allows researchers to study the labour market. A key motivation for using online labour market data is their granularity and detail; together with a large number of observations, this allows researchers to access detailed information on the demand and supply of skills in the labour market, and generate insights on important topics. These include the skill mismatch (Beblavý, Kureková, and Haita 2016), school-to-work transition (Buchs and Helbling 2016), the skills and educational characteristics of new occupations (Acemoglu et al. 2020; Beblavý et al. 2016; Rios et al. 2020), the evolution of skill demand over time (Blair and Deming 2020), and lifelong learning (Kotzab et al. 2018). Box 2 presents a discussion of the importance of skills as a variable in the analysis of online labour market data, and provides selected findings about skills, tasks and occupations based on the analysis of online data.

### ▶ Box 2: Skills, tasks, occupations: What do we see in online vacancies?

An online vacancy typically advertises an occupation. An occupation may be broadly understood as a grouping of jobs involving similar tasks, which require similar skill sets (ESCO 2015); for instance, "plumber", "teacher", or "computer programmer". Importantly, the tasks performed within an occupation – and thus the skill requirements – vary greatly over time, as well as between different labour markets (Kureková et al. 2016; Tijdens 2010).

Several well-known occupational classifications exist, including the International Standard Classification of Occupations (ISCO), European Skills, Competences, and Occupations (ESCO), Standard Occupational Classification (SOC), People's Republic of China Grand Classification of Occupations (CGCO) and O*NET, which define tasks and (in some cases) also skills taxonomies. Although these taxonomies are regularly updated, the gap between them can be years, if not decades. This creates a window of opportunity for utilizing online job postings, which contain the required skills and, in many cases, also the task content of the occupation.

Significant efforts have been made to utilize machine learning and artificial intelligence to extract skills requirements, particularly from vacancies (Boselli, Cesarini, Marrara, et al. 2018; Boselli, Cesarini, Mercorio, et al. 2018; Colombo, Mercorio, and Mezzanzanica 2018; Khaouja, Kassou, and Ghogho 2021). Some novel insights into the content of occupations have been gained using online job vacancies; for instance:

- Regarding the homogeneity of occupations, it has been discovered that those considered "high-skilled" tend to be much more internally heterogenous in terms of their task content, while low-skilled occupations tend to be more standardized (Visintin et al. 2015).
- A methodology has been proposed for systematically observing new occupations (such as "data scientist"), before they are included in the existing classifications (Beblavý et al. 2016; Marrara et al. 2017).
- The empirical examination of the role of transversal skills, which are not connected to specific occupations, but rather enable a worker to function within the context of a specific firm. For example, in the case of modern, multinational companies, a command of the English language and computer literacy are required, irrespective of the actual position advertised (Beblavý, Kureková, and Haita 2016; Drahokoupil and Fabo 2022; Fabo, Beblavý, and Lenaerts 2017).
- The role of experience in labour market demand (Beblavý, Kureková, and Haita 2016).

An important characteristic of online labour market data is their real-time availability. Real-time labour market data have been used to anticipate, for example, unemployment trends (Askitas and Zimmermann 2009; Simionescu and Zimmermann 2017) or GDP growth. Real-time availability implies that data can be analysed with a much shorter time delay than survey-based data about labour markets, while it can also capture the impact and dynamics of unexpected events and shocks, such as the COVID-19 pandemic (Campos-Vazquez, Esquivel, and Badillo 2021; Fang et al. 2020). For instance, an OECD study identified not only the sizable dip in vacancy rates at the onset of the pandemic (March–April 2020), but also the variance between countries, sectoral differences in fluctuations, and the change in skill requirements connected to the switch to working from home (OECD 2021); this implies that the impact of the COVID-related shock varied across a set of dimensions. Moreover, timely information is useful not only for research, but particularly for policy makers and educators. For instance, it is well known in economics that unsuccessful school-to-work transition has long-term implications for career outcomes (Bloom, Freeman, and Korenman 1988). It is, therefore,

important to be able to advise young people which sectors are growing despite the recession, and which are the most demanded skills that might improve their chances in a difficult labour market.

While real-time availability is a key advantage, in some cases, past online labour market data can also be reconstructed (for example, the job portal Profesia.sk stores past vacancies and has provided researchers with vacancy and CV data over a retrospective time period: see e.g. Beblavý et al. 2016). Such data can in principle be used as longitudinal data, to study trends in skills supply and demand, the emergence of new occupations and their skill requirements, or transformations such as skill-biased technological change. This enables cost-effective access to longitudinal and sometimes cross-sectional data about skills (such as via the EURES platform, Skills Panorama). Moreover, this longitudinal aspect allows the study of changes within occupations, which is rarely possible with other data sources.

Large quantities of online labour market go hand-in-hand with comprehensiveness that online labour market data possibly entail. While data are typically unstructured (Gandomi and Haider 2015), in principle the information that can be extracted from the content of job vacancies or from individual CVs is very comprehensive. In terms of online job vacancies, information about educational or qualification requirements, skills or tasks, and required experience is often detailed, and can be systematically analysed (Beblavý, Kureková, and Haita 2016). Likewise, online job applicant data in fact contain detailed professional and educational experience, and personal information (i.e. key socio-demographic characteristics). With rising competition for jobs, applicants increasingly highlight key competences or skills – especially IT and language skills – but also soft skills; this provides comprehensive input for studying profiles of job candidates (Haddad and Mercier-Laurent 2021; Kureková and Žilinčíková 2018).

It is also important to consider trends that are likely to positively impact the quality and usability of online labour market data. First, the recruitment market is moving online in a dynamic way, and in some developed economies, labour market matching is organized fully online. In an ILO report, Van Loo and Pouliakis (2020) reviewed online job markets in the EU28 countries during 2019, and concluded that in countries such as Estonia, Sweden and Finland, the proportion of vacancies published online approached 100 per cent, while in others such as Denmark it accounted for around 50 per cent. Secondly, in part driven by mismatches, firms are forced to broaden their recruitment processes to include other occupations and regions, thus in effect enlarging their candidate base (which in turn mitigates demand for higher wages). Both these trends work in favour of increasing representativeness, and reducing measurement error (van Loo and Pouliakas 2020). There is now also strong evidence that the Internet has become an effective tool for matching workers to jobs (Kuhn and Mansour, 2014), and that poorer strata of society are increasingly turning to the Internet to search for jobs because their social capital is weaker (Kuhn, 2014). It is also worth mentioning that academic research has already influenced concrete policy initiatives, some of which are summarized in Box 3.

### ▶ Box 3: Policy initiatives using online labour market data

In addition to the academic-focused research, there have been several important applied initiatives to utilize online job market data for commercial or policy purposes. For instance, in the US, the O*Net project utilized the online job vacancies collected by the Burning Glass company to determine "hot technologies" on the basis of employers' job postings (Lewis and Norton 2016). The payroll company ADP publishes monthly employment statistics based on Internet data before official numbers are made available by the Bureau of Labor Statistics. (Einav and Levin 2014). Furthermore several skill-matching platforms have been developed for the commercial needs of HR departments, such as Burning Glass (for a list, see Boselli, Cesarini, Marrara et al. 2018, 482).

In Europe, an agency of the European Commission, CEDEFOP, has piloted a cross-country system of scraping and systematizing vacancies across the EU member states (The Skills Panorama), to support discussion on the use of online vacancy data for policy purposes (Boselli, Cesarini, Marrara, et al. 2018; van Loo and Pouliakas 2020). A more recent initiative is the ESSnet big data project, which pilots the use of online vacancy data for official statistics. Some countries in the EU have already experimented with online labour market data as an auxiliary source of labour market statistics (such as the Experimental Online Job Vacancy Index in Sweden). A good example of a non-Western application is the SkillsFuture government project in Singapore (www.myskillsfuture.gov.sg/), which uses the data from job postings, combined with insights from stakeholders' interviews, to support policy and programme design.

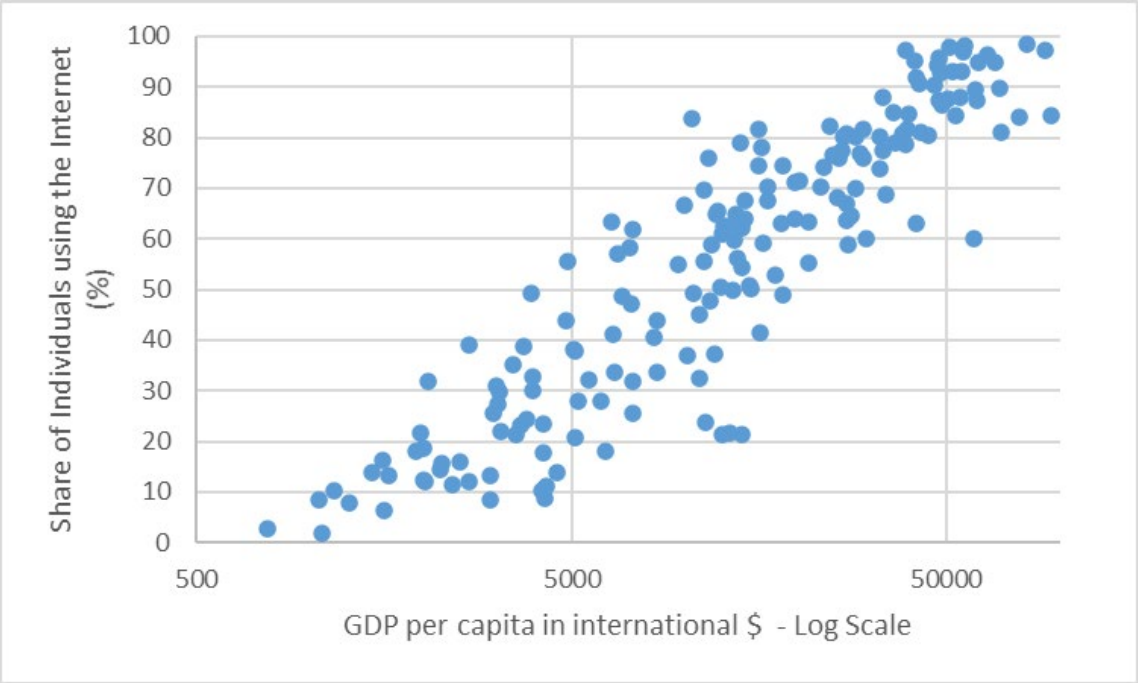
# 4 Sources of biases in online labour market data

Notwithstanding the evident advantages discussed in the previous chapter, the online labour market data suffer from various biases, particularly a lack of representativeness. Non-representativeness in applicant data might have different causes to that affecting vacancy data. With respect to job applications and job searches, the main source of non-representativeness is linked to the fact that the universe of jobs intermediated online is not equal to the universe of new jobs that exist. Internet access is an important driver of this, as individuals' ability to access the Internet remains unequal across and within countries, and varies by socioeconomic status, age or skills. Other aspects that intervene in decisions about online job seeking include a sector's level of informality, as well as the level of social capital that sustains referrals, which are more widely used in lower-skilled jobs and in smaller enterprises.

Regarding vacancy data, the individual labour market segments are unlikely to advertise open positions to an equal extent. While Internet access has become less of an issue for firms, factors such as the intensity of labour demand, nature of work, level of informality in a given sector, or aspects such as firm size, all affect the likelihood of a vacancy being published online in the first place (Sostero and Fernandez-Macias 2021). Sectors such as construction or agriculture are in some countries less amenable to the use of online labour intermediation platforms, while micro and small enterprises are more likely to rely on informal and non-advertised hiring processes. We now turn to discussing these aspects in greater detail.

The extent to which the population is connected to the Web varies greatly between different countries, but also within them. As evident from Figure 1, Internet access is close to universal in high-income countries, and is available to a majority in most upper-middle income countries and some lower-middle income countries. However, the majority of the population in low-income countries and many lower-middle income countries are still without access to the Internet. The poorest, least educated and the most distant from the labour market, even in high-income countries, are typically digitally disconnected (Warschauer 2003; van Dijk 2006; 2020; Scheerder, van Deursen, and van Dijk 2017). Furthermore, other specific groups such as females, older workers and rural populations are likely to be unable to go online in countries where Internet access is not widespread (Birba and Diagne 2012). Hence, information about online labour market matching is likely to have biases in less developed and emerging economies, due to limited or skewed Internet access.

▶ **Figure 1: Share of Internet users per population, correlated with GDP in 2017 on a log scale.**




Data source: World Bank: World Development Indicators (extracted 25 February 2022)

In countries where Internet access is close to universal, the bias of analysis using Internet labour market data might be less significant (Askitas and Zimmermann 2015); however, there is still an observable bias in the data, leading to an over-representation of tertiary educated workers and job opportunities for better-educated workers (Muller and Safir 2019; Štefánik 2012). For example, Carnevale, Jayasundera and Repnikov (2014) studied Burning Glass Technologies data for the US labour market, and estimated that 80–90 per cent of job vacancies requiring a tertiary degree (bachelor's and higher) are posted online, compared to about 40–60 per cent of job advertisements requiring a high-school diploma. In spite of this limitation, some researchers have used online labour market data to understand demand in low-skilled occupations or in unstable, typically less skilled jobs (Beblavý, Kureková, and Haita 2016; Kureková and Žilinčíková 2016).

Another important dimension to consider is that of informal labour. From the existing ILO analyses, we know that six workers out of ten work in the informal economy. Unlike some past predictions, we know that this number is not necessarily decreasing. The issue is not limited to emerging countries, and particularly affects vulnerable populations such as women, uneducated people, or migrants (ILO 2021). The reasons for informal employment vary; for instance, enterprises might opt to operate informally to avoid regulations applying to a formal employment relationship. Additionally, even formal firms might employ workers informally; in some cases this reflects the preference of workers, as in the case of online crowdworkers preferring to make some quick money on the side. Thus, informal employment might be associated with lower numbers of vacancies being published either online or offline. Nevertheless, we see that some online job portals also cover the informal labour market, such as Babajobs in India (see Nomura et al. 2017).

Next, particularly in the developing and emerging markets, a major part of the workforce finds itself in a self-employment arrangement, due to necessity or choice; even though their work is similar to that performed by employed workers (ILO 2016; Poschke 2019). Self-employed work does not typically generate vacancies (Dunlop 1966), and people might be particularly prone to being inaccurate when describing their self-employment experience in CVs (Jones 1984).

In the formal economy, there are also reasons for not advertising jobs publicly. Enterprises and job seekers might opt for an informal (internal) approach for multiple reasons, including lower search costs, ability to avoid initial screening, and because seeking workers or work through informal networks is likely to result in opportunities and applicants located in the near vicinity (van Ours 1989). The sheer size of the online job markets demonstrates that there are many situations where a formal job search is nonetheless initiated; but we need to be mindful of the limited generalizability of any patterns identified in the job postings, even in countries with a high share of Internet users and an insignificant informal economy. That being said, the relatively low cost of advertising job vacancies (or of finding a worker via a CV posted online) might empower actors who would not have initiated formal recruitment in the pre-Internet era (Sodhi and Son 2010).

In addition to non-representativeness, validity and reliability might also be of concern when using online data. Both vacancy and CV data are self-reported, and there are no tools embedded to check the validity and reliability of information provided. For example, Internet job boards can be flooded by resumés that in fact no longer correspond to people who are searching for a job – known as "stale" resumés (Kuhn 2014) – while the same might be true in the case of vacancies. Nonetheless, some researchers consider online information about job applicants to be more truthful and accurate (van Loo and Pouliakas 2020). Moreover, vacancies might be posted online even after the position has been filled, or one posted vacancy can in practice mean more job openings. These specificities warn of a measurement error due to duplicates and the lifetime of a vacancy. Research has also identified that firms might use vacancies as an advertising or company branding tool, which is likely to affect the choice of vacancy content (Winzenried, 2020).

Particular concerns might also arise for cross-country comparative research. Existing studies have shown that employers in different countries seem to use very different strategies in terms of their expressed expectations for skills or education; this might be due to underlying differences in the functioning and institutional underpinning of national labour markets across Europe (Kureková et al. 2016). Similar differences have been identified among formally identical jobs advertised in different sectors, occupations and skill levels (Beblavý, Kureková, and Haita 2016; Brandas, Panzaru, and Filip 2016; van Loo and Pouliakas 2020). Brandas et al. (2016), who studied academic vacancies worldwide, also pointed out a lack of semantic and structural compatibility of data mined from different sources. Winzenried (2020) provided examples of job vacancies that greatly vary in the "density" of skills they require, and emphasized the importance of "implicit" knowledge in vacancy posting, which can be country-, sector- or occupation-specific (such as the significance of education or experience).

# 5 Methodological aspects of online labour market data

In this chapter we present how the methodological weaknesses of online data have been addressed in existing research, by covering a range of statistical and other approaches. First, we describe data processing techniques, and explore more conceptual questions regarding the philosophy of research and its aims; we then discuss these in light of the research objectives, using online labour market data.

## A. Describing data processing techniques

Online job vacancy data research is heavily focused on text classification, with the aim of making sense of the content of vacancies in order to identify skills, tasks or education requirements. Relatedly, research has tried to advance label classification, through matching vacancies to existing occupational standards (ISCO, ESCO), or national standards such as CGCO, the Chinese occupational classification (Kotu and Deshpande 2019; Xu et al. 2017). Finally, research has attempted to advance skill or task classification. Refer again to Box 2 for a list of findings based on skills analysis using online labour market data.

Evidently, skills are also analysed from the perspective of job seekers, and the skill sets they attain. A typical processing strategy is to systematize data from CVs into respective categories. For some parameters, data can be easily turned into tabular and numerical data (such as education level), while textual analysis can be applied to process other parts of CVs. In essence, CVs include work histories and can be transformed into longitudinal data. Information in the CV can also be used to derive variables not directly present in the CV, such as foreign work experience (Kureková and Žilinčíková, 2018). Automated processes using machine learning techniques have also been developed to identify patterns in CVs.

Within the typology of big data, online labour market data – vacancies and applicant data – belong to semi-structured data (Gandomi and Haider, 2015; Blasquez and Domenech, 2018). Across job portals or social media platforms, information that can be found with respect to a vacancy or a CV includes predictable and similar categories (such as education, sector, experience), which can then be organized into a structured format by Web scraping. Commercial websites often organize their content in a structured way, which indirectly supports and facilitates potential analytics on the basis of such data (for instance, the online job portal Profesia.sk in Slovakia; or the EURES portal which aggregates public employment services (PES) vacancies across the EU).

## B. Key conceptual starting points: Deductive versus inductive science

Prior to outlining the dimensions of non-representativeness and fluctuations in online labour market data, we will highlight several conceptual points. These are based on discussions in existing studies that have theoretically (rather than empirically) engaged with online data. They also partly stem from observations derived from our review of various empirical studies using online labour market data. First, the character and quality of online data and big data generally, not just with respect to the labour market, have influenced the methodologies that researchers use to analyse them (Blazquez and Domenech 2018; Einav and Levin 2014; Mezzanzanica and Mercorio 2019; Varian 2014). Importantly, methodological developments are linked to different stages of data processing. Among the principal newer methods for accessing data are Web data mining and machine learning. Given the large amount of text, Natural Language Processing has progressed; this includes techniques such as Sentiment Analysis, Latent Semantic Analysis and Word

Embedding (Blazquez and Domenech 2018). It is beyond the scope of this paper to discuss the respective methodological advances in relation to big data analytics extensively, and we refer the reader to other studies that have engaged at length with this topic (also more generally, beyond the online labour market data) (e.g. Blazquez and Domenech 2018; Einav and Levin 2014; Gandomi and Haider 2015; Kotu and Deshpande 2019; Mezzanzanica and Mercorio 2019; Varian 2014).

Second, we consider the distinction between the different motivations and techniques of research, which can broadly be categorized as deductive and inductive. For the most part, empirical research using online labour market data is predominantly inductive and bottom-up, and often exploratory; rather than deductive and top-down, in the sense of aiming to test existing theories, concepts or relationships.[1] Specifically, data are used to understand the underlying qualities of labour markets, skill characteristics, or trends identified through the longitudinal collection of online labour market data. This is also reflected in the analytical methods applied, and implicitly in a lesser concern with data characteristics – in particular, their representativeness, and whether there is a normal distribution.

We find the inductive approach reflected in the aims and methodology of many studies that we covered in our unrestricted search of diverse studies (Phase 1). Table 1A summarizes key features from the studies reviewed in the first round of the analysis. It is evident that descriptive statistics, frequencies and correlations are frequently used methods, irrespective of the studied country. Furthermore, many studies do not discuss any aspect of bias of their data, and are not concerned with broader generalizability, beyond briefly acknowledging the issue in footnote or a short remark. In summary, unlike theory-testing and theoretically driven research that relies on probabilistic and inferential statistics, based on the assumption of normal distribution and requiring representativeness of data, inductive research is less concerned with representativeness and generalizability.

Furthermore, inductive research using online labour market data also takes account of the fact that online data have no intrinsic value; rather, data value is extrinsic, given by the analyst who applies her/his knowledge in designing research with the respective dataset (Mezzanzanica and Mercorio, 2019; Gandomi and Haider, 2015). An example of such research is what has been termed the "KDD process – knowledge discovery in databases", which inductively studies underlying features of large datasets to create patterns (Kotu and Deshpande 2019). Another example are studies that have used online labour market data to create frameworks or systems. For example, a study by Xu et al. (2017) used online vacancy data from Chinese labour recruitment websites to develop a framework for systematizing vacancies into Chinese occupational categories (CGCO). Likewise, Brandas et al. (2017) exploited global academic jobsites to set up a "Labour Market Decisions Support System".

## C. Mapping the degree of discrepancies between online data and representative data

Several recent studies have attempted to map out the scope of discrepancies between online data and representative data by comparing the sectoral and/or occupational distribution in the online data to an alternative source. An important observation is the fact that representativeness adjustments on the basis of representative data appear more appropriate for online applicant data than for online vacancy data, as we explain below.

Firstly, in advanced economies, representative sources to understand the structure of the labour market are collected on an annual basis; these include the Labour Force Survey (LFS) and its alternatives (e.g. German Socio-Economic Panel (SOEP) data, Current Population Survey (CPS) in the US). From the perspective of labour supply, requiring the analysis of online applicant data, representative surveys such as LFS provide a

[1] Blasquez and Domenech (2018) present these different approaches to research as Supervised Learning and Non-supervised Learning.

good source to compare biases, and potentially then employ weighting on the basis of an underlying representative structure. For example, in their study of young return migrants (below 35 years of age), Kureková and Žilinčíková (2018) compared online CV data with LFS data in Slovakia, and found that the online sample had an unbiased gender and age distribution, but a bias towards people with tertiary education. Štefánik (2012), who studied representativeness and skill demand for graduates in the Slovak labour market, compared online CVs data with the structure of university graduates, and found a surprisingly good fit of online data and representative data.

From the perspective of labour demand, however, labour force surveys capture the stock of jobs that exist in an economy at any given moment, while the online labour market data are a source for understanding the flows in the labour market. Online job vacancies do not capture the stock of matched or unmatched jobs, and represent only the labour market demand.[2] In other words, online vacancy data depend on turnover rates, whereas survey data often represent a cross-section of workers. To illustrate this, there might be many civil servants in a country, but far fewer civil service openings, because public service workers tend to stay in their job for a long time. Occupations with a high turnover, such as odd jobs, tend to be advertised very often, because job holders in these occupations tend to move on to more lucrative jobs as soon as they can. Sostero and Fernandez-Macias (2021) showed that the ratio between the number of job holders and job vacancies can range from nearly 1:1 to 1:100, and even 1:1,000. We therefore do not consider labour force survey data to be an appropriate source for making adjustments to online job vacancy data in particular (Kureková et al. 2015). We failed to find papers which appeared to use firm-level data to adjust online job vacancies, but there are examples of research in which online CV data (work histories) were linked with representative business data: for instance, to study interfirm mobility and innovation (Masso et al. 2012)

Secondly, making adjustments to online vacancy data is a strenuous task, because the population of vacancies is inherently unknown in most countries (Kureková et al. 2015). This is due to the reasons previously described – because hiring processes in firms have different underlying motivations, and hiring often takes places internally or informally. Moreover, an enterprise will advertise vacancies (online or offline) not only when it requires labour due to growth, but also due to replacement needs, such as in response to workforce turnover or retirement. It is therefore important to differentiate between the stock of demand for skills (a company hiring a replacement worker to compensate for attrition) and a flow of skill demand (a company reacting to IT innovations by hiring ICT specialists, creating demand for new skills). Furthermore, because filling a vacancy takes time, enterprises are likely to post vacancies when they anticipate requiring workers with a certain skill some time before they are actually needed (Ferber and Ford 1966). Finally, a need for new skill can be addressed by hiring (which will create a vacancy) or by retraining existing staff (which will not create a vacancy). Therefore, assessing skill demand only on the basis of vacancies, without considering the training investments in the companies, will not provide full information on skill demand (Holt and David 1966).

To describe the degree of discrepancies between online data and representative sources, we have identified several studies that compare the properties of online job data to some measure of labour market flows, typically vacancies (Table 1). The availability of appropriate comparator data varies between countries, as some surveys have been used quite extensively, such as the Job Openings and Labour Turnover Survey (JOLTS) in the US, or Office for National Statistics (ONS) Vacancy Survey in the UK. Nonetheless, JOLTS has its own representativeness issues,[3] as do vacancy databases maintained by public employment agencies in Europe (Drahokoupil and Fabo 2022; Hershbein and Kahn 2016). In most countries, however, firms' reporting of vacancies to public employment services is voluntary, and as such, there is no readily available source for comparing the structure of labour demand. In summary, the key challenge is the problematic

[2] Matching can to some extent be proxied by, for example, the number of clicks on a particular vacancy, which suggests interest in a given position; or alternatively, number of views of a particular CV (Kureková and Žilinčíková 2018).

[3] Robust evidence exists that many hires (perhaps as high as 20 per cent), and thus the job openings, are not mediated through vacancies, which seems to result in systematic under-reporting of vacancies in JOLTS (Davis, Faberman, and Haltiwanger 2013)

accessibility of such data – which is also one of the reasons for using the online labour market data in the first place. This is particularly the case for developing countries.

In Table 1 we review several studies, and provide the basic conclusions of the comparison. The key finding is that while there are some discrepancies in the structure of vacancies, the general picture painted by the online job vacancies largely corresponds to other data sources. Importantly, some papers that have explored the biases of widely used online sources, such as Burning Glass data, are used by subsequent studies as a reference to understand the nature of discrepancies. For example, reference to the paper by Hershbein and Kahn (2016) appears widely in papers that study US labour market with Burning Glass data. Other studies refer to past methodological discussions (such as Kureková et al. 2015) in their brief acknowledgement of online data's limitations (see Table 2).

With respect to biases, the published studies largely concur that the share of online vacancies is over-represented in sectors such as ICT or finance; while those in hospitality, food service, manufacturing, and particularly public service, tend to be under-represented. Interestingly, some difference is observable between variants of capitalism – the healthcare sector tends to be over-represented in the US vacancies but under-represented in Europe, possibly due to the public sector's much stronger role in Europe than in the US. Furthermore, white-collar, skilled jobs tend to be over-represented, while trades and manual positions tend to be under-represented. However, there is no clear line between white-collar vacancies being more posted online than blue-collar openings, as public jobs are prevailingly white-collar, but tend to be less advertised online. In addition to the type of work, hiring practices and turnover of jobs can be additional factors that shape the probability of vacancies and their volume appearing online. Overall, while some types of jobs might be under-represented in respective online labour markets, sample sizes nevertheless tend to be sufficient to conduct an effective analysis. Moreover, as explained above, for certain (typically exploratory) questions, these biases are of secondary importance and do not prohibit further analyses.

**▶ Table 1: Nature of discrepancies between online data and representative data: selected studies focusing on skills analysis**

| Study (reference) | Location of relevant information | Source of online job vacancies (OJVs) | Alternative data source | Discrepancies identified |
|---|---|---|---|---|
| Hershbein and Kahn (2016) | Section A.1 of the Internet appendix contains a lengthy discussion of representativeness of the OJV data. This source is cited by many empirical papers. | Burning Glass US data | Sectoral structure comparison with JOLTS<br><br>Occupation distribution comparison with CPS New Jobs and OES, including in time | Sectors match reasonably well; Burning Glass (BG) is over-represented in healthcare and social assistance (+2 pp), finance and insurance, and education. It is under-represented in accommodation and food services (-5 pp), public admin/government, and construction.<br><br>BG has a much larger representation of computer and mathematical occupations (four times higher than shares in OES and CPS), as well as management, healthcare practitioners, and business and financial operations. On the other hand, BG data are under-represented in transportation, food preparation and serving, production and construction, among others. Representativeness in terms of occupation structure is found to be largely constant across time. |

| Study (reference) | Location of relevant information | Source of online job vacancies (OJVs) | Alternative data source | Discrepancies identified |
|---|---|---|---|---|
| Burke et al. (2020) | Section A.1.2. of the Internet appendix | Burning Glass US data | Sectoral structure, including over time.<br>Occupation based on Minnesota Job Vacancy Survey | The BG data are over-represented in finance and insurance (+7 pp), healthcare and social assistance, education services. Meanwhile, they are under-represented in food services and accommodation (-7 pp), public administration and government, and construction. Changes over time are very small.<br>Occupation-wise, BG data are over-represented in mathematical and computer professions (+9 pp), management (+7 pp), and business and financial (+6 pp). They are under-represented in food preparation and serving (-7 pp), healthcare (-5 pp), and transportation (-4 pp). |
| Acemoglu et al. (2020) | Figure A1 | Burning Glass US data | Sectoral and occupation comparison with JOLTS and OES | Findings closely match Hershbein and Kahn (2016) and Burke et al. (2020). |
| Turell et al. (2019) | Section 3.2."Coverage and representativeness bias" | Reed co.uk – a leading UK job site | Sectoral breakdown based on ONS Vacancy Survey | The mean annual ratios of the individual sectors in the two data sources are compared. For professional and scientific activities, ICT and administration, the Reed data are of comparable magnitude to the ONS estimates of vacancies. The largest differences were identified for public administration and manufacturing. |
| Sostero and Fernandez-Macias (2021) | Section 5 Online job ads as representation of the labour market | Burning Glass UK data | Economic survey by ONS (job stocks, not vacancies) | Professional, associate professional and administrative occupations show a ratio approaching one advertisement for one person employed. Whereas in elementary occupations and skilled trades, there is one vacancy for 100 or even 1,000 people employed. |
| Drahokoupil and Fabo (2022) | Tables 3 and 4 | Profesia.sk – a leading job site in Slovakia | Sectoral and occupational comparison with vacancies registered by the public employment agency (UPSVAR), including over time | Professionals over-represented in OJVs (+8 pp to +15 pp), crafts and trades under-represented {-3 pp to -10 pp). Occupations-wise, ICT is very over-represented (+12 pp to +17pp), while manufacturing are under-represented {-6 pp to -12 pp), transportation (-9 pp to<br>-10 pp), and education, healthcare and culture (-6 pp to -14 pp). The pattern is largely stable over time. |
| Marinescu (2017) | Discussion in text, p.17 | CareerBuilder.com | Sectoral structure comparison with JOLTS<br>Geographical distribution of vacancies with JOLTS | IT, finance and insurance, real estate, rental and leasing are over-represented, while local government, accommodation and food services, other services, and construction are under-represented.<br>CareerBuilder has the same geographic distribution as JOLTS.<br>Similar overtime trends are also identified, with a correlation of 0.57 (statistically significant at 90 per cent) |
| Kureková and Žilinčíková (2018) | Table 2, descriptive statistics | Profesia.sk – a leading job site in Slovakia | Comparing online CVs to Labour Force Survey data, examining demographic characteristics | Online data are unbiased by gender and age, but are biased towards individuals with higher education. Analysis is restricted to young people aged up to 35 years. |

## D. Fluctuations in online labour market data

The fluctuations in online labour market data intake have been studied extensively, in particular for online job vacancies. It appears that the online labour vacancy intake fluctuates for a variety of reasons. Fluctuations as such are unproblematic when they reflect actual changes in the labour market, but they might be a methodological concern if they include changes in biases over time. However, we are not aware of any literature discussing the representativeness implications of flows changing over time. Most studies we identified used fluctuations in online job vacancy data to identify and measure actual labour market changes, and found online job vacancy data to be an accurate representation of real shifts.

Fluctuations can reflect economic cycles and macroeconomic trends, while they might also be indicative of structural changes within or between occupations. De Pedraza and his co-authors (2019), building on standard labour economics literature, decomposed the variation into three components: trend cycle, seasonal, and irregular (reflecting macroeconomic shocks). They identified similar patterns in all three components, in online data and a National Statistical Office dataset. They also found an underlying trend of increase in the number of online job vacancies over time, which is likely to reflect the increased importance of the Internet as the "labour market matchmaker".

Importantly, the online job vacancies data are capable of capturing fluctuations not just in the number of vacancies but also in their structure, which is possibly relevant for understanding skill demand. For example, Beblavý, Kureková and Haita (2016) pooled job vacancy data for a number of years (2007–2011) to enlarge the underlying sample. They noted that low- and medium-skilled vacancies grew in their relative share among all vacancies in 2010 and 2011, and interpreted this as a (structural) rise in demand for less-skilled jobs in the Slovak labour market, during the post-2008 recovery phase. They focused their analysis on identifying skill intensities of traditional jobs (electrician, cook, driver), as well as "new" occupations in the context of structural and technological advances, such as courier or porter, to identify how these are seen in the Slovak labour market. Hence, in their case, shifts across occupations were a focus of their analysis, and not a problem of the data. Nonetheless, there is a precedent for an online data source being found unreliable, that was once thought to be robust,. Google Flu Trends is a good example: for a considerable time, it predicted the actual doctor visits fairly well, but strongly overestimated the growth of infections, in reaction to a flurry of media reports about a flu pandemic causing people to search for flu symptoms even when not feeling sick (Lazer et al. 2014).

Overall, the general picture appears to be that the differences between online job vacancies and alternative vacancy data, in terms of sectoral and occupational structure, remain quite stable over time (Burke et al. 2020; Drahokoupil and Fabo 2022; Hershbein and Kahn 2016; Lovaglio, Mezzanzanica, and Colombo 2020). This is an important consideration from the analytical perspective, and some researchers (e.g. Hershbein and Kahn, 2016) have used this longitudinal stability as a justification for research using online data.

Long-term trends notwithstanding, the important strength of online labour market data lies in the ability to identify the "irregular" movements in skills demand caused by macroeconomic shocks. For example, the COVID-19 pandemic represented a major shock for the labour market that had a very uneven impact in different sectors: while some segments of the labour market, such as the ICT industry, seamlessly shifted to working from home and saw demand for their services increase, areas such as hospitality or tourism were devastated (Kahn, Lange, and Wiczer 2020). Box 4 summarizes some of the key findings from this literature regarding changes in online data. The overall observation is that labour market shock was widespread across sectors, and that online labour market data well described real trends in the respective economies. Discussions about fluctuations in online data related to the COVID-19 pandemic further stress that online data need to be interpreted in context, and with knowledge of the particularities of specific labour markets.

**▶ Box 4: Fluctuations in online labour market data during the COVID-19 pandemic: Selected findings**

- The OECD used the Burning Glass dataset to observe the impact across industries. The report also shows a strong correlation between the stringency of the pandemic measures and the decline in posted online job vacancies in the US (OECD 2021).
- A study using data from a popular Swedish private online job board found that together with a sharp decrease of vacancies, the average number of clicks per vacancy also declined because fewer people were actually looking for jobs. The decline in demand was most severe in the sectors particularly affected by the pandemic (Hensvik, Le Barbanchon, and Rathelot 2021).
- A similar trend was observed in Slovakia with the dominant online job portal Profesia.sk, with a marked dip in both demand and supply for labour after March 2020. The labour market recovered over the next months, and growth fully resumed in 2022, overtaking the 2019 levels. The greatest demand in March 2022 existed in occupations which were hit the most by the lockdown measures, specifically hospitality and catering (Profesia 2022).
- Another study focusing on the US labour market, based on Burning Glass data, found evidence of down-skilling caused by the pandemic, observing that firms cut back on hiring for high-skilled more than low-skilled jobs (Campello, Kankanhalli, and Muthukrishnan 2020). Nonetheless, another study using LinkedIn data found that workers from affected industries tended to apply for jobs in sectors such as ICT or healthcare, resulting in reallocation of labour and, potentially, changes in skills development (Bauer et al. 2021).
- A study which combined online data (Burning Glass) with representative data (unemployment claims and employment statistics data) in the US found that a decline in economic activity in the online labour market was reflected in representative sources of data. No difference between "essential" and "non-essential" sectors was found, in terms of a decline in online job postings (with the exception of essential retail) (Forsythe et al. 2020)

As regards the supply side more broadly, compared to job vacancies, we do not find the same evidence that fluctuations in online CVs availability match the trends existing in other data sources. Job search intensity is an important predictor of labour market developments (Mukoyama, Patterson, and Şahin 2018), which would make such an indicator very useful; but we did not discover any published research that attempted to estimate the fluctuations of CVs being posted on job portals over time. Some job portals publish raw trends in their data. For example, according to Profesia.sk data, while job applicant data fell abruptly in response to the first lockdown in March 2020, applicant data fairly quickly recovered. This is most likely due to a structural shift in demand for labour (Profesia 2022).

## E. Techniques and approaches to address non-representativeness and other biases

While many empirical studies are not concerned with the issues of representativeness or other related biases, we identified a set of studies that take a rigorous approach in their use of online labour market data. In the process of our mapping exercise, we found a variety of approaches to address biases of the data. In this section we discuss these, providing examples of concrete studies which have adopted a particular approach. As we came across far fewer studies using online applicant data, most of the discussion is based on studies using online job vacancies.

The approaches identified in the literature can be broadly divided into three categories: (1) *statistical techniques*, such as weighting and data cleaning techniques; (2) the *research design approach*, which involves adapting research questions and the research focus to issues well covered by online labour market data; and (3) the *mixed-methods approach*, where online job market data are complemented by other research strategies, including qualitative methods. In some cases, these approaches are used in combination and are seen as mutually exclusive.

### i. Statistical techniques

Statistical techniques at two stages of the analytical process are relevant: (1) the data cleaning and data preparation stage; and (2) the data analysis phase. In some instances, to avoid bias in online job vacancies (such as from duplications), data cleaning techniques employ job vacancy aggregators, which then provide

cleaned and structured datasets for external analytical use on a commercial basis (e.g. Burning Glass). A third category of approach that we categorize under statistical techniques is the treatment of sample size as a self-correcting mechanism.

First, standard data cleaning techniques, including rule-based and statistical (i.e. outlier approaches, de-noising data) are also applicable in the analysis of big data sources, such as the online labour market data (Mezzanzanica and Mercorio 2019). Here, we highlight that for assessing the toolkit available and used for de-biasing online vacancies and applicant data, the whole process of data management is relevant. For example, data matching or de-duplication are tools that can systematically de-noise underlying labour market data at the data access and preparations stage; that is, before any further analytical and econometric methodologies are applied (Blasquez and Domenech, 2018). These steps are taken to increase the veracity of data before any analytics takes place, in the pre-processing stage (Branco 2020).

Second, using large-scale representative surveys and calculating weights to correct the structure of the labour force to online demand (a sectoral, occupational focus) is perhaps the most rigorous approach that we identified for de-biasing online data. Turrell and his co-authors (2019) set out to transform online job vacancy data from a leading UK job site (Reed co.uk) into economic statistics. When comparing the mean annual ratios of the individual sectors in the online data and the ONS Vacancy Survey, they did not identify biases in professional and scientific activities, ICT and administration, whereas the largest differences appeared for public administration and manufacturing. In their study of interfirm labour mobility and innovation, Masso et al. (2012) compared CV Keskus data in Estonia to the national Labour Force Survey, and used sample weights to adjust biases which they identified (in gender and nationality).

Štefánik (2012) explored the suitability of online CV and vacancy data for tertiary-educated applicants by comparing the occupational (ISCO) and sectoral (NACE) structure of online CVs and online job vacancies in Slovakia, to the structure of the workforce in a representative Labour Force Survey (LFS). His first step was to run chi-squared goodness of fit statistics: he found that in terms of occupational match, online job vacancy data fit the overall tertiary-educated population in the Slovak LFS, but significant differences existed when comparing the sectoral composition of LFS and online data. In selecting segments for further analysis, he composed a matrix of ISCO–NACE cells and studied them with online data; this mapped the LFS cells well (i.e. technicians in public services, professionals in construction). He compared online CV data for tertiary-educated jobseekers with university graduates' data, and found a surprisingly good fit.

The key problem to highlight with respect to weighting, beyond the general difficulty in finding an alternative dataset as a basis for reweighting, is that online job vacancies tend to be richer than the alternative sources. For example, it is generally rarely the case that the representative dataset would contain information about sector and occupation, coded in a way that allows direct comparison with the online data. It follows that a reweighting strategy based on sectoral representation will align the online data closely with the representative source, in terms of the representation of individual sectors; but this will only address the difference in the studied occupations, to the extent that this difference is caused by different sectoral coverage (Turrell et al. 2019). In the previous parts of this paper we emphasized other problems with applying weights to online job vacancy data: mainly, the essential difference between measuring stocks/matched jobs versus flows/unmatched jobs, in (representative) employment statistics and in online vacancy data, respectively.

In spite of these issues, there appears to be a push towards using weighted data, at least for online job vacancy analysis. For example, a methodology for calculating weights has been recently proposed by OECD researchers, for six of the economies covered by the Burning Glass data (Cammeraat and Squicciarini 2021). Furthermore, specialized weighting strategies have been developed for specific issues, such as vacancies appearing online only for a short time, which thus may be missed in the Web scraping process (Marconi 2022).

Third, in addition to weighting, we have identified some studies where the sheer size of the online data has been exploited as a "strategy". In principle, the richer the data, the better models can be developed. Some researchers believe that due to the large number of observations typically present in big data about labour markets, the data are "self-corrective", in the sense that it is possible to include a very large number

of control variables; thus, any noise remaining in the data will be irrelevant given the sheer size of the sample (Mezzanzanica and Mercorio 2019). The large number of observations also makes it possible to remove even a significant share of observations if required by research design, and retain sufficient statistical power for the analysis (for example, Drahokoupil and Fabo 2022; or Hershbein and Kahn 2016, who removed all observations where the firm posting the vacancy could not be properly matched).

Lastly, from the perspective of data representativeness, the relatively stable industrial and occupational distribution of online job vacancies alleviates concerns about the reliability and validity of statistical inference on the basis of this data source. This is probably why most studies do not propose "hard" statistical counters, such as weighting or controlling for vacancy posting over time. Instead, studies rely on tools such as benchmarking against an established data source (Lovaglio, Mezzanzanica, and Colombo 2020), or robustness checks using an alternative data source (Forsythe et al. 2020). In principle, other aspects of volatility can be addressed by statistical means, such as dummy variables for time, or using averages over segments of time if and when appropriate.

## ii. Research design approach: Adapting the research questions and research focus

Second, adapting the research questions and research focus to the quality of data has also been an explicit strategy of researchers working with online data. A multitude of studies focus on selected aspects of the labour market by sector, occupation, or educational level. Commonly covered segments are the ICT and software industry (Bilal et al. 2017; Capiluppi and Baravalle 2010), the academic job market (Brandas et al. 2016), entry-level labour markets for students (Kureková and Žilnčíková, 2016), professional or more educated job seekers (Deming and Kahn, 2018, Hemelt et al. 2021), and IT skills (Fabo and Drahokoupil, 2020, Fabo and Kahanec, 2020). For instance, Kureková and Žilinčíková (2016) worked with the population of vacancies of a dominant job portal in the Slovak labour market, with a market coverage of 80 per cent. Their question was well-suited to their data, as they sought to understand to what extent students are represented in the low-skilled segment of the labour market, which implied the substitution of low-skilled workers by students.

This research strategy has also appeared in studies that have used online vacancy data to predict aggregate trends (de Pedraza et al. 2019; Lovaglio, Mezzanzanica, and Colombo 2020). For example, Štefánik et al. (2022) argue that at such an aggregate level, non-representativeness is lesser concern. The authors use online job vacancy data from the Profesia job portal in Slovakia to now-cast and to test the predictive power of online vacancy data for GDP growth, unemployment and employment trends, as well as working time. They find strong evidence for the robustness of online data to accurately predict these aggregate trends in Slovakia.

## iii. Multimethod research

Lastly, a mixed-methods approach essentially aims to validate or understand the biases of online data with the use of alternative research methods, including qualitative interviews (such as with HR managers) to correctly interpret the biases. Triangulation of data sources to study a particular question, when online data represent only one source, has also been applied (Huang et al. 2009; Masso et al. 2016). Another example of a non-Western application is the SkillsFuture government project in Singapore (www.myskillsfuture.gov.sg/), which uses the data from job postings combined with insights from stakeholders' interviews to support policy and programme design.

# ▶ Conclusion and implications

This paper has studied how online data on labour markets can be used to describe, analyse, understand, refine or predict labour market trends. This relates to the more general aspect of labour demand, but also more specifically to skills demand and skills changes, in light of the existing deficiencies of online data: specifically, a lack of representativeness.

The research field that relies on online labour market data has expanded rapidly in recent years. Our review revealed that this expansion is characterized by several features. First, research continues to focus on single countries, with a small number of attempts at comparative work. Second, research using online vacancy data appears more frequently than research relying on job applicant/CV data. This might be caused by the generally easier access to information about vacant jobs, than to applicant information, which some portals offer on a paid basis. However, it could also reflect a greater interest in research on the demand for labour, for which fewer alternative survey data sources are available. Third, while this research field continues to be driven by a focus on advanced economies, mainly the US and EU, there is also an evident trend towards expanding this research to developing countries´ labour markets. These studies, on average, tend to be less concerned with data biases than research focusing on advanced economies. Fourth, somewhat to our surprise, the data limitations of online sources used to study labour markets often remain undiscussed in terms of the biases, non-representativeness, or other potential pitfalls of these data. Moreover, this situation does not seem to be improving with time – in contrast to those studies that have taken a rigorous approach to understanding the qualities of online labour market data, and have addressed them using various methodological or research design approaches.

This paper advances the current debate by offering a mapping of biases recognized in online labour market vacancies and CV data, and an overview of approaches and techniques to address the identified biases. We highlight that legitimate research approaches exist, which are inductive in nature, focused on discovering patterns and trends in underlying data. These methods are by definition less concerned with generalizability of findings, as they have different objectives. For this body of research, online labour market data open new avenues for understanding developments in labour markets. (Near) real-time availability, granularity, relative affordability and size represent some of the key qualities which make online labour market data uniquely suitable for many forms of analyses – traditional as well as novel ones.

Biases in online labour market data emerge due to a myriad of factors, including populations' varying levels of Internet access; different resources, motivations or opportunities for advertising a vacancy, among different sectors or firms of different size; as well as higher levels of informality in some economies or sectors. Most evaluations of biases pertain to developed countries, and these have identified over-representation of some sectors (ICT, finance) and under-representation of others, mainly the public sector and manufacturing. While there are more skilled than non-skilled vacancies found in the online world, there is no clear evidence of white-collar vacancies' over-representation in relation to blue-collar, as public jobs are prevailingly white-collar. In addition to the type of work, additional factors such as hiring practices, job turnover or the levels of informality can shape the probability of vacancies appearing online.

With respect to the nature of discrepancies, however, these are typically not paramount to hinder research and reliability of findings. Different techniques have been adopted to deal with the non-representativeness problem. These include statistical approaches such as weighting, and tools applied at the data preparation phase (de-duplication, data matching). Dummy variables can be used to statistically account for the effect of fluctuations in data over time. An alternative approach to using this data has relied on adapting the research focus and research objectives to the quality of data. Hence, a multitude of studies focus on studying a narrower aspect of the labour market, such as academic jobs, the IT sector, or students, as these groups are well covered in the online segment. Lastly, mixed-method approaches essentially aim to validate or understand the biases of online data with the use of alternative research methods, including qualitative interviews (such as with HR managers), in order to correctly interpret the biases. Triangulation

of data sources has also been applied to study a particular question, when online data or Web-based data represent only one source.

In conclusion, addressing the biases of online labour-market vacancy data is a strenuous task, as the population of vacancies is inherently unknown. Nevertheless, representativeness problems are likely to vary in different contexts; they should be evaluated at the country level, and with respect to specific research objectives, research questions, and existing alternatives, in terms of accuracy, granularity, costs and timeliness. In essence, research ethics and transparency are key. Any analysis using online labour market data should be embedded in a particular context, and analytical steps and decisions need to be clearly defined and specified. Ideally, analytical "protocols" should be recorded and made available upon request, and also to enable replication of the analysis, as is possible with any survey data.

Although from the point of view of representativeness and generalizability, representative surveys are the first-best option in many developed countries, the accessibility in many developing countries is very limited. For these countries, online data, including web-collected surveys, represent the preferred alternative, opening new horizons and opportunities in research. Web-collected data also provide information that is difficult to gather even from representative data, such as wages (real and expected) or aspects of gender biases in labour markets. Not least, it is important to realize that representative data also suffer from biases, such as non-response and coverage bias.

Importantly, the nature of online labour market data analysis, which lies in the intersection of research, policy and industrial applications, makes it possible to pool substantial resources to increase the pace of progress. CEDEFOP (2019), for example, is committed to making training sets and ontologies public, under creative commons licences. Leading data vendors are working closely with the academics, even co-authoring papers, to a far greater extent than in other domains of social science. As a result, online labour vacancy research is one of the key areas where there is rapid advancement in the application of machine learning and artificial intelligence in social science. Furthermore, in the recent past, systematic efforts have emerged to consider biases, such as the ESSnet Big Data project,[4] which will assist further efforts to address biases.

Finally, beyond the issue of representativeness, there are other key challenges of working with big data in the labour market, which we did not discuss but would like to raise in conclusion. These are linked to issues of privacy and confidentiality (Einav and Levin 2014). While the sheer number of collected vacancies and CVs ensures on the one hand a form of anonymization, on the other hand, ethical concerns emerge regarding the (explicit) consent of firms and individuals that their information can be used for analytical and research purposes. While data are typically anonymized, monitoring of data management and research ethics appears more problematic for online (labour market) data than with surveys, as access to the online data is much less centralized, and *ad hoc*. More general regulations, such as the GDPR regulation adopted and enforced within the European Union countries, might be considered a gold standard case, as it provides specific guidelines for the protection of individual personal details (Mezzanzanica and Mercorio 2019).

[4] https://ec.europa.eu/eurostat/cros/content/essnet-big-data-1_en#WPB_Online_job_vacancies

# Annex

▶ **Table 2: Overview of studies using online labour market data**

| Short reference | Country and time coverage | Type of online data (CV or vacancy) | Source of data (specific website) | Methods applied (single or multiple; statistical method) | Focus of the analysis/ research question | Biases identified | Approaches to address biases | Representativeness discussed (Y/N, and conclusion) | Generalizability of findings discussed (Y/N, and conclusion) |
|---|---|---|---|---|---|---|---|---|---|
| Almaleh et al. 2019 | Saudi Arabia | Job postings, curriculum websites – King Abdulazis University | Bayt.com, gulftalent.com, naukrigulf.com, linkedin.com/jobs | Naïve Bayes algorithms | Matching between curricula and job postings, focused on job title and job description | Not explicitly defined | Not explicitly defined | None | Analysis is done on English jobs; Arabic jobs are excluded, but the authors consider the findings accurate to inform curricular development |
| Xu et al. 2017 | China | Job vacancy data | Chinese labour recruitment websites: ChinaHR, 51jobs, Zhaopin | Deep learning methods (AI) – unsupervised learning, supervised learning algorithm, and human assessment | Develop a framework for systematizing vacancies into the Chinese occupational categorization (CGCO) | None | None | None | The authors inductively (from scraped data), and with input from automated (AI) techniques, construct a "corpus", i.e. a classification for sorting vacancies into categories, that can be used beyond the data they built it on |
| Bilal et al. 2017 | India | Job vacancies | Rosee.pk, the most popular online job platform for software and IT jobs | Frequency distribution | Industrial demand in the Indian software industry by different characteristics | None | None | None | Results are presented as a useful input for students to gain skills which are demanded in the sector |
| Brandas et al. 2017 | Worldwide academia job market | Academic job vacancies | A number of academic job websites, worldwide coverage of jobs<br>e.g. www.academicpositions.eu | Web content Mining<br>Data mining analysis (WEKA), k-means clustering algorithm<br>Data spatialization | Empirically analyse and process vacancies in order to show ways to set up a "Labour Market Decisions Support System" | Highlight the lack of semantic and structural compatibility of data | In the future, rely on Web-semantics, Web 3.0. | None | Not discussed |

| Short reference | Country and time coverage | Type of online data (CV or vacancy) | Source of data (specific website) | Methods applied (single or multiple; statistical method) | Focus of the analysis/ research question | Biases identified | Approaches to address biases | Representativeness discussed (Y/N, and conclusion) | Generalizability of findings discussed (Y/N, and conclusion) |
|---|---|---|---|---|---|---|---|---|---|
| Fabo and Kahanec 2020 | Netherlands, August to December 2016 | Vacancy | Textkernel (vacancy aggregator) | Descriptive comparison with online Web survey and OECD PIAAC data | IT skills requirements per occupation | Lack of explicit mention of skill demand | None | None | Yes, it is mentioned that IT skills are more commonly explicitly mentioned in skilled manual occupations |
| Drahokoupil and Fabo 2020 | Slovakia (2011–2017) | Vacancy | Profesia.sk combined with public company registry | Logit | IT skills requirement per occupation, in foreign and domestic-owned firms | None | None | Comparison of occupational structure with official statistics | None |
| Beblavý et al. 2016 | US September 2013 – August 2014 | Vacancy | Burning Glass (vacancy aggregator) | Correlation | Skills requirement in the 30 most common occupations in the US | None | None | Acknowledgement that data might not be representative | Acknowledgement that data are not generalizable |
| Fabo et al. 2017 | Visegrad 4 | Vacancy | Profesia.sk, jobs.cz, profession.hu and pracuj.pl | Correlation, OLS | English and German language skills and wages | None | None | Acknowledgement that data might not be representative | None |
| Ecleo and Galido 2017 | Philippines | CVs | 100 LinkedIn profiles of data scientists in Philippines | | What skills do data scientists have? | None | None | | |
| Kotzab et al. 2018 | Germany | Vacancy | 1000 job vacancies posted on top German job portals | Sematic analysis using Bayesian techniques | Key competences in logistics | None | None | None | None |
| Rios et al. 2020 | US | Vacancy | 142,000 vacancies from two US job portals | Descriptives | Skills requirements of twenty-first century jobs | Acknowledged – a selection of firms who advertise ads to more affluent or motivated individuals | Ads which required post-secondary level credentials | Might not be representative due to omission of offline ads, and scraping of selected websites | Interpret findings as generalizable |
| Acemoglu et al. 2020 | US | Vacancy | Burning Glass | Regression | AI-exposed occupations | Demonstrate that data used are close to the population of vacancies | Data size and coverage | Yes | Yes |
| Chowdhury et al. 2018 | India | Vacancy | 800,000 vacancies on Babajob | Regression | Gender bias for low- and high-skilled occupations | None | None | None | None |
| Matsuda 2019 | Pakistan | Vacancy and CV | 5 million CVs and 412 vacancies on Rozee.pk | Descriptives | Labour market matching in Pakistan | None | None | None | None |

| Short reference | Country and time coverage | Type of online data (CV or vacancy) | Source of data (specific website) | Methods applied (single or multiple; statistical method) | Focus of the analysis/ research question | Biases identified | Approaches to address biases | Representativeness discussed (Y/N, and conclusion) | Generalizability of findings discussed (Y/N, and conclusion) |
|---|---|---|---|---|---|---|---|---|---|
| Skhvediani et al. 2021 | Russia | Vacancy | 100 job adverts on Headhunters | Descriptives | Skill requirements for data scientists in Russia | None | None | None | None |
| Debortoli et al. 2014 | US, Canada, Australia, UK | Vacancy | Monster.com | Singular value decompos-ition | Comparing business intelligence and big data skills | None | None | None | None |
| Muller and Safir 2019 | Ukraine | Vacancy | 2500 job vacancies from headhunter | Descriptives | Job requirements in Ukraine | None | None | None | Mentioned that jobs published on the website mostly target high-skilled workers |
| Nomura et al. 2017 | India | Vacancy data and job applicant data (job search behaviour) | Babajobs – one of the leading job portals | Econometric analysis<br>Text/content analysis<br>Econometric analysis/<br>probit model<br>Predictive analysis with longitudinal data<br>Randomized control trials | Gender differences in wage offers<br>Skills demand<br>Job search behaviour<br>Wage trends – real wage offers by occupation and location<br>Reduce information asymmetries | Informal labour market in India, high youth un-employment<br>Self-selection and sample error | Babajobs covers less skilled and informal markets better than other portals (phone-based access to job seekers) | Yes, reference to Kurekova et al. 2015; external data that could be used to counter some of the biases are very limited in India | Yes, reference to Kurekova et al. 2015 |

# References

Abraham, Katharine G. 1983. "Structural/Frictional vs. Deficient Demand Unemployment: Some New Evidence." The American Economic Review 73 (4): 708–24.

Acemoglu, Daron, and David Autor. 2010. "Skills, Tasks and Technologies: Implications for Employment and Earnings." Working Paper 16082. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w16082.

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo. 2020. "AI and Jobs: Evidence from Online Vacancies." Working Paper 28257. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w28257.

Álvarez, Andrés, and Marc Hofstetter. 2014. "Job Vacancies in Colombia: 1976–2012." IZA Journal of Labor & Development 3 (1): 15. https://doi.org/10.1186/2193-9020-3-15.

Anastasopoulos, L. Jason, George J. Borjas, Gavin G. Cook, and Michael Lachanski. 2021. "Job Vacancies and Immigration: Evidence from the Mariel Supply Shock." Journal of Human Capital 15 (1): 1–33. https://doi.org/10.1086/713041.

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D. Shapiro. 2014. "Using Social Media to Measure Labor Market Flows." Working Paper 20010. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w20010.

Apaza, Honorio, Américo Vidal, and Josimar Chire. 2021. "Job Recommendation Based on Curriculum Vitae Using Text Mining." In , 1051–59. https://doi.org/10.1007/978-3-030-73100-7_72.

Askitas, Nikolaos, and Klaus F. Zimmermann. 2009. "Google Econometrics and Unemployment Forecasting." Applied Economics Quarterly 55 (2): 107–21.

———. 2015. "The Internet as a Data Source for Advancement in Social Sciences." International Journal of Manpower 36 (1): 2–12. https://doi.org/10.1108/IJM-02-2015-0029.

Autor, David H. 2001. "Wiring the Labor Market." Journal of Economic Perspectives 15 (1): 25–40. https://doi.org/10.1257/jep.15.1.25.

Azar, José, Emiliano Huet-Vaughn, Ioana Marinescu, Bledi Taska, and Till von Wachter. 2019. "Minimum Wage Employment Effects and Labor Market Concentration." Working Paper 26101. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w26101.

Azar, José, Ioana Marinescu, Marshall Steinbaum, and Bledi Taska. 2020. "Concentration in US Labor Markets: Evidence from Online Vacancy Data." Labour Economics 66 (October): 101886. https://doi.org/10.1016/j.labeco.2020.101886.

Barron, John M., and John Bishop. 1985. "Extensive Search, Intensive Search, and Hiring Costs: New Evidence on Employer Hiring Activity." Economic Inquiry 23 (3): 363–82. https://doi.org/10.1111/j.1465-7295.1985.tb01773.x.

Bauer, Anja, Tobias Hartl, Christian Hutter, and Enzo Weber. 2021. "Search Processes on the Labor Market during the Covid-19 Pandemic." CESifo Forum 22 (04): 15–19.

Beblavý, Miroslav, Mehtap Akgüc, Brian Fabo, and Karolien Lenaerts. 2016. "Occupations Observatory-Methodological Note." CEPS Special Report, no. 144.

Beblavý, Miroslav, Lucia Kureková, and Corina Haita. 2016. "The Surprisingly Exclusive Nature of Medium- and Low-Skilled Jobs: Evidence from a Slovak Job Portal." Personnel Review 45 (2): 255–73. https://doi.org/10.1108/PR-12-2014-0276.

Beblavý, Miroslav, Ilaria Maselli, and Marcela Veselková, eds. 2014. Let's Get to Work!: The Future of Labour in Europe. Brussels: Centre for European Policy Studies.

———, eds. 2015. Green, Pink and Silver?: The Future of Labour in Europe. Brussels: Centre for European Policy Studies.

Bilal, Muhammad, Nadia Malik, Maham Khalid, and M. Ikram Ullah Lali. 2017. "Exploring Industrial Demand Trends in Pakistan Software Industry Using Online Job Portal Data." University of Sindh Journal of Information and Communication Technology 1 (1): 17–24.

Birba, Ousmane, and Abdoulaye Diagne. 2012. "Determinants of Adoption of Internet in Africa: Case of 17 Sub-Saharan Countries." Structural Change and Economic Dynamics 23 (4): 463–72. https://doi.org/10.1016/j.strueco.2012.06.003.

Blair, Peter Q., and David J. Deming. 2020. "Structural Increases in Skill Demand after the Great Recession." Working Paper 26680. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w26680.

Blazquez, Desamparados, and Josep Domenech. 2018. "Big Data Sources and Methods for Social and Economic Analyses." Technological Forecasting and Social Change 130 (May): 99–113. https://doi.org/10.1016/j.techfore.2017.07.027.

Bloom, David E., Richard B. Freeman, and Sanders D. Korenman. 1988. "The Labour-Market Consequences of Generational Crowding." European Journal of Population / Revue Européenne de Démographie 3 (2): 131–76.

Boselli, Roberto, Mirko Cesarini, Stefania Marrara, Fabio Mercorio, Mario Mezzanzanica, Gabriella Pasi, and Marco Viviani. 2018. "WoLMIS: A Labor Market Intelligence System for Classifying Web Job Vacancies." Journal of Intelligent Information Systems 51 (3): 477–502. https://doi.org/10.1007/s10844-017-0488-x.

Boselli, Roberto, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. 2018. "Classifying Online Job Advertisements through Machine Learning." Future Generation Computer Systems 86 (September): 319–28. https://doi.org/10.1016/j.future.2018.03.035.

Brancatelli, Calogero, Alicia Marguerie, and Stefanie Brodmann. 2020. "Job Creation and Demand for Skills in Kosovo: What Can We Learn from Job Portal Data?" Working Paper. Washington, DC: World Bank. https://doi.org/10.1596/1813-9450-9266.

Brandas, Claudiu, Ciprian Panzaru, and Florin Gheorghe Filip. 2016. "Data Driven Decision Support Systems: An Application Case in Labour Market Analysis." Romanian Journal of Information Science and Technology 19 (1–2): 65–77.

Buchs, Helen, and Laura Alexandra Helbling. 2016. "Job Opportunities and School-to-Work Transitions in Occupational Labour Markets. Are Occupational Change and Unskilled Employment after Vocational Education Interrelated?" Empirical Research in Vocational Education and Training 8 (1): 17. https://doi.org/10.1186/s40461-016-0044-x.

Burke, Mary A., Alicia Sasser, Shahriar Sadighi, Rachel B. Sederberg, and Bledi Taska. 2020. "No Longer Qualified? Changes in the Supply and Demand for Skills within Occupations." Working Paper 20–3. Working Papers. https://doi.org/10.29412/res.wp.2020.03.

Cammeraat, Emile, and Mariagrazia Squicciarini. 2021. "Burning Glass Technologies' Data Use in Policy-Relevant Analysis: An Occupation-Level Assessment." Paris: OECD. https://doi.org/10.1787/cd75c3e7-en.

Campello, Murillo, Gaurav Kankanhalli, and Pradeep Muthukrishnan. 2020. "Corporate Hiring under COVID-19: Labor Market Concentration, Downskilling, and Income Inequality." Working Paper 27208. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w27208.

Campos-Vazquez, Raymundo M., Gerardo Esquivel, and Raquel Y. Badillo. 2021. "How Has Labor Demand Been Affected by the COVID-19 Pandemic? Evidence from Job Ads in Mexico." Latin American Economic Review 30 (May): 1–42.

Capiluppi, Andrea, and Andres Baravalle. 2010. "Matching Demand and Offer in On-Line Provision: A Longitudinal Study of Monster.Com." In 2010 12th IEEE International Symposium on Web Systems Evolution (WSE), 13–21. https://doi.org/10.1109/WSE.2010.5623576.

CEDEFOP. 2014. "Briefing Note - Skill Mismatch: More than Meets the Eye." March 19, 2014. https://www.cedefop.europa.eu/en/publications/9087.

———. 2019. "Online Job Vacancies and Skills Analysis: A Cedefop Pan-European Approach | VOCEDplus, the International Tertiary Education and Research Database." https://www.voced.edu.au/content/ngv:82496.

Chowdhury, Afra Rahman, Ana Carolina Areias, Saori Imaizumi, Shinsaku Nomura, and Futoshi Yamauchi. 2018. "Reflections of Employers'Gender Preferences in Job Ads in India: An Analysis of Online Job Portal Data." SSRN Scholarly Paper ID 3150092. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3150092.

Colombo, Emilio, Fabio Mercorio, and Mario Mezzanzanica. 2018. "Applying Machine Learning Tools on Web Vacancies for Labour Market and Skill Analysis."

Comyn, Paul, and Olga Strietska-Ilina. 2019. Skills and Jobs Mismatches in Low- and Middle-Income Countries. Geneva, Switzerland: ILO. https://www.semanticscholar.org/paper/Skills-and-jobs-mismatches-in-low-and-middle-income-Comyn/1a2c34a7e6c337b46fc0b004291c2c8d27f5c85f.

Das, Jishnu, and Quy-Toan Do. 2014. "US and Them: The Geography of Academic Research." VoxEU.Org (blog). February 11, 2014. https://voxeu.org/article/geographical-bias-top-journal-publication.

Das, Marcel, Peter Ester, and Lars Kaczmirek. 2018. Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies. Routledge.

Davis, Steven J., R. Jason Faberman, and John C. Haltiwanger. 2013. "The Establishment-Level Behavior of Vacancies and Hiring *." The Quarterly Journal of Economics 128 (2): 581–622. https://doi.org/10.1093/qje/qjt002.

Debortoli, Stefan, Oliver Müller, and Jan vom Brocke. 2014. "Comparing Business Intelligence and Big Data Skills." Business & Information Systems Engineering 6 (5): 289–300. https://doi.org/10.1007/s12599-014-0344-2.

Deming, David, and Lisa B. Kahn. 2018. "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals." Journal of Labor Economics 36 (S1): S337–69. https://doi.org/10.1086/694106.

Dijk, Jan van. 2006. "Digital Divide Research, Achievements and Shortcomings." Poetics, The digital divide in the twenty-first century, 34 (4): 221–35. https://doi.org/10.1016/j.poetic.2006.05.004.

———. 2020. The Digital Divide. 1st edition. Cambridge, UK ; Medford, MA: Polity.

Dörfler, Laura, and Herman van de Werfhorst. 2009. "Employers' Demand for Qualifications and Skills." European Societies 11 (5): 697–721. https://doi.org/10.1080/14616690802474374.

Drahokoupil, Jan, and Brian Fabo. 2016. "The Platform Economy and the Disruption of the Employment Relationship." ETUI Research Paper-Policy Brief 5.

———. 2022. "The Limits of Foreign-Led Growth: Demand for Skills by Foreign and Domestic Firms." Review of International Political Economy 29 (1): 152–74.

Dunlop, John T. 1966. "Job Vacancy Measures and Economic Analysis." In The Measurement and Interpretation of Job Vacancies, 27–47. NBER. https://www.nber.org/books-and-chapters/measurement-and-interpretation-job-vacancies/job-vacancy-measures-and-economic-analysis.

Ecleo, Jerina Jean, and Adrian Galido. 2017. "Surveying LinkedIn Profiles of Data Scientists: The Case of the Philippines." Procedia Computer Science, 4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia, 124 (January): 53–60. https://doi.org/10.1016/j.procs.2017.12.129.

Einav, Liran, and Jonathan Levin. 2014. "Economics in the Age of Big Data." Science 346 (6210): 1243089. https://doi.org/10.1126/science.1243089.

Ernesto, Caroleo Floro, and Pastore Francesco. 2016. "Overeducation: A Disease of the School-to-Work Transition System." In Youth and the Crisis. Routledge.

ESCO. 2015. "European Skills, Competences, Qualifications and Occupations." (https://ec.europa.eu/esco/portal/home#modal-one.

Fabo, Brian, Miroslav Beblavý, and Karolien Lenaerts. 2017. "The Importance of Foreign Language Skills in the Labour Markets of Central and Eastern Europe: Assessment Based on Data from Online Job Portals." Empirica 44 (3): 487–508. https://doi.org/10.1007/s10663-017-9374-6.

Fabo, Brian, and Martin Kahanec. 2018. "Can a Voluntary Web Survey Be Useful beyond Explorative Research?" International Journal of Social Research Methodology.

———. 2020. "The Role of Computer Skills on the Occupation Level." European Journal of Business Science and Technology 6 (2): 87–99. https://doi.org/10.11118/ejobsat.2020.006.

Fang, Hanming, Chunmian Ge, Hanwei Huang, and Hongbin Li. 2020. "Pandemics, Global Supply Chains, and Local Labor Demand: Evidence from 100 Million Posted Jobs in China." Working Paper 28072. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w28072.

Ferber, Robert, and Neil Ford. 1966. "The Time Dimension in the Collection of Job Vacancy Data." In The Measurement and Interpretation of Job Vacancies, 447–61. NBER. https://www.nber.org/books-and-chapters/measurement-and-interpretation-job-vacancies/time-dimension-collection-job-vacancy-data.

Fondeur, Y., and Frédéric Karamé. 2013. "Can Google Data Help Predict French Youth Unemployment?" Economic Modelling 30 (C): 117–25.

Forsythe, Eliza, Lisa B. Kahn, Fabian Lange, and David Wiczer. 2020. "Labor Demand in the Time of COVID-19: Evidence from Vacancy Postings and UI Claims." Journal of Public Economics 189 (September): 104238. https://doi.org/10.1016/j.jpubeco.2020.104238.

Gandomi, Amir, and Murtaza Haider. 2015. "Beyond the Hype: Big Data Concepts, Methods, and Analytics." International Journal of Information Management 35 (2): 137–44. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Gortmaker, Jeff, Jessica Jeffers, and Michael Lee. 2021. "Labor Reactions to Credit Deterioration: Evidence from LinkedIn Activity." SSRN Scholarly Paper ID 3456285. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3456285.

Haddad, Rabih, and Eunika Mercier-Laurent. 2021. "Curriculum Vitae (CVs) Evaluation Using Machine Learning Approach." In Artificial Intelligence for Knowledge Management, edited by Eunika Mercier-Laurent, M. Özgür Kayalica, and Mieczyslaw Lech Owoc, 48–65. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-80847-1_4.

Hensvik, Lena, Thomas Le Barbanchon, and Roland Rathelot. 2021. "Job Search during the COVID-19 Crisis." Journal of Public Economics 194 (February): 104349. https://doi.org/10.1016/j.jpubeco.2020.104349.

Hershbein, Brad. 2016. "Is College the New High School? Evidence from Vacancy Postings." https://research.upjohn.org/projects/169.

Hershbein, Brad, and Lisa B. Kahn. 2016. "Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings." Working Paper 22762. National Bureau of Economic Research. https://doi.org/10.3386/w22762.

Holt, Charles, and Martin David. 1966. "The Concept of Job Vacancies in a Dynamic Theory of the Labor Market." In The Measurement and Interpretation of Job Vacancies, 73–110. NBER. http://www.nber.org/chapters/c1599.pdf.

Horton, John J. 2011. "The Condition of the Turking Class: Are Online Employers Fair and Honest?" Economics Letters 111 (1): 10–12. https://doi.org/10.1016/j.econlet.2010.12.007.

Huang, Hayan, Lynette Kvasny, K.D. Joshi, Eileen Trauth, and Jan Mahar. 2009. "Synthesizing IT Job Skills Identified in Academic Studies, Practitioner Publications and Job Ads." In Proceedings of the Special Interest Group on Management Information System's 47th Annual Conference on Computer Personnel Research. Ireland: ACM.

ILO. 2016. "Disguised Employment / Dependent Self-Employment." Document. November 11, 2016. http://www.ilo.org/global/topics/non-standard-employment/WCMS_534833/lang--en/index.htm.

———. 2021. "Transition from the Informal to the Formal Economy - Theory of Change." Briefing note. January 29, 2021. http://www.ilo.org/global/topics/employment-promotion/informal-economy/publications/WCMS_768807/lang--en/index.htm.

Jackson, Michelle. 2007. "How Far Merit Selection? Social Stratification and the Labour Market1." The British Journal of Sociology 58 (3): 367–90. https://doi.org/10.1111/j.1468-4446.2007.00156.x.

Japec, Lilli, and Lars Lyberg. 2020. "Big Data Initiatives in Official Statistics." In Big Data Meets Survey Science, 273–302. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118976357.ch9.

Jones, Lyndon. 1984. "Lies, Damned Lies, and CVs." Education + Training 26 (4): 124–26. https://doi.org/10.1108/eb002125.

Kahn, Lisa B., Fabian Lange, and David Wiczer. 2020. "Labor Supply in the Time of COVID19." 06–2020. Cahiers de Recherche. Cahiers de Recherche. Centre interuniversitaire de recherche en Ã©conomie quantitative, CIREQ. https://ideas.repec.org/p/mtl/montec/06-2020.html.

Khaouja, Imane, Ismail Kassou, and Mounir Ghogho. 2021. "A Survey on Skill Identification From Online Job Ads." IEEE Access 9: 118134–53. https://doi.org/10.1109/ACCESS.2021.3106120.

Kobayashi, Vladimer, Stefan T. Mol, Gábor Kismihók, and Maria Hesterberg. 2016. "Automatic Extraction of Nursing Tasks from Online Job Vacancies." In Professional Education and Training through Knowledge, Technology and Innovation, 51. Siegen, Germany: Universitätsverlag Siegen.

Kotu, Vijay, and Bala Deshpande. 2019. "Chapter 9 - Text Mining." In Data Science (Second Edition), edited by Vijay Kotu and Bala Deshpande, 281–305. Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-814761-0.00009-5.

Kotzab, Herbert, Christoph Teller, Michael Bourlakis, and Sebastian Wünsche. 2018. "Key Competences of Logistics and SCM Professionals – the Lifelong Learning Perspective." Supply Chain Management: An International Journal 23 (1): 50–64. https://doi.org/10.1108/SCM-02-2017-0079.

Kuhn, Peter. 2014. "The Internet as a Labor Market Matchmaker." IZA World of Labor, May. https://doi.org/10.15185/izawol.18.

Kuhn, Peter, and Hani Mansour. 2014. "Is Internet Job Search Still Ineffective?" The Economic Journal 124 (581): 1213–33. https://doi.org/10.1111/ecoj.12119.

Kuhn, Peter, and Kailing Shen. 2013. "Gender Discrimination in Job Ads: Evidence from China." The Quarterly Journal of Economics 128 (1): 287–336. https://doi.org/10.1093/qje/qjs046.

Kureková, Lucia, Miroslav Beblavý, Corina Haita, and Anna Elisabeth Thum Thysen. 2016. "Employers' Skill Preferences across Europe: Between Cognitive and Non-Cognitive Skills." Journal of Education and Work 29 (6): 662–87. https://doi.org/10.1080/13639080.2015.1024641.

Kureková, Lucia, Miroslav Beblavý, and Anna Elisabeth Thum Thysen. 2015. "Using Online Vacancies and Web Surveys to Analyse the Labour Market: A Methodological Inquiry." IZA Journal of Labor Economics 4 (1): 18. https://doi.org/10.1186/s40172-015-0034-4.

Kureková, Lucia, and Zuzana Žilinčíková. 2016. "Are Student Jobs Flexible Jobs? Using Online Data to Study Employers' Preferences in Slovakia." IZA Journal of European Labor Studies 5 (1): 20. https://doi.org/10.1186/s40174-016-0070-5.

———. 2018. "What Is the Value of Foreign Work Experience for Young Return Migrants?" International Journal of Manpower 39 (1): 71–92. https://doi.org/10.1108/IJM-04-2016-0091.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." Science 343 (6176): 1203–5. https://doi.org/10.1126/science.1248506.

Leitner, Sandra M., and Oliver Reiter. 2020. "Employers' Skills Requirements in the Austrian Labour Market: On the Relative Importance of ICT, Cognitive and Non-Cognitive Skills over the Past 15 Years." Working Paper 190. wiiw Working Paper. https://www.econstor.eu/handle/10419/240633.

Lenaerts, Karolien, Miroslav Beblavý, and Brian Fabo. 2016. "Prospects for Utilisation of Non-Vacancy Internet Data in Labour Market Analysis—an Overview." IZA Journal of Labor Economics 5 (1): 1. https://doi.org/10.1186/s40172-016-0042-z.

Lewis, Phil, and Jennifer Norton. 2016. "Identification of 'Hot Technologies' within the O*NET® System." https://www.onetcenter.org/reports/Hot_Technologies.html.

Loo, Jasper van, and Konstantinos Pouliakas. 2020. "Cedefop and the Analysis of European Online Job Vacancies." In The Feasibility of Using Big Data in Anticipating and Matching Skills Needs. Geneva, Switzerland: ILO.

Lovaglio, Pietro Giorgio, Mario Mezzanzanica, and Emilio Colombo. 2020. "Comparing Time Series Characteristics of Official and Web Job Vacancy Data." Quality & Quantity 54 (1): 85–98. https://doi.org/10.1007/s11135-019-00940-3.

Mamertino, Mariano, and Tara M. Sinclair. 2019. "Migration and Online Job Search: A Gravity Model Approach." Economics Letters 181 (August): 51–53. https://doi.org/10.1016/j.econlet.2019.05.005.

Marconi, Gabriele. 2022. "Content Removal Bias in Web Scraped Data: A Solution Applied to Real Estate Ads." SSRN Scholarly Paper ID 4031466. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=4031466.

Marinescu, Ioana. 2017. "The General Equilibrium Impacts of Unemployment Insurance: Evidence from a Large Online Job Board." Journal of Public Economics 150 (June): 14–29. https://doi.org/10.1016/j.jpubeco.2017.02.012.

Marinescu, Ioana, and Roland Rathelot. 2018. "Mismatch Unemployment and the Geography of Job Search." American Economic Journal: Macroeconomics 10 (3): 42–70. https://doi.org/10.1257/mac.20160312.

Marinescu, Ioana, and Ronald Wolthoff. 2020. "Opening the Black Box of the Matching Function: The Power of Words." Journal of Labor Economics 38 (2): 535–68. https://doi.org/10.1086/705903.

Marrara, Stefania, Gabriella Pasi, Marco Viviani, Mirko Cesarini, Fabio Mercorio, Mario Mezzanzanica, and Marco Pappagallo. 2017. "A Language Modelling Approach for Discovering Novel Labour Market Occupations from the Web." In Proceedings of the International Conference on Web Intelligence, 1026–34. WI '17. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3106426.3109035.

Masso, Jaan, Raul Eamets, Pille Mõtsmees, and Kaia Philips. 2012. "The Impact of Interfirm Labor Mobility on Innovation: Evidence from Job Search Portal Data." In Innovation Systems in Small Catching-Up Economies: New Perspectives on Practice and Policy, edited by Elias G. Carayannis, Urmas Varblane, and Tõnu Roolaht, 297–321. Innovation, Technology, and Knowledge Management. New York, NY: Springer. https://doi.org/10.1007/978-1-4614-1548-0_16.

Masso, Jaan, Lucia Kureková, Maryna Tverdostup, and Zuzana Žilinčíková. 2016. "Return Migration to CEE after the Crisis: Estonia and Slovakia." https://style-handbook.eu/contents-list/migration-and-mobility/return-migration-to-cee-after-the-crisis-estonia-and-slovakia/.

Matsuda, Norihiko, Tutan Ahmed, and Shinsaku Nomura. 2019. "Labor Market Analysis Using Big Data: The Case of a Pakistani Online Job Portal." SSRN Scholarly Paper ID 3491253. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3491253.

Maurer-Fazio, Margaret, and Lei Lei. 2015. "'As Rare as a Panda': How Facial Attractiveness, Gender, and Occupation Affect Interview Callbacks at Chinese Firms." International Journal of Manpower 36 (1): 68–85. https://doi.org/10.1108/IJM-12-2014-0258.

McLaren, Nick, and Rachana Shanbhogue. 2011. "Using Internet Search Data as Economic Indicators." Bank of England Quarterly Bulletin 51 (2): 134–40.

Mezzanzanica, Mario, and Fabio Mercorio. 2019. "Big Data for Labour Market Intelligence: An Introductory Guide." European Training Foundation. https://www.etf.europa.eu/en/publications-and-resources/publications/big-data-labour-market-intelligence-introductory-guide.

Modestino, Alicia Sasser, Daniel Shoag, and Joshua Ballance. 2020. "Upskilling: Do Employers Demand Greater Skill When Workers Are Plentiful?" The Review of Economics and Statistics 102 (4): 793–805. https://doi.org/10.1162/rest_a_00835.

Mukoyama, Toshihiko, Christina Patterson, and Ayşegül Şahin. 2018. "Job Search Behavior over the Business Cycle." American Economic Journal: Macroeconomics 10 (1): 190–215. https://doi.org/10.1257/mac.20160202.

Muller, Noël, and Abla Safir. 2019. "What Employers Actually Want: Skills in Demand in Online Job Vacancies in Ukraine." Working Paper. Washington, DC: World Bank. https://doi.org/10.1596/31884.

Nomura, Shinsaku, Saori Imaizumi, Ana Carolina Areias, and Futoshi Yamauchi. 2017. "Toward Labor Market Policy 2.0: The Potential for Using Online Job-Portal Big Data to Inform Labor Market Policies in India." Working Paper. Washington, DC: World Bank. https://doi.org/10.1596/1813-9450-7966.

OECD. 2021. "OECD Skills Outlook 2021: Learning for Life." https://www.oecd.org/education/oecd-skills-outlook-e11c1c2d-en.htm.

Ours, Jan van. 1989. "Durations of Dutch Job Vacancies." De Economist 137 (3): 309–27. https://doi.org/10.1007/BF02115697.

Ours, Jan van, and Geert Ridder. 1992. "Vacancies and the Recruitment of New Employees." Journal of Labor Economics 10 (2): 138–55.

Pedraza, Pablo de, Stefano Visintin, Kea Tijdens, and Gábor Kismihók. 2019. "Survey vs Scraped Data: Comparing Time Series Properties of Web and Survey Vacancy Data." IZA Journal of Labor Economics 8 (1). https://doi.org/10.2478/izajole-2019-0004.

Pejic-Bach, Mirjana, Tine Bertoncel, Maja Meško, and Živko Krstić. 2020. "Text Mining of Industry 4.0 Job Advertisements." International Journal of Information Management 50 (February): 416–31. https://doi.org/10.1016/j.ijinfomgt.2019.07.014.

Piróg, Danuta. 2016. "Job Search Strategies of Recent University Graduates in Poland: Plans and Effectiveness." Higher Education 71 (4): 557–73.

Pitukhin, Eugene, Marina Astafyeva, and Irina Astafyeva. 2020. "Methodology for Job Advertisements Analysis in the Labor Market in Metropolitan Cities: The Case Study of the Capital of Russia." In Intelligent Algorithms in Software Engineering, edited by Radek Silhavy, 413–29. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-51965-0_37.

Poschke, Markus. 2019. "Wage Employment, Unemployment and Self-Employment across Countries." https://www.iza.org/publications/dp/12367/wage-employment-unemployment-and-self-employment-across-countries.

Profesia. 2022. "Covid19 Recruitment Report." 2022. https://public.tableau.com/app/profile/profesia.analytics4840/viz/ProfesiaReport_V2/Covid?publish=yes.

Rios, Joseph A., Guangming Ling, Robert Pugh, Dovid Becker, and Adam Bacall. 2020. "Identifying Critical 21st-Century Skills for Workplace Success: A Content Analysis of Job Advertisements." Educational Researcher 49 (2): 80–89. https://doi.org/10.3102/0013189X19890600.

Scheerder, Anique, Alexander van Deursen, and Jan van Dijk. 2017. "Determinants of Internet Skills, Uses and Outcomes. A Systematic Review of the Second- and Third-Level Digital Divide." Telematics and Informatics 34 (8): 1607–24. https://doi.org/10.1016/j.tele.2017.07.007.

Simionescu, Mihaela, and Klaus F. Zimmermann. 2017. "Big Data and Unemployment Analysis." 81. GLO Discussion Paper Series. GLO Discussion Paper Series. Global Labor Organization (GLO). https://ideas.repec.org/p/zbw/glodps/81.html.

Skhvediani, Angi, Sergey Sosnovskikh, Irina Rudskaia, and Tatiana Kudryavtseva. 2021. "Identification and Comparative Analysis of the Skills Structure of the Data Analyst Profession in Russia." Journal of Education for Business 0 (0): 1–10. https://doi.org/10.1080/08832323.2021.1937018.

Smyk, Magdalena, Joanna Tyrowicz, and Lucas van der Velde. 2018. "A Cautionary Note on the Reliability of the Online Survey Data: The Case of Wage Indicator." Sociological Methods & Research, July, 0049124118782538. https://doi.org/10.1177/0049124118782538.

Sodhi, M S, and B-G Son. 2010. "Content Analysis of OR Job Advertisements to Infer Required Skills." Journal of the Operational Research Society 61 (9): 1315–27. https://doi.org/10.1057/jors.2009.80.

Sostero, Matteo, and Enrique Fernandez-Macias. 2021. "The Professional Lens: What Online Job Advertisements Can Say About Occupational Task Profiles." JRC Working Papers on Labour, Education and Technology 2021–13. Joint Research Centre (Seville site). https://econpapers.repec.org/paper/iptlaedte/202113.htm.

Štefánik, Miroslav. 2012. "Internet Job Search Data as a Possible Source of Information on Skills Demand (with Results for Slovak University Graduates." In , 258–72. Thessaloniki: CEDEFOP.

Štefánik, Miroslav, Štefan Lyócsa, and Matúš Bilka. 2022. "Using Online Job Vacancies to Predict Key Labour Market Indicators." Social Science Computer Review. DOI: https://doi.org/10.1177/08944393221085705

Su, Zhi. 2014. "Chinese Online Unemployment-Related Searches and Macroeconomic Indicators." Frontiers of Economics in China 9 (4): 573–605. https://doi.org/10.3868/s060-003-014-0027-3.

Tambe, Prasanna. 2014. "Big Data Investment, Skills, and Firm Value." Management Science 60 (6): 1452–69.

Tambe, Prasanna, Lorin Hitt, Daniel Rock, and Erik Brynjolfsson. 2020. "Digital Capital and Superstar Firms." Working Paper 28285. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w28285.

Tijdens, Kea. 2010. "Measuring Occupations in Web-Surveys: The WISCO Database of Occupations." https://pure.uva.nl/ws/files/1491455/118626_1000_WP86_Tijdens_Measuring_occupations_WISCO_database.pdf.

Tijdens, Kea, and Stephanie Steinmetz. 2016. "Is the Web a Promising Tool for Data Collection in Developing Countries? An Analysis of the Sample Bias of 10 Web and Face-to-Face Surveys from Africa, Asia, and South America." International Journal of Social Research Methodology 19 (4): 461–79. https://doi.org/10.1080/13645579.2015.1035875.

Turrell, Arthur, Bradley Speigner, David Copple, Jyldyz Djumalieva, and James Thurgood. 2021. "Is the UK's Productivity Puzzle Mostly Driven by Occupational Mismatch? An Analysis Using Big Data on Job Vacancies." Labour Economics 71 (August): 102013. https://doi.org/10.1016/j.labeco.2021.102013.

Turrell, Arthur, Bradley J. Speigner, Jyldyz Djumalieva, David Copple, and James Thurgood. 2019. "Transforming Naturally Occurring Text Data Into Economic Statistics: The Case of Online Job Vacancy Postings." Working Paper 25837. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w25837.

Vankevich, Alena, and Iryna Kalinouskaya. 2020. "Ensuring Sustainable Growth Based on the Artificial Intelligence Analysis and Forecast of In-Demand Skills." E3S Web of Conferences 208: 03060. https://doi.org/10.1051/e3sconf/202020803060.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." Journal of Economic Perspectives 28 (2): 3–28. https://doi.org/10.1257/jep.28.2.3.

Visintin, Stefano, Kea Tijdens, Stephanie Steinmetz, and Pablo de Pedraza. 2015. "Task Implementation Heterogeneity and Wage Dispersion." IZA Journal of Labor Economics 4 (1): 20. https://doi.org/10.1186/s40172-015-0036-2.

Warschauer, Mark. 2003. "Demystifying the Digital Divide." Scientific American 289 (2): 42–47.

Xu, Haoyu, Chongyang Gu, Han Zhou, Sengpan Kou, and Junjie Zhang. 2017. "JCTC: A Large Job Posting Corpus for Text Classification." ArXiv:1705.06123 [Cs], June. http://arxiv.org/abs/1705.06123.

Zhu, Chen, Hengshu Zhu, Hui Xiong, Pengliang Ding, and Fang Xie. 2016. "Recruitment Market Trend Analysis with Sequential Latent Variable Models." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 383–92. KDD '16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939689.

# Acknowledgements

We would like to thank Verónica Escudero, Hannah Liepmann and Janine Berg for their excellent comments on earlier versions of this working paper. The authors acknowledge that this publication has received financial support from ILO, and L. M. Kureková also acknowledges financial support from the VEGA 2/0079/21 project, provided by the Ministry of Education, Science and Sports of the Slovak Republic and the Slovak Academy of Sciences.

## Advancing social justice, promoting decent work

The International Labour Organization is the United Nations agency for the world of work. We bring together governments, employers and workers to improve the working lives of all people, driving a human-centred approach to the future of work through employment creation, rights at work, social protection and social dialogue.


**Contact details**

**Research Department (RESEARCH)**

International Labour Organization
Route des Morillons 4
1211 Geneva 22
Switzerland
T +41 22 799 6530
research@ilo.org
www.ilo.org/research



ISBN 9789220372814