Towards an Understanding of Job Matching Using Web Data

Brian Fabo

CentER

Tilburg University

A thesis submitted for the degree of

Doctor of Philosophy

August 2017

Towards an Understanding of Job Matching Using Web Data

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. E.H.L. Aarts, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 3 november 2017 om 14.00 uur door

BRIAN FABO

geboren op 29 april 1986 te Kosice, Slowakije.

Promotiecommissie:

Promotores:

Prof. dr. J.W.M. Das

Prof. dr. M. Kahanec

Em. Prof. dr. K.G. Tijdens

Overige leden:

Prof. dr. A.H.O. van Soest

Prof. dr. ir. J.C. van Ours

Dr. S.M. Steinmetz

Dr. N. Askitas

Acknowledgements

I would like to thank my supervisors: Marcel Das, Kea Tijdens and Martin Kahanec. Marcel is probably the most optimistic person I have ever known, making every problem seem solvable. Kea is the personification of the word 'brilliance', with the ability to seamlessly harness the power of emerging technologies and understand new developments far faster than anyone else. Martin is someone with whom I have worked on a daily basis for five years and I still feel I could learn new things from him for another fifteen years.

Along with my supervisors, I feel indebted to Eduworks colleagues and friends: Gábor Kismihók, Maarten van Klaveren, Stefan Mol, Pablo de Pedraza, Stefano Visintin, Magdalena Ulceluse, Christian Weber, Stéphanie Gauttier, Jovana Karanović, Sofia Pajić, Vladimer Kobayashi, Sisay Adugna, Raquel Sebastián Lago, Sudipa Sarkar, Scott Harrison and all the rest of the team. Being part of this group has been the biggest achievement and the best time of my life.

I would like to extend my thanks to(in no particular order):

- My CEU comrades: Levente Littvay (for taking me on as a TA), Anil Duman (for helping me kickstart my academic career back in the day), Sharon S. Belli (for everything), Adela Danaj, Arthur Nogacz (for being my closest friend these last two years), Katerina Dukova (for being able to always count on her), Garrett Jones, Alexandru Moise, Tatiana Rogovich (for all the Pythonic stuff), Olga Löblová , Sanja Hajdinjak, Daniel Izsak, Alina Poliakova (for being the best possible research assistant), Jakub Kostolný, Sára Kende (for being there when it counted), Donát Szűcs, Iryna Koval, Riham Wahba, Jasmin Gamez, Mikhail Guliaev and all the rest. #IstandwithCEU.

- My CEPS colleagues and co-authors: Miroslav Beblavý and Karolien Lenaerts, Gabriele Marconi and Mikkel Barslund. It has been a pleasure and an honour!

- My CELSI folks: Marta Kahancová, Mária Sedláková, Tomáš Mamrilla, Michal Mudroň and Katarína Gandžalová.

- The amazing people who do not fall into any of the categories above: Mina Sumati, Martin Myant, Jan Drahokoupil, Michal Polák, Zoltán Pogátsa, Agnieszka Piasna, Lucia Mýtna-Kureková, Vladimír Kvetan, Juraj Draxler, Mario Sante Belli, Edward Branagan (for being so amazing at being my first boss ever), Július Horváth, Francesco Nicolli, Paulien Osse, Dirk Dragstra, Tendayi Matimba, Dani Ceccon, Janna Besamusca, Huub Bouma, Duko Dokter, Wietze Helmantel, Ernest Ngeh Tingum (and his son Brian), Klára Brožovičová, Roman Vido, Vít Hloušek, Petr Kaniok, Jan Řezáč, Hana Delsoir, Michal Lehuta, Jakub Jošt, Silvia Hudáčková, Andrej Svorenčík, Marek Hlaváč, Eva Liberda, Hana Janderová, Braňo Slávik, Richard Golier, Janka Kušnírová, Daiva Repečkaitė, Monika Kokštaitė, Zoltán Egeresi, Aliona Romaniuk, Márton Bárta, Soňa Mikulíková, my MCC students and all those special people I inexcusably forgot to mention.

Above all, huge thanks goes to my family: Mum, Dad, my little Sister and all the rest. Love you forever.

Contents

List of Tables

List of Figures

General Introduction

"Data are not taken for museum purposes; they are taken as a basis for doing something. … The ultimate purpose of taking data is to provide a basis for action or a recommendation for action"

W. Edwards Deming[1]

"It is, in fact, amazing how little labor economists know about the actual mechanics of how workers get assigned to jobs."

Peter J. Kuhn[2]

The quotations above have been chosen to introduce this dissertation because they represent the motivation and purpose of this dissertation. Even though they were pronounced in two very different periods with over seventy years between them, taken together they convey one message: There is a lot that we do not know (but need to know) about how matching on the labour market works, which data can show us. At the same time, the sole purpose of collecting data is to learn more about the world and potentially take action to make it a better place by addressing pressing challenges that hinder societal progress, in our case by ensuring that workers are equipped with the skills they need on the job market.

Labour market matching is one of the most salient challenges in terms of research as well as policy. This is particularly true in Europe, where the issue of equipping workers with right skills for employment has been considered an important policy priority for quite some time (CEDEFOP 2014, 2015). This policy discourse reflects the important debate about the "future of work" in the literature, with Tyler Coven's *Average Is Over*, being perhaps the most well-known example. According to Coven, "Quality labor with unique skills" (Cowen 2013) is one of three crucial resources needed in the modern economy, where an increasing number of tasks traditionally performed by humans will be conducted by robots.

[1] American statistician (1900 – 1993). Cited from presentation of director DG ESTAT Emanuele Baldacci on Big Data (Baldacci 2016)

[2] American economist, University of California, Santa Barbara. Quote from his IZA World of Labour piece, *Internet as a labor market matchmaker* (Kuhn 2014).

The debates outlined above cannot be separated from the rise of the Internet phenomenon. The Internet changed the way how the hiring process is organized, how the work itself looks like and allowed us to collect immense volumes of very detailed data on nearly any aspect of human life including work. Given these developments, it comes as no surprise that the push for an understanding of the role of skills on the labour market, going beyond the limitations of the canonical models such as Beveridge curve, is rising fueled by the new reality on the labour market and the unprecedented access to innovative data sources.

At the same time, while it is widely accepted that changes are afoot, we are still quite unsure how deep are they. Is the fact that web is becoming so crucial in labour market matching – up to the point of work being organized online through online platforms – going to improve the labour market matching? Do the robots and artificial intelligence increasingly present at the workplace alongside humans represent just a "tactical mutation" or something more fundamental with regards to how society is organized. Are the "big" web data the future of research. These are important complex puzzles, which can not all be answered in depth within a thesis.

Being aware of the limitation outlined above, the thesis aims for a pragmatic approach through making steady advances exploring the methodological issues but also showcasing the potential of web data to understand the labour market role of specific skills, such as language and computer proficiency from both supply and demand perspectives. As such, the presented research represents an ahead of the curve exploration and is intended to set the stage for the future research. At the same time, the thesis very much aims to place the web-based research of labour matching within the cyclopean scope and diversity of the applications of web data in modern labour studies and in the wider scientific enquiry. The presented research draws from my research collaborations focusing on labour market matching and the use of web data, which took place within the framework of the InGRID and Eduworks international collaborative research projects, funded under the 7th financial program of the European Commission in the period 2013-2017.

Each of the individual chapters presents an essay on a distinct topic and, as such, is structured on the basis of a standard research paper, with an introduction, literature review, description of methodology, presentation of analytical results and a concluding section. Chapter 3 is a slight

exception, because it discusses three separate approaches to the same methodological issue, each constituting an individual piece of work. The chapters have been published, or are in the process of being published, in scientific journals or as online discussion papers.

The research presented in this dissertation shows that web data has a potential to contribute significantly to improving our understanding of labour market matching by enabling research into the role of specific skills from both supply and demand perspectives. Nonetheless, rather than displacing traditional data sources, the web data appears to be useful in supplementing them either through enabling initial exploratory research as well as offering additional insights and more up-to-date coverage, which can be quite robust as long as there is a representative data source to benchmark the findings again.

This is particularly true for the online job vacancies, which have developed into a full-fledged industry with companies such as Burning Glass or Textkernel producing regularly updated, high-quality datasets. With the further development of text mining technologies, the use of this source of data will no doubt become more prominent in the future. Nonetheless, given that employers do not explicitly list all skill requirements, the information obtained will likely continue to be of limited use without knowledge of the context. The study of metadata, such as the occupational classification (if provided) or any other information, which might be present in the standardised form, represents another promising avenue for future research. Finally, the online labour platforms, represent yet another potentially rich source of data, enabling the researchers to observe job matching as it happens.

In addition to the online job vacancies, online surveys are likely to gain in importance. Already the web has become an important tool for survey dissemination, efforts will likely continue to use the Internet without any 'offline' component to survey various populations. This will drive demand for the study of Internet behaviour to understand how people are contacted by general invitation to participate in a survey and what makes them accept or decline participation.

The variety of other web data sources, such as social media, search engine searches or CVs – although not explored in detail in this thesis - will undoubtedly present additional avenues for further research. The undergoing transformation of the labour markets, requiring stakeholders to understand changing skill requirements in a timely and robust manner, will result in a growing

interest in using web data for labour market research purposes.

The text of the dissertation is divided into five chapters. With the exception of the first one, each is meant to represent my contribution to the literature of web data application in social science research (chapters 2 and 3) and labour market matching (chapters 4 and 5). The opening chapter sets the stage for the later chapters by providing a detailed overview of the ongoing state of the art. Nonetheless, while the individual chapters may be read as standalone pieces, there are synergies between them and they are all part of a single research agenda. Their composition is as follows:

The first chapter aims to map the current state of the art in the relevant research from two angles. Firstly, it summarises the structural ways of thinking about work analysis through concepts such as occupation, job, task and skill and then proceeds to discuss interactions between these concepts as well as the important changes to the labour market in the 21st century, such as polarisation, or technologically-driven transformation of employment to justify the need for detailed and up-to-date datasets for the purpose of developing adequate policies to respond to these changes. In the second part of the chapter, online data sources are presented as a possible solution to this need. A large number of different applications of web data is presented, although perhaps the two most relevant ones (web surveys and online job vacancies) are only discussed in passing, because I focus on those in chapters 2 and 3.

The second chapter discusses the applicability of large-scale, voluntary web surveys in social science research beyond exploratory research. By complementing the existing body of literature, which found significant biases in the data obtained through such surveys, we argue that as the survey becomes well-established in a country, the structure of respondents tend to stabilise in time, despite their being self-selected. Even though this, by itself, does not suffice to result in dependable estimates, we show that in some cases it might. As long as a quality data source applicable for benchmarking is available, low-cost web surveys may potentially generate nearly real-time and relatively robust measurements of relationships between variables.

The third chapter discusses online job vacancies, covering both vacancies posted online for ‘offline’ jobs and online platforms, through which workers can provide services to other platform users. The three applications consider the straightforward, text-focused approach to vacancy

analysis, the metadata-based analysis and finally an analysis of data obtained through online platforms. The main finding is that each of these approaches comes with specific trade-offs: Vacancy text analysis is the most comprehensive, but also resource-intensive approach. Metadata analysis is more straightforward and can also offer additional insights, but the available data is very dependent on a particular website. Finally, online labour platforms are potentially revolutionary for labour matching research, because they enable observation of both supply and demand dynamics, however, so far they only represent a niche labour market and the generalisability of results obtained from this source is uncertain.

The fourth chapter focuses on the specific application of web data to study the role of foreign language knowledge in Europe. Using metadata (language requirement tasks) from key job vacancies websites in the four 'Visegrád' countries: Czechia, Hungary, Poland and Slovakia combined with web survey data, we were able to identify and especially quantify the strong demand for the English language in non-manual occupations across these countries.

The fifth and final chapter again combines the application of job vacancy and web survey data to analyse the use of computers at work in the Netherlands. Here, we identified a major issue due to computer skill requirements not being explicitly requested for many occupations, which clearly required the use of computers, in particular, for professional and managerial occupations, where it was likely simply assumed that candidates would have computer skills. The web survey data, meanwhile, proved to be robust and not significantly different from representative survey data.

Chapter 1: State of the Art[3]

Introduction

The last few decades have been characterised by an astonishing advancement of technology, which has substantially transformed many aspects of life. Perhaps the most visible manifestation of this change has been the rise of the Internet. The number of Internet users has skyrocketed. A recent report by the International Telecommunications Union shows that at the end of 2014, almost three billion people had access to the Internet (ITU 2014). Globally, close to 44% of households have Internet access. The economic and societal changes that result from these developments are considerable and have attracted the attention of academics and policymakers. Along with the advancement of the Internet, researchers have increasingly shown interest in the Internet, not only as a research subject but also as a potential data source. This interest has not been limited to a single field but stretches out across many different domains. In the discipline of economics, labour economics has been identified as a field for which web-based data is particularly promising. In their seminal papers, Kuhn and Skuterud (2004), as well as Askitas and Zimmermann (Askitas and Zimmermann 2009, 2015), argue that web data could be very valuable for research on the labour market.

Just as these new, web-based data sources are emerging, the need for a more in-depth understanding of labour market matching is increasingly pressing. The labour markets around Europe have undergone important changes, manifesting themselves through dynamic processes such as job creation and destruction, skill upgrading, unemployment and wage inequality. The underlining force driving these symptoms is the fact that the structure of employment is constantly changing, and new jobs and skills are frequently arising (Goos et al. 2009).

The aim of this chapter is to provide an overview of the state of art in both aspects discussed

[3] This paper represents the author's contribution to the InGRID collaborative research project, developed in collaboration with Miroslav Beblavý and Karolien Lenaerts and has previously been made available online (Beblavý et al. 2016a). Furthermore, the discussion of the utilization of web-based data in labour research has been published in the IZA Journal of Labor Economics (Lenaerts et al. 2016).

above, the changes in the labour market and the rise of the Internet as a data source, and thus provide the background for research activities presented in chapters two to five of this dissertation.

The structure of this chapter follows the following logic: Firstly, the conceptualisation of key terms such as occupation, job, skill and task, which are heavily used throughout the entire dissertation. Building on this basis, the academic discourse regarding interactions between these key concepts is presented, setting the stage for a structured discourse on the ongoing changes in the labour market, which in turn encourage the need for increased use of innovative, web-based data sources. The state of the art with regards to the use of these data sources is discussed throughout the remaining part of the chapter. Arguably, the two most crucial data sources, web surveys and online job vacancies, are not discussed in detail because they are the focus of Chapters two and three respectively.

Occupations, Jobs, Tasks and Skills and the Complex Relationships between Them

Occupation is the crucial concept in labour market analysis, lying at the heart of the academic debate on labour topics (Tijdens 2010). An occupation may be defined as ‘a grouping of jobs involving similar tasks, which require similar skills set’. It includes multiple jobs or job titles that have common characteristics. A job, which is more commonly used in the context of labour market practice rather than research, ‘is bound to a specific work context and executed by one person’ (ESCO 2015). In the latest edition of the International Standard Classification of Occupations (ISCO), the International Labor Organization (ILO) identifies a job as a ‘set of tasks and duties performed, or meant to be performed, by one person, including for an employer or in self-employment’ while an occupation is ‘a set of jobs whose main tasks and duties are characterized by a high degree of similarity’ (Hunter 2009). Tijdens (2010) uses the following definition: ‘An occupation is a bundle of job titles, clustered in such a way that survey respondents in a valid way will recognize it as their job title; an occupation identifies a set of tasks distinct from another occupation; an occupation should have at least a non-negligible number of jobholders and it should not have an extremely large share in the labor force’ . Elias (1997) goes back to the history of an ‘occupation’, which still has an impact on how the concept is regarded today. He maintains that occupations have clear space and time dimensions which

can extend beyond the job that one holds. Damarin (2006) explains that occupations are generally regarded as a mechanism for dividing, allocating and directing labour. This view builds on the work of Abbott (1995), which lists three crucial occupational features: ‘a particular group of people, a particular type of work and an organized body or structure other than the workplace itself’. This group of people may be distinguished by their skills, experience, culture, gender or race, while the group of tasks may be split according to products, activities, tools or customers (and other categories). Occupations, however, are considered as relatively stable across time and organisations. Occupations are typically presented in an occupational classification, in which they are grouped on the basis of similarity in terms of tasks, responsibilities, education and skill level.

Although the difference between an occupation and a job is clear in theory, it is not straightforward to disentangle these two concepts from one another in practice. In fact, the two concepts may even coincide. An example is ‘project manager’, which may refer to a broader occupation or a specific job (e.g. project manager in an IT firm). In addition, in some cases, it is rather difficult to infer any information about a worker’s occupation from the job title: a project manager in an IT firm and one working for a charity can have a very different set of tasks (depending on the work context). At the same time, two workers with the same occupation may have completely different job titles, e.g. astronaut, cosmonaut or taikonaut, in spite of the fact that they perform similar tasks. Another issue is that given that an occupation is a group of jobs with similar tasks and skills, one may wonder how similar these actually have to be. Damarin (2006) further finds that when workers are asked to describe their jobs, many of them list multiple roles (that often vary across jobs and organisations). Some occupations may be distinguished through differences in education requirements or earnings. Because the distinction between occupations and jobs is not always clear, the concepts are sometimes regarded as ‘interchangeable’. Moreover, there are many studies on the labour market and work that start from the concept of ‘jobs’ (without referencing to occupations), while the term ‘occupation’ appears to be particularly important in specific strands of literature. Tomaskovic-Devey (1995) explains that the concept is relevant for comparative work at the national or international level, because occupations are independent of the specific work context, unlike jobs. Levenson and Zoghi (2010) maintain that occupations have a central role in the labour market. Both formal education and on-the-job training are often aimed towards a set of skills useful in different

categories of jobs (i.e. occupations). Moreover, occupation-specific experience appears to be valuable in the labour market.

In each of the definitions of occupations and jobs listed above, the concepts of 'tasks' and 'skills' are present. These concepts are therefore clearly important building blocks in the literature as well. Acemoğlu and Autor (2011) define a task as a 'unit of work activity that produces output (goods and services)' (p.2) and a skill as a 'worker's endowment of capabilities for performing various tasks' (p.2). In exchange for a wage, workers apply their skill endowments to tasks and generate output. Commonly, tasks are divided into routine and non-routine tasks (Baumgarten, 2015). Another definition for skills is given by the ILO; where a skill is 'the ability to carry out the tasks and duties of a given job' (Elias 1997). In ISCO, both the skill level and skill specialisation are considered. The European Commission uses 'skills' and 'competences' (ESCO 2015). Both are defined according to the European Qualifications Framework. Skills are 'the ability to apply knowledge and use know-how to complete tasks and solve problems'. Competences refer to 'the proven ability to use knowledge, skills and personal, social and/or methodological abilities in work or study situations and in professional and personal development'. From a review of the concept and measurement of skills in the social sciences, Spenner (1990) concludes that skills are increasingly measured directly either via expert systems (e.g. Dictionary of Occupational Titles) or self-report measures. Correlations between both measures are high. Initially, skills were commonly assessed on a case-by-case basis but later large-scale surveys of employers and employees were used instead (Gallie et al. 2007). Furthermore, the skill level of an occupation was often derived from the occupational classification. Occupational classifications, however, are not stable over time and reflect different bundles of tasks from one period to another (due to technological or organisational change) (Gallie et al. 2007). For this reason, skill levels are often proxied by learning requirements in more recent work.

There are two caveats to this approach, however: it primarily focuses on initial knowledge acquisition and it ignores the issue of mismatch (Borghans et al. 2001; Gallie et al. 2007). Other ways to measure skills are standardised tests (PISA), wages, experience or other proxies (Borghans et al. 2001; Elias and McKnight 2001). Similarly to tasks, skills are commonly separated into groups (generic and occupation-related skills (Tijdens et al. 2012, 2013b),

cognitive and non-cognitive skills (Brunello and Schlotter 2011; Mýtna-Kureková et al. 2016). Tijdens et al. (2012) indicate that in contrast to generic skills, which are commonly measured via surveys, occupation-related skills are hardly ever measured in this way. In addition, they find that it is difficult to measure mismatch by comparing educational attainment and skill requirements of occupations.

Making the distinction between tasks and skills can be rather complicated. Workers of a given skill level can carry out a range of tasks and at the same time workers with the same skill level can perform tasks of different levels of complexity. As workers need to possess the right set of skills to be able to do the tasks associated with their job, employers emphasise skills in the hiring process (Winterton 2009). Additionally, there is a clear link between skills, tasks, jobs and occupations. Occupations are grouped on the basis of tasks and responsibilities, education and skills. Moreover, skills are often proxied by occupations or derived from the occupational classifications. This implies that when doing research on one of these concepts, one also has to account for the other concepts.

Use of occupations, tasks and skills in social science is very widespread. For instance, Tijdens et al. (2012) analyse how work activities and skill requirements are measured on the basis of occupations. For comparative research on this topic, a sufficiently detailed occupational classification is required (one going beyond the four-digit level). Other types of work deal with a single occupation or a set of occupations. Recent work has concentrated on STEM (science, technology, engineering and mathematics) occupations (Brunello and Schlotter 2011; Rothwell 2014). An additional stream of work analyses sociological or psychological topics, such as the gender dimension, socio-economic or ethnic gaps and stereotypes in specific occupations (Byars-Winston et al. 2015; Pan 2015; Daniels and Sherman 2016).

Nonetheless, the main research focuses on skills matching with occupations. Fitzenberger and Lickleder (2016) consider school-to-work transitions, skill formation and career guidance of students graduating from lower-track secondary schools in Germany. Most students with poor grades appear to continue with pre-vocational training despite the fact that career guidance appeared to be effective (as students became more aware of their desired occupation). Virolainen and Stenström (2014) compare the system of vocational training in Finland with the systems of

Norway, Denmark, Sweden, Germany and the United Kingdom. They report that completion of upper secondary education is highest in Sweden and Finland, which could be due to the fact that in both countries both vocational and upper secondary education students are eligible for and proceed to higher education. In other words, vocational training is not a 'dead end' in these countries. The massification of higher education, however, complicates the transition of vocational education graduates to the labour market: there is increasing competition between higher education graduates (in all countries except for Germany, the completion of third-level education has increased). Tyler et al. (1999) focus on the cognitive skills of young high school dropouts in the United States. They find that annual earnings are higher for young dropouts with higher levels of basic cognitive skills.

Equally important is the topic of mismatch. Caroleo and Pastore (2015) survey the literature on educational and skills mismatch. The mismatch can be of a horizontal (level of schooling is appropriate, the type of schooling is not) or vertical (over- or under education) nature. These issues have mostly been investigated from the supply rather than the demand side of the labour market. Theoretical work explains over-education on the basis of a set of models: the human capital theory (over-education results from a lack of skills gained through work experience), the job competition model hugely in demand for highly educated labour encourages students to acquire more education, which could be more than that requested), the assignment theory, job search models and career mobility models. Allen and van der Velden (2001) put the assignment theory to the test. Educational mismatches do not necessarily imply skill mismatches. Furthermore, educational mismatches have a clear impact on wages, and only a small part of this effect is accounted for by skill mismatches. Skill mismatches, on the other hand, are important for job satisfaction and on-the-job search, in contrast to educational mismatches. For skills, there seems to be an extensive literature covering mismatch, over-education, educational attainment, skill measurement and a variety of other subjects.

Occupations and Skills in the 21st Century

In this section, occupational and skill changes are discussed in depth, with the aim of identifying their drivers and consequences. This analysis is not limited to the changes that occurred in the 21st century; it also considers historical trends. This allows us to better understand whether these

labour market dynamics are completely new or whether they are part of a longer history of similar changes. As Katz (1972) notes, there has been a vast increase in the number of distinct occupations between the 19th and 20th centuries. Abbott (1995), in contrast, noted that occupations are relatively stable across time and organisations. The section covers job polarisation, skill-biased technological change and other hypotheses. At the end of the section, an outlook towards the future is presented.

Studies on the history of occupational and skill change often depend on case studies to illustrate how occupations and skills were affected by specific factors in the past. One of the main reasons for this is that information on this period is difficult to come by.

Chin et al. (2006) focus on the Second Industrial Revolution at the end of the 19th century. The basis for their work is the literature on technological change and its implications during this period. Early work had reached a broad consensus that technological progress was skill-replacing (i.e. de-skilling). This is confirmed by Frey and Osborne (2017), who explain that technologies increasingly substituted for skills (by task simplification) as artisan shops were replaced by factories and steam power was adopted. The introduction of steam power along with major developments in continuous-flow production –(production parts became identical and interchangeable), also gave rise to assembly lines. A well-known example is the Ford Motor Company assembly line, where work that was previously performed by one person was now divided among many workers. Frey and Osborne (2017) conclude that in the 19th century, physical capital was a relative complement to unskilled labour but acted as a substitute for relatively skilled labour.

Along with the technological developments that shape the labour market, education has changed dramatically. In the US, this was first reflected in an increasing number of workers with a secondary school degree. Recently, it became even more prominent during the ‘computer revolution’ (which started in 1960 with the first commercial uses of computers and rapidly grew in the 1990s with the introduction of the Internet). Since the 1980s, educational wage differentials and wage inequality have grown a lot. This is often explained by the stronger complementarity between capital and skills that results from the computer revolution. Computerisation raises demand for clerical skills, similar to the introduction of office machines

in the beginning of the 20th century, but it can also lead to automatisation (Autor et al. 2003; Frey and Osborne 2017). Frey and Osborne, therefore, conclude that computers have caused a shift in the occupational structure of the labour market: 'the result has been an increasingly polarised labour market, with growing employment in high-income cognitive jobs and low-income manual occupations, accompanied by a hollowing-out of middle-income routine jobs' (p. 12).

As indicated above, one of the most remarkable characteristics of new jobs and new skills, at least in the context of developed economies, is their polarised nature. The polarisation of labour (or jobs) is a phenomenon where the demand for labour does not rise linearly with the skill level but rather resembles a U-shaped function (as illustrated on Figure 1). Instead, there is a polarisation in favour of both low-skilled and high-skilled jobs. Evidence of job polarisation has been found around the world (Autor et al. 2006; Goos et al. 2009; Fernández-Macías 2012; Ikenaga and Kambayashi 2016). In their work, Gallie et al. (2004) discuss the polarisation of skills. Skill polarisation may occur at the occupation level: where workers in lower occupational classes face skill stagnation or depreciation, the opposite holds for a worker in higher occupational classes because their employers tend to invest in on-the-job training. Skill polarisation could also arise on the basis of contractual status, in a core-periphery setting. At the core, we find full-time permanent workers, who are offered skill training; at the periphery we find part-time and temporary workers.

Job polarisation has received attention from academics and policy-makers. Many studies have been conducted on the causes and consequences of the phenomenon. Among the possible causes of job polarisation are technological change and globalisation, both of which mainly impact routine jobs (Autor 2001; Blinder 2009). Importantly, jobs, particularly in the middle of the distribution, are affected (Maselli 2012). The demand for high-skilled employment has been on the rise for many years, and a similar trend is detected for low-skilled jobs (especially in the service sector) (Maxwell, 2006). Low-skilled service jobs, however, commonly offer minimal levels of job quality and job security, low wages and few possibilities for advancement. These jobs, therefore, tend to come attached with negative selection stigma and are difficult to fill (as many of the unemployed try to avoid such positions, particularly if before they held low-skilled manufacturing jobs that offered a higher wage) (Lindsay and McQuaid, 2004). Furthermore,

while low-skilled service jobs are sometimes referred to as de-skilled due to the very low barrier to entry; in many cases, they tend to be quite demanding in terms of social and language skills and – in some cases – even in terms of formal education.

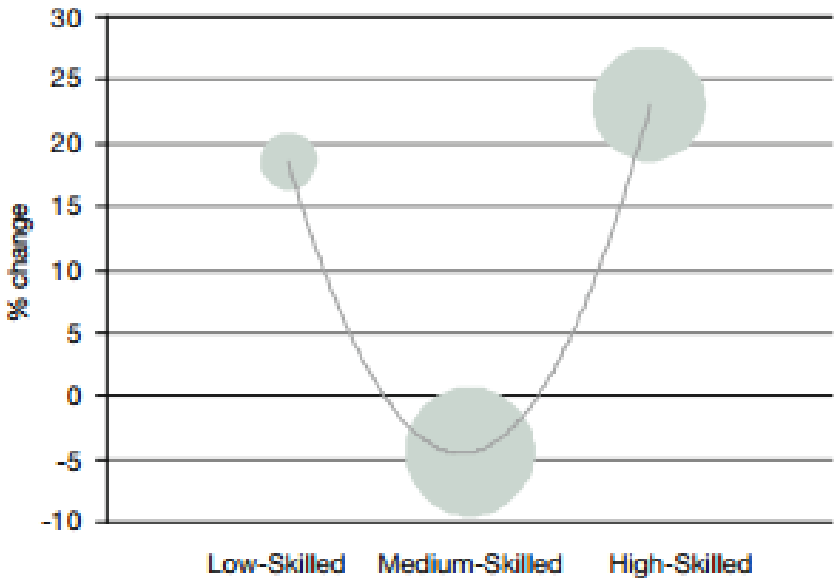

Figure 1: Changes in demand for jobs per ISCO skill level.
Source: (Maselli, 2012)

The up-skilling of some occupations combined with de-skilling associated with many new jobs (especially those in the low-skilled sector) complicates thinking about the labour market structure. This problem, however, is difficult to address. The issue also became clear in the EUROCCUPATIONS project, which measured the internal consistency of a wide range of occupations and found little grounds to assume that workers in the same occupation groups actually perform similar tasks (Tijdens et al. 2012). For these reasons, efforts have been made to include more dimensions into the way we think about jobs (e.g. computer use). In addition, there is a separate stream of literature that is focused on skills rather than jobs (Tijdens et al. 2012) as well as work that aims to bridge the gap between jobs and skills on the basis of novel datasets (Tijdens 2010; Fabo and Tijdens 2014).

What is driving occupational and skill change? Oesch (2013) explores different possibilities on the basis of a supply-demand-institutions framework. This framework is embedded in the

canonical model of the labour market, which attributes changes in the skill premium and skill-upgrading to shifts in the relative demand or supply for skilled workers and to labour market institutions. Oesch detects an occupational 'upgrading', i.e. the average occupation has become higher-skilled and better paid. This upgrading could be driven by demand-related factors (skill-biased) technological change, routinisation), supply-related factors (changes in skill supply, immigration) and institution-related factors (de-standardisation of work contracts). In another paper, Oesch and Rodríguez (2011) explore the drivers of polarisation on the basis of the same framework. In the remainder of this section, several of these theories are explored in more detail.

Before many researchers set out to address the issue of job polarisation, the literature focused on another very closely-related issue. This was the global increase of wage and employment inequalities between skilled and unskilled workers that has been documented in several contributions. Many of the early contributions have attributed these rising inequalities to skill-biased technological change (among others Chin et al. 2006; Oesch and Menés 2011). This view was challenged in later work, as we will see later on. Most work has focused on developed economies (e.g. Juhn et al. 1993; Nickell and Bell 1996), but there are also studies that cover developing countries. One example of the latter is Conte and Vivarelli (2011), who examine the occurrence of skill-enhancing technology import and find that this significantly increases the demand for skilled workers (while it does not affect the demand for unskilled workers). Katz and Murphy (1991) mainly attribute the increasing wage inequality in the US between 1963 and 1987 to skill-biased technological change within sectors (resulting from computerisation; see Krueger, 1993). Alternative drivers, such as labour allocation shifts between sectors and globalisation, appear to be less important. In a recent paper, Weiss and Garloff (2011) relate skill-biased technological change to unemployment and wage inequality in Europe and the US. Whereas skill-biased technological change is associated with increasing wage dispersion in the United Kingdom and the United States, it increases the level of unemployment in continental Europe. Antonietti (2007) reviews the literature on the relationship between skills and technology. He concludes that technology is a complement of non-routine, non-manual tasks and a partial substitute for repetitive manual tasks.

Skill-biased technological change, however, cannot explain the phenomenon of labour market polarisation that characterises many economies today. In fact, the evidence of job polarisation in

favour of high-skilled and low-skilled jobs is inconsistent with the hypothesis of skill-biased technological change (Wright and Dwyer 2003; Autor et al. 2003; Goos and Manning 2007; Jung and Mercenier 2014). These papers suggest that employment growth has taken place in low-paid personal service jobs and in well-paid professional and managerial jobs, while employment in average-paid production and office jobs has disappeared (Oesch and Menés 2011). This is why a number of alternative explanations have been put forward since the early 1990s. Chin et al. (2006), for example, consider labour market frictions and computerisation. From his review of the empirical literature, Antonietti (2007) concludes that early studies have relied mostly on the sector- and firm-level data, while more recent work used work-level data or even job-level data. The latter appears to present a more complex picture of the underlying dynamics.

In their seminal paper, Autor et al. (2003) propose an alternative theory of technological change to explain job polarisation: 'routinisation-biased technological change'. Routinisation-biased technological change entails that technology (computers in particular) can replace labour in routine tasks but not in non-routine tasks. Routine tasks are defined as codifiable tasks that involve a step-by-step procedure. One of the main features of this theory is that it shifts the focus from skills to tasks. In the model, technology affects the returns to tasks rather than skills. The plausibility of this theory as an explanation for job polarisation has been confirmed by Goos and Manning (2007), who show that routine tasks are indeed concentrated in the centre of the distribution (using data for the UK). Moreover, Acemoğlu and Autor (2011) demonstrate that the variance in the growth of US wages since the early 1980s may be attributed to changes in inter-occupational wage differentials. Other work also stresses the importance of this phenomenon during the First Industrial Revolution (2013).

Another potential explanation for the recent labour market polarisation is related to changes in international trade (or globalisation - see Jung and Mercenier 2014). In this regard, work on two dimensions has been particularly important: trade in intermediates (Feenstra and Hanson 1999) and offshoring (delegating work to other countries) (Blinder 2009; Grossman and Rossi-Hansberg 2012; Ebenstein et al. 2013; Helpman et al. 2017). In this literature, the idea of trade in tasks has also been highly relevant, with digital platforms increasingly enabling the offshoring of tasks without intermediaries (Drahokoupil and Fabo 2016). In addition to these factors, demand-side factors may also have an impact on the labour market (Jung and Mercenier 2014). Shifts in

the composition of demand (e.g. because of population aging, non-homothetic preferences) have been investigated by Manning (2004) and Autor and Dorn (2009). The latter two examine employment growth in service occupations.

Institutional factors are also of high importance. Labour market deregulation, the decline of trade unions and general preference of governments for job creation over social considerations[4] has contributed to shaping a world, in which employment is much more fluid and uncertain than it used to be over much of the post-war period. Technological progress can accelerate these trends by increasing the relative power of employers vis-à-vis workers or make some occupations more easily standardizable and outsourceable (that is, transferable to external contractors, rather than done internally) (Huws 2014; Drahokoupil and Fabo 2017). At the same time, the technology also drives the competition for the best talent, which is increasingly "footloose" and able to choose the place of its work (Huws 2014). Such qualified workers are have guaranteed good working conditions on the basis of their skills and do not need to rely on traditional labour market institutions.

In an interesting contribution by Jung and Mercenier (2014) compare the impact of these different explanations on the distribution of employment and wages. Firstly, the authors model a 'closed economy' to examine the impact of skill-biased technological change, reutilisation-biased technological change and demand shifts. The model only provides empirical support for the second hypothesis. When an 'open economy' model is used, Jung and Mercenier (2014) conclude that labour market polarisation is likely to be jointly induced by reutilisation-biased technological change and by globalisation. Nevertheless, the authors find that the within-group and overall wage inequalities – which are changing disproportionately– can only be accounted for by reutilisation-biased technological change. Another paper that compares several potential explanations for polarisation is Goos et al. (2009). These authors study job polarisation in 16 European countries in the period 1993-2006, with a focus on three hypotheses: reutilisation, globalisation and offshoring, and wage inequality. They find clear evidence of reutilisation, while the results for offshoring and inequality are less convincing.

[4] These aspects have been presented to a very different extend in different economies, resulting in both heterogenity of institutions and heterogenity of outcomes.

Oesch and Rodríguez (2011) point to the role of institutions in Britain, Germany, Spain and Switzerland. In all four countries, they detect a pattern of occupational upgrading, as the strongest employment growth is found at the top of the distribution. Furthermore, in all countries employment declined more in average-paid than in low-paid jobs. Importantly, wage-setting institutions do appear to filter the pattern of occupational change: countries only experience a trend towards polarisation if wage-setting institutions facilitate the creation of low-paid interpersonal service jobs. This may be more the case in Britain and Spain than Germany and Switzerland.

A final hypothesis to take into account is that of Schumpeterian creative destruction (Immergluck 1999; Mendez 2002). The emergence of new highly-skilled jobs can result in creative destruction. An example of this is the finance industry. This industry used to employ many clerks focused on interacting with clients, but many of these jobs has been disappeared due to increased automatisation and a stronger focus on areas such as secondary mortgage markets (Immergluck 1999).

Technological change in the past and the present has clearly had its labour market implications, as evidenced by many studies. Frey and Osborne (2017) summarise the conclusions of this literature as follows: technological progress has been accompanied by substantial changes in the occupational structure throughout history, but it has not resulted in widespread technological unemployment as predicted by Keynes (2010). This is due to the fact that technological progress has two opposing effects on employment: a capitalisation effect (employment grows in highly productive sectors) and a destruction effect (technology and labour are substitutes) (Aghion and Howitt, 1994). Frey and Osborne (2017)) argue that in the past the former has been dominant. The impact of capital deepening on the relative demand for skilled labour has changed substantially throughout history. In the 19th century, manufacturing technologies and skilled labour were substitutes. The 20th century was characterised by job polarisation, caused by computerisation. Other work has related these conclusions to education and training. For the United States, Rauscher (2015) explores the link between educational expansion and occupational change over the period from 1850 to 1930. Results suggest that compulsory school attendance laws and the educational expansion are associated with skill-biased technological change, a higher average occupational standing and an expanded occupational distribution.

Education may change the occupational structure.

A question that still remains unanswered, however, is what the future will look like? In their paper, Frey and Osborne (2017) suggest that although the capitalisation effect historically has been dominant, this does not necessarily apply to the future. In fact, computerisation is no longer limited to manual and cognitive routine tasks but it is being extended to non-routine tasks as well. This development is supported by the recent advancements related to 'big data' and robotics (e.g. robot senses and dexterity). Frey and Osborne (2017) estimate the probability of computerisation for 702 occupations in the United States. They find that about 47% of total US employment is at a high risk of computerisation. Transport, logistics, office and administrative support and production occupations are at high risk. Interestingly, a vast share of the service occupations is likely to be computerised in the future as well. Furthermore, they document a negative relationship between the probability of computerisation and wages and educational attainment. In a related paper, Beaudry et al. (2015) show that the demand for skills is decreasing. For the 21st century, Frey and Osborne predict a curb in the current trend towards polarisation: further computerisation is limited to low-skill and low-wage workers, who will switch to tasks that are not susceptible to computerisation. To this end, workers will have to acquire social and creative skills. Education and skills will remain important in the future for all workers. Another example of this is the incredible growth in STEM jobs in the past decade and the clear emphasis from policy-makers on STEM skills. As a result, educational institutes worldwide have introduced STEM-oriented training programmes and are encouraging students to opt for STEM training.

Finally, Autor (2015) predicts that polarisation will not continue indefinitely. He argues that although some of the tasks in middle-skill jobs are susceptible to automation, many of the jobs in this segment of the distribution will still demand a mixture of tasks from across the skill spectrum. Moreover, the tasks bundled into the middle-skill jobs cannot be unbundled so easily, without causing a substantial decline in job quality. Autor therefore maintains that many middle-skill jobs will combine routine with non-routine tasks in the future. These jobs are not off-shorable. In these jobs, workers continue to have the comparative advantage (e.g. problem-solving skills, interpersonal interactions). Autor concludes that the emphasis of human capital investment should be on the production of skills that are complemented rather than substituted by

technological change.

Researching the Labour Market Using Web Data

Internet has risen as an important potential source of valuable data for labour market research. The literature therefore distinguishes between studies that cover the Internet and studies that make use of the Internet to conduct research but note that these two domains are actually strongly connected (Hooley et al. 2011). The earliest studies were mostly of the first type, with research that focused on the social dimension of the web (Freeman 1984; Finholt and Sproull 1990). Shortly after these first studies, work that used the Internet to do analyses emerged (Kiesler and Sproull 1986; Foster 1994; Kehoe and Pitkow 1996). As the field expanded, new approaches and data collection methods were developed, which were often strongly embedded in the existing methodological framework and enriched with insights from technological progress.

The universe of various web data-based methods is vast and quite offten used intecconectively or interchangibly with the concept of "Big Data", refering to the fact that data collected online tend to surpass threshold where theirs sheer size makes the difficult to process using standard social science methodologies and equipment. But web data do not necessarily need to be "big". For instance, in their book, Hooley et al. (2011) delimit online research using found categories of research, where datasets tend to be relatively small: surveys, interviews and focus groups, ethnographies, and experiments. Of course modern web data-driven research goes far beyond these found approaches, as evidenced by a large number different research applications of diverse web data discussed in the remaining part of this chapter. While this thesis only focuses on a number of selected web data sources, the used datasets reflect diversity of the field being markantly different in orgin and size, united only by their web-based origins.

Nonetheless, the "Big Data" conceptualization can be quite useful due to its core concepts of the "Vs" such as volume, velocity and variety at the core of it (http://tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf) What makes the potential of web data particularly excisting and potentially fundamental shift vis-à-vis traditional data sources is the potential to provide (a) large number of observations allowing for detailed cross-tabulation of many variables, (b) enabling data collection in near real-time (c) including variables covered poorly or not at all by traditional data sources, such as detailed information about tasks

performed by the workers and their skills .

Overview of Existing Web-Data-based Research

In this section, we provide an overview of the data variety of web data sources available. Generally, we first describe the source. Then, we discuss which information can be collected from the source, with a focus on labour-related features. We continue with an overview of applications that either cover the source itself or use it for research on other topics. Because the field is rapidly developing, we do not limit our analysis to articles published in academic journals but also consider works-in-progress and other contributions. We aim for an extensive coverage going going beyond the data sources used in the following chapter of this dissertation in line with our ambition to locate our research within the "big pricture".

Web Surveys, Interviews and Focus Groups

Surveys were among the first research activities performed online. In fact, the first recorded email survey was done in 1986 (Kiesler and Sproull 1986) and the first recorded web survey in 1994 (Kehoe and Pitkow 1996). Compared to traditional paper-and-pencil methods, online surveys have the advantage of being flexible, fast, cheap and easy to set up. Data may be collected from a larger and more diverse sample, including hard to survey groups such as the undocumented migrants, which has a positive impact on data accuracy. At the same time, the respondents' anonymity is perhaps more easily ensured, because fewer people are involved in data processing. Web surveys also possibly contain more information to analyse than traditional surveys, because they offer additional information, such as meta-data. Disadvantages of online surveys include sample bias, measurement error, non-response and dropout, as well as other technical and ethical issues.

Of particular interest are the Internet panel surveys, which are well on the way to becoming the dominant method of collecting panel data in social science (Das et al. 2011; Das 2012; Blom et al. 2016; Bosnjak et al. 2016). Several Internet panel surveys exist currently. In the USA, there are two Internet panel surveys that we wanted to bring to the reader's attention: RAND's American Life Panel (ALP with 6,000 participants) and the Understanding America Study (UAS) panel

(2,500 participants, CESR, University of Southern California). Both panels represent the US population of ages 18 and up. In the Netherlands, the important CentERPanel started in 1991, representing 2,000 households and the more recently established Longitudinal Internet studies for Social Science (LISS) panel, consisting of 5,000 households. In Germany, there is the German Internet Panel (GIP, University of Mannheim, nearly 3500 participants) and the Gesis panel (4,900 participants, mixed method, partly collected offline). In France, there is the Longitudinal Internet Studies for Social Sciences (ELIPSS) with 2,500 participants. Beyond academia, many commercial operators exist and these facilitate Internet panel surveys (Ipsos, 'Contribute' by Survey Monkey).

Other web-based data sources buildig on an interaction with respondents, such as online interviews and focus groups have developed more slowly. This research mainly concerns asynchronous email interviews, although limited work does consider synchronous interviews and focus groups (Mann and Stewart 2000; Fielding et al. 2008). Online interviews and focus groups are more flexible as well as cost- and time-effective. They do, however, require technical competence from the participants, shift the power balance in their favour and prevent the researcher from observing any non-verbal communication. Online ethnographers examine how humans live and interact online; research commonly deals with social interactions on online communities, networks, gaming, discussion groups, bulletin boards and social media (Papacharissi 2009; Guo et al. 2012).

Online Experiments

Online experiments have been extensively used beyond the boundaries of social science. There are many examples in the literature of using experimental web data in economic research (Krantz and Dalal 2000; Musch and Reips 2000; Horton et al. 2011; Pallais 2014; Zheng et al. 2014; Dukova 2016; Horton 2017). Pallais (2014) performs an experiment on oDesk—an online labour platform—to test the hypothesis that young workers have a higher probability of being unemployed than older workers because of barriers to labour market entry. oDesk is also used by Pallais and Sands (2016), to examine why referred workers have a better chance of being hired. Horton (2017) uses oDesk to investigate the role of recommendations (which appear to increase the probability of hiring). Horton et al. (2011) perform a set of experiments on Amazon

Mechanical Turk (MTurk). More specifically, the authors replicate three classic experiments online and prove that such experiments are valid and beneficial to researchers. Aside from oDesk and MTurk, there is another platform on which researchers can do online experiments: TESS (Time-sharing Experiments for the Social Sciences, http://www.tessexperiments.org/). Researchers may submit proposals for experiments, which are peer-reviewed. When a proposal is approved, TESS does the experiment free of charge on a representative sample of US-based adults. These demonstrate that online experiments are more flexible so they are faster, cheaper and easier to conduct than real-life experiments and they allow for a broader scope, being able to potentially recruit a much larger number of participatns. The sample of participants that one can reach is also larger and more diverse. These features improve the quality of the study.

Observing online activity

In recent years, the analysis of labour demand, supply and matching became increasingly based on online data (Askitas and Zimmermann 2009; Kuhn 2014; Kuhn and Mansour 2014; Askitas and Zimmermann 2015). Internet data can contribute to our understanding of the labour market by filling the gaps that currently exist in the literature (Mýtna-Kureková et al. 2015). A first web data source that may be particularly useful in this case is a dataset extracted from online job portals. Many job boards are not limited to vacancies but also collect CVs and résumés. They also offer wage comparisons, employer evaluations and career advice. Some examples of this source use are: Marinescu (2016), who uses CareerBuilder to investigate geographical mismatch in the USA, while Marinescu and Rathelot (2016) rely on this portal to study unemployment insurance. Hershbein (2015) provides evidence for upskilling with a sample of vacancies from Burning Glass: employers require more in areas with higher unemployment rates, Tijdens et al (2015a) looks at the required educational attainment per occupation while Beblavý et al. (2013) analyse the values of different study fields. Mamertino and Sinclair (2016) explore the connection between job search and migration. Using a similar approach and dataset, Sasser Modestino et al. (2014) examine changes in employers during the Great Recession (2007–2012) and the subsequent recovery (2010–2012). In ‘bad’ labour markets, employers are more demanding, in terms of both education and experience. Kudlyak et al. (2014) rely on matched applicant-vacancy data from SnagAJob to assess how job seekers’ behaviour changes during their job search. In another contribution based on SnagAJob, Faberman and Kudlyak (2014)

report that job seekers' search effort declines with search duration. Brenčič (2014) looks into information acquisition through portals and résumé databases, suggesting that users only access a small portion of available information. Agrawal and Tambe (2014) use online résumés to track workers' career paths, focusing on workers previously employed in firms acquired through leveraged buyouts.

Note that there is a distinction between online jobs portals, which connect workers to traditional 'offline' jobs and other online labour market platforms like oDesk, Amazon Mechanical Turk (MTurk), CoContest and TaskRabbit, which directly connect workers and beneficiaries of labour. Much of the work on these market intermediaries has focused on MTurk, an online marketplace through which employers offer tasks that require human intelligence (i.e. that computers are unable to do). Horton (2011), for example, examines the fairness of MTurk employers. Buhrmester et al. (2011) evaluate MTurk's potential as a data source in the field of psychology and the behavioral social sciences. We have used the CoContest and Listminut platforms to study labour market matching and income-generating potential of the platforms (Maselli and Fabo 2015; de Groen et al. 2016; Fabo et al. 2017b, c). Other studies cover a wide range of diverse labour platforms (Ghani et al. 2012; Pallais 2014; Berg 2016; Mandl 2016).

Google Trends

Google Trends (https://www.google.com/trends/) was launched by Google in 2006 and has been established as an excellent tool to detect new trends (Askitas and Zimmermann 2009; Askitas 2015, 2016). The service is based on Google Search and analyses a percentage of these searches. In particular, Google Trends allows users to check how often search terms, or combinations thereof, are entered relative to the total number of searches done (by region, across time). When multiple search terms are entered, their relative popularity is compared. More precisely, on the web page, there is a search button which allows users to type in their search term of interest. Google Trends displays the interest in this search term over time (on the rise or declining) and by region (at global, national or regional levels) and related searches, spilt out into topics and queries. Regional searches are possible because a user's location can be identified via his/her IP address. Google Trends provides a normalised volume of queries for a specified term which is presented in a graph. Spikes in the graph are associated with news headlines, when possible. However, to protect the privacy of its users, Google does not publish results when there are

insufficient observations. Trends data excludes searches made by very few people, duplicate searches and special characters. Searches and search outcomes can be manipulated by companies operating the search engines, because they are the ones capable of modifying the search algorithms. One has to keep in mind that it is a company that develops content, sells advertisements and promotes its sub-brands (e.g. Yahoo Finance). This may particularly affect small- and medium-sized firms, which see their search ranks worsen and lose significant amounts of traffic. Moreover, organisations and large companies are able to manipulate search results as well, in order to maximise traffic and exposure.

Google Trends further has lists of 'hot searches' and 'hot topics'. The former tracks the most rapidly increasing searches at that given moment, while the latter captures trending terms in the news and on social media. Google Trends also features top stories that can be filtered by region and topic (eg business or health). In 2008, Google launched Google Insights, an extension to Google Trends, which allowed users to track words and phrases entered into search boxes, analyse results and structure data. The tool was integrated into Google Trends in 2012. All data can be downloaded in .csv format. Because Google Trends enables users to verify which (combinations of) search terms are on the rise, the platform provides us with more insight into the type of positions that job seekers are looking for, the types of skills that are in great demand, the industries that are booming and many other aspects. Google Trends presents information on labour demand and supply.

Nevertheless, there are some caveats to Google Trends. Because only a sample of searches is used and searches for which there are insufficient observations are excluded, Google Trends data may be affected by sample bias; in small samples, only random draws with enough observations are shown (Kearney and Levine 2015). A second issue is sampling variability (problematic for standard error calculations when data is treated as fixed rather than random variables). To address these issues, the authors repeat their searches on Google Trends several times and calculate the average of the indices (to reduce the sampling variability). As temporal and geographic variations are sources of variation that labour economists typically rely on, the above issues are important to account for.

Google Trends serves as a data source in a large number of contributions. One of the most well-

known applications is Google Flu Trends. In an influential article published in Nature, Ginsberg et al. (2009) explain how Google Trends may be used to improve the early detection of seasonal influenza by monitoring search engines like Google. This approach seemed to work well because of the high correlation between the percentage of doctor visits and the relative frequency of specific queries on Google, nonetheless eventually failed due to being unable to differentiate between searches related to the actual doctor visits and searches caused by media panic (Lazer et al. 2014). The authors can predict weekly influenza activity in the USA (with a time lag of one day). Other studies have used Google Trends to examine health-related topics as well (e.g. papers that extend or improve Ginsberg's method or focus on other diseases).

The strand of literature that relies on Google Trends for forecasting and now-casting is also extensive. Choi and Varian (2012) show that Google Trends is a useful tool for predicting the 'present' (in the form of subsequent data releases, i.e. the short-term future) due to the correlation between queries and economic indicators. They illustrate this result with the examples of travel, retail sales, home sales and automotive sales. Carriere-Swallow and Labbé (2013) work on a related topic, focusing on automobile purchases in Chile. Preis et al. (2013) and Preis and Moat (2015) relate Google queries to stock market dynamics and show that losses are often preceded by a growing volume of specific stock market search terms. In a recent publication, Chen et al. (2015) evaluate the extent to which Google search queries can be used to 'now-cast' business cycle turning points during 2007–2008. Schmidt and Vossen (2012) use Google Trends to account for special events in economic forecasting. In another paper, Preis et al. (2012) link queries, specifically whether they refer to the future or past, to countries' economic success. Constant and Zimmermann (2008) use Google Trends to measure economic and political activities, while Askitas and Zimmermann (2009), Foundeur and Karamé (2013) and Choi and Varian (2009) use it to predict unemployment. Yang et al. use search engine traffic to predict tourist traffic in China (2015), while Bangwayo-Skeete et al. extend this application more broadly (2015).

Other studies use Google Trends for behavioral analysis. Moat et al. (2014) use Google Trends to predict collective behaviour. In a series of articles, Stephens-Davidowitz (Stephens-Davidowitz 2014) uses Google Trends to explore topics such as racism, religion, prejudice and health. Rode and Shukla (2013) use a Google search query to examine racial differences in

labour market outcomes in the USA. The authors provide evidence for racial prejudice: in metropolitan areas with more racially charged searches, black-white gaps in annual income, hourly wage and annual hours worked are wider. This result appears to be somewhat stronger for less-educated workers. Another relevant paper is Kearney and Levine (2015), who combine data from Google Trends, Twitter and two other sources to examine how media images affect adolescents' attitudes and outcomes for the case of MTV's reality TV show *16 and Pregnant*. Interestingly, the TV series appeared to increase the amount of Google searches and tweets on birth control and abortion. Moreover, the show is associated with a 5.7% reduction in teen births in the 18 months after its introduction. Kearney and Levine (2015) do point to potential endogeneity: the interest in *16 and Pregnant* is likely higher in areas where the teen birth rate is higher or where it is rising or falling more slowly. While the former may be tackled via geography-fixed effects, the latter is addressed with an instrumental variables (IV) strategy in which ratings are instrumented with ratings of any MTV show broadcasted during the same time in the previous period.

The ability to 'nowcast' economic indicators using Google can be quite important in overcoming the gap between research and policy cycles and thus potentially contribute to the smoother implementation of evidence-driven policy solutions.

LinkedIn

In the last few years, many studies have appeared that concern social networking websites. Social networks commonly have large user bases comprising individuals, firms and other organisations. User profiles often contain detailed information about current employment, experience, educational attainment and other qualifications (labour supply). Information about individuals' behaviour and preferences can easily be obtained from these sites. In addition, firms and organisations often have profiles on these networks as well, through which they can interact with current employees and interested job applicants, as well as share vacancies (labour demand). Information is often publicly available. A vast majority of companies use social networks to look for candidates. Social networks can reduce search frictions. Acquisti and Fong (2015) investigate how information available on job applicants' profiles affects their interview invitation rates. A third of employers searched online for information on the candidates. Results

also suggest that employers in the Republican parts of the USA have a significant bias against Muslim candidates and in favour of Christian applicants.

Of the social networks discussed in this article, LinkedIn (www.linkedin.com) is the most obvious candidate to serve as a data source for labour market analysis because of its focus on professional networks, which are now being used for labour market research as well (Barslund and Busse 2016). By enabling users to set up profiles, connect with other users and find or list job openings, LinkedIn aims to 'connect the world's professionals to make them more productive and successful' and to 'transform the ways companies hire, market and sell' . LinkedIn was founded in 2002 and became available online in the spring of 2003. About 4,500 users signed up during the first month. Since then, LinkedIn has developed into the largest global online professional network, connecting over 364 million users (individuals and organisations) in over 200 countries and territories. In the first quarter of 2015, over 75% of LinkedIn's new users were not US-based. LinkedIn currently supports 24 languages. In Europe, LinkedIn has more than 89 million users. Two new users sign up every second. In the USA, 28% of the adult Internet users use LinkedIn. The website is particularly popular among college graduates, higher-income households and the employed. LinkedIn is the only platform where people aged 30–64 are more likely to be users than those aged 18–29 (Duggan et al. 2015).

Because LinkedIn holds profiles of companies, job seekers and recruiters, it is an interesting platform to analyse labour market dynamics. Firms can use LinkedIn to set up a 'Company Page' on which they can post job opportunities or create dedicated 'Career Pages' for this purpose. LinkedIn users can go through company pages to find job opportunities, use the general search options that LinkedIn offers, or connect with recruiters. Since 2011, users can even apply for jobs directly by using their profile as a résumé when they click on the 'Apply with LinkedIn' button on job listing pages. Another feature is the option to establish or become a member of an interest group (e.g. the group 'Java Developers'). User profiles of employees and job seekers further comprise valuable information on their education level and skills. In the autumn of 2012, LinkedIn added a feature through which users could comment on each other's profiles and endorse each other's skills. From this, it is clear that LinkedIn reduces search frictions, as job seekers (employed or unemployed) can easily find positions in their organisation or sector of interest and employers can easily browse through a large set of profiles to find their ideal

candidate (passive candidates are also available). LinkedIn therefore is a good starting point for labour market analysis. For the 94% of job candidates (two-thirds of recruiters), LinkedIn is the most important social network for job hunting (candidate sourcing) (University of Kent 2015).

Currently, the majority of the work on LinkedIn covers the platform itself. For example, there are several studies that examine how LinkedIn can be used in selection, recruitment or other business processes taking the perspective of job seekers and employers (Caers and Castelyns 2011; Bonsòn and Bednárová 2013; Rangel 2014; Zide et al. 2014). Jarrow et al. (2011) discuss LinkedIn's stock price. On the other hand, there are only a few contributions in which LinkedIn serves as a data source. An interesting example is State et al. (2014), who examine migration to the USA among professional workers of different education levels with a database of geo-located career histories from LinkedIn, a line of research further pursued for IT works by Barslund and Busse (2016). Boucher and Renault (2015) use a dataset compiled by hiQ Labs, which comprises many job titles and LinkedIn profile summaries, to construct a job classification. Gee (2014) takes vacancies published on LinkedIn to do an online experiment that covers 2.3 million job seekers. She demonstrates that reporting the number of previous applications increases the likelihood of application, especially among female job seekers. Tambe (2014) examines how labour market factors shape early returns to investment in big data technologies such as Hadoop and Apache Pig on the basis of LinkedIn. Other studies, of which several are related to the analysis of jobs and skills, can be found on the LinkedIn website (http://data.linkedin.com/publications).

A final set of applications worth mentioning are embedded in LinkedIn's 'Economic Graph' challenge. This challenge was launched in 2012 and sets out to create an 'economic graph' within a decade (i.e. to digitally map the world economy). For this challenge, teams were invited to propose how they would use LinkedIn data to research a range of topics related to the job market. 11 teams were selected (see http://economicgraphchallenge.linkedin.com/). Unfortunatell, outside of such events, LinkedIn is generally quite unkeen on providing access to data for researchers, hindering the use of this data source.

Facebook

Facebook (www.facebook.com) is a well-known online social network that was launched in

2004. On Facebook, users can set up a profile on which they can post messages, photos and videos; update their status; and use other features. User profiles can be public or private. On their profile, users can share their employment status or occupation, education level, family situation, skills, interests and hobbies, and other information (labour supply). Users can connect with others by becoming 'friends', in which case they receive notifications when a friend updates his/her profile (via the 'news feed'), and are able to send messages or chat. Since 2004, users have the possibility to create or become a member of a (private) Facebook group. As of 31 March 2015, Facebook had 1.44 billion monthly active users. The average number of daily active users during March 2015 was 936 million. About 83% of the daily active users do not reside in the USA or Canada. This number reveals that Facebook has an extremely large global user base. The website has a much larger network of users than any of the other social networks discussed. Duggan et al. (2015) find that Facebook is the most popular social network: it is used by 71% of online American adults. Women are particularly likely to use Facebook compared to men (66% of men; 77% of women have a profile).

Companies, recruiters and other organisations can also create a Facebook profile. This possibility was introduced in 2007 and is known as 'Facebook Pages'. Facebook Pages are public profiles held by celebrities, businesses, organisations and brands. On their profile, firms can present themselves to users, interact with them, introduce new products and post job vacancies (labour demand). On this note, there are also many job portals that have their own Facebook page through which they look for new talent, share job opportunities and offer career advice (labour demand, matching). Some examples are Indeed and Monster. In 2012, Facebook launched a job board, the 'Social Jobs Partnership Application' , which is the result of a collaboration with the Department of Labor, the National Association of Colleges and Employers, the Direct Employers Association, and the National Association of State Workforce Agencies. The introduction of the job board was motivated by the outcome of a survey performed by the National Association of Colleges and Employers (NACE), which targeted 530 employers and recruiters in the spring of 2012. This survey revealed that (i) 50% of the employers used Facebook in the hiring process, (ii) almost 90% of the companies claimed that recruiting via Facebook is more cost-effective than through other channels and (iii) especially networking and referrals are key features to find new employees (e.g. engaging in a network with a candidate who 'liked' the company's Facebook page). Via the application, recruiters can post job opportunities which can be sorted by

location, industry and skills. When the application was first launched, it combined offers from BranchOut, Direct Employers Association, Jobvite, Monster and Work4 Labs. The goal of the project is to support finding and sharing jobs via Facebook. A final option is to exploit Facebook Graph Search (Headworth 2014). Note that Facebook lowers search frictions (users can easily connect with an employer of interest or join a group; employers can browse through profiles and discover interesting candidates more easily via groups).

There are many studies on the topic of Facebook, but at first glance, only a few studies exist that use Facebook data to analyse labour market dynamics. Overall, the literature is extensive and covers many fields. Some examples are health-related issues, network analysis and education. Wilson et al. (2012) examined the research on Facebook in the social sciences. Their analysis is based on 412 articles published up until 1 January 2012. Wilson et al. (2012) classified these articles into five categories that reflect five research questions: (i) Who is using Facebook and what are users doing while on Facebook? (ii) Why do people use Facebook? (iii) How are people presenting themselves on Facebook? (iv) How is Facebook affecting relationships among groups and individuals? and (v) Why are people disclosing personal information on Facebook despite potential risks? The bulk of the articles addressed the fourth question. These questions are also particularly interesting to get more insight into labour market dynamics, recruitment and selection. Some work has been done on these issues, from the perspective of both job seekers and that of employers and recruiters. Research confirms that Facebook is a popular tool to screen applicants (Kluemper and Rosen 2009; Karl et al. 2010). This, however, implies that employers and recruiters can also evaluate job applicants on other criteria, such as their gender or race or ‘inappropriate’ material on their profile (Kluemper and Rosen 2009; Bohnert and Ross 2010). Kelkar and Kulkarni (2013) discuss the usefulness of Facebook from the labour supply side. While pointing out Facebook’s possible advantages for job seekers, They do find that only a small number of members actually use Facebook to look for jobs and for networking. The authors criticise the Social Jobs Application, which tends to yield inconsistent and ineffective results and does not appear to add much to what is already available on job portals. One can conclude that Facebook has the potential to serve as a source for recruitment and selection, but this has not yet been fully exploited. Two other examples of work on Facebook are Gee et al. (2016) on the link between job transmission and Facebook and Gee et al. (2017 on weak ties), based on the influential theory of Mark Granovetter (1973). Facebook appears to be an

interesting and valuable data source that has been used for labour market research. Wilson et al. (2012), however, do note that data crawling techniques are becoming less effective because of stricter privacy settings.

Twitter

Twitter (www.twitter.com) is a micro-blogging website through which users can read and send short messages (of no more than 140 characters) called 'tweets'. Whereas these messages can be read by anyone, only registered users can send tweets. Users are 'connected' to each other when they follow or are followed by other users. Moreover, messages sent out by one user can be re-tweeted by others. Tweets can cover any topic and can be grouped by topic or via hash tags. Twitter also tracks 'trending topics' (global and regional, via an algorithm that accounts for the location and interests of users). Twitter was launched in 2006 and has grown substantially ever since. About 500 million tweets are sent each day, most of which are accessible for public view as tweets are publicly visible by default. As of 31 March , 2015, Twitter has 302 million monthly active users. Twitter supports 33 languages. 77% of Twitter's accounts are held outside the USA, which illustrate the website's global outreach. Twitter's mission is to 'give everyone the power to create and share ideas and information instantly, without barriers' . This idea is put into practice through following and followers, re-tweeting and the public nature of the service.

Twitter can be used by job seekers, employers or companies as well as recruiters. It is, therefore, a useful tool for analysing the labour market. Job seekers can use Twitter to get more information on companies, discover job openings and find out more about the qualifications required by following these firms. They can also interact with current employees (labour supply). Moreover, job seekers can also find vacancies via the general 'search' function. Companies, on the other hand, do not only turn to Twitter for marketing purposes and sales; they may also use the website to tweet about vacancies and look for employees (labour demand). Such tweets often consist of a job title, a brief description of the position and a link to a webpage with more information. Companies can also use Twitter to strengthen their profile and spread information to clients and (potential) employees. A further option that companies have is to work with third parties, for example, 'Tweet My Jobs', to share their vacancies (Schawbel 2012). Recruiters can also rely on Twitter to find talent, through their own account, by becoming a member of groups

or by following other users. Another feature that is particularly useful in this regard is the option to embed a web link in the Twitter profile page of the company or recruiter, which directs job seekers to their website. Furthermore, many job portals, such as CareerBuilder, Indeed, Simply Hired and Monster, have their own Twitter accounts through which they share job listings and offer job search and career advice. As Twitter messages are fairly short, employers will generally not use Twitter as their main recruiting tool, but rather as a part of a whole recruitment strategy (Larsen 2011). Duggan et al. (2015) suggest that 23% of the US adults online use Twitter. The site is particularly popular among those younger than 50 and those who went to college.

Although Twitter was only introduced in July 2006, there already exists an extensive literature on the micro-blogging website. Academics and other researchers from many different fields have taken an interest in Twitter, which resulted in a high number of studies on a variety of topics (including but not limited to the field of computer and information sciences, physics, linguistics and economics). This interest is motivated by the scale of the database (many users and tweets) and its time dimension. In a recent article, Williams et al. (2013) focus on Twitter-based research to identify and classify the types of studies that are being done. From their review of the literature, they conclude that the following four elements are generally covered: the message, user, technology and concept. Other elements that are considered in some papers are the domain (e.g. education, health, business and security), data and research method. For a sample of 575 papers on Twitter published between 2007 and 2011, the authors show that most research deals with the content of Tweets followed by work on the users (together they represent 80% of the papers). Some more specific examples of recent work on Twitter are the paper by Achrekar et al. (2011) on the prediction of flu trends, by Murth (2015) on elections and by Yu and Wang (2015) on sentiments expressed in tweets during the World Cup of 2014. Despite a large number of topics covered, research on the relationship between Twitter and labour market dynamics and the use of Twitter data to identify or classify occupations and skills is limited. Kearney and Levine (2015) use Google Trends and Twitter but find that it is more difficult to access the latter than the former. Historical data cannot be accessed nor is there information on the frequency of tweets on the site. A library of past tweets can be obtained, but this library is difficult to manipulate due to its size and format. Data can be obtained through third-party vendors. Other limitations are that geographical information is difficult to obtain (no IP address) and that information on demographics is unavailable. Only a few papers seem to use Twitter for behavioural analysis.

In February 2014, Twitter launched a pilot project called ‘Twitter Data Grants’ . The project consisted of a call for proposals for research institutions to collaborate with Twitter staff and obtain historical and public data (see https://blog.twitter.com/2014/twitter-datagrants-selections). Six teams were selected. Before the Data Grant programme was launched, free access to data was limited to the previous seven days. The difficulty in accessing Twitter data also spurred several papers on the topic. Kwak et al. (2010), for example, describe crawling the platform to examine its topological characteristics. Twitter, in contrast to other platforms, is mainly relevant for labour demand as user profiles are fairly limited. One option, however, is that job seekers can spread their résumé via Twitter in the hope of attracting the attention of companies and recruiters.

Glassdoor

Web-based surveys, such as WageIndicator and Glassdoor, have already been introduced in the first sections of this article. In this section, we present Glassdoor (www.glassdoor.com), a popular career community website that was first launched in 2008. Although the website operates as a job board, Glassdoor goes beyond traditional job portals as the company also targets employers, recruiters, career centres and libraries. The company is US-based and has managed to develop into one of the largest websites there. Its user base is rapidly expanding towards a more global audience. In fact, Glassdoor currently has over 30 million members in over 190 countries worldwide. The Glassdoor website is organised into six categories. The first four categories are oriented towards job seekers and employees (who are invited to post their résumé), while the last two target firms and recruitment agencies.

The first out of the six categories is ‘jobs’. Similarly to the traditional portals, Glassdoor functions as a job board on which millions of vacancies are listed. Job seekers can browse through these job openings to discover which firms are hiring and which positions are available. On the website, job seekers are able to look for positions by location, title, occupation or keywords. Glassdoor also presents them with a list of popular and related searches.

The second category is ‘companies’. Job seekers can look for firms in a specific location and are redirected to a detailed company page when they select a firm. On a company page, job seekers can upload or read company reviews, find CEO approval ratings, discover the salaries that the

firm offers for specific roles, ask questions to current or former employees, find office photos and other information, read interview tips and so on. One of the innovative features of Glassdoor is that all this information is provided by companies' former or current employees. The website covers over 400,000 firms worldwide. To write a company review, users need to have an account. Users are only allowed to write one review per firm worked at per year. Reviews are published anonymously. Firms have no information on the identity of the employee that posted the review and cannot manipulate or remove reviews. Before a review is published, it has to be approved by Glassdoor. Reviews that do not meet the guidelines are not published (e.g. reviews should be balanced, cannot disclose trade secrets). Employers can respond to reviews and flag reviews that do not meet the guidelines, are inappropriate or fraudulent.

The third category on Glassdoor is 'salaries', which can be viewed for specific positions. The fourth category targeted towards job seekers is 'interview'. Job seekers can find information on the interview style, level of difficulty and sample questions. Since much of the information provided on Glassdoor is provided by former and current employees, the site clearly has a 'web survey' dimension.

Glassdoor also aims to attract employers. The two remaining categories are 'employers' and 'post a job'. On the website, employers can set up an 'Enhanced Employer profile' on which they can share their history and discuss their activities, promote vacancies or post a link to their Facebook account. Glassdoor offers employers recruiting and branding solutions via 'Glassdoor for Employers'. Employer branding implies that employers can track the candidates that are looking into their company and the reputation that their firm has on the website. Job advertising involves listing a single or all available jobs, targeting job seekers and using analytics to improve matching. Glassdoor has over 2,000 clients and partners for which they do employer branding promotion, job advertising (especially to candidates who may not have been aware of the position) or both. Glassdoor also offers a solution for companies in specific sectors, such as tech, telecom and SaaS companies; banking and finance companies; and consulting firms. Finally, Glassdoor also reaches out to career centres and libraries, which can offer unlimited access to the website without having to post information.

Glassdoor is an excellent source for the analysis of labour market dynamics as the website

collects a very large amount of real-time information on several dimensions. The database comprises millions of vacancies, company reviews, CEO approval rates, and salary and benefit reports. For employers, the website provides analytics of search behaviour, targeted job advertising and employer branding solutions. According to Glassdoor, users of all ages and backgrounds are on the website. Moreover, the average company rating is 3.4 (on a 1–5 scale). Of the employees, 70% indicate 'OK' or 'satisfied' when asked about their employer. A survey of over 4,600 Americans revealed that 2,201 of them use Glassdoor (Osterhaus 2014). About 50% consult Glassdoor at the start of their job search to identify top employers. Osterhaus (2014) reports that especially job seekers aged 55–64 and those earning between $25,000 and $49,999 annually are active on the site. Most respondents live in urban or suburban areas. Glassdoor seems to attract users of all ages and income levels. As Glassdoor engages more users and companies, representativeness may increase further.

There are numerous studies that use data extracted from Glassdoor to study the labour market. In many cases, data on wages and other information posted on Glassdoor are used to complement a more extensive analysis. An example of this is a study on the opportunities of females in IT by Thiele (2014), who uses Glassdoor to find the average wages for several IT professions (including software trainers, system programmers and network technicians). Massimino et al. (2015) use Glassdoor to find employee satisfaction rates. Other publications refer to Glassdoor as a source of information on interview techniques, skill requirements and other useful tips. Kaplan (2014), for example, refers to Glassdoor in her article on how to prepare for job interviews. Another example is Lauby (2013), who considers Glassdoor's growing popularity as an indication of the rising importance of employer and career branding. Chandra (2012) uses Glassdoor data on the work-life balance ratings across firms to compare Eastern and Western perspectives on work-life balance. American and European companies rank higher than Indian companies as they pay substantial attention to this issue.

Glassdoor also has its own research team of economists and data scientists: Glassdoor Economic Research (http://www.glassdoor.com/research/). Because Glassdoor gathers an enormous amount of real-time data on different labour market aspects, the website is a unique and rich data source. This data has been difficult to collect in the past, especially on such a large scale. Recent studies published on the site deal with hiring times, jobs affected by the introduction of a minimum

wage, and the link between salary and employee satisfaction. The site will offer downloadable data in the future and supports 'Job Tools'. The latter are two interactive map-based tools that can be used by job seekers nationwide. The first tool is 'Job Explorer', through which job seekers can find where their skills are in demand. The second tool is 'Apprenticeship Finder' that can be used to explore apprenticeship and career opportunities. Job seekers can view apprenticeship opportunities (high to low) on a map of their states with the aid of this second tool. For example, in the state of California, 5,704 apprenticeship jobs were available at the end of July 2015, most of which were in the areas around San Francisco, Sacramento, San Diego, Los Angeles and Silicon Valley. With the first tool, the job explorer, Job seekers can select a job category or type in a keyword to view on a map where these skills or jobs are concentrated with the aid of the first tool, Job Explorer. For California, this resulted in 94 'programming skills' jobs and 44,556 'programmer' jobs. The tool indicates 'top cities' for the jobs and lists 'other jobs you should consider' (for 'programmer', these jobs are senior programmer, consultant, software engineer, programmer analysts).

Pros and Cons of Using Web Data

Although several of the advantages and limitations of traditional and web-based data sources have already been pointed out above, we devote more attention to this topic in this paragraph. Labour market data is typically initiated and made available by governments or international organisations. This data is generally considered to be more accurate, better structured and more complete than data from other sources. Traditional sources such as the Current Population Survey in the USA or Census further have the advantage of being based on a randomly selected sample of the populatio. A disadvantage of traditional sources, however, is that statistics are usually distributed with a lag. Some databases are not regularly updated or even frequently revised. Data is often gathered from administrative sources or surveys, which could result in a small sample size or data unavailability for sectors or regions with limited coverage (Wright 2012). For many less-developed countries, labour market data is simply absent or of low quality (Friederici et al. 2016).

By contrast, online data is available in real time (no lags, no revisions) and is thus an excellent tool to detect current trends and dynamics. Web data allows researchers to fill gaps where

traditional data sources are absent or weak (e.g. due to low coverage or limited quality) (Autor 2001; Shapiro 2014). Online, researchers can assemble large, diverse and potentially more representative datasets. As an increasing part of the population is active online, some claim that sampling may become unnecessary in the future (Askitas and Zimmermann 2015), because the internet access has become nearly universal in many developed countries. Other advantages are that data collection and analysis are easy, fast, flexible and relatively inexpensive and that logistical issues associated with traditional sources can be avoided (e.g. tedious data entry prone to errors) (Benfield and Szlemko 2006). Another advantage of web data is that it facilitates research on self-employment, which is a key driver for entrepreneurship and job creation and may become even more relevant in the future labour market, which will make it even more important research topic. This is also highlighted in a recent OECD report: the web is a catalyst for business innovation, across all sectors of the economy, but it is not easy to study these dynamics (OECD 2014). At the same time, self-employment is difficult to measure on the basis of traditional data sources, because data is lagged and the definition of self-employment differs across data providers (Fairlie and Robb 2009). Many datasets cover information on either the owner or the business, but not on both. Web data may be a solution, as users can indicate on their profiles whether or not they are self-employed and information on the business can be found online.

Furthermore, the web may potentially be tremendously useful in capturing hard to identify/hard to reach groups, such as undocumented workers, workers in the informal economy or very young workers (Tijdens and van Klaveren 2011; Tijdens et al. 2015b). Another advantage is that online portals and social networks facilitate on-the-job search. Stevenson (2008) reports that 77% of the people that use the web for job search are employed. These employed job seekers are more likely to leave their current job and have better negotiation positions opposite their employer. Stevenson (2006) already concluded that the Internet leads to better matches for the employed (e.g. higher wage growth when changing jobs). Kuhn and Skuterud (2004) assess which unemployed workers look for a job online and whether they became re-employed more quickly. Online job search seems ineffective in reducing unemployment duration. However, an alternative explanation is negative selection on unobservable variables (e.g. poor networks).

Web data, however, is also characterised by some limitations (Benfield and Szlemko 2006;

Carnevale et al. 2014; Shapiro 2014). There are ethical and technical issues (e.g. privacy, familiarity with a computer), online data and vacancy data, in particular, are incomplete and biased towards specific regions or sectors, when compared to the actual existing jobs (Wright 2012). For example, Carnevale et al. (2014) find a bias towards high-skilled, white-collar workers and STEM positions in the vacancies published online. Autor (2001) further points to an adverse selection of job applicants (applying for a job is cheap and easy; therefore, job seekers apply for many jobs, for which they could be over- or underqualified).

Selection bias is an important issue for web data (see Bethlehem 2010; Carnevale et al. 2014; Kearney and Levine 2015). Websites or online platforms may attract specific users, which affects our ability to generalize on the basis of this data source. For this reason, we describe the demographics of the social networks below. Kureková et al. (2014) examine the representativeness of vacancy data. This data is not missing by sheer coincidence; this results from sampling. The quality, consistency, accuracy and volatility of web data should be examined (Carnevale et al. 2014).

Conclusion

In the chapter, we have shed some more light on the academic discourse on jobs and skills and focused on how occupations and skills have been transformed recently as well as the ways in which web data can help us shed light on these transformations. This section clearly shows that occupational and skill changes are not new phenomena. By contrast, they appear to have a permanent nature and are (predominantly) driven by technological progress. The technologies at the core of these advancements appear to differ (e.g. the steam engine, computers and robots). What also differs is the labour market impact that technological progress has. In this regard, the issues of de-skilling, up-skilling and job polarisation have been discussed. This conclusion is supported by Gray (2013), who states that the electrification episode in the beginning of the 20th century ‘mirrors the more recent polarization of the US labor force associated with computerization’ (Gray 2013).

One may wonder whether the concept of an occupation is still relevant, given that most of the literature appears to emphasise tasks and jobs. In the past, occupations were regarded as the dominant form of work organisation (Damarin 2006). This view has been challenged by the

emergence of large multi-divisional firms and Fordist mass production, which caused many occupations to become part of a large organisational structure. However, organisations are also subject to change (more flexible work forms, web-based work), which in turn could affect occupations. Damarin focuses on web labour, which has a modular structure; i.e. multiple roles are combined in ways that vary substantially across projects or organisations. In a way, this is similar to the flexible practices in other sectors such as job rotation. Damarin argues that the existing literature largely neglects this issue. Nevertheless, the concept of occupations is still relevant as occupations remain task-based mechanisms to divide labour.

From this overview of the academic literature, we conclude that occupations, jobs, tasks and skills are strongly related and intertwined. This notion is also reflected in the many theoretical and empirical contributions that deal with several of these concepts simultaneously. In the literature, occupations, skills, jobs and tasks are all important (in some branches more so than in others) and rather clearly defined. Nevertheless, some of these concepts are used as synonyms rather than as distinct concepts (e.g. occupations and jobs).

Furthermore, we have shown the great diversity and growing importance of labour market research based on web data. Most work in the field relies on vacancy or CV data extracted from online job boards or online labour market intermediaries or online survey. Nevertheless, there are many other web-based data sources that could be valuable for labour market analysis. These sources have not been overlooked completely, but their use is more limited. In this chapter, we have therefore explored the potential of Google Trends, social networks and Glassdoor.

From our survey, it is also clear that there are many possible avenues for future research. Some first examples are the use of web data for research on self-employment and on-the-job search. Research could also focus in more depth on the demographics of the users of the different platforms, from the perspective of both the job seekers and the companies or recruiters. Although social networks are used by employers to share vacancies, little work seems to exist on the nature of these vacancies (e.g. in which sectors, types of jobs). In addition, only few studies seem to examine the effectiveness of the different platforms for job search, selection and matching. The future of Internet-based labour market research, therefore, seems to be bright.

Chapter 2: Using Voluntary Web Surveys Beyond Exploratory Research[5]

Introduction

The main objective of this chapter is to contribute to the debate among social scientists about the usefulness of web data (Baltar and Brunet 2012; Edwards et al. 2013; Tijdens and Steinmetz 2016). To this end, we use ten years of data from the WageIndicator (WI) survey in the Netherlands, which is probably the most comprehensive and largest web-based labour force survey available. Specifically, we show that this data source can provide salient estimates of some key relationships between labour market variables compared to the estimates obtained from the same econometric model run on a random sample of Dutch households from the European Survey of Income and Living Conditions (SILC).

Our motivation is related to the general need to collect labour market data more timely, efficiently and comprehensively. Web-based techniques have this very potential, but due to the non-randomness of the produced samples, they are often deemed unsuitable for salient statistical analysis. As a rule, surveys strive to obtain representative samples of the respective populations, as researchers typically aim to generalise their empirical findings to the entire studied population (Couper 2000; Steinmetz et al. 2013). However, such efforts are getting increasingly difficult due to growing difficulties associated with contacting respondents and ensuring their collaboration (Kitchenham and Pfleeger 2002; Tijdens et al. 2005; Kohut et al. 2012; Baltar and Brunet 2012; Sloane 2014). Various techniques, including web-experimenting, non-reactive data collection, web-based testing and, in particular, web-surveys have been developed to counter the negative effects of social changes by harnessing the power of technology. (Reips 2006; Stieger and Reips 2010; Lenaerts et al. 2016).

The WI survey, which we examine in this paper, represents a clear case of what is called, in the survey literature, a convenience sample. It is an online survey with a set of questions modeled on the basis of the EU Labour Force Survey (LFS) available in about 100 countries around the world in a large number of languages. Respondents are reached by general invitation to

[5] This work has been submitted to the International Journal of Social Research Methodology as a paper co-authored by Martin Kahanec.

participate in the survey posted on a network of websites offering a veriety labour market information, with response rate being determined by who gets attracted to the website and who decides to accept the invitation to participate in a survey. As such, attempts to use it for social science research are naturally treated with caution. While a degree of convenience sampling has been commonly included in qualitative studies (Marshall 1996), quantitative research has historically frowned upon convenience samples. Nonetheless, the prospect of wider usage of WI is quite tempting, due to a very large sample covering many countries, including those not covered by traditional surveys, with data available nearly in the real time.

Our aim in this paper is certainly not to argue that prudence in regards to sampling methods is overrated or that it should be abandoned in favour of convenience offered by the new data sources. Neither do we claim that non-representative data in general or WI in particular are necessarily usable as a source of data for reliable analysis in the same way as the probabilistic sample data. Rather, we intent to demonstrate that statistical results based on non-probabilistic survey data can be possibly similar and even statistically not significantly different from a probabilistic sample. To our knowledge, this chapter is the first to present such evidence.

Literature Review

The strengths and weaknesses of web-surveys have been discussed extensively in the literature (Couper 2000; Fricker and Schonlau 2002; Steinmetz et al. 2013; Tijdens and Steinmetz 2016). The comparative advantage of web-based surveys lies in cost-effectiveness, versatility, speed, and the opportunities for innovation; and hence the potential for generation of large samples. While by itself, a large sample does not equate quality, it allows for research into specific subgroups, which might not be present in sufficient quantity in smaller samples. Important methodological challenges, primarily linked to the lack of Internet access of some sub-populations and selection into survey samples, relate to sources of survey error associated with coverage, non-response and measurement, as well as inference from non-probability samples (Couper 2000; Dillman and Bowker 2001; Bandilla et al. 2003; Bethlehem 2010).

The challenges of web-based surveys regularly lead to two outcomes: discouragement of researchers and thus underuse of available web-based datasets or inappropriate use of such datasets and misinterpretation of the analytical results. Nonetheless, there has been effort made

to address shortcomings of this data source. Ex-ante design-based methods alleviate the representativeness issues by means of auxiliary sampling methods, such as provision of Internet access if needed, or ensuring respondents do not become too 'trained' in filling questionnaires (Toepoel et al. 2008; Das et al. 2011) or by focusing on optimising the questionnaire design and user interface (Toepoel et al. 2009). The model-based approach uses weighting techniques to correct for representation bias of non-probability web-surveys ex-post (Schonlau et al. 2009; de Pedraza et al. 2010). The non-coverage problem has decreased, to a certain degree, as Internet access became nearly universal in developed countries (Askitas and Zimmermann 2015). The position that web-based survey data are potentially a very rich resource that we should learn to use wisely rather than discard a priori is gaining footing among an increasing number of researchers, as evidenced also by concrete academic projects based on web-based data (van der Laan and van Nunspeet 2009; Reips 2012; Mýtna-Kureková et al. 2015).

Topics as diverse as job security (de Bustillo and de Pedraza 2010), satisfaction of workers (Guzi and de Pedraza 2015), education requirements (Steinmetz et al. 2013) and skill mismatch for migrants (Visintin et al. 2015b) have been explored using WI data. A unique trait of WI data is that question have been modeled according to the LFS, and thus results can be easily benchmarked against LFS itself or other surveys adhering to a comparable question structure or other datasets following the same variable structure.

Efforts made in that direction so far have found that WI data can be used to correctly identify significant relations between variables, as well as the direction of the relation. Nonetheless, this did not appear to be true for the size of the estimates. The strength of a relationship was found to be under/over-estimated when statistical inference is run on WI data, as opposed to the situation when the same model is run on a data source generally seen as representative (de Pedraza et al. 2010; Guzi and de Pedraza 2015). The main potential sources of bias in WI data are as follows: (1) self-selection, due to targeting only users interested in WI content; (2) limitation of sample to people who are on the Internet; and (3) non response, as many people from WI audience choose not to answer the survey, or start answering but drop out (Guzi and de Pedraza 2015). Tijdens and Steinmetz (2016) therefore conclude that WI is mainly useful as a tool for exploratory research, to determine whether there appears to be a relationship between two or more variables, but not solid enough to test hypotheses.

Data and Empirical Strategy

Data and Research design

WI is an international survey covering 85 countries, the quality of the data varies among countries. The largest data sample by far is collected in the Netherlands. About 18% of all data collected by WI worldwide in 2015 came from this one country. In terms of the audience, the number is even greater – one in four visitors of the WI sites globally comes to the Dutch website. In 2015, the site recorded 7.5 million visitors, more than 5 million of them unique[6]. These numbers represent a very large segment of the country's total population of 17 million.

Nonetheless, not everyone invited to participate in the survey did so. The response rate was about 1%, which is around 50,000 people who accepted the invitation to participate in the survey in 2015. Additionally, not every participant actually completed the survey: a significant portion was lost along the way due to dropout, particularly due to questions enquiring about sensitive personal information, such as wage (Steinmetz et al. 2013). In 2015, a total of 15,000 Dutch people – 30% out of the survey respondents and 0.3% of the unique visitors provided wage information.

To benchmark the data, we will use the EU Statistics on Income and Living Conditions survey (SILC). SILC is a comprehensive source of income statistics in the EU. In the Netherlands it is collected since 2005 on an annual basis, the most recent covered year being 2014. In general, SILC data is collected by national statistical offices in line with the best practices for obtaining a high-quality sample in each covered country (European Commission 2015). Furthermore, the SILC dataset contain weights to ensure representativness of frequency statistics callculated on the basis of the survey. Unfortunatelly, such weights are not available for WI, thus we had to compare unweighted samples. In the Dutch case, the data is collected as a four-year rotational panel. Households were randomly chosen based on a stratified selection of 40 NUTS3 regions in the country, to ensure data quality comparable with the Labour Force Survey (Statistics Netherlands 2013).

[6] According to Google Analytics over 95% of the visitors originate from the Netherlands.

The choice of the Netherlands as the focus of the analysis is also based on the fact that it is rather homogenous in economic development. Analysing wages in the Netherlands does not entail dealing with issues such as the Flemish – Wallonian divide in Belgium, or federal structure and legacy of country division in Germany, to name two other countries, which have long been covered by WI and come with a large number of observations.[7] Additionally, Internet access has become nearly universal among the working age population in the Netherlands (Statistics Netherlands 2016).

To avoid outlier effect from both of these data sources, we select only respondents who fulfill the following conditions:

- Are employees;
- Reported a wage higher than the minimum wage;
- Work between 10 and 60 hours a week.

Observations containing a missing value in any of the variables included in the model are removed. In a typical year, about half of the observations with wage information are retained after data selection, the other half consisting of the economically inactive, the unemployed and people who do not satisfy the conditions listed above.

The number of observations in WI varies more over the years than in EU SILC. This is caused by factors such as the advertising budget for respondents recruitment, time of year (very few people fill the survey during summer holidays or major holidays). Nonetheless, more observations are typically collected in WI than in SILC, years 2007 and 2009 being exceptions (see Table 1). The dropout is substantial, with only about a fourth of observations retained for EU-SILC and even less for WI. The reasons for the large-scale elimination of observations is that all non-employees are removed from the dataset and removal of observations with missing values. The later was much more widespread in the WI survey, which suffers from high dropout rates (Tijdens 2014), but given the sensitivity of the wage variable, affects also the EU SILC sample considerably.

[7] For comparison for 15,000 ‘valid’ observations in Netherlands, there were 7,500 in Germany and 2,500 in Belgium.

Table 1: Number of observations per year after data cleaning

	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
SILC	2,093	2,233	2,501	2,476	2,173	2,160	2,130	2,068	1,894	1,795
WI	13,186	4,387	2,008	9,154	1,006	5,140	3,517	3,428	2,876	2,185

We observe that in the Netherlands, a country where WI has a fifteen-year tradition and a large number of website visitors, there is a stabilised audience (see the Descriptive statistics section of this chapter). These are the people who are interested in WI content. They therefore either regularly visit the WI website or are likely to respond to an invitation to visit the site and, at the same time, accept the invitation to fill the survey either due to a desire to contribute to the WageIndicator project in line with the theory of gift exchange reciprocity (Falk and Fischbacher 2006) or to be in for a chance to win a cash prize[8].

In this article, we test whether these two data sources – WI and SILC – produce comparable wage estimates. Unfortunately, we can only look at the unweighted data, because there are no weights for the WI survey available. To determine this, we run the same economic model, checking for common determinants of wages separately on these two data sources and then comparing the estimates. The similarity of estimates is evaluated statistically using a test of joint equality of coefficients. The test in effect verifies that the coefficient on the dummy variable denoting the dataset (SILC or WI) and the coefficients of interaction terms of all other independent variables with that dummy variable are jointly zero[9]. In effect, this test tells us whether estimated parameters from the two regressions differ.

Model

We evaluate the usefulness of WI data based on the (dis-)similarity of estimates based on WI and EU LFS data on a key question of labour economics – the wage equation. This issue has been covered in a large body of literature. There is a standardised way to measure wage determination that has been widely adopted by the economic literature (Heckman et al. 2006; Lemieux 2006;

[8] Equal to the amount of the weekly minimum wage in the Netherlands, adjusted annually, which replaced the original prize of a trip to South Africa in the first three years..

[9] See https://www.stata.com/support/faqs/statistics/test-equality-of-coefficients/ for details about the test.

Beblavý et al. 2013). The Mincer earnings function (Mincer 1974) represents a formalisation of economic thinking about earning differentials, mainly based on human capital theories. In other words, wages – just like all earnings in general – are seen as payoffs from investments in acquisition of skills either through education or employment.

The standard formalisation of Mincer's equation is as follows:

$$\ln y = \ln y_0 + \beta_1 S + \beta_2 X + \beta_3 X^2$$

with y denoting earnings, S years of education and X years of potential work experience. For the purpose of this study, we use gross hourly earnings calculated from the wages reported by the WI respondents[10] and annual gross earnings from the EU – SILC data from which we obtained hourly earnings in line with the methodology developed by Berger and Schaffner (2012).

Years of schooling are calculated based on a difference between the current age of the respondent and the Dutch standard of beginning primary school education at 4 years of age as:

$$S = Y_s - (Y_b + 4)$$

With Y_s representing year of graduation and Y_b year of birth. While in the Dutch system, children start primary school at the age 4 or 5, we ignored the first two years, the so-called group 1 and group 2. We stick to the international standard of counting years of education from the age of 6. The reasons are that group 1 is not mandatory. Before 1995, the first two groups were not considered a part of elementary school and only in group 3 did children actually start learning reading, writing and arithmetics.

The years of potential working experience are calculated as the difference between the year in which the respondent started working in the first job and the year of the survey:

$$X = Y_x - Y_f$$

We amend the Mincer equation with several control variables based on theoretical literature on wages, although limiting ourselves to those feasible given the available variables in the dataset.

[10] Respondents can choose the reporting period, in the case of the Netherlands, 95% of the respondents report monthly wage

Firstly, we consider the gender dimension by running the regressions separately for men and women. This is in line with the vast literature on gender pay gap literature as reviewed by Altonji and Blank (1999), which recognises that key wage determinants affect men and women in different ways. Secondly, we include variables measuring the permanency of the contract and whether the respondent holds a supervisory position. The literature overwhelmingly confirms that workers in permanent jobs, as well as those in positions of responsibility, are better paid than those in more precarious work arrangements. The likely reason being that permanent contracts are offered to workers whose replacement is seen as associated with high fixed cost, which is also the case for supervisors (Averett and Hotchkiss 1996; Mitra 2003). Thirdly, we consider the size of the firm, because bigger firms are generally able to attract higher quality (and thus better paid) labour (Oi and Idson 1999). Finally, we check for marital status to account for the 'marriage wage premium' (Antonovics and Town 2004).

The resulting model is as follows:

$$InW_i = \beta_0 + \beta_1 wex + \beta_2 wex^2 + \beta_3 edu + \beta_4 firmsize + \beta_5 contract + \beta_6 svisor + \beta_7 marstat + \varepsilon_i$$

where natural logarithm of gross wage (W_i) is the dependent variable and the independent variables are wex (years of potential working experience), wex² (years of potential working experience squared), edu (years of education), and dummies for firm size (number of co-workers), contract (has a permanent contract), svisor (is in a supervisory position) and marstat (marital status). The error term represents the deviation between model-based expectation and the observed values.

Descriptive statistics

For the period 2005-2014, respondents in the WI survey generally reported lower gross wages (range: €15-18 per hour) than the workers tested by EU SILC (range: €16-20) (see Figure 2). Meanwhile, the wage gap between WI and SILC data increased over time up until the year 2013. A possible reason could be that the younger workers (mainly covered by WI) were disproportionally affected by the Great Recession and saw their wage stagnated for quite some time as a result (Bell and Blanchflower 2011; Kahanec and Fabo 2013; Beblavý and Fabo 2015).

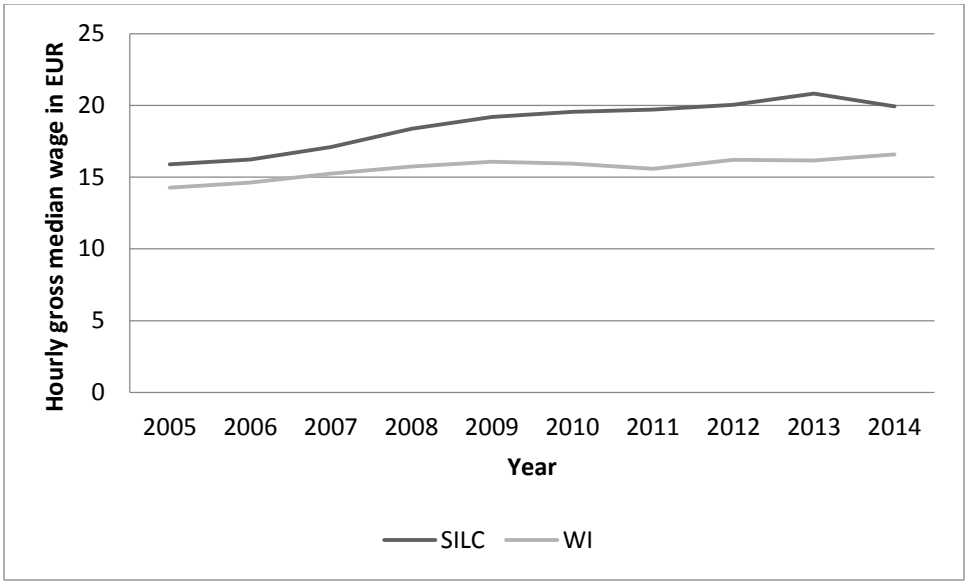

Figure 2: Comparison of median wages between WI and SILC data.
Own calculation.

Respondents in the WI sample work longer hours – 35 to 36 hours a week on average, compared to 32-33 hours for the EU SILC respondents. However, WI respondents are less experienced, which is in line with the finding of de Pedraza et al. (2010) for Spain. In contrast with the Spanish situation, however, we do not see any difference between the educational attainment of the SILC and WI respondents. This is possibly due to a wider spread of internet connection in the Netherlands compared to Spain, which probably means that the internet is also widely accessible for people without a particularly high education attainment (see Table 2).

The younger audience might be due to focus of the WI websites on the provision of wages and labour market information, which might be more relevant for more junior workers, rather than ones already established in their career field.

Table 2: Comparison of non-wage continuous variables between WI and SILC in %

		2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
SILC	mean workhours	32.8	31.5	31.4	31.5	31.9	31.8	31.7	31.6	31.6	31.8
	sd workhours	8.1	8.8	8.8	8.7	8.7	8.5	8.5	8.6	8.5	8.4
	mean experience	17.8	18.5	19.1	19.8	19.9	20.2	20.2	20.3	21	21.4
	sd experience	9.9	9.7	9.7	9.8	9.8	9.6	10	10	10	10
	mean education	15.2	15.3	15.3	15.4	15.6	15.6	15.6	15.7	15.8	15.9
	sd education	2.6	2.5	2.5	2.4	2.3	2.3	2.4	2.4	2.5	2.5
WI	mean workhours	36.2	35.3	36.1	35.9	36.2	36	35.6	35.4	35.8	35.7
	sd workhours	5.9	6.7	6.5	6.5	6.2	6.3	6.6	6.7	6.3	6.5
	mean experience	15.6	15.8	17.1	17.9	18.4	18.3	18.5	18.6	17.8	17.8
	sd experience	9.9	10	10	9.9	10	9.9	10	10	10	10
	mean education	15.6	15.6	15.4	15.4	15.2	15.4	15.5	15.6	15.8	15.9
	sd education	2.4	2.4	2.4	2.5	2.4	2.5	2.5	2.4	2.5	2.4

Own calculation

When looking at the categorical variables, we see a big difference in gender composition. While the SILC sample consists of about 50% of women, with the exception of the year 2005, in the case of the WI sample, the share of women varies and is generally about 45% since 2009. Before that, it was even more unbalanced because it greatly oversampled women in the first couple of years, possibly a consequence of the historical focus of the Dutch WI project on women. They are also more likely to be single, rather than married and also more likely to be employed with a temporary, rather than a permanent contract (see Table 3).

Table 3: Comparison of categorical variables between WI and SILC

			2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
SILC		Woman	48%	52%	52%	53%	52%	52%	53%	54%	54%	54%
	Number of coworkers	<10	13%	13%	14%	14%	13%	12%	12%	12%	11%	11%
		10 to 19	10%	10%	11%	11%	11%	11%	10%	10%	11%	9%
		20 to 49	20%	19%	18%	18%	17%	17%	18%	17%	16%	17%
		>50	57%	57%	57%	57%	59%	60%	60%	61%	62%	63%
	Family status	Never Married	34%	31%	33%	34%	35%	36%	36%	36%	35%	36%
		Married	57%	61%	58%	57%	55%	54%	54%	54%	55%	54%
		Divorced/Widowed	9%	8%	9%	9%	10%	10%	10%	10%	10%	11%
	Job specifics	Permanent contract	48%	52%	52%	53%	52%	52%	53%	54%	54%	54%
		Supervisor	13%	13%	14%	14%	13%	12%	12%	12%	11%	11%
WI		Woman	53%	54%	41%	38%	44%	43%	47%	48%	46%	45%
	Number of coworkers	<10	16%	16%	17%	14%	18%	16%	16%	17%	17%	18%
		10 to 19	12%	12%	11%	11%	14%	13%	14%	13%	13%	14%
		20 to 49	17%	17%	18%	17%	16%	16%	18%	18%	16%	17%
		>50	54%	55%	54%	57%	53%	55%	53%	52%	53%	51%
	family status	Never Married	49%	51%	44%	45%	43%	47%	44%	42%	45%	46%
		Married	41%	40%	47%	46%	46%	44%	49%	49%	47%	47%
		Divorced/Widowed	10%	10%	9%	8%	10%	9%	7%	8%	8%	7%
	Job specifics	Permanent	53%	54%	41%	38%	44%	43%	47%	48%	46%	45%
		Supervisor	16%	16%	17%	14%	18%	16%	16%	17%	17%	18%

Results

We ran the regression for all years pooled together with very large samples (47,000 for WI and 20,000 for SILC). We found all relationships to be highly significant and the coefficients largely similar (Table 4). Nonetheless, the model run on SILC data explains more variance in wages for both genders (33 % for men and 24% for women), the same model run on WI data explains only 29% and 23% respectively. Although the coefficients look largely similar between the two

samples, when tested statistically, the difference between coefficients is significant. The statistical test of joint equality of coefficients produced the resulting F – statistic of 13.31 for men and 22.03 for women, which are both very far from the critical value of 1.57 for the standard 5% significance threshold.

Table 4: Pooled OLS run on WI and SILC datasets covering the period 2005-2014

	Men SILC	**Men WI**	**Women SILC**	**Women WI**
Years of potential experience	0.0165***	0.0193***	0.0135***	0.0193***
	(0.00142)	(0.000849)	(0.00125)	(0.000758)
Years of experience squared	-0.000209***	-0.000236***	-0.000189***	-0.000295***
	(3.25e-05)	(2.05e-05)	(2.97e-05)	(1.91e-05)
Years of education	0.0431***	0.0353***	0.0442***	0.0305***
	(0.00130)	(0.000870)	(0.00124)	(0.000848)
Firm size 10 – 20	0.0501***	0.0405***	0.100***	0.0564***
	(0.0135)	(0.00777)	(0.0112)	(0.00661)
Firm size 20 – 50	0.0799***	0.0805***	0.117***	0.0905***
	(0.0120)	(0.00707)	(0.00993)	(0.00629)
Firm size 50 +	0.201***	0.187***	0.188***	0.165***
	(0.0103)	(0.00599)	(0.00828)	(0.00503)
Permanent contract	0.126***	0.0982***	0.0653***	0.0771***
	(0.0118)	(0.00602)	(0.00994)	(0.00472)
Supervisory position	0.124***	0.118***	0.0663***	0.0816***
	(0.00680)	(0.00420)	(0.00742)	(0.00438)
Married	0.129***	0.108***	0.0111	0.0387***
	(0.00730)	(0.00491)	(0.00680)	(0.00463)
Divorced or widowed	0.0741***	0.0612***	0.0131	0.0154**
	(0.0129)	(0.00895)	(0.01000)	(0.00659)
Year dummies	Yes	Yes	Yes	Yes
	(0.0139)	(0.00907)	(0.0130)	(0.00874)
Constant	1.596***	1.655***	1.686***	1.745***
	(0.0279)	(0.0173)	(0.0260)	(0.0161)
Observations	9,553	24,739	10,851	22,146
R-squared	0.325	0.287	0.238	0.225

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

The pooled regression is not an ideal choice for estimation technique, mainly due to a major external shock represented by the Great Recession that took place during the covered period and severely affected the labour market. Thus, we broke the sample per year and ran separate regressions for each year. The predictors are again highly significant and quite similar at the first sight. Nonetheless, the explanatory power of the model estimated on WI data appears to be lower

and moreover varies from year to year, while the explanation power of models run on the SILC data appears to be quite stable. One potential explanation for this phenomenon is that the younger workers in the WI sample might see their wages affected more by factors related to the general well-being of the economy, which goes beyond the presented model, such as the highly disproportionate effect the Great Recession had on the younger generation. Indeed, when trying to see whether removing individual variables from the model can help us identify individual variables driving the result, we find the differences between estimates on the basis of the two data sources becoming larger, rather than smaller.

Furthermore, we once again applied the equality of coefficients test, as explained in the empirical strategy section. What we see is that the coefficients tend to be significantly different for most year-gender combinations, just like it is the case with the pooled data from all years. Nonetheless, in some cases, the coefficients not only appear similar, but also pass the statistical test, which means the test does not provide sufficient evidence to rule out their similarity. Specifically, in 2007, the difference between the coefficients estimated on both samples is not significant for both genders. Furthermore, when we take 5% significance threshold (which makes the critical value around 1.79 due to the decreased number of variables in the model as a result of the year dummies not being present), the estimates for men are not significantly different in 2007, 2009 and 2014. (see Table 5).

Table 5: Statistical test results of equality of estimates generated from WI and SILC data (F = F-ratio).

Year	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
Men	3.27	2.15	**1.21**	3.21	**1.71**	3.37	6.96	5.72	3.35	**1.45**
Women	5.27	4.5	**1.72**	3.36	2.46	5.05	8.55	6.53	8.93	4.47

Note: **Bold** coefficients are NOT significantly different at the 5% significance threshold.

Having confirmed similarity of coefficients only in four out of twenty regression pairs, we position our findings in line with the literature, reaffirming that while the WI survey can be used to gain a general idea about relationships between variables, it produces different results compared to the high quality SILC sample. At the same time, we have shown, that there are estimates based on the WI data which are not significantly different from the ones obtained through a widely used, traditional survey. By showing that these coefficients can be similar in a

standard model, commonly used in labour economics, we strengthen the case for future methodological research on the WI data in particular and web surveys in general.

Conclusion

Based on statistical analyses of the WI and SILC data, we have shown that while the WI data (most likely) does not represent the Dutch working-age population, the estimates are at the first sight quite comparable with the SILC–based ones. While this similarity is not necessarily also confirmed by statistical tests, we managed to identify occasions, where similar coefficients were indeed produced. While such coefficients were only identified in a minority of cases, being the first to show it is possible in such a robust model makes our expectations somewhat more optimistic in regards to the usability of non-probability based web surveys than most of the literature published so far.

Nonetheless, it is important not to overstate the importance of our results. Indeed, in most tests, we have identified significant differences and thus our findings do not in any way contradict the consensus in the literature, that WI usefulness in mainly limited to exploratory research. However, we believe it is possible to further build on our findings by exploring the underlining reasons behind the similar results in cases when such a matching coefficients are produced. One interesting finding is that this appears to be the case when looking at male respondents and less of the case when looking at women. This gender aspect represents one avenue that ought to be looked into more closely.

Additionally, we believe that our choice of the Netherlands as the country of focus was an important reason for getting a WI survey sample which is not too dissimilar to SILC in important respects, such as educational attainment of the respondents. That is because internet access in the Netherlands is much more widespread than in previously analysed countries, such as Spain (de Pedraza et al., 2010). Additionally, we analysed a survey which has been accessible for a very long time from a website, which has its established fairly large audience.

As a result, we are analysing a sample, which generally is not comparable with the actual population of the country, but which is typically similar enough to determine the determinants of wages comparable to estimates from a representative sample. Although the socio-demographic

composition of respondents has changed over the years – the share of women has decreased as the project moved away from its initial focus on women towards being a survey of general population – the share of workers in supervisory positions declined. But in terms of average wages, working hours, educational attainment and working experience, the WI surveys cover quite a stable sample, which does not change more year to year than the samples collected through standard sampling techniques. Consequently, we see potential for exploring whether the sample is also comparable to representative samples in matters other than wage determinants.

Chapter 3: Using Online Job Vacancies to Better Understand Labour Market[11]

Introduction

This chapter examines the potential of online job portals as a data source for labour market research and particularly for research on the rise of new occupations and skills. The aim is to establish an overview of the main existing approaches to online vacancy analysis, contextualize the analytical approach within the broader literature, and to showcase some empirical applications to demonstrate the advantages and drawback of the individual approaches.

Thee distinct methodologies for approaching vacancy analysis are presented in this chapter.The first two methodologies assess the potential job boards in the classical sense. Firstly, we discuss the possibility of using the text of vacancies and secondly, we will focus on so-called metadata. The third methodological discussion revolves the developing phenomenon of 'online labour platforms', that is, spaces offering the general public the chance to sell a wide range of services on an open market to customers, with work being organised entirely online.

We find that each of those methodologies has different uses and come with different downsides. More specifically, the analysis of the text of the vacancy is the most universally applicable at the expanse of possibly quite noisy data, that might not be in possible to interpret in a straightforward way (see Chapter 5 for an example of this shortcoming potentially having a major impact on the usability of the data). The metadata are much more simple to draw conclusion from, but typically only allow us to ask a limited set of questions. Finally, the online job platforms allow researchers to observe both supply and demand side of matching to an unprecedented extend, but findings might be impossible to generalize beyond individual platforms.

The remainder of this chapter follows this structure: Firstly, existing applications of job

[11] This paper represents the author's contribution to the InGRID collaborative research project, developed in collaboration with Miroslav Beblavý and Karolien Lenaerts and has been published as working papers (de Groen et al. 2016; Beblavý et al. 2017). The platform analysis was used as a basis for a paper published in Transfer: European Review of Labour and Research (Fabo et al. 2017c).

vacancies as a data source are presented. Secondly, the three methodologies we propose for using online job vacancies as a data source are discussed, along with a short example of their practical implementation. The chapter ends with a discussion of pros and cons of the individual methodologies.

Literature Review

A vacancy is a key concept in labour economics, being studied since the famous Holt paper *The concept of job vacancies in a dynamic theory of the labor market* in the mid 1960s. Holt was mainly concerned about the quantities – numbers of openings by the companies and the number of workers available on the market. Nonetheless, already Holt was already interested in the content of the vacancy, namely skill requirements and salary offered (Holt and David 1966).

Historically, attempts to use vacancies on labour market analysis on printed job advertisements. Jackson et al. (2005) and Jackson (2001), for instance, work with job advertisements published in newspapers to test a number of sociological theories. Their work focuses on the merit selection hypothesis and on the link between social mobility and education in the UK. This work does not appear to be at all concerned with the questions of data quality and generalizability. This shortcoming was to an extent addressed by Dörfler and van de Werfhorst (2009), who examine the merit selection hypothesis for the case of Australia, on the basis of 47 advertisements published in newspapers between 1985 and 2005. Further, they attempt to increase the robustness of their findings econometrically, opting for a multivariate regression approach to account for the fact that some skills levels may be under-represented in their database. van Ours analyses the duration of vacancies being advertised with connection to the required educational attainment (Ours 1989).

Most newer studies, however, focus on analysis of online job advertisements instead of the printed ones. In an early study, Backhaus (2004) examined how firms describe themselves to job seekers in their recruitment materials (or, in other words, how firms go about company branding and which marketing materials are used with that objective in mind). To this end, she obtained a sample of job advertisements from Monster.com. Other authors developed the topic further: Kuhn and Shen (2013) use a data sample that includes millions of job listings extracted from the third-largest Chinese job portal (through web ‘spidering’) to analyse gender discrimination in the

recruitment of workers. The database is supplemented by firm-level data. They detect gender discrimination but note that it is less problematic for positions that require highly-skilled workers. In other studies, Maurer-Fazio and Lei (2015) consider discrimination based on gender and facial attractiveness in the Chinese labour market, while Maurer-Fazio (2012) looks into ethnic discrimination. Martínek and Hanzlík (2014) combine data from job portals with data obtained from the Ministry of Labour and Social Affairs to study labour market dynamics in Czechia. Masso et al. (2014) take a different approach and use job board data to study the occupational mobility of return migrants in Estonia. The sample that these authors have is extensive, but not completely representative due to a bias towards younger, highly-educated workers and towards private sector jobs with higher remuneration levels. Yet, no action has been taken to address these issues.

Skills and matching research is perhaps the most developed research direction using online job vacancies. Mýtna-Kureková et al. (2012), for example, discuss the formal qualifications and other skills requested for low- to medium-level skilled occupations in Slovakia. These authors maintain that their results may be considered generalisable to the Slovak labour market, because the job portal from which the data is assembled covers a substantial market share. In another recent paper, Mýtna-Kureková and Žilinčíková (2016) explore whether low-educated workers and student workers are competitors for the same positions. Their results suggest that this does not appear to be the case, as they have different skill sets that are in fact complementary.

Other papers focus on the skill requirements in the IT industry. Wade and Parent (2002), for example, look into the relationship between performance and job skills. More specifically, they compile a database of job advertisements for webmasters, to which data gathered via a web survey are added. They also apply multivariate regression techniques. Meanwhile, Huang et al. (2009) differentiate between business, humanistic and technical IT skills. Capiluppi and Baravalle (2010) explore the popular website Monster.com to investigate the potential mismatch between the skills required for IT staff and those developed in education or training programmes in the UK. Kennan et al. (2006) compare skills requirements for librarians in Australia and the US, on the basis of a set of advertisements published online and in the printed media. They find that skill requirements vary greatly across these two countries and through time. Hershbein and Kahn (2015) provide evidence for upskilling with a sample of vacancies from Burning Glass:

employers require more in areas with higher unemployment rates.

Using a similar approach and dataset, Sasser Modestino et al. (2014) examine changes in employers' demands during the Great Recession (2007-2012) and the subsequent recovery (2010-2012). Employers are more demanding in bad labour markets, both in terms of education and experience. Kudlyak et al. (2014) rely on matched applicant-vacancy data from SnagAJob to assess how job seekers' behaviour changes during their job search. In another contribution based on SnagAJob, Faberman and Kudlyak (2014) report that job 48 seekers' search efforts decline with search duration. Finally, in a recent report, Rothwell (2014) uses data from Burning Glass to study the duration and skill requirements of job advertisements, focusing in particular on Science, Technology, Engineering and Mathematics (STEM) positions.

In these studies, too, methodological issues do not appear to receive much attention, due to their mainly explorative nature. However, it is important that all vacancies are posted online. Not every job opening creates a vacancy and not all vacancies are actually (new) jobs. In addition, online job listings appear to be targeted towards more highly-educated applicants looking for white-collar and STEM jobs in sectors with high skill requirements (Carnevale et al. 2014). Furthermore, job advertisement data is highly volatile and may be inconsistent. In this respect, it has been argued that job advertisements are combined with other labour market information and educational data sources. The combined data approach has already been established by Dunlop (1966), who compares the advantages and disadvantages of vacancy analysis. According to this author, job vacancy data may be regarded as the counterpart of unemployment series and can be used as a measure of economic fluctuations. Vacancy data may support labour allocation, labour administration and the development of (re-)training programmes. Nevertheless, job opportunities are made available in a variety of different ways. Moreover, vacancy data does not capture the internal labour market and self-employment. As such, future advancement in using online vacancy data is necessarily connected to an increase of methodological sophistication of the conducted analysis.

Methodological Aspects of Using Vacancy Data

Online job portals

At the heart of the online job portals are the vacancies or job advertisements published by employers looking for qualified applicants to fill a position in their firm. Many of these websites, however, also allow job seekers to post their CV and résumé, provide job-search or career advice and other information (such as average wages by sector or legal advice on employment contracts). Some well-known examples of job portals are Monster.com, Careerbuilder.com and Glassdoor.com. Job portals can list domestic positions and/or jobs abroad. Although many portals cover all sectors and occupations, there are also a lot of job boards that specifically target a narrow selection of sectors or jobs (some examples are itjobs.com and euroeconomistjobs.com) or ones that focus on a specific region (e.g. jobsinberlin.eu).

Job seekers benefit from using online portals as they can browse through a high number of positions, use search criteria to find a position that matches their profile and get a better understanding of employers' requirements. For firms and recruiters, job boards offer many advantages too, such as the ability to list job openings on targeted websites while keeping advertising costs low. Online job boards also serve as a useful tool for workforce agencies and colleges, as they facilitate the identification of emerging occupations or new education and skill requirements (Carnevale et al. 2014).

Job portals are typically structured in a way that allows job seekers to easily find similar jobs to the one they are looking for. To this end, portals rely on an occupational classification to structure their database and facilitate job search. In some cases, advertisements are assigned to a specific category ('tagged') by the advertiser. The list of tags can be published online or stored in a library which is called by a search Application Processing Interface (API). Job portals generally update these tags quite regularly to capture changes in the labour market.

The Slovak job portal profesia.sk, for example, included about ten new occupations every year between 2011 and 2014. In other cases, job portals do not use the tags system but associate occupations purely on the basis of keywords in the job description. This approach, however, is more prone to errors. The underlying occupational classification is a good data source to capture

the occupational structure in a region at a certain point in time. These classifications can easily be compared with occupational classifications from other sources, such as ISCO.

Job advertisements themselves also contain a lot of information that can be used to detect new occupations and skills. An advertisement typically includes a job title, description (e.g. responsibilities, tasks), requirements (e.g. level and type of education, skills) and other information (e.g. details about the position, firm or industry, such as salary, company name and field of activity). In their study based on online advertisements, Burning Glass (2014), a company which collects job vacancies posted on the American and British markets, detects more than 70 'data fields' in a single job post.

The appearance of a new type of platforms, which are no longer limited to serving as notice boards for "offline jobs", but rather facilitate online work directly opens up additional options for research. Job matching on those platforms is quite often transparent and capturable through techniques such as data crawling and data scrapping. Furthermore, the rules governing matching of jobs and workers are also often publically available. As a result, researchers are able to observe characteristics of jobs as well as individual applicants and see the matching to a much greater detail than it is the case with 'traditional' job vacancies, allowing us to for instance better understand the dynamics of freelancing.

Job portals of all sorts connect the demand and supply side of the labour market. As finding and responding to an advertisement is often the first step, job portals are a valuable source for labour market research, especially in light of technological progress (Carnevale et al. 2014; Kuhn 2014; Kuhn and Mansour 2014; Mýtna-Kureková et al. 2015) Job advertisements shed more light on the qualifications and skills that employers are looking for. Compared to traditional data sources, web data such as job vacancies have the advantage of being more detailed and providing information that may not have been available before.

Using online job portals for research

Using online job vacancy data for labour market analysis comes with several strengths and weaknesses. In general, they should be understood not as a replacement for traditional data sources, such as the Labour Force Survey (LFS) or census data, but rather as complimentary

resources, where traditional methodologies fall short.

The first set of benefits relates to the data collection process (Wade and Parent 2002; Kennan et al. 2006; Benfield and Szlemko 2006; Mang 2012). Internet data can be collected in real time. In contrast to traditional sources, there are no lags or revisions. Another advantage is that web data allows researchers to collect large and diverse samples in a fast, flexible, easy and inexpensive way. Web data also makes some of the tedious and complicated steps of collecting data from traditional sources redundant (e.g. data entry). Similarly, even though printed vacancies are a rich source of data, they are more difficult to manage and manipulate than job advertisements published online.

Secondly, web data can be used to fill research gaps (Shapiro 2014). Internet data is particularly useful for covering areas where traditional sources do not exist or are only limited (e.g. low quality or a lot of missing information). In that way, web data can also enable research on topics or concepts that are difficult to grasp or measure with traditional data sources. An important example is self-employment. Self-employment is a driver of job creation and entrepreneurship. It has therefore received a lot of attention from policymakers. Nonetheless, it is not well represented in traditional data sources (Fairlie and Robb 2009). More specifically, the traditional data sources on self-employment commonly only provide details on the business or on the owner and only few datasets appear to combine both. Other issues are that data is lagged and self-employment is defined in different ways in different sources.

Another advantage is that web data is often publicly available and relatively detailed. Moreover, job advertisements in particular typically have a clear structure and contain real job titles, descriptions and requirements (Shapiro 2014). They have clear space and time dimensions and reflect what is demanded in the labour market. As such, vacancies provide an excellent tool for studying new occupations and skills as they arise, which is why they are at the heart of our work - along with other data available on the job boards that publish them. Aside from vacancies and job portals, other web-based data sources may be relevant.

Nonetheless, as online job vacancies are becoming increasingly established as a data source for labour research, some are hesitant about using them. A variety of authors (Carnevale et al. 2014; Mýtna-Kureková et al. 2015; Kearney and Levine 2015) have drawn attention to the

methodological issues related to using data obtained from online job boards (vacancies in particular) and intermediaries. The main issue is data representativeness. First of all, not all job openings entail a vacancy. Not all vacancies are published online and vacancies may not refer to (new) jobs. Secondly, it is unlikely that all vacancies would be collected, even if all would be published online. Thirdly, vacancies may appear to be targeted towards highly-educated white collar workers and STEM jobs, at least in the US (Carnevale et al. 2014). Fourthly, statistics generated on the basis of this data source tend to be quite volatile in time and may be inconsistent; vacancies may be duplicate and poorly structured. In particular, vacancies may not explicitly contain all requirements demanded. Fifthly, there remain some ethical issues connected with the mass download of vacancies and the potential bandwidth cost data collection might include on job board operators.

Mýtna-Kurekova et al. (2015) suggests two venues possibly suitable for increasing robustness of findings based on online job vacancies. Firstly, they argue combining a higher number of various data sources might be a suitable way forward. Alternatively, they suggest utilizing advanced econometric methods such as Bayesian inference or maximum likelihood. In the remaining part of this chapter, we point out another possible venue, connected with careful consideration of all information presented in the vacancy, including the metadata and also inclusion of vacancies posted on online labour platforms.

Vacancy Data Collection Methods

Method I: Vacancy Content

Job vacancies and CVs may be assembled via web crawling or ‘spidering’. With a ‘spider’ or web bot, a substantial sample of job advertisements may be obtained and stored in a database. The data collection process is relatively fast and easy. Vacancies are typically detailed and contain crucial information such as a job title, description (e.g. responsibilities, tasks), requirements (e.g. level and type of education, skills) and other information (e.g. details on the position, firm or industry, such as salary, company name and field of activity).

One of the main difficulties of using vacancy data is that data cleaning, processing and management is relatively complicated (Carnevale et al. 2014). Data is assembled in a database,

extracted and divided into smaller fragments. These fragments are then coded and analysed. This process can be very straightforward or very difficult, depending on the structure and content of the job vacancy. To carry out this process, a comprehensive taxonomy of variables and words is essential. In addition, semantic analysis and text mining are often required to support data coding. These issues are difficult to address, but nonetheless have to be overcome when one uses vacancy or CV data.

An example of this research is an analysis of about two million job vacancies from the US, obtained from the Burning Glass company, in thirty common occupations, which was undertaken in 2016. In this exercise, we broke down the large body of vacancies, identified common words in the vacancy text using an algorithm and selected those, which could potentially constitute a skill indicator (see Table 6). Based on this exercise, we calculated the share of vacancies containing one of the specified list of keywords.

Table 6: Main skill-relevant keywords identified in the vacancies

Education & formal qualifications		
	Formal education	Diploma, GED, bachelor, college, university, degree, high school
	Specialised training & licenses	Apprenticeship, training, card, certificate, certification, certified, CDA, ASE, CPR, AED
Cognitive skills		
Specific	**Computer skills**	Computer, PC, CRM, Microsoft (Office, Word, Excel, Power Point, Outlook, and Access), Windows, SAP
	Analytical skills	Mathematics, analytical, logic, quantitative
	Language skills	Bilingual, English, Spanish
	Driving licence	Driving licence
Generic	**Ability to learn**	Ability to learn, able to learn
Non-cognitive skills		
Social	**Communication**	Communication, speak, write, articulate, verbal,

	skills	interact, communicate
	Service skills	Client, customer, guest, needs and attention
	Team-work skills	Team, attitude, focus, leader work, environment, spirit
Personal	**Creativity**	Creative
	Flexibility	Flexible, stay overnight, travel, shifts, work evenings
	Independence	Self-motivated, takes initiative, independent
	Pleasant demeanour & manners	Attitude, positive, mature, helpful, confident, enthusiastic, professional
	Reliability	Attention to detail, dependable, reliable
	Stress-resistant	Calm, stress, crisis situation
	Timeliness & punctuality	Timely manner, deadlines, act quickly, punctual, prioritise
Experience		
	Experience	Experience combined with year, month, work, preferred, desirable
Appearance / look		
	Appearance	well-groomed, clean, neat, professional appearance
Other		
	Citizenship	US citizen, citizenship
	Criminal record	No criminal record, clean criminal record
	Drug testing	Drugs, drug testing

On the basis of this exercise, we found that 67% of the vacancies examined include formal education requirements. 49% of these vacancies demand service skills and 38% of them refer to

experience. These three factors constitute the main 19 filters that US employers use to select job applicants. Interestingly, other non-cognitive skills are also relevant: 30-40% of the vacancies refer to a pleasant demeanour (manners) and flexibility, while 20-30% contain communication skills, timeliness and creativity. As expected, there are huge differences among the 30 occupations. In general, vacancies in the United States prove to be relatively demanding in terms of the requirements that applicants have to meet. This also applies to low-skilled and mid-skilled jobs. Nevertheless, employers tend to be more demanding as the level of complexity of an occupation goes up (at least for some requirements).

As a following step, we perform a factor analysis. This technique is widely used in the social sciences because it allows researchers to identify the underlying relationships between a set of variables (which may not be directly measured). With this technique, a large set of variables can be replaced by a smaller number of 'factors', which reflect what the variables have in common with one another. In other words, factor analysis is a data reduction technique that can be used to remove redundancy or duplication from a set of correlated variables and replaces these with a smaller set of factors. This smaller set of factors may be easier to interpret and derive conclusions from. In our work, we identify the underlying relationships between the education, skills and qualifications demanded in the job advertisements. By identifying the underlying factors – which can be regarded as 'profiles' - we can substantially reduce the list of qualifications and skills that we have been using so far to simplify the subsequent analysis. In this way, we can get more insight into the general characteristics and the combinations of requirements that employers have in mind when publishing vacancies.

How was this factor analysis executed? In a factor analysis, observed variables are modeled as linear combinations of the potential factors plus error terms. The information gained about the interdependencies between observed variables can be used to decrease the number of variables in a dataset. In our analysis, we start from a set of 22 skills and end up with 5 factors (which now combine 18 skills and qualifications). Initially, we obtained 15 factors, but only 5 of them had an eigenvalue that was at least equal to 1. These five factors explain 85.3% of the cumulative variance. When the eigenvalue of a factor is equal to one or higher, it explains more variance than a single observed variable. We then excluded 10 factors, by restricting the eigenvalue to be at least equal to one when re-running the analysis; also restricting the number of factors to five

and the factor loadings to at least 0.6 (after rotation) (Field 2013). We also considered the uniqueness of the variables. The results that were obtained from the factor analysis and skills and other qualifications that they comprise are depicted in Figure 3.

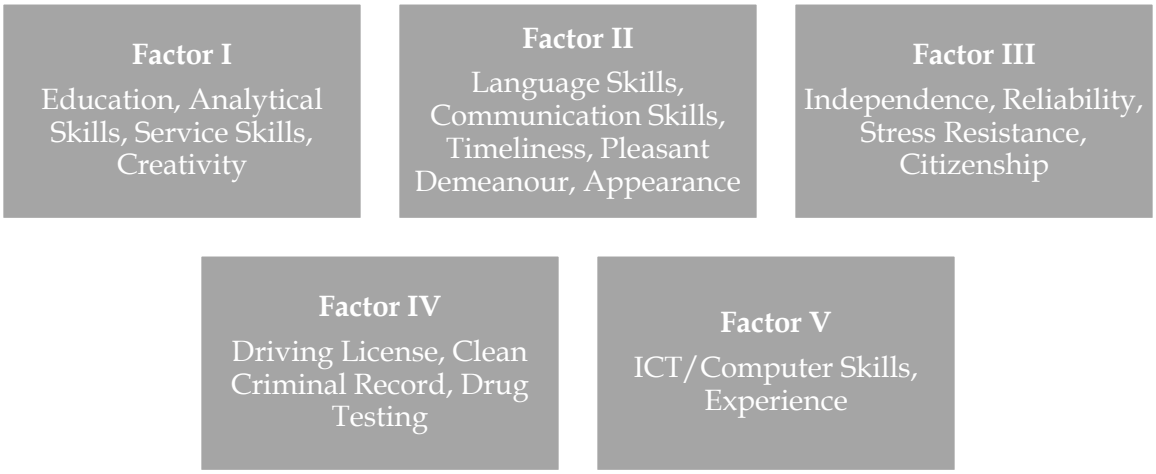

Figure 3: Overview of the factor analysis outcomes

As indicated above, we started our analysis with a set of 22 skills and requirements. As Figure 3 shows, five factors were identified that group 18 individual qualifications and skills. Factor I includes education, analytical skills (cognitive), service skills and creativity (both non-cognitive). Factor II comprises language skills (cognitive), communication skills, timeliness and a pleasant demeanour (all three non-cognitive), and appearance/look. Factor III is composed of three non-cognitive skills, namely independence, reliability and stress-resistant, and citizenship. Factor IV covers driving license, no criminal record and drug testing. Finally, Factor V is composed of ICT / computer skills and experience. From the factor analysis it is clear that specialised training and licenses, team-working skills, flexibility and the ability to learn are not positively correlated with any of the factors. Some of these skills and requirements can be expected to be relevant in a broad sense, and may, therefore, be not really related to any of the other skills or requirements.

While these factors might sound logical, they differ from the classification that one would arrive at based on the literature. Following the literature (Mýtna-Kureková et al. 2012), one may be

inclined to separate cognitive from non-cognitive skills, to distinguish between experience and education, to put the other factors in a distinct category, and so on. As such, this exercise shows that while text analysis of vacancies gives a lot of information and can be very useful to identify general paterns, the results can be influenced by noise in the data.

Method II: Vacancy Metadata

An alternative methodology is to rely on the metadata of online job portals rather than the text of vacancies and CVs published. Job boards are typically structured in such a way that a job seeker is offered a range of similar vacancies (rather than one perfect match) when looking for job offers. This implies that portals arrange their content or advertisements within a clear framework, often based on an occupational classification. Job portals often contain tens of thousands of vacancies, which have to be organised in a structured way in order to keep them manageable.

How does a job board set up an occupational classification? Again, there are several approaches that can be followed. A first approach is to create an occupational classification on the basis of a system of ‘tags’. Job advertisements can be assigned to specific categories or ‘tagged’ by the advertiser or the portal. Tags can refer to many things, such as job type, employer type, location, sector, wage bracket, etc. On most job portals, but not all, each vacancy is also assigned one or more occupation tags. Some job boards publish their list of tags online, so that job seekers can select their occupation of interest from the list. Other portals store the list of tags in a library that can be accessed by an application program interface (API) search. In this case, when job seekers start typing into a search box, the system generates options for automatic completion (as illustrated in Figure 4).

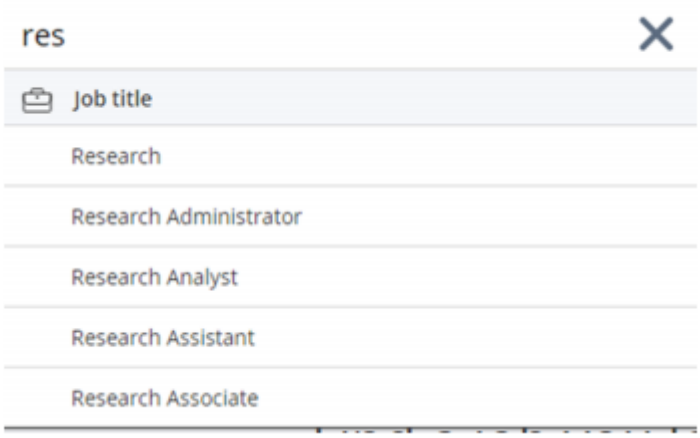

Figure 4: Example of autocomplete functionality from job board reed.co.uk.

The job tags that appear when one types ‘res’ are displayed

One should not think of such an occupational classification as a ‘static’ list. Instead, the occupational classification in both cases is regularly updated to account for changes in the labour market (e.g. 75 occupations that appear or disappear, new skills that are being requested, etc.). The Slovak job board profesia.sk is a good example: between 2011 and 2014, the job board added about 10 new occupations each year. Job portals’ occupational classification is a good data source to capture the occupational structure in a region at a certain point in time. These classifications may also be compared with those from other sources, such as ISCO and ESCO, to identify new occupations (e.g. what is available on the job portals but missing in the official lists).

An example of this approach is the so-called ‘occupational observatory’, we set up in 2015 as a method for detection of emerging new occupations across Europe. In this we regularly collected all the occupation tags from job portals in 11 EU countries. Each time we found a new tag, we analysed the associated job vacancies and determined, looked at comparable labour markets and determined, whether the occupation is potentially new. After having been examined, the new tag was added to the benchmark list of tags and as such no longer classified as new (see the visualisation of the analytical strategy on Figure 5)

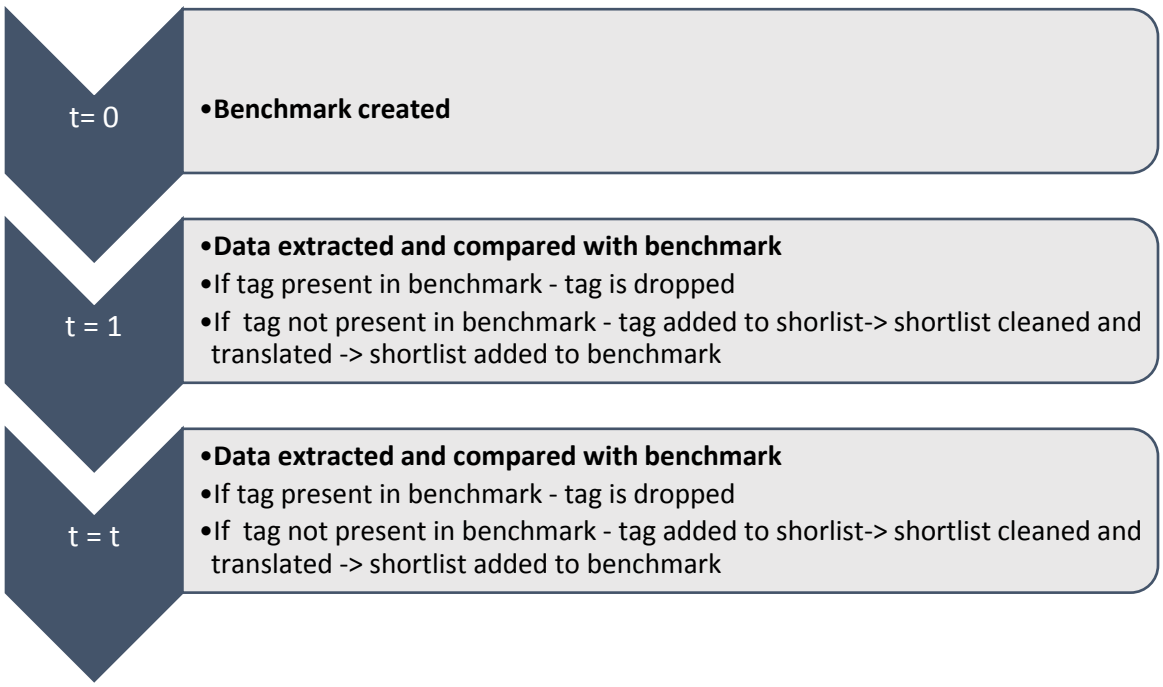

Figure 5: Illustration of the data collection steps

As shown in Table 7, only a small percentage of the new tags correspond to potentially new occupations. The remaining tags generally appeared to refer to traditional occupations, which were not advertised online or for which no vacancies were available when the benchmark was created. Another reason might be that some tags are associated with seasonal jobs, which are only advertised during a specific time of the year.

Table 7: Number of tags in the benchmark and on the shortlist for each country

Country	Job Portal	Number of tags in benchmark	Number of new tags	Number of valid new tags
BE	vacature.com (Dutch-speaking part) (www.vacature.com/vacature/bladeren/functienaam/), references.be (French-speaking part) (www.references.be/job/parcourir/fonction/)	415	714	NA
CZ	jobs.cz (www.jobs.cz/)	424	0	0
DE	jobsnet.dk (https://job.jobnet.dk/CV)	1,507	0	0
DK	keljob.com (www.keljob.com/emploi/metiers)	2,163	0	0
ES	jobs.de (www.jobs.de/regional/taetigkeitssuche)	233	2	1
FR	profession.hu (www.profession.hu/)	1,087	32	11
HU	lavoro.corriere.it (http://lavoro.corriere.it/)	369	91	11
IT	jobs.pl (www.jobs.pl/stanowiska-pracy)	962	0	0
PL	profesia.sk (www.profesia.sk/)	795	15	7
SK	careerbuilder.es (www.careerbuilder.es/)	468	30	16
UK	jobsite.co.uk (www.jobsite.co.uk/jobs/all/job-titles/)	6,063	0	0

Note" The number of valid tags for Belgium is missing, because of the peculiar way in which the job portal adds new tags (which, as explained, not only refer to actual occupations but also to other items and words).Therefore, there were many tags that indeed carried valuable information, but a manual inspection of these

The metadata analysis has a big advantage over text analysis connected to the lack of need to explore a large number of data points to identify relevant information. As a downside, the information we get is much more limited and we depend to a much greater extent to the structure of functioning of the individual portals.

Method III: Online Labour Platforms 'Vacancies'

The data collection methods so far discussed have been limited to the situation when the Internet serves merely as a notice board for jobs, which exist 'offline'. However, that is not the case anymore. Online labour platforms have developed globally overnight, offering companies and even individuals the opportunity to outsource all sorts of tasks to an online labour force of 'seervice providers' (Drahokoupil and Fabo 2016). A sizable body of literature has been recently produced, dealing with the rise of these platforms and their labour market impact (Codagnone et al. 2016a; Codagnone and Martens 2016; Maselli et al. 2016; Goudin 2016; Degryse 2016; de Groen and Maselli 2016; Codagnone et al. 2016b; Schmid-Druner 2016).

Unlike other data sources such as surveys or online job vacancies, which normally only provide information about labour supply or labour demand, the platforms can be used as a source of data for both supply and demand and enable us to observe labour market matching. That is because the matching algorithm of the platforms is typically made public and many platforms similar to social networks are organised. Service providers have their profiles online along with references from the customers and indeed the customers themselves have profiles, which contain information about tasks previously commissioned on the platform.

While this data is not always available and not always complete, when presented they can often be obtained by means of web scrapping. As such, it is possible to reconstruct entire functioning of the 'labour market' of a particular platform. One such exercise was carried out on Belgian local tasks platform Listminut, which we conducted in December 2015 and January 2016. On the platform, customers post small jobs, such as fence repair or pet sitting, potential workers apply and customers choose from among the applications. After scrapping over 14,000 workers' profiles and nearly 9,000 tasks advertised on the platform, we found over 75% of tasks were not met.

Based on this finding, we could determine matching dynamics for individual categories of tasks advertised on the platform, given that all workers have information about fields they are available to work in. One simple exercise is to compare the share of advertised tasks per field with the number of workers available to work in that field. As we can see on Figure 6 , three clusters are evident in the data – common housework tasks and computer assistance tasks which are quite often fulfilled, wellness and event organisation tasks which are often not fulfilled (wellness, events Organisation), because there is an insufficient supply of qualified labour and finally tasks involving a heavy degree of trust (babysitting, tutoring, household and animal care), which are often unfulfilled in spite of a large supply or workers (see de Groen et al. 2016 for detailed discussion of methodology and results of this exercise)

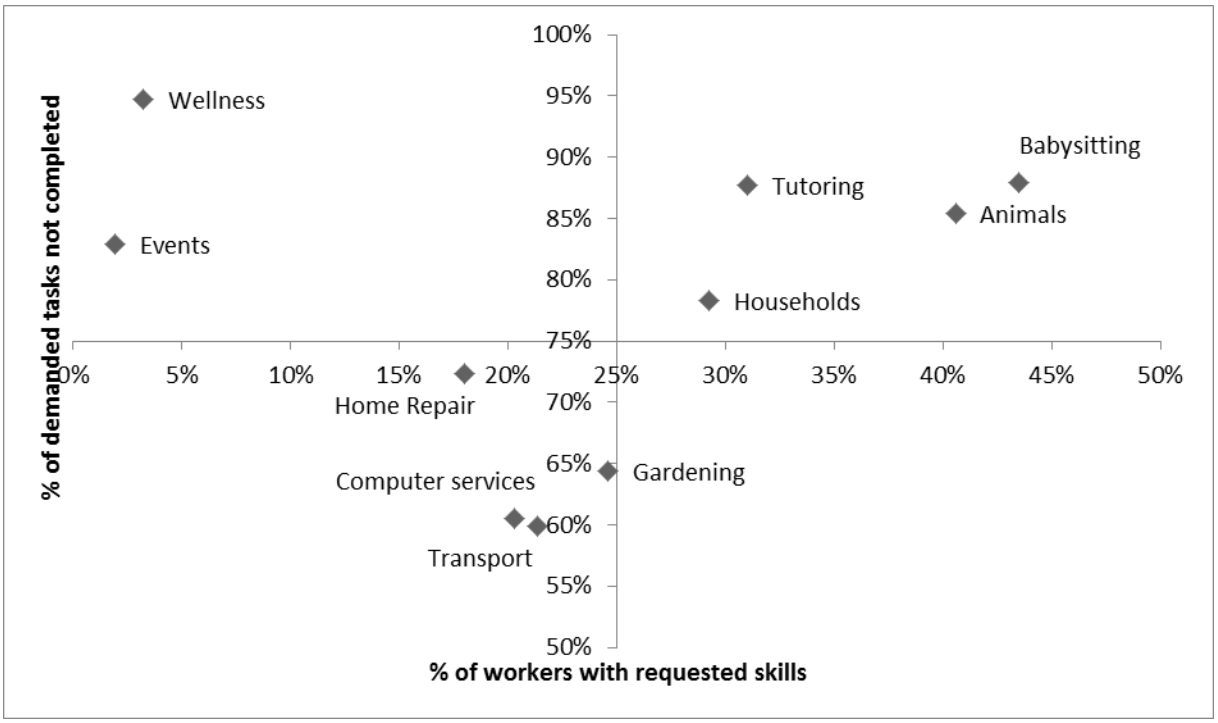

Figure 6: Share of tasks not completed on the Listminut platform contrasted with the share of workers with requested skills available for each of these categories

Conclusion

The body of literature using job vacancies as a data source is quickly growing, nonetheless methodological aspects of using the data source are often treated as an afterthought. In this chapter, we propose three distinct methodological approaches to using online job vacancies to learn about the labour market, each with its own advantages and disadvantages (summarised in Table 8).

Table 8: Overview of the advantages and limitations of using vacancies, job portal metadata and to study labour market

	Vacancies	Metadata	Data from Platforms
+	Detailed information, clear structure (job title, job description, requirements, etc.), easy to collect large and diverse sample	Easy and fast to collect, manage, process and analyse data due to smaller samples, and to use and interpret	Provide both sides of the equation (supply and demand), complete dataset (for the specific platform)
-	Data processing, coding and analysis is very resource intensive. The resulting data tend to remain quite noisy.	Depending on the portal, data may not be very detailed or complete, more limited potential for research	Platforms so far only small part of the labour market, hard to generalise findings beyond a single platform.

The future for this data source appears to be bright. Undoubtedly, in the near future, more and more countries will be covered with persistent and comprehensive data crawling algorithms of the kind applied by Burning Glass and Textkernel. Furthermore, advancement in Big Data analytics will enhance our ability to gauge meaning from the data, including effective use of metadata. Finally, the rise of platforms and digital work in general is likely to enhance our ability to observe labour market dynamics in real time.

At the same time, analysing this data source requires learning to deal with issues such as processing big volumes of data and finding creative solutions to gauge meaning from datasets, which are often structured very differently from traditional datasets commonly used in economics. Consequently, methodological considerations should become a much more prominent stream in literature dealing with job vacancies as a data source.

Chapter 4: Analysing Skill Supply: 'Pricing of Skills'[12]

Introduction

In a globalised world where countries are firmly engaged in international trade, investment and migration, knowledge of foreign languages is an important skill. Foreign language skills are considered a major asset particularly in Europe – not only because of globalisation but also due to the continent's inherent linguistic diversity. Therefore, a lot of effort has been made to promote foreign language acquisition and multilingualism and to better understand how this can be facilitated (Eurydice and Eurostat 2012). Initiatives such as the Erasmus+ programme and the Creative Europe programme are only a few examples. This focus on foreign language skills is motivated by the role of language skills for everyday life (as a means of communication, interaction with others, bridging intercultural gaps etc.), and their significance in Europe's labour markets more specifically. Esser (2006), for example, examines the link between language and human capital. Other work has analysed the relationship between language skills and employability, and reports that this relationship is positive (ET2020 2011). Moreover, foreign language skills are crucial for labour mobility and labour market integration.

Although the importance of foreign language skills has been widely recognised, studies on the supply and demand of these skills have only started to emerge in the last two decades (Antonietti and Loi 2014). Economics literature has mostly discussed the topic of language in relation to migration and trade, from both aggregate and individual perspectives (Antonietti and Loi 2014; Isphording 2014). The *supply of foreign language skills* is typically measured on the basis of interviews and surveys. According to a Eurobarometer survey conducted in 2012, 54% of Europeans speak at least one other (foreign) language, in addition to their mother tongue, while 25% can speak at least two foreign languages and 10% at least three (Eurydice and Eurostat 2012). These results further show that at least half of the population cannot speak any foreign

[12] This paper has been published in Empirica: Journal of European Economics, as a co-authored paper with Karolien Lenaerts and Miroslav Beblavý (Fabo et al. 2017a)

language well enough to hold a conversation.[13] Meanwhile, foreign language proficiency appears to be even less common in the Visegrad countries than in the EU (as reported in Table 9).

Table 9: Share of people able to have a conversation in English or German in the EU27 and the V4

	English	**German**
EU27	**38%**	**11%**
Czechia	27%	15%
Hungary	20%	18%
Poland	33%	19%
Slovakia	26%	2%

Source: Eurostat, 2012: Special Eurobarometer Report 386

Nevertheless, even in those countries, foreign language skills are required for the labour market. In fact, when it comes to *labour demand*, previous research has confirmed that the demand for foreign language skills is on the rise and that this trend is likely to continue in the future (see a.o. ET2020 2011; Antonietti and Loi 2014; Isphording 2014). This development is driven by the multilingual and multicultural environment in which firms increasingly operate, the global competition that they face and their aim to broaden their market access (ET2020 2011). It is also an important factor facilitating mobility of workers on the single EU labour market (Barslund and Busse 2016).

In our paper, we contribute to this literature and focus on the importance of foreign languages in the labour markets of Central and Eastern Europe. Specifically, we examine how common it is for employers to make foreign language skills a requirement in their job advertisements and whether proficiency in foreign languages is rewarded with a wage premium.

We discuss the case of the Visegrád Four (V4), which is composed of Czechia, Hungary, Poland

[13] The ten countries in which over half of the population cannot speak any foreign languages are: Hungary, Italy, Portugal, the United Kingdom, Ireland, Spain, Bulgaria, Romania, Czechia and Poland.

and Slovakia. We selected the Visegrád Four because we are interested in the role of foreign language skills on the labour markets in countries that are open to international trade and foreign investors, but where the population's language skills may not be very well developed. During the last two decades, the economies of the V4 countries have grown considerably, attracting substantial amounts of foreign direct investment (FDI). The FDI-driven nature of growth resulted in a deep integration of the V4 economies in European and global capitalism. The regional bank sector is dominated by foreign-owned banks (typically Western European), multinational companies established shared service centres in all major towns of the region, and the V4 countries are very active in international trade[14]. The openness of the V4 economies creates demand for workers capable of communicating in foreign languages, particularly English – the *lingua franca* of international business (Sanchez et al. 2011).

In addition to English, German and Russian have historically been widely taught and used in the region. The importance of the Russian language has diminished somewhat since the end of the Cold War, but Germany remains the most important trading partner for all four V4 countries. Additionally, Germany and Austria are major sources of FDI into the region. FDI transfers from German-speaking countries (Germany, Austria) have been very visible in the field of manufacturing and, while initially outsourcing entailed mainly manual work, it now also includes outsourcing of skill-intensive activities (particularly since the 2004 EU enlargement) (Marin, 2010). Given the strong mutual integration of the four Visegrád countries, languages spoken in one of the other V4 members can also be relevant (e.g. for cross-border labour mobility). Yet, out of the ten countries in which at least 50% of the population is unable to hold a conversation in a foreign language, five are located in Central and Eastern Europe and three are part of the Visegrád Four. For three of the countries, the main national language is not among the five most widely spoken languages in Europe (Polish shares the fifth position with Spanish). Our focus on the V4 countries is further motivated by the strong embeddedness of the region in the global division of labour and their position in global value chains, which allows us to gain new insights

[14] In Slovakia, exports in goods and services as a percentage of GDP amounted to 93.5% of GDP according to Eurostat data for year 2015. For Hungary, the share was 90.7% and for Czechia it was 83%. Only Poland is a partial exemption here, due to its relatively sizable internal market. For Poland, the share of exports in goods and services represented 49.6% of GDP in 2015.

into the role of languages in the contemporary economy.

In order to examine which foreign language skills are demanded on the V4 labour markets and whether knowledge of these languages would result in a wage premium, we combine data obtained from online job portals with data from the WageIndicator survey. The WageIndicator survey (WI) is a continuous, voluntary web-based survey of wages and working conditions. For our purposes, it is important that the wage data be available at occupational level, coded at the 4-digit ISCO level, which makes it possible to connect data at occupation level with the job portals data. Furthermore, since September 2015, the WageIndicator survey has been gradually enlarged by variables on language proficiency, which allows us to perform limited exploratory statistical analysis on an individual level as well.

The four job portals that we use for our analysis each have an interesting feature: each of them offers the option to select vacancies according to the demand for individual languages (e.g. the Slovak job portal also allows its users to select vacancies that require English language skills). This feature considerably facilities data extraction. Most job portals, however, do not have this feature. Instead, selection is normally limited to criteria such as income bracket, region, minimum education attainment etc. Because we were able to use the information about language requirements for each vacancy, normally coded by the prospective employer, we could avoid the need to work with the entire text of the vacancies (i.e. the job description) and having to deal with identification of skill requirement, which is normally a painstaking process (Beblavý et al. 2016b; Mýtna Kureková et al. 2016).

The remainder of this chapter is structured as follows. In the first section, we review the literature on the role of foreign language skills on the labour market. In the second section, we describe our data sources' composition, limitations and elaborate on the methodology used. Thirdly, we provide a descriptive and statistical analysis of our data. The final section presents a discussion of the results and conclusion.

Literature Review

Our analysis of employer demand for foreign language skills in the Visegrád region is embedded

in literature on the role of language in the economy, which was surveyed fairly recently by Zhang and Grenier (2013). Their main argument is that traditional research on the relationship between economics and languages may be divided into three strands: the first strand deals with language and economic status, the second strand covers the dynamic development of languages (from an economic perspective) and the third strand addresses language policy and planning (also from an economic perspective). Since the 1990s, these three strands have been supplemented by applications of game theory to model linguistic issues (Zhang and Grenier 2013). The game theory approach has become very popular, particularly for studying bilingual labour markets (Armstrong 2015). Antonietti and Loi (2014) organise this literature into three subject areas: immigration, international trade and firm performance.

In our work, we are especially interested in empirical work that bridges the gap between globalisation, labour and the historical context, as all these factors come into play in the Visegrád Four. Many recent contributions are dedicated to the relationship between language skills and *trade or foreign direct investment*. Kim et al. (2015) investigate the link between language and FDI, finding a robust, significant relationship between them. Oh et al. (2011) explore the transaction costs that exist between country pairs that do not speak the same language - with a focus on English, French, Spanish and Arabic - and compare the transaction costs related to trade and FDI for these four languages. Their results suggest that speaking a common language raises both trade and FDI, but appears to have a larger impact for the former than the latter. They also detect a common hierarchy in transaction costs for both phenomena: transaction costs are the lowest for English, followed by French, Spanish and Arabic. Another important contribution on this topic is the work of Melitz, which looks into the channels through which foreign language skill influences trade (Melitz 2008; Melitz and Toubal 2014).

The relationship between language and *migration* has also received a lot of attention. Yao and van Ours (2015) examine the importance of Dutch language skills for immigrants in the Netherlands. They only find an effect on working hours for female immigrants. In another article, Budria and Swedberg (2012) assess how immigrants' Spanish language proficiency affects their earnings. The discovered effect is very high for highly-educated workers, but negligible for uneducated workers. Rooth and Ekberg (2006) relate language skills to occupational mobility in Sweden. The results suggest that upward mobility accelerates among

those who have invested in a Swedish academic education or destination-specific language skills. Bleakley and Chin (2004) indicate that language is a social and economic barrier that separates immigrants from natives. Migrants with poor language skills often face discrimination and social isolation. Finally, Chiswick and Miller (2010) investigate the demand for English in the United States. They find a stronger relationship between proficiency in English and intra-occupational earnings differences than between proficiency in English and inter-occupational differences.

Yet another set of papers considers *historical and cultural ties* to explain why certain language skills are (still) relevant on the labour market. As an example of this work, we list two papers that cover Russian language skills in Central and Eastern European countries. Other authors discuss the issue of language proficiency in labour market entry among young workers in Estonia and Ukraine (Constant et al. 2012; Lindemann and Kogan 2013). In Ukraine, being able to speak Russian - as well as the national language - appears to be much more important than in Estonia. This difference may be attributed to disparities in the immigration history of the Russian minority population and the language prevalence in both countries. In another contribution, Duncan and Mavisakalyan (2015) consider the importance of Russian language skills in the labour markets of Armenia, Azerbaijan and Georgia for the period 2008-2010. In these three countries, the Russian language is still commonly used in everyday life. The authors confirm that speaking Russian is a valued skill in their labour markets.

Closer to our research is the literature that examines the importance of language skills in *specific occupations***.** Maxwell (2010), for example, analyses the role of English language skills in low-skilled jobs arguing that workers with a limited knowledge of the English language have less employment opportunities and lower wages. In the case of Germany, Stöhr (2015) examines the wage premiums of occupational language requirements and finds a systematic premium for English only, not for other languages. Another example is the paper by Coombs and Cebula (2010), who study the language skill rewards for registered nurses in the United States but find mixed results. In a related study, Beblavý et al. (Beblavý et al. 2016b) examine the skill demand of U.S. employers on the basis of a sample of advertisements extracted from Burning Glass, identifying the explicit demand for language skills in 16% of the vacancies. Another interesting article is Mýtna-Kureková et al. (2012). These authors consider the demand for a range of skills for 23 low- to medium-skilled occupations in Slovakia. They find that knowledge of foreign

languages is explicitly demanded in 38% of the vacancies, making it the second-most-demanded requirement. Interestingly, this skill is important, even for low- and medium-skilled jobs performed on the domestic market.

Data and Methodology

Our main data sources are popular online vacancy portals in the four analysed countries. In fact, we do not analyse the content of these vacancies as such but instead work with the metadata available on the job portals. Online job portals are increasingly used as a data source for the labour market research (Carnevale et al. 2014; Kuhn 2014; Kuhn and Mansour 2014; Askitas and Zimmermann 2015; Mýtna-Kureková et al. 2015). While analysing online vacancies is a relatively new trend, it builds upon an existing empirical literature based on an analysis of traditional, printed job advertisements (Jackson et al. 2005; Jackson 2007; Dörfler and Werfhorst 2009). Nonetheless, the literature warns that this web-based source may not be completely representative, given that not all job vacancies are published online and white-collar jobs vacancies tend to be overrepresented on online job boards (Carnevale et al. 2014; Mýtna-Kureková et al. 2015). For that reason, job vacancies are a good way of understanding skill requirements for individual occupations, but not necessarily of estimating the number of vacancies on the market (Slane 2013). In other words, online vacancy analysis may provide valuable insights for a specific research setting (such as skill demand for a set of occupations), yet its results cannot be generalised. While data representativeness has been put forward as a major concern in this young though rapidly developing discipline, few solutions have been put forward (Štefánik 2012; Mýtna-Kureková et al. 2015; Beblavý et al. 2016a)

Our approach differs from the common methodology in that we focus on the metadata – the tag system that job portals use to structure the vacancies – rather than perform an analysis of the vacancies themselves. This approach is explained in more detail below. While an analysis of the text of the vacancies provides more detailed information, tags are quicker and easier to collect, process and analyse than vacancies (because sample sizes are smaller, yet essential information is still available). More specifically, for each of the Visegrád countries, we extracted data from a major online job portal (all are presented in Table 2). These job boards were not randomly

chosen but instead selected after a careful analysis of their labour market coverage. We looked for job portals that are dominant players in their respective national labour markets and have been previously used as a source of data by local scholars (Szabó 2011; Chmielecki 2013; Bohmová and Pavlíček 2015; Beblavý et al. 2016e). For example, the Slovak portal www.profesia.sk that we use has a market share of over 80% (Štefánik 2012).The Czech board www.jobs.cz belongs to the dominant player in the online vacancy boards sector (Švihel, 2016).

The four job boards that we used are listed in Table 10, together with the total number of vacancies that was available on each of them when we first extracted our data (July 2015). In total, the four portals contained approximately 74,000 vacancies. The Czech job board counted close to 15,300 advertisements, the Hungarian portal covered about 11,200 vacancies, the Polish job board published about 36,000 advertisements and the Slovak portal had close to 11,300 vacancies. For reference, we added the number of job advertisements available in Q3 of 2015 as reported by Eurostat (the data was not seasonally adjusted, derived from national statistical offices). Apart from Czechia, for which a very large number of vacancies was reported on Eurostat, it appears that the share of vacancies available on the job portals we use is quite substantial. Note, however, that this data is not fully comparable as the Eurostat data refers to a three-month period while the vacancy counts were obtained at a single point in time. More details on these vacancy counts and the extent to which they may be influenced by seasonal dynamics is provided below.

Table 10: Overview of the online job portals used and the number of job advertisements available for the four countries in our sample (in July 2015).

	Online job board on which the vacancies were found	**Total number of vacancies available on the job board (July 2015)**	**Number of vacancies available in 2015 Q3 (Eurostat)**
Czechia	www.jobs.cz	15,269	102,141
Hungary	www.profession.hu	11,231	46,559
Poland	www.pracuj.pl	36,079	73,154
Slovakia	www.profesia.sk	11,344	17,288

Each of the four job portals uses a system of tags to organise individual job advertisements into

clusters. Clustering facilitates the search process, as this allows users to select, for example, only vacancies requiring specified language skills or only those for specified occupations. It is important to note here that tags are not disjunctive groups: a single job advertisement could indeed be mapped into several occupations or may, for instance, call for knowledge of multiple foreign languages (e.g. a vacancy that would call for English and German language skills would carry two foreign language-related tags: ‘English’ and ‘German’ – by tracking these tags, we can obtain information about skill demand more easily than if we would have to examine the full text of the vacancy instead). We therefore considered these tags as a proxy.

Our analysis is composed of two important steps. Firstly, we consider the total number of vacancies offered on each of the four job portals as well as the number of vacancies offered with a tag that refers to foreign language skills. For each country, we calculate the share of vacancies requiring individual language skills out of the total number of vacancies. Here, we do not focus on individual occupations but rather examine the aggregate demand for foreign language skills.

As a next step in our empirical analysis, we concentrate on the foreign language requirements for individual occupations in each of the V4 countries. Our aim is to arrive at a set of occupations that are present in all Visegrád countries and for which we could find a sufficient number of advertisements. To this end, we refine our analysis in the following way: For each Visegrad country, we remove the occupations for which fewer than 30 vacancies are published on the job portal. Thus, we no longer consider 11% of the job advertisements published on the Czech portal; for the Hungarian, the Polish and the Slovak job boards the corresponding percentages are 7%, 12% and 9%. As a second step, we matched the occupations across the countries, which proved to be a rather difficult task. Firstly, we translated the occupations (their occupational titles) to English and then compared them across the Visegrád group. Occupations that were not represented in each of the four countries were dropped from the sample immediately. Then, we mapped the occupations that were present in all countries into each other. For some occupations, this was relatively easy because there was only one suitable match in each country. For other occupations, more than one possible match was found. The reason this can occur is that an individual occupation or tag in one country (portal) may be split into several occupations or tags in the other countries (portals). For example, the tag ‘teacher’ - which is represented by a single tag in Czechia, Hungary, and Slovakia - is split into two tags, ‘teacher’ (Nauczyciel) and

'instructor' (Wykładowca), in Poland. Another example is that of programmer, for which a job portal can use multiple tags to distinguish between JAVA programmers, Python programmers and other programmers (whereas other job portals simply attach the 'programmer' tag to all sub-groups). In these cases, where multiple possible matches arose, a weighted average was calculated (weights equal to the number of job advertisements by occupation).

Altogether 59 occupations were identified as being sufficiently represented in all four countries (both in terms of the number of available vacancies and presence across the countries, although in some cases combining several tags). After identifying the 59 occupations, we return to the four job portals and extract the amount of vacancies available for each of these occupations as well as the amount of vacancies that demand certain foreign languages (i.e. that are tagged). From these numbers, we calculated the share of English and German languages, because we only find these languages relevant in the analysis of the entire sample of job vacancies. In the hypothetical example that the Slovak job portal would have 500 job advertisements for the 'teacher' occupation (i.e. 500 vacancies are tagged 'teacher'), of which 150 would also be tagged 'English' (i.e. 150 carry the tags 'teacher' and 'English'), the share would be 0.3.

Finally, we analyse the relationship between foreign language demand and wages. Unfortunately, there is no publically accessible representative data source that links proficiency in foreign languages to wage variables (on the occupation or individual level). While the Adult Education Survey, organised by Eurostat, has information on foreign language proficiency and wages, the wage variable contains only deciles, as opposed to specific amount, and the data is only collected in five-year intervals. More importantly, the occupation variable is not coded on the four-digit level. Meanwhile, the online-based WageIndicator (WI) survey has the required variables, but out of the V4 countries, it only covers Czechia and Slovakia with a sufficient sample. While this data source is not representative, it generally deemed reliable for exploratory research and, therefore, serves our purposes (de Pedraza et al. 2010; Tijdens and Steinmetz 2016). Our wage data are based on the median hourly gross wages, obtained from the WI web surveys collected from the beginning of 2013 until the end of the first quarter of 2016. We correlate the median wage in the occupation with the demand for foreign languages in that occupation.

From September 2015, the WI survey also contains the question 'In which languages are you

skilled?' with answers 'Not at all' , 'Barely skilled' , 'Rather skilled' , 'Skilled' and 'Mother language' . The question is asked for the English, Spanish, Russian, Chinese and Arabic languages, thus it does not allow us to study other languages. Nevertheless, the information we can gather for English enables us to increase the robustness of our findings by analysing the relationship between knowledge of foreign languages and wages on the individual level. Unfortunately, the sample size does not allow estimating this relation for individual occupations, which is why we run our analysis for the entire labour market. We base our test on the standard Mincerian earnings function, which predicts the logarithm of earnings (y) based on years of potential working experience and years of education (Mincer 1974). Formally, the Mincer equation is represented as follows, with S being the year of schooling and X years of potential working experience:

$$\ln y = \ln y_0 + rS + (\beta_1 X + \beta_2 X^2)$$

On top of the Mincer's predictors, we added dummies for English proficiency (with categories 'skilled' and 'native speaker' merged due to the low number of native speakers), gender, residency in the capital city, sector of employment, supervisory position, size of the company, occupation[15] and whether the respondent is in Slovakia or Czechia (more details on what variables are included in what model are provided in the next section). Using the Mincer function is a technique that has been used in the context of estimating value added of language skills particularly in the research of focusing on migration and multilingual countries (see e.g. Chiswick and Miller 1994, 1999; Constant et al. 2012). To the best of our knowledge, we are not aware of any study that used this technique in the context of the foreign language skills in the Visegrad region.

Results

Firstly, we looked at the number of vacancies that do *not* require foreign language knowledge. Of the job advertisements published on the Czech portal, 38% that are tagged 'does not require any foreign languages' . This share is equal to 25% for Hungary and 43% for Slovakia. Nonetheless, the relatively low number in the Hungarian case may be explained by the fact that

[15] Represented by 1-digit International Classification of Occupations 2008 version occupational groups.

not all vacancies are tagged for languages. If approximately 20% of vacancies with no language tags were tagged as without language requirement, the share would be similar to the Czech and Slovak portals. As no such tag is available on the Polish job board, we do not have information on the share of job advertisements that do not require any foreign language skills in this case (note that we cannot simply calculate this number, as a single job advertisement may have multiple tags). From these percentages, it is clear that there are considerable differences between the four countries that make up the Visegrád group. Foreign language skills are more demanded explicitly on the Hungarian and Czech labour markets than in Slovakia. We can only capture job advertisements that are explicitly tagged, not those where foreign language skills are implicitly demanded. This is an important caveat that one has to be aware of in any research that is based on vacancy-based analysis. In addition, foreign language skills do appear to matter to employers, as they are explicitly requested in more than half of the vacancies. The tags, therefore, do appear to have a signalling function.

We then focus on the ***individual languages****, first at the regional level and then at the country level*. When we consider all job advertisements at once (for the Visegrád region as a whole), we notice that 52% of them require English language skills, 12% demand German language skills, 2% list French language skills and less than 2% request Italian, Spanish or Russian language skills. However, these aggregates likely hide differences between the countries and could certainly be affected by the number of vacancies for each country. For these reasons, we also look at the percentages of job advertisements within each country that comprises these language demands.

Interestingly, *English* is the most frequently demanded language in all four countries. Nonetheless, there is significant variation between countries. 28% of the Czech vacancies refer to English language skills, 64% of the Polish advertisements demand English language skills. In Hungary, the share is 39%, while in Slovakia, it reaches 49%. *German* is the second most demanded language in all countries of the V4. For German, the shares seem to differ to a smaller extent. Across the four countries the share of advertisements with German language demands ranges from 10% in Czechia to 15% in Slovakia (again, this means that on the four job portals, between 10% and 15% of all advertisements published is tagged 'German'). Even though French, Italian and Spanish are used extensively in the European Union, these three languages

are in low demand in the Visegrád labour markets.

These percentages should be interpreted with some caution, as they might reflect *seasonal dynamics* (as indicated above, our data was collected in the month of July). To rule out seasonal dynamics, we contacted each of the job boards to inquire about the statistics for the entire year 2015. The Slovak job board informed us that 46% of the vacancies required English, while 14% required German. In Czechia, the 10% share of vacancies for German was confirmed, but the share for English was reported to vary between 28% and 59% depending on whether the vacancies were supplied by public employment agencies or employers. The Hungarian portal reported that 55% of the vacancies demanded English or German, 10% specifically requested English and 2% specifically called for German. We could not verify whether or not German was meant to substitute English. We unfortunately did not receive any further information from the Polish job board. Overall, these numbers are largely in line with our initial results.

Let us now focus our attention on the demand in ***individual occupations***, for which results are presented in Table 11 and Table 12. Table 11 shows the share of vacancies for each country that carry the tag for English language skills (ISCO codes 1-3). Table 12 depicts similar results, but for the low- and medium-skilled occupations (ISCO codes 4-9). In both tables, the five occupations in each country with the highest percentages are marked with a shaded background (i.e. we use country share rather than regional shares).

From the tables, we derive that it is relatively high in occupations where ISCO codes start with 2 or 4, the professional and administrative occupations. Meanwhile, we detect low demand for positions involving manual work. Another interesting result, particularly in Poland, is that some of the ISCO 7 occupations, denoting craftsmen, show a relatively high demand for English-language skills. This may indicate that these workers are employed by foreign employers or that they work abroad.[16]

[16] With regard to the latter option, we explored the share of vacancies that refer to positions abroad for the Slovak job portal. Overall, less than 10% of the positions advertised were positions outside the country (with the vast majority in Czechia). Vacancies were advertised for positions in Germany and Austria in the case of only a few occupations.

Table 11: Percentage of vacancies for high-skilled occupations that require English-language skills in each of the V4 countries. The five occupations with the highest shares in each country are indicated in grey.

ISCO	Occupation	Czechia	Hungary	Poland	Slovakia
1135	HR Manager	87%	42%	67%	67%
1211	Financial Manager	63%	54%	62%	50%
1213	Operations Manager	53%	48%	65%	59%
1221	Marketing Manager	71%	95%	79%	86%
1221	Sales Director	53%	45%	49%	51%
1321	Production Director	69%	50%	55%	74%
2139	Consultant	44%	64%	49%	66%
2141	Industrial Engineer	86%	81%	85%	83%
2142	Civil Engineer	23%	24%	22%	34%
2149	Tester	88%	87%	76%	89%
2164	Architect	66%	50%	64%	79%
2211	Doctor	31%	56%	44%	26%
2221	Nurse	27%	30%	64%	21%
2330	Teacher	51%	15%	55%	27%
2411	Tax Advisor	81%	50%	86%	98%
2411	Financial Controller	83%	78%	72%	93%
2412	Financial Analyst	81%	78%	72%	88%
2431	Marketing Professional	68%	85%	55%	71%
2431	Key Account Manager	63%	90%	45%	71%
2431	Business Analyst	74%	97%	81%	70%
2433	Product Manager	83%	94%	74%	74%
2511	IT Analyst	81%	87%	74%	88%
2512	Programmer	82%	84%	77%	85%
2519	Project Manager	81%	68%	75%	84%
2519	Computer Specialist	82%	83%	74%	80%
2611	Lawyer	72%	42%	72%	90%
3112	Technician	56%	35%	34%	46%
3119	Quality Control	72%	61%	57%	50%
3122	Foreman	33%	15%	34%	54%
3323	Buyer	84%	86%	68%	81%
3341	Office Manager	76%	91%	58%	78%
3341	Team Leader	73%	83%	42%	76%
3511	IT Administrator	81%	67%	72%	87%
3512	Customer Support Worker	67%	66%	58%	69%
3538	Account Manager	53%	93%	48%	67%

Table 12: Percentage of vacancies for low- and medium-skilled occupations that require English-language skills in each of the V4 countries. The five occupations with the highest shares in each country are indicated in grey. ISCO

	Occupation	Czechia	Hungary	Poland	Slovakia
4226	Receptionist	87%	86%	67%	87%
4232	Transport Clerk	70%	69%	57%	65%
4311	Accountant	74%	51%	73%	79%
4419	Office Worker	61%	52%	61%	70%
4419	Assistant	73%	70%	71%	68%
5120	Cook	26%	10%	27%	12%
5222	Sales Team Leader	37%	36%	28%	46%
5223	Retail assistant	25%	15%	29%	15%
5230	Cashier	36%	11%	10%	15%
5242	Regional Sales Representative	48%	27%	35%	68%
5242	Sales Representative	41%	36%	29%	38%
5242	Merchandiser	39%	33%	15%	13%
5242	Financial advisor	10%	59%	18%	10%
5244	Telesales/Call Centre	29%	16%	23%	50%
7212	Welder	3%	6%	39%	10%
7231	Mechanic	20%	10%	32%	16%
7233	Locksmith	4%	7%	44%	8%
7411	Electrical Mechanic	44%	3%	26%	32%
7411	Electrician	23%	6%	42%	16%
8121	Labourer	5%	8%	12%	5%
8322	Driver	21%	3%	19%	18%
8344	Forklift Operator	9%	5%	7%	4%
9321	Packer/Auxiliary Labourer	4%	14%	11%	11%
9333	Warehouse Worker	11%	7%	10%	13%

To further show the diversity of language, we split the occupations into three groups, where the first one contains occupations that most often require English-language skill, while the third group contains occupations where this requirement is rare (see Figure 7). More specifically, we first gathered all available vacancies for each occupation (thus bringing together the vacancies available in all four countries) and calculated the share of vacancies that required English-language skills. This gives us a single number for each occupation and allows us to rank occupations according to their demand for English. We rank the occupations from high to low, identifying two cut-off points (on the basis of percentiles). This enables us to distinguish between three groups, which we have labelled ‘highest’, ‘intermediate’ and ‘lowest’. Note that

within each group, occupations are still ranked on the basis of their share, meaning that the occupations towards the top of each bar report higher shares than the occupations towards the bottom.

Once again, we show that the demand for English-language skills is the highest in managerial, professional, technical, administrative and IT-related occupations. On the lower side of the distribution, we mainly find manual labour jobs (Figure 7).

Figure 7: Classification of occupations into three classes depending on the demand for English-language skills across the V4

We also calculated requirements for the German language, using the same strategy as for English-language skills. In this case, the results are noisier because the variance is much smaller, as no occupation in either of the analysed countries was found to contain at least 50% of vacancies requiring German. Nonetheless, we have discovered some interesting results. Firstly, the demand for English is higher than the demand for German across countries and occupations. More advertisements request German than English in only 14 of the 236 occupation-country combinations. These are all artisan occupations, particularly welders and locksmiths. The only exceptions are Slovak nurses, where the demand for German is 13% higher than the demand for English. This largely reflects the high level of circular migration of Slovak nurses to German-speaking countries, in particular Austria (Bahna, 2014). In general, the demand for German does not seem so dependent on occupation complexity as it is in the case of demand for English. We find several ‘manual’ occupations in the group with the highest demand for language skills (Figure 8).

Figure 8: Classification of occupations into three classes depending on the demand for German language skills across the V4

For three of the Visegrád countries, Czechia, Hungary and Slovakia, we were able to find data on the hourly wages (expressed in the national currency) for some occupations. This wage data is based on the median hourly gross wages, obtained from the WageIndicator web surveys collected from 2013 to 2016. Our data covers 29 occupations for Czechia, 25 occupations for Hungary, and 27 occupations for Slovakia[17]. There is a clear positive correlation between the

[17] Only occupations with at least 30 observations were taken into consideration.

shares of vacancies that require English and wages in all three countries (Figure 9). It is the strongest in Czechia, where each additional 10% demand for English is connected with an additional 50 cents of hourly income, followed by Slovakia, where the value is about 35 cents and Hungary, where more linguistically-demanding occupations result in a premium of only 20 cents per additional 10% for vacancies demanding English. The correlation between the German-language demand and wages is much weaker in Czechia and even negative for Slovakia. Interestingly, it is very strong for Hungary, resulting in a premium of €1.15 for an additional 10% of vacancies demanding German (Figure 10). Following up on our previous line of thought that much of the demand for German speakers might be for jobs located in German-speaking countries, the lack of monetary premium associated with speaking German might be unobservable on the Slovak labour market because German speakers are collecting it abroad.

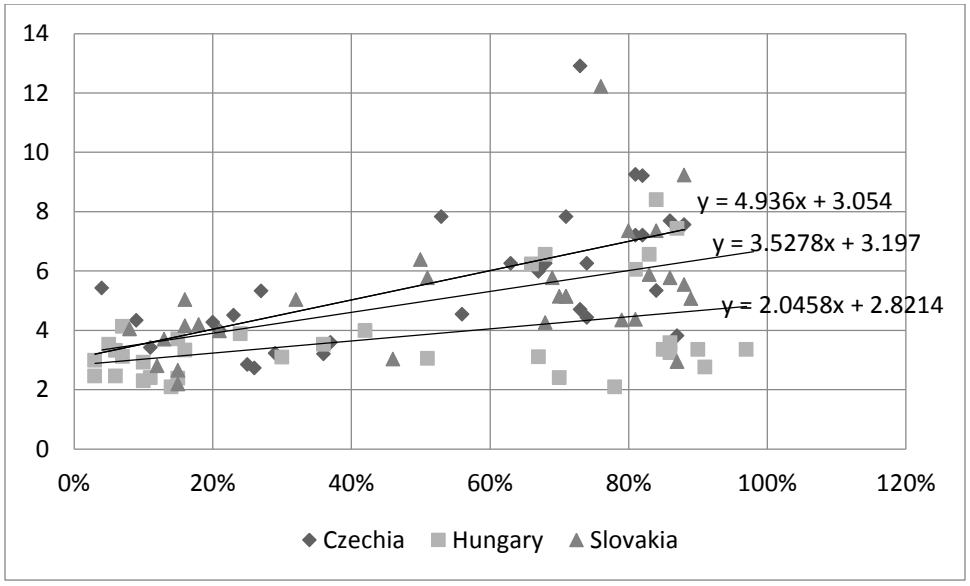

Figure 9**:** Correlation between the share of job advertisements that require English and the hourly log wages in Czechia, Hungary and Slovakia

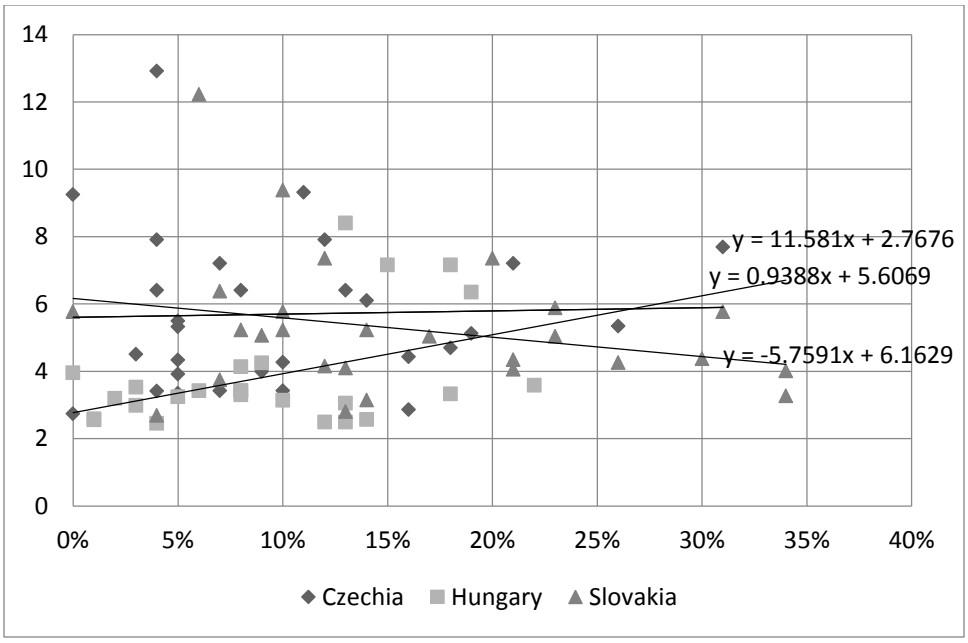

Figure 10: Correlation between the share of job advertisements that require German and the hourly log wages in Czechia, Hungary and Slovakia

Finally, with the question about English proficiency recently added to the WI survey, we examine the relationship between English proficiencyand wages on individual levels as well. The analysis is conducted on data collected between September 2015, when the language question has been added to the survey[18] and the end of 2016. An important caveat of using WageIndicator data is the potential measurement error in survey-based occupation data. This is a well-known issue that affects surveys in general. Bound et al. (2001) explain that survey responses are typically not perfectly reliable and that even relatively common variables may not be reported without error. For the 'occupation' variable, for example, response errors as well as coding errors occur quite frequently, as occupational classifications can be fairly detailed (e.g. 4-digit level) or may differ across institutes. As the WageIndicator survey is widely used in several disciplines, the issue of measurement error has also been assessed (de Pedraza et al. 2010; Guzi and de Pedraza 2015). Although it is clear that the WageIndicator survey may not be fully representative, it has been found to produce estimates of labour market outcome determinants

[18] December 2015 for Hungary.

that are consistent with representative data sources.

The results of our estimation exercise are presented in Table 13. We consider three estimation models. In the simplest model 1, there is no relation between having a basic command of English and earnings (when compared to no knowledge at all, which is the reference category). However, being rather skilled already appears to result in a 15% earnings premium, while being highly skilled increases the salary by 37%. This model relates wages to experience and experience squared, gender, education, and English proficiency and further includes country and year dummies.

When controlling for the occupation dummies in model 2, by adding a full set of ISCO dummies to the variables considered in model 1, we see that the coefficient remains statistically significant, yet the size of the premium decreases to 13% and 31% respectively. Finally, when more detailed occupation variables, such as firm size, sector, supervisory position and job being located in the capital city are taken into account, the coefficient decreases to 6% for rather skilled English speakers and 20% to fluent English speakers. Model 1 explains 18.5% of variation in wages, model 2 22% and model 3 already 30%, which already indicates that our results are somewhat robust, especially considering that the coefficients of control variables are in line with expectations.

Table 13: OLS analysis of the relationship between English proficiency and wages in Czechia, Slovakia and Hungary

	(1)	**(2)**	**(3)**
Years of Experience	0.0295***	0.0252***	0.0212***
	(0.00419)	(0.00416)	(0.00404)
Years of Experience Squared	-0.000582***	-0.000511***	-0.000397***
	(0.000110)	(0.000109)	(0.000106)
Education	0.0444***	0.0330***	0.0273***
	(0.00491)	(0.00508)	(0.00492)
Women	-0.181ource***	-0.191***	-0.165***
	(0.0238)	(0.0249)	(0.0241)

English skill level (reference category: no English)			
- Basic	0.0313	0.0214	-0.0186
	(0.0337)	(0.0332)	(0.0320)
- Rather skilled	0.154***	0.128***	0.0618*
	(0.0377)	(0.0374)	(0.0360)
- Skilled	0.364***	0.306***	0.197***
	(0.0417)	(0.0418)	(0.0409)
Country and year dummies	YES	YES	YES
Basic occupation dummies	NO	YES	YES
Extended occupation dummies	NO	NO	YES
Observations	1,988	1,988	1,947
R-squared	0.185	0.217	0.300

* p < 0.1; ** p < 0.05; *** p < 0.01

Given the limitations of the data source used, the significance of the analysis should not be overstated. Nonetheless, the signs of the estimated coefficients are in line with our initial expectations (and that coefficients are found to be statistically significant across the three models), we can see our analysis as an additional piece of support suggesting that English language skills are important to employers in the countries examined.

Conclusion and Policy Implications

In this chapter, we used metadata associated with approximately 74,000 job advertisements published on leading online job boards in the Visegrád countries to analyse demand for foreign language skills. We found that demand is widespread, but limited to two languages: English and, to a lesser extent, German. The exact degree of demand varies between individual countries and occupations, although we do observe some common trends.

Most importantly, our analysis suggests that English is a language of professionals and white collar workers. For these professions, it is demanded almost everywhere. For working-class professions, English language skills seem less important. These results, however, necessarily apply to the German language. We report a high demand for German in some professional occupations, but also in artisan and nursing occupations. Additionally, both individual-level and occupational-level analyses show that English-language skill comes with a wage premium. We did not find a similar correlation for German.

While we found that learning German can be beneficial for some occupations (in addition to benefits unrelated to the labour market), English is by far the foreign language most in demand. Furthermore, we found evidence that English proficiency is associated with wage premium, suggesting that employers are willing to pay wage premium for English-language proficiency.

We believe that our research provides some tentative evidence for policy-makers in the V4 region to invest in English teaching. Proficiency in English in the V4 countries currently ranges from 20% in Hungary to 33% in Poland and is thus well below the EU average of 38%. Meanwhile, the knowledge of German varied between 15% in Czechia and 22% in Slovakia and was thus above the EU average of 11%. In this regard, it seems that governments are becoming aware of the importance of foreign language skills. Eurostat data on the number of foreign languages that students learn in school reveals that well over 90% of them learn at least one foreign language in upper secondary education (and in general secondary education students seem to even learn two or more foreign languages) in all four Visegrád countries. In Czechia, Poland and Slovakia, over 70% of primary school students learn at least one foreign language.

Nonetheless, it is important to reiterate the caveats of our work. Firstly, it is important to acknowledge that job vacancies are an incomplete measure of labour demand, for several reasons (e.g. not all job openings are advertised as recruitment; they may occur through internal or informal channels). Secondly, online job boards have additional limitations: not all vacancies may be advertised online and those that are published online may be biased towards specific regions, industries or applicants (Carnevale et al. 2014). Our results and conclusions, therefore, have to be weighed against these limitations. In other words, we are aware that the structure of the sample may have been biased by its limited focus on online job search and vacancies, so that

an extension of our findings the ‘offline’ labour market may not be clear-cut. At the same time, the use of online vacancy data has been advocated by Kuhn and Skuterud (2004) and Askitas and Zimmermann (2009, 2015), who have pointed to its potential as a real-time data source to capture ongoing trends. For an in-depth discussion of the potential and caveats of web data, we also refer to Benfield and Szlemko (2006), Wright (2012), Carnevale et al. (2014) and Beblavý et al. (2016b).

Our chapter represents a first effort to exploit innovative data sources to better understand this important topic. While our results are both internally consistent and in line with the literature, we believe that additional reinforcement checks are needed. One issue of our research is that it is based only on a cross-section of crawled data from job portals. Efforts such as the one currently spearheaded by the European Centre for the Development of Vocational Training (Cedefop) to establish a system for continuous data collection across the EU28 are, therefore, crucial. Secondly, the newly-added language questions in the WageIndicator survey will eventually result in a valuable source of data on labour market effect of language skills among job holders across the EU. Future research that strengthens the methodological framework in which a sample of vacancies published online is used as an input for labour-related analysis is needed.

Chapter 5: Analysing Skill Demand: Measurements of Skills Intensity of Occupations[19]

Introduction

Skill matching in the labour market is an important conceptual and policy issue. But to better measure skill mismatch and provide for improved labour market matching, we need to understand the interconnections between tasks, skills and occupations. However, mapping these connections is costly when conducted by occupations experts, as it requires considerable expertise and time. Some attempts have been made to measure this mapping using new data collection techniques using data originating from the Internet (Fabo and Tijdens 2014; Visintin et al. 2015a; Beblavý et al. 2016c). Each of these attempts was based on a single web-based data source, mapping either the supply side or the demand side. A pioneering and thus far, to the best of our knowledge, single attempt to compare the supply and demand side using web data looked at education requirements only (Tijdens et al. 2015a). In this chapter, we look specifically at the highly policy relevant aspect of computer skills supply and demand across occupation using two distinct data sources – the WageIndicator web survey and job vacancies posted online.

Our research is potentially very salient policy wise, due to its potential to help tackle skill mismatch. According to the European Skills and Jobs (ESJ) survey, 25% of highly-qualified employees were overqualified for the job in 2014. While an applicant may prefer to take a job requiring a lower level of qualification than they hold, the same survey also found that 27% of employees are stuck in so-called 'dead end' jobs: positions that do not allow workers to develop their skills and improve their productivity (CEDEFOP 2015). Furthermore, a large proportion of Europeans, particularly young Europeans, struggle to even enter the labour market, according to Eurostat.

Economists have long considered the role of skills to be central to the understanding of the matching between employees and employers in the labour market, but also from a policy

[19] This work has been submitted to the International Journal of Manpower as a paper co-authored with Martin Kahanec.

perspective. Of particular importance is the human capital approach, which has become widely accepted in economics (Becker 1962; Schultz 1971; Benhabib and Spiegel 1994). According to this approach, more skilled workers are more productive and thus more valuable to employers. Unskilled workers, on the other hand, are not as productive and thus are offered only lower wages or less favourable working conditions. They may not be hired at all, if, for example, a minimum wage policy makes their employment economically unsound.

But how can a situation where workers underuse their skills on a large scale co-exist with skill shortages perceived by employers? Empirical research shows that the likely reason is a shortage of specific skills especially required by employers (CEDEFOP 2014). Being able to identify the demand for skills is thus crucial for informing policy makers in particular in areas such as education and training. Beblavý, Fabo and Lenearts (2016) study IT skill demand based on a large sample of job vacancies posted online throughout 2013 for 30 common occupations in the USA. They find that (i) demand for computer skills is high across occupations and growing with increasing complexity of an occupation, (ii) while there are many different computer skills, only a relatively small number of them is relevant for workers outside of the IT industry itself, (iii) computer skills determined on the basis of job vacancies are highly in line with the computer skills inferred from tasks defined for the individual occupations in occupational classification systems O*NET and ISCO.

These findings potentially open doors for using web data to better understand the mapping between skills, tasks and occupations. They also inform policy on a wide range of areas connected to skill acquisition and use. In particular, web-based data collection techniques may help us gauge information about the usefulness of skills across occupations and thus provide enhanced policies aimed at improved skill matching in the labour market. However, the scope and usefulness of web-based data for classification of jobs and tasks pertaining thereto has yet to be tested. In this chapter, we extend Beblavý, Fabo and Lenearts (2016) in this direction by (i) benchmarking based on vacancy data, benchmarking web-based survey data about the association of tasks to occupations with respect to expert mapping of tasks to occupations and by (ii) looking at a wider scope of occupations at all skill levels. We aim to show to which degree these different web-based data sources can represent the demand for IT skills on the labour market. Our main finding is that the web surveys are potentially powerful tool to analyse

computer skills demand, while online job vacancies need to be interpreted carefully, due to requirements not being necessarily explicitely listed.

This chapter is structured as follows: After a literature review section describing the state of the art with regards to analysing occupations using web data, we describe data sources used in the analysis in detail, including a debate on the representativeness of these surveys. Building up on this, we follow up with a methodology section and results. The chapter concludes with a discussion.

Literature review

Studying labour market matching is a complicated matter, because the universe of jobs and tasks that workers do, as well as their skills, is large, complex and dynamic (Fabo and Tijdens 2014; Visintin et al. 2015a; Beblavý et al. 2016a). In this section, we focus on the current state of the art with regards to ways on how to systematically study the phenomenon. Jobs may be aggregated into smaller clusters by abstracting from the context of work and focusing purely on the tasks performed by the workers in those jobs. These groups, called occupations, can be defined as follows: 'An occupation is a bundle of job titles, clustered in such a way that survey respondents in a valid way will recognise it as their job title; an occupation identifies a set of tasks distinct from another occupation; an occupation should have at least a non-negligible number of jobholders and it should not have an extremely large share in the labour force' (Tijdens 2010).

This aggregation of jobs is done both for practical reasons – to better organise labour (Damarin 2006) and for research purposes – to understand skill demands on the labour market (Levenson and Zoghi 2010). Employers and employees use occupations to characterise and refer to jobs. The skill dimension is inseparable from tasks, because a skill is 'worker's endowment of capabilities for performing various tasks' (Acemoglu and Autor 2011). Consequently, a salient aggregation of jobs into occupations is necessary for our understanding of the matching between employers and employees in the labour market.

Skills can be generally split into two categories: *job-specific* and *transferable*. For example, the O*Net classification of occupations, used widely in the US considers the 'content' and 'process'

skills to be job specific and 'social skills', 'technical skills', 'complex problem-solving skills', 'systems skills' and 'resource management skills' to be transferable. Computer skills may be placed into both categories (BGT 2015) as they may act as the 'content' of the job particularly in IT occupations, but they may also be regarded as a technical skill increasing productivity in non-IT workers. Recently, empirical evidence has underlined the importance of the transferable dimension of computer skills across the entire labour market, including some of what had traditionally been considered low-skill occupations (Beblavý et al. 2016e, d). Scholars along with the policy makers started recognising exclusion from the labour market faced by workers lacking computer skills (CEDEFOP 2015; Smith 2015; Horrigan 2016; Bührer and Hagist 2017). Consequently, it appears pertinent to analyse the role of computer skills in non-IT occupations as well.

There are several approaches for determining the demand and supply for IT skills on a labour market:

The most obvious approach is connected with the occupations themselves. Because occupations are defined through tasks and tasks imply skills, it is possible to infer skills demand from the tasks associated with the individual occupations. Individual classifications of occupations, such as the International Standard Classification of Occupations (ISCO) or the American O*NET classification, define each occupation on the basis of a list of tasks assigned to it by labour market experts (Elias 1997). The downside is that such systems are only updated infrequently. For example, the last update of ISCO took place in 2008, which was preceded by the 1988 update; overall there hsve only been four updates of the classification since the first version in 1958. Meanwhile, the nature of the work is quickly changing due to numerous factors, including technological progress, outsourcing/offshoring and change in labour organisation (Acemoglu and Autor 2011; Cowen 2013; Beblavý et al. 2016a)

Fortunately, it is not necessary to always rely on experts. Skill and task questions have been increasingly included in surveys such as the OECD Adult Skills survey PIAAC and the web-based WageIndicator survey (Fabo and Tijdens 2014). Importantly, the evaluation of job requirements by the job holder was found to be largely consistent with the expert estimates, even in the case of non-representative web surveys (Tijdens et al. 2013a, see the data section for our own robustness check). Such web survey-based estimates of computer skill applicability can,

therefore, be used for economic analysis (Mýtna-Kureková et al. 2015; Visintin et al. 2015a; Tijdens and Visintin 2016). In particular the large scale, continuous web-based surveys provide up to-date-information and are usable for reliable exploratory analysis, especially if the results can be benchmarked against a representative data source on a regular basis (Steinmetz et al. 2014; Tijdens and Steinmetz 2016).

Additionally, it is possible to learn from employers. That is, because when hiring, employers tend to think of the list of tasks the worker needs to be able to perform (Autor 2001; Winterton 2009). Furthermore, the Internet is increasingly becoming the place where job vacancies are advertised and thus an important source of data about skill demand (Askitas and Zimmermann 2015; Mýtna-Kureková et al. 2015; Lenaerts et al. 2016; Beblavý et al. 2016e, d). Just like in the case of web surveys, there is some controversy associated with the use of these new data sources for social science research. Nonetheless, in spite of significant concerns pertaining to mainly the representativeness and the potential for generalisation of results created on the basis of this data source, there is a growing consensus among scholars that web-based data is going to be an increasingly important source for the research of the labour market (Askitas and Zimmermann 2015).

These two approaches to mapping tasks to occupations – web survey-based and vacancy-based – are evaluated to test these methodologies in terms of their precision as benchmarked to expert-based mapping. This sheds more light on possible approaches that can help to address the need for a more precise and up-to-date efficient mapping of tasks to occupations in a rapidly changing and developing labour market.

Data

This chapter analyses skill requirements inferred from three data sources: (i) the tasks assigned to the individual occupations by labour market experts in the ISCO2008 classification of occupations, (ii) the self-reported use of computer skills as measured by the WageIndicator survey and (iii) the share of an explicit IT requirement in vacancies posted on the Internet. We will focus on the Netherlands because of the availability of data, as explained later in this section.

We obtained survey-based data from the WageIndicator survey, which is a continuous, large-scale, voluntary web survey covering more than 90 countries on all continents. Nonetheless, the quality of the data varies among countries. 18% of all data intake is collected in the Netherlands, where the survey originally started and where the local website hosting the survey gets approximately five million unique visitors every year. The high number of respondents to the survey in the Netherlands and the fact that Internet use is very widespread in the country result in a sample which is quite similar to a general population sample and suitable for exploratory analysis (see Chapter 2).

In the WI survey, we added a module asking employees about their use of computers at work. More specifically, we asked 'When do you use a computer or tablet?' with possible answers being (i) Only during working hours (ii) Both during working hours and free time (iii) Only during free time (iv) Never. We recorded these questions such that if a respondent selected either option (i) or (ii they would be considered as someone using a computer at work, otherwise we would consider them a non-user.

We launched the module on 18 August 2016 and collected the data until the end of the year. The respondents were not obliged to answer all the questions in the questionnaire. We only used occupations for which at least ten respondents answered the IT module. Within less than five months, we managed to cover 62 occupations (out of a total 436 occupations defined in ISCO2008), with a total of 1644 responses, covering all aggregated occupations groups (one digit ISCO), except military occupations and skilled agriculture workers. As a result, our analysis is limited to common occupations and might not be representative of occupations with few holders, which is a widespread general problem in occupation research (Tijdens et al. 2013a).

To get a sense of the representativeness of our data, we compare it against a representative data source, the Dutch Survey of Adult Skills (PIAAC), which is organised by OECD. One limitation is that the PIAAC data was collected in the years 2011 and 2012, so there is a 5-year asynchrony. A further limitation is that when we applied the same criterion as with the WageIndicator data and eliminated all occupations with less than 10 observations, we could match only 49 out of the 62 occupations for which we had data in both datasets. Both datasets exhibit a largely similar picture for nearly all office jobs holders who show a very high intensity of computer use at work,

while the percentage of those using computers is lower among the holders of manual jobs. The correlation between the two indicators is highly significant and very strong (r=0.922, also see Figure 11). Statistically, we can also compare the two sets of estimates using a paired t-test for two sets of estimates. This test is statistically not significant with a t value of t = -0.8719 (p = 0.806). Where we see differences, these typically reflect a large growth of use of computer at work by some manual occupation holders, such as electrical mechanics, truck and taxi drivers, cooks or carpenters.

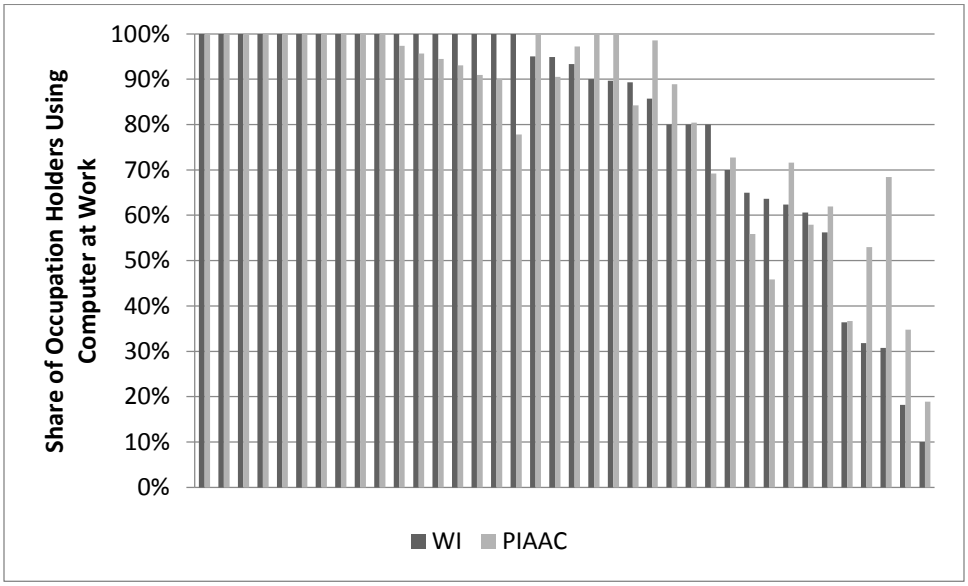

Figure 11: Comparison of self-reported computer use per occupation between WageIndicator and PIAAC datasets

Source: Own calculation using the WageIndicator and OECD PIAAC dataset

Vacancy-based data on the share of computer skills in occupation were obtained from job vacancies posted online between August and December 2016. The dataset was provided by the company Textkernel, which is the market leader in the collection, processing and analysing job vacancy data in western Europe using a large number of advanced algorithms to get as representative sample of job vacancies as possible, with by far the most complete dataset being collected in the Netherlands (Zavrel 2016). The sample covers 60 of the analysed occupations and is based on nearly 300,000 unique job vacancies posted online. In line with our previous analysis (Beblavý et al. 2016b), we calculate the share of vacancies containing at least one keyword associated with computer use within occupations defined at the most detailed (4-digit)

level of ISCO occupations classification[20].

The Dutch labour market represents an ideal environment to explore the use of online data because it contains very high quality online survey as well as a near-exhaustive database of online job vacancies. Furthermore, the Netherlands has a large high-skilled workforce and a high degree of computerisation, which makes the market for computer skills broad and deep enough to get sufficient number of observations for a large-enough numbers of occupations. Finally, the Dutch labour market is relatively homogenous with limited degree of regional variations, making it particularly suitable for an analysis of occupations within the country.

Analytical Strategy

The aim of our analysis is to evaluate what the two online data sources can tell us about the intensity of computer skills use across occupations in the Dutch labour market. A key question is how we can decide which of the mappings is more accurate. After al two things are being measured – demand by employers and self-reported use of skills by workers. Our strategy rests on the ISCO occupational classification itself and on the information it contains about skill use intensity across individual occupations, there are two distinct pieces of information to be identified:

Firstly, we looked at the relevance of computer skills for the tasks associated with individual occupations in the ISCO classification. We coded a total of 466 individual tasks such that each one was either classified as clearly requiring a computer (for instance 'Developing and implementing software and information system testing policies, procedures and scripts') or not necessarily requiring computer skills, but having use for them (such as 'Designing and modifying curricula and preparing courses of study in accordance with requirements' , which can

[20] After having tried many of various IT skills from different sources while working with the US data (Beblavý et al. 2016c), we found the regular expression ((MS|Microsoft) Excel)|((MS|Microsoft) excel)|((MS|Microsoft) Word)|((MS|Microsoft) word)|((MS|Microsoft) Office)|((MS|Microsoft) office)|(PC)|(Computer) the best to estimate the IT skills requirement. Luckily, the Dutch word for computer is the same as the English one, so the regular expression was also usable with Dutch vacancies.

still be performed using pen or paper, but is likely done using a text processing software by most educators) and those that have no use for computer skills at all (such as 'Maintaining discipline and good working habits in the classroom'). For coding, we used two typical uses of computer applications we identified in previous job vacancy analysis (Beblavý et al. 2016c) - general computer use, including using job-specific software and office application such as spreadsheet and text processor.

Secondly, we used the information about occupation complexity. The ISCO classification also associates occupations with the concept of task complexity involved by sorting each civilian occupation into one of nine aggregated occupation groups plus an additional group for military occupations, which are not a subject of analysis in this chapter. Eight out of those groups can be connected to degrees of task complexity associated with them. The most complex, professional, occupations belong to Group 2. Groups 3-8 contain occupations with intermediate degrees of complexity, with groups 3 and 4 containing office jobs and groups 5-8 containing manual labour. Finally, Group 9 contains elementary occupations with low complexity of associated tasks. Group 1 is more heterogeneous than the other ones, because it contains managerial occupations which are generally associated with an intermediate-to-high degree of complexity (Hunter 2009; Mýtna-Kureková et al. 2013), which makes it impossible to exclusively assign this group to a specific complexity level.

In the analysis, we used these occupation traits determined by labour market experts when constructing the ISCO classification as a guide. More precisely, we defined three distinct groups of occupations: Those that require computer skills to perform at least one of the tasks associated with them, those that do not have any apparent use of the computer skills and finally those that do not necessarily require the computer skills, but entail tasks, which might be performed more efficiently using a computer. Similarly, we looked at the occupation complexity to see the demand and use of computer skills across manual and office/service occupations with various levels of complexity.

We then compared the (i) measured demand for computer skills determined on the basis of actual job vacancies, (ii) the use of computers reported by occupation holders and (iii) the expert-based mapping of skills to occupations based on the ISCO classification. Based on these three sources,

we established a multi-dimensional picture of the role of the IT skills in the contemporary Dutch labour market and shed light on their use and measurement options.

Results

Having described our analytical tools and data, we can now present findings. This section is organised on the basis of the ISCO classification, first looking at occupations on the basis of the applicability of computer skills inferred from tasks associated with them.

Through task coding, we identified 17 occupations, a mixture of professional and administrative occupations, which require the use of a computer for at least one task associated with them. As can be seen in Table 14, all or nearly all holders of these occupations use computers at work. Nonetheless, when looking at the vacancies, we see these high-skill occupations (including several in the IT, such as computer network professionals, which cannot be performed without computer skills) do not explicitly ask for the computer skills in the majority of the cases. This likely follows from the fact that not all skill requirements are explicitly specified because they are sometimes taken for granted. That would explain why the demand for computer skills surges for administrative occupations, such as secretaries, where those skills are not necessarily taken for granted. Nonetheless, if we accept this explanation, it strengthens the case for skepticism towards the use of job vacancies to determine the applicability of specific skills, as the tendency of employers to leave out some skill requirements they take for granted is clearly a major concern (Mýtna Kureková et al. 2016).

Table 14: Web-based measurement of applicability of computer skills for occupations requiring computer skills.

Occupation	Vacancies	WI
Secretaries (general)	25%	94%
General office clerks	17%	99%
ICT professionals	16%	100%
Accounting and bookkeeping clerks	16%	100%
Personnel clerks	15%	96%
Contact centre information clerks	14%	100%

Systems analysts	10%	100%
Advertising and marketing professionals	9%	100%
Computer network professionals	9%	100%
Industrial and production engineers	8%	100%
Accountants	8%	100%
Applications programmers	8%	100%
Software and applications developers	8%	100%
Statistical, finance and insurance clerks	8%	98%
Accounting associate professionals	7%	100%
Draughtspersons	6%	100%
Graphic and multimedia designers	4%	96%

Source: Own calculation based on the WageIndicator dataset and Textkernel

Shifting our attention to thirteen manual occupations, both skilled and unskilled, which have no apparent use for computer skills according to tasks associated with them, many of the workers (but certainly not all) in these occupations use computers (

Table 15). Of particular interest are Healthcare assistants, out of whom nearly 90% use computers, as well as Stationary plant and machine operators, Motor vehicle mechanics and repairers, which are all above 70%. This suggests a major shift in the nature of those occupations as they now require skills not foreseen when constructing the 2008 update of the ISCO classification. The appearance of new tasks in existing occupations has been observed in labour studies literature and is a common manifestation of a shifting skill demand (Barley and Tolbert 1991; Crosby 2002; Beblavý et al. 2016a). Turning to job vacancies, we see that demand for computer skills in these occupations is in general lower than for occupations that entail tasks associated with computer skills.

Meanwhile, the information we can learn from job vacancies is much more limited and does not match the story inferred from the WI survey. The explicit requirement for computer skills in vacancies recruiting for these jobs is negligible and does not allow us to identify the uptick in demand for computer skills in some occupations, which is evident from the WageIndicator data.

Table 15: Web-based measurement of applicability of computer skills for occupations with no apparent use for computer skills.

Occupation	Vacancies	WI
Stationary plant and machine operators	*5%*	76%
Health care assistants	*3%*	89%
Electrical mechanics and fitters	*3%*	56%
Freight handlers	*3%*	65%
Motor vehicle mechanics and repairers	*2%*	80%
Fitness and recreation instructors	*2%*	73%
Child care workers	1%	31%
Cooks	1%	61%
Waiters	*1%*	32%
Car, taxi and van drivers	1%	64%
Domestic cleaners and helpers	1%	10%
Carpenters and joiners	*0%*	18%
Heavy truck and lorry drivers	0%	36%

Source: Own calculation based on the WageIndicator dataset and Textkernel

This leaves us with a group of 32 occupations, containing professional and administrative occupations, service and sales occupations and even some skilled manual occupations. Occupations in this group do not require computer skills but entail tasks which can potentially be performed more efficiently (

Table 16). In most such occupations, the use of computers is universal or nearly universal. While there are still some occupations, such as Shop assistants, Primary school teachers or Building and related electricians, which still employ a substantial portion of workers who do not use computers at work, even in those occupations a vast majority of workers benefit from computers. Interestingly, the requirement for computer skills in posted vacancies is widely explicitly stated only for a few administrative positions. Consequently, on the basis of the vacancy data, we would not be able to see how widespread the use of computer skills is in these occupations.

Table 16: Web-based measurement of the applicability of computer skills for occupations with possible, but not necessary, use for computer skills.

Occupation	Vacancies	WI
Administrative and executive secretaries	27%	100%
Receptionists (general)	*24%*	90%
Production clerks	19%	100%
Buyers	14%	100%
Business services and admin managers	11%	90%
Human resource managers	10%	100%
Product graders and testers (except food)	10%	91%
Sales and marketing managers	9%	100%
Management and organisation analysts	9%	100%
Commercial sales representatives	9%	95%
Journalists	8%	100%
Stock clerks	8%	80%
University and higher education teachers	7%	100%
Managing directors and chief executives	6%	93%
Personnel and careers professionals	6%	100%
Environmental and occupational health inspectors	6%	100%
Legal professionals	5%	100%
Employment agents and contractors	5%	100%
Nursing professionals	4%	95%
Policy administration professionals	4%	100%
Construction supervisors	4%	100%
Shopkeepers	4%	100%
Agricultural and industrial machine mechanics	4%	80%
Vocational education teachers	3%	100%
Social work and counselling professional	3%	100%
Shop supervisors	3%	81%
Shop sales assistants	3%	62%
Building and related electricians	3%	70%

Physiotherapists	1%	100%
Primary school teachers	1%	86%
Clerical support workers	N/A	92%

Source: Own calculation based on the WageIndicator dataset and Textkernel

Finally, we consider the complexity of tasks (see the discussion in the empirical strategy section) associated with the individual occupations, rather than specifically the applicability of computer skills. The results based on the WI survey and vacancy data largely match the previous results as far as the division between office and manual occupations is concerned. Regardless of task complexity, office and service jobs require computer skills more often than manual jobs.

However, new insights emerge when looking at the details. Based on the WI data we see that nearly all holders of office/service jobs use computers, while only about 60% of skilled manual workers do. Low-skilled employment is a specific category with less than 40% of workers using computers at work. The detailed story told by vacancies is different from the ones inferred from the WI survey, or indeed can be expected on the basis of task complexity associated with the occupations. In all categories, less than 20% of vacancies demand computer skills. Most commonly, they are demanded in administrative occupations (16%). Paradoxically, the lowest explicit demand for computer skills among office and service occupations is in highly complex, professional occupations (7%). The only non-office occupational group with non-negligible demand for computer skills is crafts and trades workers (4%).

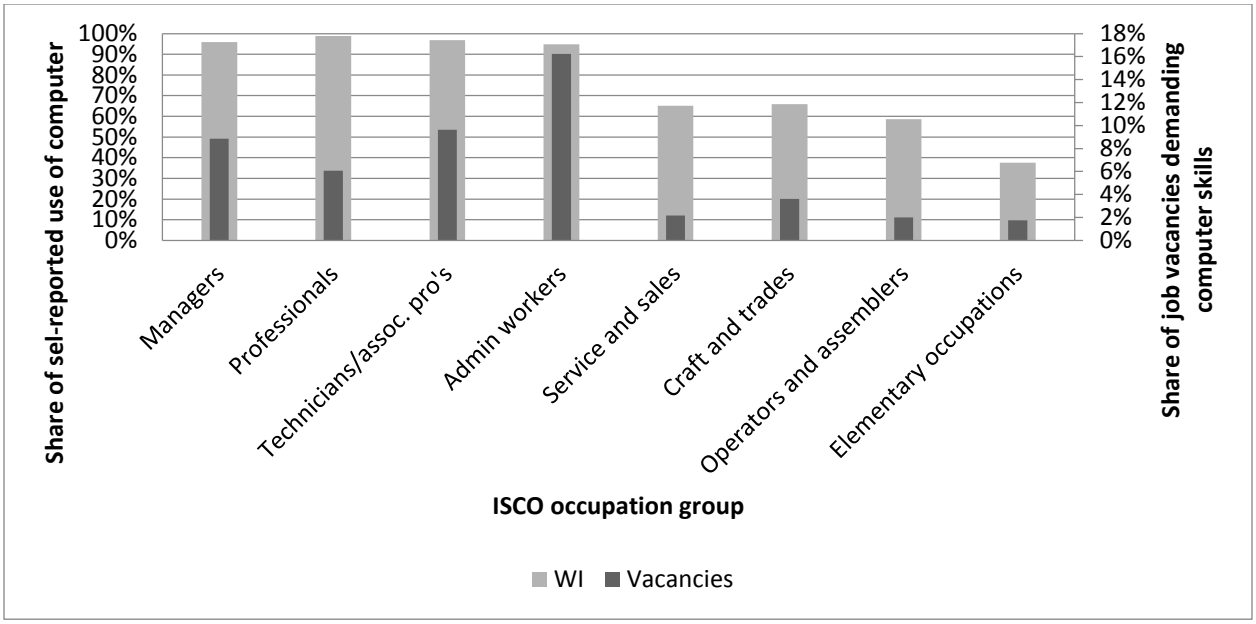

Figure 12: Average computer skills applicability per ISCO occupation group in the WI dataset and in job vacancies

Source: Own calculation using the WageIndicator and OECD PIAAC dataset

Notes: Inferred from self-reported rates in the WI survey (left axis) and vacancies containing IT skill requirement (right axis). Source: Own calculation based on the WageIndicator dataset and Textkernel data

Finally, we review our findings from the three data sources in Table17. What we see based on the comparison of tasks associated with individual occupations in the ISCO classification contrasted with WI data is that while computer skills are typically necessary in particular for high-skill, professional occupations and only rather useful for medium-skill office/service occupations, nearly all workers in this category use computers at work. Indeed, when looking at vacancy data, we often see that the explicit demand for computer skills is the highest among medium-skill non-manual occupations, in particular service and sales. It seems that for high-skilled occupations, computer skills are assumed, because they are taken for granted by employers. Among the manual occupations holders, we see it is often the case that occupations in which computers are rarely used are limited to low-skill labour.

Table 17: Overview of computer skills applicability based on different data sources

	Computer skills requirements inferred from tasks	Job vacancy demand for computer skills	Use of computers reported by occupation holders
High Skill Occupations	Typically necessary, in some cases useful	Medium	High
Medium Skill Office Occupations	Typically useful, in some cases necessary	Medium to high	High
Medium Skill Manual Occupations	Useful or not relevant	Low	Medium
Low-skill Occupations	Typically not relevant	Low	Low to medium

Own elaboration

Conclusion

Our analysis shows the potential of web data as a source of information for scholars and professionals in need of information about the relevance of specific skills for the labour market. In particular, we show that web surveys can be used to explore computer skill applicability across detailed occupations. This is an important finding, because traditional data sources containing skills information are rarely up to date, making use of web data a must when researching such a fast-changing environment as computer skills requirements. We also show, however, that there are limitations connected to using vacancy data. The main cause for concern is connected to the low incidence of explicit requirements for general computer skills even in some types of vacancies, where use of such skills is clearly required for fulfillment of tasks associated with the respective occupations.

One way forward with this would entail a wider application of categorical metadata (such as the foreign language skills indication discussed in Chapter 4), whereas employers will indicate their requirement for common skills by checking a box. Such a system would allow researchers to more easily understand potential barriers to matching between a job and a candidate. Nonetheless, it is not very realistic to expect such as system to be introduced for all skills in all countries due to increased workload on the side of the employers. Therefore, a more feasible alternative is an exploration of job portals for cases, where such a system is already introduced (likely because the skill is already an important bottleneck, such as foreign language skills in the Visegrad countries) and increased focus on such sources.

We show that computers have not only become nearly universal in almost any office job, but they are increasingly common in manual occupations as well. Jobs not entailing any use of a computer appear increasingly concentrated in the low complexity, elementary occupations category. This represents a challenge particularly for older skilled workers lacking the computer skills, who might be excluded from labour market if not provided with appropriate IT training.

Looking at job vacancy data, we see that while for professional jobs computer skills are not explicitly mentioned, which is likely due to expectation of employers that all candidates for some position are able to use computers, they are relatively often specified for administrative positions. Among skilled manual occupations, IT requirements appear to be often stressed for craft and trade workers. One way to interpret this finding is that in these occupations, there perhaps remains greater friction between the demand for computer skills and supply of qualified candidates with those skills, encouraging employers to stress the importance of computer proficiency as a requirement for employment.

Further research is needed to establish the link between the explicit listing of a skill requirement and expectation of it being instrumental for labour market matching. Nonetheless, there is a potential for interesting synergy between the two web-based data sources: An online survey can be deployed to gain a detailed account of the computerisation of work across occupations, while job vacancies can tell us where there is a particular risk of mismatch between job requirements and the skill of job candidates.

References

Abbott A (1995) Things Of Boundairies. Soc Res 62:857–882.

Acemoglu D, Autor D (2011) Chapter 12 - Skills, Tasks and Technologies: Implications for Employment and Earnings*. In: Ashenfelter DC and O (ed) Handbook of Labor Economics. Elsevier, pp 1043–1171

Achrekar H, Gandhe A, Lazarus R, et al (2011) Predicting Flu Trends using Twitter data. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). pp 702–707

Acquisti A, Fong CM (2015) An Experiment in Hiring Discrimination Via Online Social Networks. Social Science Research Network, Rochester, NY

Allen J, van der Velden R (2001) Educational mismatches versus skill mismatches: effects on wages, job satisfaction, and on-the-job search. Oxf Econ Pap 53:434–452. doi: 10.1093/oep/53.3.434

Altonji JG, Blank RM (1999) Race and gender in the labor market. In: Economics B-H of L (ed). Elsevier, pp 3143–3259

Antonietti R (2007) Opening the "Skill-Biased Technological Change" Black Box: A Look at the Microfoundations of the Technology-Skill Relationship. Econ Polit. doi: 10.1428/25821

Antonietti R, Loi M (2014) The demand for foreign languages in Italian manufacturing.

Antonovics K, Town R (2004) Are All the Good Men Married? Uncovering the Sources of the Marital Wage Premium. Am Econ Rev 94:317–321.

Armstrong A (2015) Equilibria and efficiency in bilingual labour markets. J Econ Behav Organ 112:204–220. doi: 10.1016/j.jebo.2015.01.011

Askitas N (2015) Calling the Greek Referendum on the nose with Google Trends.

Askitas N (2016) Trend-Spotting in the Housing Market.

Askitas N, Zimmermann KF (2009) Google econometrics and unemployment forecasting. Appl Econ Q. doi: 10.3790/aeq.55.2.107

Askitas N, Zimmermann KF (2015) The internet as a data source for advancement in social sciences. Int J Manpow 36:2–12. doi: 10.1108/IJM-02-2015-0029

Autor D, Dorn D (2009) This Job is "Getting Old:" Measuring Changes in Job Opportunities Using Occupational Age Structure. National Bureau of Economic Research

Autor DH (2001) Wiring the Labor Market. J Econ Perspect 15:25–40.

Autor DH (2015) Why Are There Still So Many Jobs? The History and Future of Workplace Automation. J Econ Perspect 29:3–30. doi: 10.1257/jep.29.3.3

Autor DH, Katz LF, Kearney MS (2006) The Polarization of the U.S. Labor Market. National Bureau of Economic Research

Autor DH, Levy F, Murnane RJ (2003) The Skill Content of Recent Technological Change: An Empirical Exploration. Q J Econ 118:1279–1333. doi: 10.1162/003355303322552801

Averett SL, Hotchkiss JL (1996) Discrimination in the Payment of Full-Time Wage Premiums. Ind Labor Relat Rev 49:287–301. doi: 10.1177/001979399604900207

Baldacci E (2016) From Data to Knowledge.

Baltar F, Brunet I (2012) Social research 2.0: virtual snowball sampling method using Facebook. Internet Res 22:57–74. doi: 10.1108/10662241211199960

Bandilla W, Bosnjak M, Altdorfer P (2003) Survey Administration Effects? A Comparison of Web-Based and Traditional Written Self-Administered Surveys Using the ISSP Environment Module. Soc Sci Comput Rev 21:235–243. doi: 10.1177/0894439303021002009

Bangwayo-Skeete PF, Skeete RW (2015) Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. Tour Manag 46:454–464. doi: 10.1016/j.tourman.2014.07.014

Barley SR, Tolbert PS (1991) Introduction: At the intersection of organizations and occupations.

Barslund M, Busse M (2016) How mobile is tech talent? A case study of IT professionals based on data from LinkedIn.

Beaudry P, Green DA, Sand BM (2015) The Great Reversal in the Demand for Skill and Cognitive Tasks. J Labor Econ 34:S199–S247. doi: 10.1086/682347

Beblavý M, Akgüç M, Fabo B, et al (2017) A methological inquiry into data generating process concerning new jobs and skills. Taxonomy.

Beblavý M, Akgüc M, Fabo B, Lenaerts K (2016a) What are the new occupations and the new skills? And how are they measured?

Beblavý M, Fabo B (2015) Students in Work and their Impact on the Labour Market. Centre for European Policy Studies

Beblavý M, Fabo B, Lenaerts K (2016b) Skills Requirements for the 30 Most-Frequently Advertised Occupations in the United States: An analysis based on online vacancy data. Centre for European Policy Studies

Beblavý M, Fabo B, Lenaerts K (2016c) Demand for Digital Skills in the US Labour Market: The IT Skills Pyramid.

Beblavý M, Fabo B, Lenaerts K (2016d) Skills Requirements for the 30 Most-Frequently Advertised Occupations in the United States: An Analysis Based on Online Vacancy Data. Social Science Research Network, Rochester, NY

Beblavý M, Lehouelleur S, Maselli I (2013) Useless Degrees of Useless Statistics? A Comparison of the Net Present Value of Higher Education by Field of Study in Five European Countries. Social Science Research Network, Rochester, NY

Beblavý M, Mýtna-Kureková L, Haita C (2016e) The surprisingly exclusive nature of medium- and low-skilled jobs: Evidence from a Slovak job portal. Pers Rev 45:255–273. doi: 10.1108/PR-12-2014-0276

Becker GS (1962) Investment in Human Capital: A Theoretical Analysis. J Polit Econ 70:9–49. doi: 10.1086/258724

Bell DNF, Blanchflower DG (2011) Young people and the Great Recession. Oxf Rev Econ Policy 27:241–267. doi: 10.1093/oxrep/grr011

Benfield JA, Szlemko WJ (2006) Internet-Based Data Collection: Promises and Realities. J Res Pract 2:1.

Benhabib J, Spiegel MM (1994) The role of human capital in economic development evidence from aggregate cross-country data. J Monet Econ 34:143–173. doi: 10.1016/0304-3932(94)90047-7

Berg J (2016) Income security in the on-demand economy: Findings and policy lessons from a survey of crowdworkers.

Berger M, Schaffner S (2012) How to Use the EU-SILC Panel to Analyse Monthly and Hourly Wages. Social Science Research Network, Rochester, NY

Bethlehem J (2010) Selection Bias in Web Surveys. Int Stat Rev 78:161–188.

BGT (2015) Digital Skills Gap in the Workforce.

Bleakley H, Chin A (2004) Language Skills and Earnings: Evidence from Childhood Immigrants*. Rev Econ Stat 86:481–496. doi: 10.1162/003465304323031067

Blinder AS (2009) How many US jobs might be offshorable? World Econ 10:41.

Blom AG, Bosnjak M, Cornilleau A, et al (2016) A Comparison of Four Probability-Based Online and Mixed-Mode Panels in Europe. Soc Sci Comput Rev 34:8–25. doi: 10.1177/0894439315574825

Bohmová L, Pavlíček A (2015) The Influence of Social Networking Sites on Recruiting Human Resources in the Czech Republic. Organizacija 48:23–31.

Bohnert D, Ross WH (2010) The influence of social networking web sites on the evaluation of job candidates. Cybertechnol Behav Soc Netw. doi: 10.1089/cyber.2010.9999

Bonsòn E, Bednárová M (2013) Corporate LinkedIn practices of Eurozone companies. Online Inf Rev. doi: 10.1108/OIR-09-2012-0159

Borghans L, Green F, Mayhew K (2001) Skills Measurement and Economic Analysis: An Introduction. Oxf Econ Pap 53:375–384.

Bosnjak M, Das M, Lynn P (2016) Methods for Probability-Based Online and Mixed-Mode Panels Selected Recent Trends and Future Perspectives. Soc Sci Comput Rev 34:3–7. doi: 10.1177/0894439315579246

Boucher E, Renault C (2015) Job Classification Based on LinkedIn Summaries.

Bound J, Brown C, Mathiowetz N (2001) Chapter 59 - Measurement Error in Survey Data. In: Leamer JJH and E (ed) Handbook of Econometrics. Elsevier, pp 3705–3843

Brunello G, Schlotter M (2011) Non-Cognitive Skills and Personality Traits: Labour Market Relevance and Their Development in Education & Training Systems. Social Science Research Network, Rochester, NY

Budría S, Swedberg P (2012) The Impact of Language Proficiency on Immigrants' Earnings in Spain. Social Science Research Network, Rochester, NY

Bührer C, Hagist C (2017) The Effect of Digitalization on the Labor Market. In: Ellermann H, Kreutter P, Messner W (eds) The Palgrave Handbook of Managing Continuous Business Transformation. Palgrave Macmillan UK, pp 115–137

Buhrmester M, Kwang T, Gosling SD (2011) Amazon's Mechanical Turk—a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci. doi: 10.1177/1745691610393980

Byars-Winston A, Fouad N, Wen Y (2015) Race/ethnicity and sex in U.S. occupations, 1970–2010: Implications for research, practice, and policy. J Vocat Behav 87:54–70. doi: 10.1016/j.jvb.2014.12.003

Caers R, Castelyns V (2011) LinkedIn and Facebook in Belgium: the influences and biases of social network sites in recruitment and selection procedures. Soc Sci Comput Rev. doi: 10.1177/0894439310386567

Capiluppi A, Baravalle A (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). pp 13–21

Carnevale AP, Jayasundera T, Repnikov D (2014) Understanding online job ads data: a technical report.

Caroleo FE, Pastore F (2015) Overeducation: A Disease of the School-to-Work Transition System. Social Science Research Network, Rochester, NY

Carrière-Swallow Y, Labbé F (2013) Nowcasting with Google Trends in an Emerging Market. J Forecast 32:289–298. doi: 10.1002/for.1252

CEDEFOP (2014) Briefing note - Skill mismatch: more than meets the eye.

CEDEFOP (2015) Skill shortages and gaps in European enterprises: striking a balance between vocational education and training and the labour market.

Chen T, So EPK, Wu L, Yan IKM (2015) The 2007-2008 US Recession: what did the real-timer Google Trends Data tell the United States? Contemp Econ Policy. doi: 10.1111/coep.12074

Chin A, Juhn C, Thompson P (2006) Technical Change and the Demand for Skills during the Second Industrial Revolution: Evidence from the Merchant Marine, 1891–1912. Rev Econ Stat 88:572–578. doi: 10.1162/rest.88.3.572

Chiswick BR, Miller PW (2010) Occupational language requirements and the value of English in the US labor market. J Popul Econ 23:353–372. doi: 10.1007/s00148-008-0230-7

Chiswick BR, Miller PW (1994) Language choice among immigrants in a multi-lingual destination. J Popul Econ 7:119–131. doi: 10.1007/BF00173615

Chiswick BR, Miller PW (1999) Language skills and earnings among legalized aliens. J Popul Econ 12:63–89. doi: 10.1007/s001480050091

Chmielecki M (2013) Rekrutacja z wykorzystaniem mediów społecznościowych. Zesz Nauk Wyższej Szk Bank We Wrocławiu 4:37–51.

Choi H (2009) Predicting Initial Claims for Unemployment Benefits. Social Science Research Network, Rochester, NY

Choi H, Varian H (2012) Predicting the present with Google Trends. Econ Rec. doi: 10.1111/j.1475-4932.2012.00809.x

Codagnone C, Abadie F, Biagi F (2016a) The Future of Work in the 'Sharing Economy'. Market Efficiency and Equitable Opportunities or Unfair Precarisation? Publications Office of the European Union

Codagnone C, Biagi F, Abadie F (2016b) The Passions and the Interests: Unpacking the "Sharing Economy." Social Science Research Network, Rochester, NY

Codagnone C, Martens B (2016) Scoping the Sharing Economy: Origins, Definitions, Impact and Regulatory Issues.

Constant A, Zimmermann KF (2008) Im Angesicht der Krise: US-Präsidentschaftswahlen in transnationaler Sicht.

Constant AF, Kahanec M, Zimmermann KF (2012) The Russian–Ukrainian earnings divide. Econ Transit 20:1–35. doi: 10.1111/j.1468-0351.2011.00428.x

Conte A, Vivarelli M (2011) Imported Skill-Biased Technological Change in Developing Countries. Dev Econ 49:36–65. doi: 10.1111/j.1746-1049.2010.00121.x

Coombs CK, Cebula RJ (2010) Are there rewards for language skills? Evidence from the earnings of registered nurses. Soc Sci J 47:659–677. doi: 10.1016/j.soscij.2010.01.009

Couper MP (2000) Review: Web Surveys: A Review of Issues and Approaches. Public Opin Q 64:464–494.

Cowen T (2013) Average Is Over: Powering America Beyond the Age of the Great Stagnation. Penguin

Crosby O (2002) New and emerging occupations. Occup Outlook Q 46:16–25.

Damarin AK (2006) Rethinking Occupational Structure: The Case of Web Site Production Work. Work Occup 33:429–463. doi: 10.1177/0730888406293917

Daniels EA, Sherman AM (2016) Model Versus Military Pilot: A Mixed-Methods Study of Adolescents' Attitudes Toward Women in Varied Occupations. J Adolesc Res 31:176–201. doi: 10.1177/0743558415587025

Das M (2012) Innovation in Online Data Collection for Scientific Research: The Dutch MESS Project. Methodol Innov Online 7:7–24. doi: 10.4256/mio.2012.002

Das M, Toepoel V, van Soest A (2011) Nonparametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys. Sociol Methods Res 40:32–56. doi: 10.1177/0049124110390765

de Bustillo RM, de Pedraza P (2010) Determinants of job insecurity in five European countries. Eur J Ind Relat 16:5–20. doi: 10.1177/0959680109355306

de Groen PW, Maselli I (2016) The Impact of the Collaborative Economy on the Labour Market. Centre for European Policy Studies, Brussels

de Groen PW, Maselli I, Fabo B (2016) The Digital Market for Local Services: A One-Night Stand for Workers? An Example from the On-Demand Economy. Social Science Research Network, Rochester, NY

de Pedraza P, Tijdens K, de Bustillo RM, Steinmetz S (2010) A Spanish Continuous Volunteer Web Survey: sample Bias, Weighting and Efficiency. Reis Rev Esp Investig Sociológicas 109–130.

Degryse C (2016) Digitalisation of the Economy and its Impact on Labour Markets. European Trade Union Institute, Brussels

Dillman DA, Bowker DK (2001) The web questionnaire challenge to survey methodologists. In: Bemad Batinic, Ulf-Dietrich Reips, Michael Bosnjak & Andreas Werner: Online Social Sciences. Seattle: Hogrefe & Huber. pp 53–71

Dörfler L, Werfhorst HG van de (2009) Employers' Demand for Qualifications and Skills. Eur Soc 11:697–721. doi: 10.1080/14616690802474374

Drahokoupil J, Fabo B (2016) The platform economy and the disruption of the employment relationship.

Drahokoupil J, Fabo B (2017) Outsourcing, Offshoring and the Deconstruction of Employment: New and Old Challenges in the Digital Economy. Social Science Research Network, Rochester, NY

Duggan M, Ellison N, Lambe C, et al (2015) Duggan M, Ellison NB, Lampe C, Lenhart A, Madden M. PEW Research Center

Dukova K (2016) Is picture worth a thousand words: experimental research on the usage of information graphics as a presentation tool for political information. Master thesis, Central European University

Duncan A, Mavisakalyan A (2015) Russian language skills and employment in the Former Soviet Union. Econ Transit 23:625–656. doi: 10.1111/ecot.12075

Dunlop JT (1966) Job vacancy measures and economic analysis. In: The measurement and interpretation of job vacancies. NBER, pp 27–47

Ebenstein A, Harrison A, McMillan M, Phillips S (2013) Estimating the Impact of Trade and Offshoring on American Workers using the Current Population Surveys. Rev Econ Stat 96:581–595. doi: 10.1162/REST_a_00400

Edwards A, Housley W, Williams M, et al (2013) Digital social research, social media and the sociological imagination: surrogacy, augmentation and re-orientation. Int J Soc Res Methodol 16:245–260. doi: 10.1080/13645579.2013.774185

Elias P (1997) Occupational Classification (ISCO-88). Organisation for Economic Co-operation and Development, Paris

Elias P, McKnight A (2001) Skill measurement in official statistics: recent developments in the UK and the rest of Europe. Oxf Econ Pap 53:508–540. doi: 10.1093/oep/53.3.508

ESCO (2015) European Skills, Competences, Qualifications and Occupations,.

Esser H (2006) Sprache und Integration: die sozialen Bedingungen und Folgen des Spracherwerbs von Migranten. Campus

ET2020 (2011) Languages for Jobs - Providing multilingual communication skills for the labour market.

European Commission (2015) EU statistics on income and living conditions (EU-SILC) methodology.

Eurydice, Eurostat (2012) Key Data on Teaching Languages at School in Europe.

Faberman RJ, Kudlyak M (2014) The intensity of job search and search duration.

Fabo B, Beblavý M, Lenaerts K (2017a) The importance of foreign language skills in the labour markets of Central and Eastern Europe: assessment based on data from online job portals. Empirica 1–22. doi: 10.1007/s10663-017-9374-6

Fabo B, Hudáčková S, Nogacz A (2017b) Can Airbnb Provide Livable Incomes to Property Owners?: An Analysis on National, Regional and City District Level. Social Science Research Network, Rochester, NY

Fabo B, Karanović J, Dukova K (2017c) In search of an adequate European policy response to the platform economy. Transf Eur Rev Labour Res 1024258916688861. doi: 10.1177/1024258916688861

Fabo B, Tijdens K (2014) Using Web Data to Measure the Demand for Skills.

Fairlie RW, Robb A (2009) Entrepreneurship, self-employment and business data: an introduction to several large, nationally-representative datasets.

Falk A, Fischbacher U (2006) A theory of reciprocity. Games Econ Behav 54:293–315. doi: 10.1016/j.geb.2005.03.001

Feenstra RC, Hanson GH (1999) The Impact of Outsourcing and High-Technology Capital on Wages: Estimates For the United States, 1979–1990. Q J Econ 114:907–940. doi: 10.1162/003355399556179

Fernández-Macías E (2012) Job Polarization in Europe? Changes in the Employment Structure and Job Quality, 1995-2007. Work Occup 39:157–182. doi: 10.1177/0730888411427078

Field A (2013) Discovering Statistics Using IBM SPSS Statistics. SAGE

Fielding NG, Lee RM, Blank G (2008) The SAGE Handbook of Online Research Methods. SAGE

Finholt T, Sproull LS (1990) Electronic groups at work. Organ Sci. doi: 10.1287/orsc.1.1.41

Fitzenberger B, Lickleder er S (2016) Career Planning, School Grades, and Transitions: The Last Two Years in a German Lower Track Secondary School. Jahrb Für Natl Stat 235:433–458. doi: 10.1515/jbnst-2015-4-507

Fondeur Y, Karamé F (2013) Can Google data help predict French youth unemployment? Econ Model 30:117–125. doi: 10.1016/j.econmod.2012.07.017

Foster G (1994) Fishing with the Net for research data. Br J Educ Technol. doi: 10.1111/j.1467-8535.1994.tb00094.x

Freeman LC (1984) The impact of computer-based communication on the social structure of an emerging social scientific speciality. Soc Netw. doi: 10.1016/0378-8733(84)90011-X

Frey CB, Osborne MA (2017) The future of employment: How susceptible are jobs to computerisation? Technol Forecast Soc Change 114:254–280. doi: 10.1016/j.techfore.2016.08.019

Fricker RD, Schonlau M (2002) Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. Field Methods 14:347–367. doi: 10.1177/152582202237725

Friederici N, Ojanperä S, Graham M (2016) The Impact of Connectivity in Africa: Grand Visions and the Mirage of Inclusive Digital Development. Social Science Research Network, Rochester, NY

Gallie D, Felstead A, Green F (2007) Skill, task discretion and new technology. Année Sociol 53:401–430.

Gallie D, Felstead A, Green F (2004) Changing Patterns of Task Discretion in Britain. Work Employ Soc 18:243–266. doi: 10.1177/09500172004042769

Gee L (2014) The More you Know: Information Effects in Job Application Rates by Gender in a Large Field Experiment. Department of Economics, Tufts University

Gee LK, Jones JJ, Burke M (2016) Social Networks and Labor Markets: How Strong Ties Relate to Job Finding On Facebook's Social Network.

Gee LK, Jones JJ, Fariss CJ, et al (2017) The paradox of weak ties in 55 countries. J Econ Behav Organ 133:362–372. doi: 10.1016/j.jebo.2016.12.004

Ghani E, Kerr WR, Stanton CT (2012) Diasporas and outsourcing: evidence from oDesk and India, NBER. Working paper 18474.

Ginsberg J, Mohebbi MH, Patel RS, et al (2009) Detecting influenza epidemics using search engine query data. Nature. doi: 10.1038/nature07634

Goos M, Manning A (2007) Lousy and Lovely Jobs: The Rising Polarization of Work in Britain. Rev Econ Stat 89:118–133. doi: 10.1162/rest.89.1.118

Goos M, Manning A, Salomons A (2009) Job Polarization in Europe. Am Econ Rev 99:58–63. doi: 10.1257/aer.99.2.58

Goudin P (2016) The Cost of Non-Europe in the Sharing Economy. European Parliament

Granovetter MS (1973) The Strength of Weak Ties. Am J Sociol 78:1360–1380. doi: 10.1086/225469

Gray R (2013) Taking technology to task: The skill content of technological change in early twentieth century United States. Explor Econ Hist 50:351–367. doi: 10.1016/j.eeh.2013.04.002

Grossman GM, Rossi-Hansberg E (2012) Task Trade Between Similar Countries. Econometrica 80:593–629. doi: 10.3982/ECTA8700

Guo Y, Shen S, Visser O, Iosup A (2012) An analysis of online match-based games. In: 2012 IEEE International Workshop on Haptic Audio Visual Environments and Games (HAVE 2012) Proceedings. pp 134–139

Guzi M, de Pedraza P (2015) A web survey analysis of subjective well-being. Int J Manpow 36:48–67. doi: 10.1108/IJM-12-2014-0237

Headworth A (2014) Recruiters: why are you not using Facebook Graph Search? http://sironaconsulting.com/2014/09/recruiters-facebook-search/.

Heckman JJ, Lochner LJ, Todd PE (2006) Chapter 7 Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond. In: Welch EH and F (ed) Handbook of the Economics of Education. Elsevier, pp 307–458

Helpman E, Itskhoki O, Muendler M-A, Redding SJ (2017) Trade and Inequality: From Theory to Estimation. Rev Econ Stud 84:357–405. doi: 10.1093/restud/rdw025

Hershbein B, Kahn LB (2015) Is College the New High School?

Hershbein BJ (2015) Is college the new high school? Evidence from vacancy postings.

Holt C, David M (1966) The concept of job vacancies in a dynamic theory of the labor market. In: The measurement and interpretation of job vacancies. NBER, pp 73–110

Hooley T, Wellens J, Marriott J (2011) What is Online Research?: Using the Internet for Social Science Research. Bloomsbury Academic, London; New York

Horrigan JB (2016) Lifelong Learning and Technology. In: Pew Res. Cent. Internet Sci. Tech. http://www.pewinternet.org/2016/03/22/lifelong-learning-and-technology/. Accessed 15 Jul 2016

Horton JJ (2017) The effects of algorithmic labor market recommendations: evidence from a field experiment.

Horton JJ, Rand D, Zeckhauser R (2011) The online laboratory: conducting experiments in a real labor market. Exp Econ. doi: 10.1007/s10683-011-9273-9

Huang H, Kvasny L, Joshi KD, et al (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. ACM, New York, NY, USA, pp 121–128

Hunter D (2009) ISCO-08 draft definitions.

Huws U (2014) Labor in the Global Digital Economy: The Cybertariat Comes of Age. NYU Press

Ikenaga T, Kambayashi R (2016) Task Polarization in the Japanese Labor Market: Evidence of a Long-Term Trend. Social Science Research Network, Rochester, NY

Immergluck D (1999) Cities and finance jobs: The effects of financial services restructuring on the location of employment. Brookings Institution, Center on Urban and Metropolitan Policy

Isphording IE (2014) Language and Labor Market Success.

ITU (2014) Measuring the information society report 2014.

Jackson M (2001) Non-Meritocratic Job Requirements and the Reproduction of Class Inequality: An Investigation. Work Employ Soc 15:619–630. doi: 10.1017/S0950017001000393

Jackson M (2007) How far merit selection? Social stratification and the labour market1. Br J Sociol 58:367–390. doi: 10.1111/j.1468-4446.2007.00156.x

Jackson M, Goldthorpe JH, Mills C (2005) Education, Employers and Class Mobility. Res Soc Stratif Mobil 23:3–33. doi: 10.1016/S0276-5624(05)23001-9

Jarrow R, Kchia Y, Protter P (2011) Is there a bubble in LinkedIn's stock price? J Portf Manag. doi: 10.3905/jpm.2011.38.1.125

Juhn C, Murphy KM, Pierce B (1993) Wage Inequality and the Rise in Returns to Skill. J Polit Econ 101:410–442. doi: 10.1086/261881

Jung J, Mercenier J (2014) Routinization-Biased Technical Change and Globalization: Understanding Labor Market Polarization. Econ Inq 52:1446–1465. doi: 10.1111/ecin.12108

Kahanec M, Fabo B (2013) Migration strategies of crisis-stricken youth in an enlarged European Union. Transf Eur Rev Labour Res 19:365–380. doi: 10.1177/1024258913493701

Kaplan K (2014) Interviews: en garde. Nature. doi: 10.1038/nj7488-397a

Karl K, Peluchette JV, Schlaegel C (2010) A cross-cultural examination of student attitudes and gender differences in Facebook profiles content. Int J Virtual Communities Soc Netw. doi: 10.4018/jvcsn.2010040102

Katz LF, Murphy KM (1991) Changes in Relative Wages, 1963-1987: Supply and Demand Factors. National Bureau of Economic Research

Katz MB (1972) Occupational Classification in History. J Interdiscip Hist 3:63–88. doi: 10.2307/202462

Kearney MS, Levine PB (2015) Media Influences on Social Outcomes: The Impact of MTV's 16 and Pregnant on Teen Childbearing. Am Econ Rev 105:3597–3632. doi: 10.1257/aer.20140012

Kehoe CM, Pitkow JE (1996) Surveying the territory: GVU's five WWW user surveys.

Kelkar A, Kulkarni S (2013) Value of facebook for job search: Languishing present to a lucrative future. In: International Conference on Information Society (i-Society 2013). pp 222–226

Kennan M, Cole F, Willard P, et al (2006) Changing workplace demands: what job ads tell us. Aslib Proc 58:179–196. doi: 10.1108/00012530610677228

Keynes JM (2010) Economic Possibilities for Our Grandchildren. In: Essays in Persuasion. Palgrave Macmillan UK, pp 321–332

Kiesler S, Sproull LS (1986) Response effects in the electronic survey. Public Opin Q. doi: 10.1086/268992

Kim M, Liu AH, Tuxhorn K-L, et al (2015) Lingua Mercatoria: Language and Foreign Direct Investment. Int Stud Q 59:330–343. doi: 10.1111/isqu.12158

Kitchenham B, Pfleeger SL (2002) Principles of Survey Research: Part 5: Populations and Samples. SIGSOFT Softw Eng Notes 27:17–20. doi: 10.1145/571681.571686

Kluemper DH, Rosen PA (2009) Future employment selection methods: evaluating social networking web sites. J Manag Psychol. doi: 10.1108/02683940910974134

Kohut A, Keeter S, Doherty C, et al (2012) Assessing the representativeness of public opinion surveys.

Krantz JH, Dalal RS (2000) Validity of web-based psychological research. In: Birnbaum MH (ed) Psychological experiments on the Internet. Academic Press, San Diego,

Kudlyak M, Faberman J, others (2014) The Intensity of Job Search and Search Duration. In: 2014 Meeting Papers. Society for Economic Dynamics,

Kuhn P (2014) The Internet as a labor matchmaker. IZA World of Labor No. 18.

Kuhn P, Mansour H (2014) Is Internet Job Search Still Ineffective? Econ J 124:1213–1233. doi: 10.1111/ecoj.12119

Kuhn P, Shen K (2013) Gender Discrimination in Job Ads: Evidence from China. Q J Econ 128:287–336. doi: 10.1093/qje/qjs046

Kuhn P, Skuterud M (2004) Internet job search and unemployment durations. Am Econ Rev. doi: 10.1257/000282804322970779

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? WWW Proceedings of the 19th international conference on world wide web, ACM New York.

Larsen M (2011) How the big boys do Twitter recruitment. https://www.recruiter.com/i/twitter-recruitment/?PageSpeed=noscript.

Lauby S (2013) The new social workplace. Hum Resour Manag Int Dig. doi: 10.1108/09670731311318352

Lazer D, Kennedy R, King G, Vespignani A (2014) The Parable of Google Flu: Traps in Big Data Analysis. Science 343:1203–1205. doi: 10.1126/science.1248506

Lemieux T (2006) The "Mincer Equation" Thirty Years After Schooling, Experience, and Earnings. In: Grossbard S (ed) Jacob Mincer A Pioneer of Modern Labor Economics. Springer US, pp 127–145

Lenaerts K, Beblavý M, Fabo B (2016) Prospects for utilisation of non-vacancy Internet data in labour market analysis—an overview. IZA J Labor Econ 5:1–18. doi: 10.1186/s40172-016-0042-z

Levenson A, Zoghi C (2010) Occupations, Human Capital and Skills. J Labor Res 31:365–386. doi: 10.1007/s12122-010-9098-x

Lindemann K, Kogan I (2013) The Role of Language Resources in Labour Market Entry: Comparing Estonia and Ukraine. J Ethn Migr Stud 39:105–123. doi: 10.1080/1369183X.2012.711050

Mamertino M, Sinclair TM (2016) Online Job Search and Migration Intentions Across EU Member States.

Mandl I (2016) Remuneration and social protection of platform workers.

Mang C (2012) Online job search and matching quality. Ifo Working Paper

Mann C, Stewart F (2000) Internet communication and qualitative research: A handbook for researching online. Sage

Manning A (2004) We Can Work It Out: The Impact of Technological Change on the Demand for Low-Skill Workers. Scott J Polit Econ 51:581–608. doi: 10.1111/j.0036-9292.2004.00322.x

Marinescu I (2016) The General Equilibrium Impacts of Unemployment Insurance: Evidence from a Large Online Job Board. National Bureau of Economic Research

Marinescu I, Rathelot R (2016) Mismatch Unemployment and the Geography of Job Search. National Bureau of Economic Research

Marshall MN (1996) Sampling for qualitative research. Fam Pract 13:522–526. doi: 10.1093/fampra/13.6.522

Martínek T, Hanzlík P (2014) Analysis of the Structure of Job Offers on the Czech Labour Market. Rev Econ Perspect 14:287–306. doi: 10.2478/revecp-2014-0015

Maselli I (2012) The evolving supply and demand of skills in the labour market. Intereconomics 47:22–30.

Maselli I, Fabo B (2015) Digital workers by design? An example from the on-demand economy. Centre for European Policy Studies

Maselli I, Lenaerts K, Beblavý M (2016) Five things we need to know about the on-demand economy. Centre for European Policy Studies

Massimino PM, Kopelman RE, Joseph ML (2015) Explaining hospital performance via the Cube One framework.

Masso J, Eamets R, Mõtsmees P (2014) Temporary migrants and occupational mobility: evidence from the case of Estonia. Int J Manpow 35:753–775. doi: 10.1108/IJM-06-2013-0138

Maurer-Fazio M (2012) Ethnic discrimination in China's internet job board labor market. IZA J Migr 1:12. doi: 10.1186/2193-9039-1-12

Maurer-Fazio M, Lei L (2015) "As rare as a panda": How facial attractiveness, gender, and occupation affect interview callbacks at Chinese firms. Int J Manpow 36:68–85. doi: 10.1108/IJM-12-2014-0258

Maxwell NL (2010) English Language and Low-Skilled Jobs: The Structure of Employment. Ind Relat J Econ Soc 49:457–465. doi: 10.1111/j.1468-232X.2010.00609.x

Melitz J (2008) Language and foreign trade. Eur Econ Rev 52:667–699. doi: 10.1016/j.euroecorev.2007.05.002

Melitz J, Toubal F (2014) Native language, spoken language, translation and trade. J Int Econ 93:351–363. doi: 10.1016/j.jinteco.2014.04.004

Mendez R (2002) Creative Destruction and the Rise of Inequality. J Econ Growth 7:259–281. doi: 10.1023/A:1020158115979

Mincer JA (1974) Schooling, Experience, and Earnings.

Mitra A (2003) Access to Supervisory Jobs and the Gender Wage Gap among Professionals. J Econ Issues 37:1023–1044. doi: 10.1080/00213624.2003.11506641

Moat HS, Preis T, Olivola CY, et al (2014) Using big data to predict collective behavior in the real world. Behav Brain Sci 37:92–93.

Murth D (2015) Twitter and elections: are tweets, predictive, reactive, or a form of buzz? Inf Commun Soc. doi: 10.1080/1369118X.2015.1006659

Musch J, Reips U-. D (2000) A brief history of web experimenting. In: Birnbaum MH (ed) Psychological experiments on the Internet. Academic Press, San Diego,

Mýtna-Kureková L, Beblavý M, Haita C, Thum A-E (2016) Employers' skill preferences across Europe: between cognitive and non-cognitive skills. J Educ Work 29:662–687. doi: 10.1080/13639080.2015.1024641

Mýtna-Kureková L, Beblavý M, Haita C (2012) Qualifications or Soft Skills? Studying Job Advertisements for Demand for Low-Skilled Staff in Slovakia. Social Science Research Network, Rochester, NY

Mýtna-Kureková L, Beblavý M, Thum-Thysen A (2015) Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. IZA J Labor Econ 4:1–20. doi: 10.1186/s40172-015-0034-4

Mýtna-Kureková L, Haita C, Beblavý M (2013) Being and Becoming Low-Skilled: A Comprehensive Approach to Studying Low-Skillness. Social Science Research Network, Rochester, NY

Mýtna-Kureková L, Žilinčíková Z (2016) Are student jobs flexible jobs? Using online data to study employers' preferences in Slovakia. IZA J Eur Labor Stud 5:20. doi: 10.1186/s40174-016-0070-5

Nickell S, Bell B (1996) Changes in the Distribution of Wages and Unemployment in OECD Countries. Am Econ Rev 86:302–308.

OECD (2014) Skills and jobs in the Internet economy. OECD Publishing, Paris

Oesch D (2013) Occupational Change in Europe: How Technology and Education Transform the Job Structure. OUP Oxford

Oesch D, Menés J (2011) Upgrading or polarization? Occupational change in Britain, Germany, Spain and Switzerland, 1990–2008. Socio-Econ Rev 9:503–531. doi: 10.1093/ser/mwq029

Oh CH, Travis Selmier W, Lien D (2011) International trade, foreign direct investment, and transaction costs in languages. J Socio-Econ 40:732–735. doi: 10.1016/j.socec.2011.08.003

Oi WY, Idson TL (1999) Chapter 33 Firm size and wages. In: Economics B-H of L (ed). Elsevier, pp 2165–2214

Osterhaus E (2014) How job seekers use Glassdoor reviews. Survey by Software Advice.

Ours JC van (1989) Durations of Dutch job vacancies. Econ 137:309–327. doi: 10.1007/BF02115697

Pallais A (2014) Inefficient hiring in entry-level labor markets. Am Econ Rev. doi: 10.1257/aer.104.11.3565

Pallais A, Sandse.g. (2016) Why the Referential Treatment? Evidence from Field Experiments on Referrals. J Polit Econ 124:1793–1828. doi: 10.1086/688850

Pan J (2015) Gender Segregation in Occupations: The Role of Tipping and Social Interactions. J Labor Econ 33:365–408. doi: 10.1086/678518

Papacharissi Z (2009) The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and ASmallWorld. New Media Soc 11:199–220. doi: 10.1177/1461444808099577

Preis T, Moat HS (2015) Early Signs of Financial Market Moves Reflected by Google Searches. In: Gonçalves B, Perra N (eds) Social Phenomena. Springer International Publishing, pp 85–97

Preis T, Moat HS, Stanley HE (2013) Quantifying trading behavior in financial markets using Google Trends.

Preis T, Moat HS, Stanley HE, Bishop SR (2012) Quantifying the advantage of looking forward.

Rangel L (2014) Writing a LinkedIn profile to get found by recruiters. Career Plan Adult Dev J 30:126.

Reips U-D (2006) Web-based methods. In: Handbook of multimethod measurement in psychology. pp 73–85

Reips U-D (2012) Using the Internet to collect data. In: APA handbook of research methods in psychology. Washington, DC: American Psychological Association.

Rode A, Shukla A (2013) Prejudicial Attitudes and Labor Market Outcomes.

Rooth D-O, Ekberg J (2006) Occupational Mobility for Immigrants in Sweden. Int Migr 44:57–77. doi: 10.1111/j.1468-2435.2006.00364.x

Rothwell J (2014) Still searching: job vacancies and STEM skills.

Sanchez JL, González CS, Alayon S (2011) Evaluation of transversal competences in the final year project in engineering. In: 2011 Proceedings of the 22nd EAEEIE Annual Conference (EAEEIE). pp 1–5

Sasser Modestino A, Shoag D, Ballance J (2014) Upskilling: Do employers demand greater skill when skilled workers are plentiful? Working Papers, Federal Reserve Bank of Boston

Schawbel D (2012) How recruiters use social networks to make hiring decisions now. Time

Schmid-Druner M (2016) The Situation of Workers in the Collaborative Economy. European Parliament

Schmidt T, Vossen S (2012) Using Internet data to account for special events in economic forecasting.

Schonlau M, van Soest A, Kapteyn A, Couper M (2009) Selection Bias in Web Surveys and the Use of Propensity Scores. Sociol Methods Res 37:291–318. doi: 10.1177/0049124108327128

Schultz TW (1971) Investment in Human Capital. The Role of Education and of Research.

Shapiro H (2014) Use of real time labour market data.

Slane R (2013) The Myth of "Real-Time Labour Market Data." In: Emsi. http://www.economicmodelling.co.uk/2013/06/12/the-myth-of-real-time-labour-market-data/. Accessed 10 Apr 2017

Sloane P (2014) Overeducation, skill mismatches, and labor market outcomes for college graduates. IZA World Labor. doi: 10.15185/izawol.88

Smith A (2015) Searching for Work in the Digital Era. In: Pew Res. Cent. Internet Sci. Tech. http://www.pewinternet.org/2015/11/19/searching-for-work-in-the-digital-era/. Accessed 15 Jul 2016

Spenner K (1990) Skill: Meanings, Methods, and Measures. Work Occup 17:399–421. doi: 10.1177/0730888490017004002

State B, Rodriguez M, Helbing D, Zagheni E (2014) Migration of professional to the US—evidence from LinkedIn data, published in Social Informatics (SOCINFO 2014, edited by Aiello and McFarland). Lect Notes Comput Sci. doi: 10.1007/978-3-319-13734-6_37

Statistics Netherlands (2013) Final qulity report. EU-SILC 2010. The Netherlands.

Statistics Netherlands (2016) ICT gebruik van personen naar persoonskenmerken, 2005-2013.

Štefánik M (2012) Internet job search data as a possible source of information on skills demand (with results for Slovak university graduates)(89). Build Ski Forecast Methods Appl 246.

Steinmetz S, Bianchi A, Tijdens K, Biffignandi S (2014) Improving web survey quality. In: Callegaro rio, Baker R, Bethlehem J, et al. (eds) Online Panel Research. John Wiley & Sons, Ltd, pp 273–298

Steinmetz S, Raess D, Tijdens K, de Pedraza P (2013) Measuring Wages Worldwide: Exploring the Potentials and Constraints. In: Advancing research methods with new technologies. p 100

Stephens-Davidowitz S (2014) The cost of racial animus on a black candidate: Evidence using Google search data. J Public Econ 118:26–40. doi: 10.1016/j.jpubeco.2014.04.010

Stevenson B (2008) The Internet and Job Search. National Bureau of Economic Research

Stevenson B (2006) The Impact of the Internet on Worker Flows.

Stieger S, Reips U-D (2010) What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. Comput Hum Behav 26:1488–1495. doi: 10.1016/j.chb.2010.05.013

Stöhr T (2015) The returns to occupational foreign language use: Evidence from Germany. Labour Econ 32:86–98. doi: 10.1016/j.labeco.2015.01.004

Szabó I (2011) Comparing the competence contents of demand and supply sides on the labour market. In: Information Technology Interfaces (ITI), Proceedings of the ITI 2011 33rd International Conference on. IEEE, pp 345–350

Tambe P (2014) Big Data Investment, Skills, and Firm Value. Manag Sci 60:1452–1469. doi: 10.1287/mnsc.2014.1899

Tijdens K (2014) Dropout Rates and Response Times of an Occupation Search Tree in a Web Survey : Journal of Official Statistics.

Tijdens K, Beblavý M, Thum-Thysen A (2015a) Do educational requirements in vacancies match the educational attainments of job holders.

Tijdens K, Besamusca J, Klaveren M van (2015b) Workers and Labour Market Outcomes of Informal Jobs in Formal Establishments. A Job-based Informality Index for Nine Sub-Saharan African Countries. Eur J Dev Res 27:868–886. doi: 10.1057/ejdr.2014.73

Tijdens K, de Ruijter E, de Ruijter J (2013a) Comparing tasks of 160 occupations across eight European countries. Empl Relat 36:110–127. doi: 10.1108/ER-05-2013-0046

Tijdens K, de Vries DH, Steinmetz S (2013b) Health workforce remuneration: comparing wage levels, ranking, and dispersion of 16 occupational groups in 20 countries. Hum Resour Health 11:11. doi: 10.1186/1478-4491-11-11

Tijdens K, Ruijter JD, Ruijter ED (2012) Measuring work activities and skill requirements of occupations: Experiences from a European pilot study with a web-survey. Eur J Train Dev 36:751–763. doi: 10.1108/03090591211255575

Tijdens K, Steinmetz S (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. Int J Soc Res Methodol 19:461–479. doi: 10.1080/13645579.2015.1035875

Tijdens K, Visintin S (2016) What do workers do? Measuring the intensity and market value of tasks in jobs.

Tijdens K, Wetzels C, Andralojc M, Michon P (2005) Measuring wages, and calculating hourly wages in the WageIndicator dataset.

Tijdens KG (2010) Measuring occupations in web-surveys: the WISCO database of occupations.

Tijdens KG, van Klaveren M (2011) Young women in service sector occupations:Bookkeepers, call centre operators, receptionists, housekeepers, IT-programmers, sales persons, secretaries, travel agency clerks. Amsterdam:

Toepoel V, Das M, Van Soest A (2008) Effects of design in web surveys comparing trained and fresh respondents. Public Opin Q 72:985–1007.

Toepoel V, Das M, Van Soest A (2009) Design of web questionnaires: The effects of the number of items per screen. Field Methods 21:200–213.

Tomaskovic-Devey D (1995) Jobs Not Occupations.

Tyler JH, Murnane RJ, Willett JB (1999) Do the Cognitive Skills of School Dropouts Matter in the Labor Market? National Bureau of Economic Research

University of Kent (2015) Using Social Media in Jobhunting.

van der Laan P, van Nunspeet W (2009) Modernising household surveys in the Netherlands: design, efficiency gains and perspectives.

Virolainen M, Stenström M-L (2014) Finnish vocational education and training in comparison: strengths and weaknesses. Int J Res Vocat Educ Train 1:81–106. doi: 10.13152/IJRVET.1.2.1

Visintin S, Tijdens K, Steinmetz S, de Pedraza P (2015a) Task implementation heterogeneity and wage dispersion. IZA J Labor Econ 4:1–24. doi: 10.1186/s40172-015-0036-2

Visintin S, Tijdens K, van Klaveren M (2015b) Skill mismatch among migrant workers: evidence from a large multi-country dataset. IZA J Migr 4:1–34. doi: 10.1186/s40176-015-0040-0

Wade M, Parent M (2002) Relationships Between Job Skills and Performance: A Study of Webmasters. J Manag Inf Syst 18:71–96. doi: 10.1080/07421222.2002.11045694

Weiss M, Garloff A (2011) Skill-biased technological change and endogenous benefits: the dynamics of unemployment and wage inequality. Appl Econ 43:811–821. doi: 10.1080/00036840802599933

Williams SA, Terras MM, Warwick C (2013) What do people study when they study Twitter? Classifying Twitter related academic papers. J Doc 69:384–410. doi: 10.1108/JD-03-2012-0027

Wilson RE, Gosling SD, Graham LT (2012) A review of Facebook research in the social sciences. Perspect Psychol Sci. doi: 10.1177/1745691612442904

Winterton J (2009) Competence across Europe: highest common factor or lowest common denominator? J Eur Ind Train 33:681–700. doi: 10.1108/03090590910993571

Wright EO, Dwyer RE (2003) The patterns of job expansions in the USA: a comparison of the 1960s and 1990s. Socio-Econ Rev 1:289–325. doi: 10.1093/soceco/1.3.289

Wright J (2012) Making a key distinction: real-time LMI & traditional labor market data.

Yang X, Pan B, Evans JA, Lv B (2015) Forecasting Chinese tourist volume with search engine data. Tour Manag 46:386–397. doi: 10.1016/j.tourman.2014.07.019

Yao Y, van Ours JC (2015) Language skills and labor market performance of immigrants in the Netherlands. Labour Econ 34:76–85. doi: 10.1016/j.labeco.2015.03.005

Yu Y, Wang X (2015) World Cup 2014 in the Twitter World: a big data analysis of sentiments in US sports fans' tweets. Comput Hum Behav. doi: 10.1016/j.chb.2015.01.075

Zavrel J (2016) Industry Strength Labor Market Web Mining.

Zhang W, Grenier G (2013) How can language be linked to economics?: A survey of two strands of research. Lang Probl Lang Plan 37:203–226. doi: 10.1075/lplp.37.3.01zha

Zheng JD, Dur R, Tijdens KG (2014) Do Workers Work More if Their Wage Compares Well to That of Their Peers in the Economy? A Survey Experiment.

Zide J, Elman B, Shahani-Denning C (2014) LinkedIn and recruitment: how profiles differ across occupations.