DISCUSSION PAPER SERIES

IZA DP No. 8555

# Using Internet Data to Analyse the Labour Market: A Methodological Enquiry

Lucia Mýtna Kureková
Miroslav Beblavý
Anna-Elisabeth Thum

October 2014

IZA

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

# Using Internet Data to Analyse the Labour Market: A Methodological Enquiry

**Lucia Mýtna Kureková**
*Slovak Governance Institute (SGI), Central European University (CEU) and IZA*

**Miroslav Beblavý**
*Center for European Policy Studies (CEPS) and Comenius University*

**Anna-Elisabeth Thum**
*Center for European Policy Studies (CEPS) and European Commission*

Discussion Paper No. 8555
October 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 8555
October 2014

# ABSTRACT

## Using Internet Data to Analyse the Labour Market: A Methodological Enquiry[*]

With the growth of the Internet, online job portals have become an important medium for job matching. This paper focuses on methodological issues arising from the usage of online job vacancy data and voluntary web-based surveys to analyse the labour market. In addition to providing a comprehensive review of the literature based on online data, we highlight the advantages and possible disadvantages of using online data and suggest strategies for overcoming selected methodological issues. We underline the difficulties in adjusting for representativeness of online job vacancies, but nevertheless argue that this rich source of data should be exploited.

JEL Classification: E4, J2

Keywords: internet, job, labour market, matching, methodology, occupation, online, representativeness, vacancy

Corresponding author:

Lucia Mýtna Kureková
SGI
Gajova 4
811 09 Bratislava
Slovakia
E-mail: kurekova@governance.sk

[*] This research was funded by the European Commission (project name NEUJOBS) through the Seventh Framework Programme (FP7), grant Agreement No. 266833. The authors wish to thank for highly valuable comments to participants of the Skillsnet Methodological Conference at CEDEFOP in Thessaloniki in October 2013, in particular to Nick Sofroniou. All errors remain our own.

## Introduction

Internet-based data collection and research are growing research areas with a strong potential to deepen and widen our knowledge about various socio-economic issues. A particular area of increasing interest is the usage of innovative data sources and analytical methods for the study of the labour market (Askitas and Zimmermann 2009; D'Amuri and Marcucci 2010). With the growth of the Internet, many aspects of job search have been transformed due to the availability of online tools for job searching, candidate searching and job matching (European Commission and ECORYS 2012). This also presents new opportunities with respect to possibilities to collect and analyse web-based data about labour market demand and supply, which can enrich our micro-level understanding of a range of pertinent issues such as occupational changes, wages and working conditions and skill and task requirements of employers.

The debate over whether these new data sources should be used more extensively and to what purpose is still open. One of the most prominent qualities of data collected or generated online is the large number of observations one can obtain from online job portals or voluntary web-based surveys focused on labour market issues. There are additional advantages, such as time and cost effectiveness and the easy variability of survey questions (Mang 2012; Wade and Parent 2001; Kennan et al. 2006). Several methodological issues persist, however, which are predominantly related to the quality and representativeness of such data, and generalisability of findings (Gosling et al. 2004; Pedraza, Tijdens, and Muñoz de Bustillo 2007; Steinmetz, Tijdens, and García 2009; Štefánik 2012a). These concerns are especially pertinent in research fields such as political science, economics and sociology where the core quantitative analytical tools are based on representative[1] data and inferential statistical analysis. Despite these possible limitations, studies using new sources of data and advancing research methods

[1] A representative sample is one that accurately reflects the characteristics of the underlying population.

with new technologies have been published in leading social science journals, suggesting that the field is likely to expand rather than decline (Edelman 2012; Sappleton 2013; Taylor, Schroeder, and Meyer 2014). Given the increasing reliance on Internet-based recruitment and the spreading access to the Internet across socio-economic groups and countries, it is highly likely that reliance on such data will grow.

This article reviews a broad selection of existing works using online data from various disciplines in order to identify ways in which researchers and analysts have been dealing with the methodological issues arising from this type of data. We look mainly at studies that have used online job advertisements and web-based voluntary survey data about wages as examples of data that cover demand and supply sides of labour markets. In addition, we discuss how principles of statistical treatment of missing data could inform adjustments of online job vacancies. We highlight the advantages and possible disadvantages of using online data and propose strategies for dealing with the issue of representativeness and generalisability of findings. This article significantly advances the debate on how some of the weaknesses could be corrected, for which types of research questions and research fields they might be of a smaller concern, and which policy areas can be informed by analysis of online data.

## Methodological and research design questions arising from the use of web-based data

Internet-based data collections that focus on labour market research and analysis differ in their objectives and scope. In addition to more traditional activities such as posting vacancies and CVs, there has been a proliferation of sites offering employer evaluations and salary comparisons in recent years.

First, online job portals gather vacancies and CVs and serve as important platforms for labour market matching (Mang 2012). While a few early efforts were made, it is surprising that these data have not been used more extensively to study aspects of labour market demand and supply. As the objective of online portals is not tied to research but rather to providing a platform on which demand and supply meet, data are seldom stored and used as an input to analyse labour market trends and developments. Research engaged with online vacancy data has typically relied on private job portals (Capiluppi and Baravalle 2010; Štefánik 2012a; Kuhn and Shen 2013; Backhaus 2004). A useful potential source of online data is the EURES website, which collects job vacancies across the EU countries in a standardised platform.[2] Its particular utility could potentially lie in enabling cross-country comparisons, which could inform policy-makers at the national and EU level (Kureková et al. 2013).

Second, there are projects whose explicit objective is the collection of data on selected aspects of the labour market, such as wages or employer evaluations. These data are gathered with a view to enhance understanding of country-specific aspects of the labour market from the supply side (e.g. the WageIndicator project[3]) (Fabo and Tijdens 2014). Commercially-oriented websites (e.g. Salary.com or Payscale) provide a similar service, but with less research-oriented focus. More ambitious services, such as Glassdoor or Vault, aim to integrate salary information into wider information about employers and working environments. Such information has already been used in research (Young and Case 2004; DeKay 2013).

These online-collected data sources in many instances cover gaps in our knowledge about various aspects of labour market. Vacancy data are scarce and online availability provides opportunities to access and analyse the content of job advertisements to better understand what employers require. Voluntary web-based surveys collect information about wages and working conditions, which are problematic areas in representative surveys where people do

[2] Created in 1993, EURES is a European 'job mobility' portal operated jointly by the European Commission and the Public Employment Services of the EEA Member States. Its purpose is to provide information, advice and recruitment/placement (job-matching) services for the benefit of workers and employers as well as any citizen wishing to benefit from the principle of the free movement of persons.

[3] Wageindicator project collects data about wages and working conditions through web-based platforms in a range of countries across the world. For more see: http://www.wageindicator.org/

not report their wages or where more detailed information about the working environment is absent. In developing countries where centralised statistical collection is less frequent, web-based surveys might be the only source of data about labour markets.

*Web-based data on labour market supply*

Research on the usage of web-based voluntary surveys appears to be methodologically more advanced. This type of data has developed a relatively strong following in the field of psychology, but it also has critics. The rapid growth of the Internet provides a range of opportunities for research in psychology, which researchers have been exploiting since the mid-1990s. In the field of psychology, data representativeness is less important than in other social science fields which rely on probabilistic sampling and inferential statistical methods. It has nevertheless also been characterised by certain 'preconceptions' related to the quality of the data, such as the lack of demographic diversity of Internet samples, social isolation and depression of Internet users, anonymity of participants, lack of motivation of respondents inter alia (Gosling et al. 2004).

In the area of labour market research, studies related to the WageIndicator project are relatively advanced with respect to testing data representativeness and ways to improve its usability for generalisable research, and also allowing comparisons across countries. The key concern behind the use of these data is the sampling error that can occur due to the fact that respondents are not randomly selected from a representative sampling frame, but the target population rather forms a 'convenience sample' of self-selected individuals.[4] This is closely related to the coverage error. The socio-economic and demographic characteristics of the pool of Internet users differs in important respects from those of the general population, e.g. the elderly or people with lower levels of education typically have lower access to the Internet.

[4] Probabilistic web-based surveys are also conducted where a proper sampling frame exists, which allows for drawing a probability-based random sample from a population in which every individual has the same probability of being selected. Examples include email requests, mixed-mode surveys or pre-recruited access panels of Internet users (Steinmetz, Tijdens, and García 2009).

Research results based on web-based data are therefore unable to provide inferences for these categories of people, or only with limited precision and reliability. An additional source of error is the non-response bias, which appears if non-respondents differ from the survey respondents in important characteristics, again making the inferences based on such data imprecise with respect to the overall target population.

Compared to online vacancy data, an important advantage of web-based individual voluntary surveys is the fact that the sample population's characteristics can typically be ascertained. Censuses or representative labour market surveys can be used to compare the characteristics structure of web survey respondents and potentially make adjustments. Two weighting techniques have been suggested to correct for the biases inherent in web surveys (Steinmetz, Tijdens, and García 2009). Post-stratification weighting can be applied to correct mainly for demographic differences in the data. Propensity score adjustment can also be applied to correct for socio-demographic as well as attitudinal or behavioural differences related to an individual's decision to take the survey. Results of these corrections found that un-weighted samples bring more consistent results between web-based survey data and representative samples than adjusted data samples. Comparisons also revealed that types of biases differ across countries and are often related to the overall income inequality in the country, and the strength of biases can vary across different variables (Pedraza, Tijdens, and Muñoz de Bustillo 2007; Steinmetz, Tijdens, and García 2009; Steinmetz et al. 2013). In developing countries, where representative surveys might not exist, web-based surveys can provide a valuable and unrivalled source for understanding these labour markets. Critiques of web-based data should also keep in mind that even probability-based samples face problems of self-selection, non-response or coverage.

### *Job advertisements as a source of data about labour market demand*

Job advertisements are the first step in a screening process that communicates an employer's views about an *ideal* candidate. While in the actual recruitment process only a subset of specified requirements might come into play in an employer's decision, they are nevertheless highly suggestive in identifying the desired skills and qualifications for a particular position.

Job advertisement research using online data sources is quite recent, but it has been preceded by studies based on printed job advertisements. These typically test large sociological theories and concepts, such as class, merit selection or characteristics of modern industrial societies. Jackson et al (2005) and Jackson (2007) study the importance of education in mediating social mobility and test the merit selection hypothesis in the UK society. No concern over a possible selection bias or the representativeness of overall labour market demand in the UK is expressed in the studies and the findings are discussed in a generalisable form. For example, Jackson (2007) took advertisements from national and local newspapers with a high circulation and the sample was chosen to be *representative of the range of occupations* in the occupational sectors. Findings were generalised to inform new trends in how occupational positions are allocated on the basis of meritocratic criteria, questioning classes as an appropriate unit of sociological research in modern industrial societies.

Dörfler and van de Werfhorst (2009) analyse the Austrian labour market and test the merit selection hypothesis over time, covering a time span of 20 years (1985, 1990 and 2005). They investigate nearly 1,000 newspaper ads and expand the operationalisation of education to include the field of study. Printed job advertisements enable retrospective historical analysis, which is not possible with surveys, which typically gather data at a particular point in time. The authors suggest that the problem of non-response, typical for survey research, is not present, although other biases might exist, such as underrepresentation of certain skill levels. They apply inferential statistical methods (multivariate regression analysis) to test inductively formulated hypotheses.

Studies using online job advertisements differ from those relying on printed job adverts typically in the number of ads they analysed, and sometimes in the techniques they are able to employ. A leading private internet recruitment site – Monster.com – has been used as a source of online job vacancy data in various studies. Capiluppi and Baravalle (2010) developed a study to investigate what is demanded of IT personnel in the UK and to what extent are the needed skills delivered by the universities. The authors developed a 'web spider' to download vacancies from the Monster.com website and then analysed the skills required. Backhaus (2004) analysed job advertisements from Monster.com from the perspective of employers. Using content analysis she studied corporate descriptions in job adverts to understand aspects of company branding and marketing in human resources, pioneering this type of research.

Among the most recent studies that have analysed a very large sample of job ads is the work of Kuhn and Shen (2013), who studied gender discrimination in the recruitment process in the Chinese labour market. Data were collected by a web-crawler from the third-largest online job portal in China. This led to a sample size suitable for sophisticated empirical analysis using advanced statistical methods. Their analysis of over a million job ads from the late 2000s, subsequently merged with firm data, revealed high levels of gender preference, although vacancies for highly skilled positions were less discriminatory. They acknowledge that their sample of job ads is not representative of the overall population of jobs in the Chinese labour market, using the 2005 census in China as a comparator dataset. They found that compared to the general demographics of employed persons in the same provinces, the analysed ads are aimed at workers who are much younger, better educated, better paid and work in the private sector. While acknowledging these limitations, authors present general findings about aspects of explicit gender discrimination in the Chinese labour market with implications for policy-making in developing countries more generally. Shen and Kuhn (2013) analyse job applications submitted in response to a selected number of job ads online in Chinese urban areas to study what effect over-qualification has on labour market entrants. Representativeness is not considered, and only the bias related to duration of vacancy posting is addressed by additional analysis. Sophisticated statistical tools are employed and the findings are generalised for Chinese urban youth.

Štefánik (2012a, 2012b) studied online data from a private job portal in Slovakia, analysing both vacancies and CV data. His study concentrated on the labour market segment of the highly skilled and examined the matching of demand and supply of university graduates, concentrating on a small number of narrowly defined highly-skilled professions. He focused on university graduates, in the belief that they would be relatively well represented among the job applicants due to characteristics of Internet users and would therefore offer a representative data sample. His representativeness test was based on comparing the structure of portal vacancies and CVs to the structure of the whole population based on the national Labour Force Survey. He then excluded occupations that were not equally represented in online portal datasets and the EU Labour Force Survey (LFS) operated by Eurostat, and only studied two relatively narrow job profiles to compare what was demanded by employers in these positions and which skills the applicants cited in their CVs. Kureková, Beblavý, and Haita (2012) use the same private online job portal data to study demand for formal qualifications and other skills in a wide range of low- and medium-skilled occupations. They consider their findings generalisable for the Slovak context due to the dominant market share and very high reputation of the portal among employers and employees.

Wider research attention has been given to IT-related professions, which have been on the rise in the past decades. Wade and Parent (2001) study the relationship between job skills and performance of webmasters looking at job vacancy data and complementing this analysis by a targeted web-based survey of webmasters to determine the required skill mix and the degree to which subjective assessments of the possession of skills affect job performance in this occupation. Vacancies were gathered from two trade journals and five online job search indices, producing a total of 800 job descriptions. These data were combined with webmaster survey responses to employ multivariate analysis techniques. The authors highlight the usefulness of researching online job ads in identifying the mix of skills sought after, especially in new professions, and in building profiles of positions, which can serve as valuable input to student counselling services or curricula development. They acknowledge that the  coverage bias can be remedied by complementing their methodology with structured interviews with employers or recruiters.

Huang et al. (2009) examine technical, humanistic and business IT skills across three genres of text: scholarly articles, practitioner literature and online job ads. In order to construct reliable profiles of job positions, study finds that it might be useful to review a wider range of sources.

Comparative studies based on job ads data are rather scarce. An exception is the work of Kennan et al. (2008), who study changing workplace demands for information specialists, looking specifically at the librarian profession to determine the required skills in Australia and the US. They build on other studies that investigated this profession, including longitudinal examinations of new trends in skills and characteristics. The study combines data from printed ads and online ads. Based on a 'dictionary' of skill categories and terms the authors develop, the study calculates frequencies, ranks the groups of skill categories and by means of cluster analysis identify 'skill clusters'. The study finds differences in the relative importance of skills across the two countries, and also variations over time. The results are presented as generalised for the librarian profession under study.

Kureková et al. (2013) pioneer the usage of the European-wide publicly administered job-vacancy portal EURES and carry out a comparative analysis of employers' skill demand in three small European economies (Czech Republic, Denmark and Ireland). They find that the mix of skills called for is very diverse across the three countries, implying that there is no universal set of requirements and also that domestic institutions and structures strongly affect how demand is formulated. The authors argue that EURES data are a well-suited source for comparative analysis due to their standardised platform and relatively wide usage across European countries (see also Ackers 2012).

To sum up, existing research using online or printed job vacancies has been characterised by a single-country focus and has grown mainly through the usage of data from private online job portals rather than publicly collected data (e.g. public employment services data). Various types of questions have been investigated ranging from testing or enriching established social science concepts and theories (gender discrimination, social stratification, merit selection, expansion of service sector and company branding) to relatively narrow and focused questions related to a particular sector or industry (IT sector, librarian profession). Interestingly, methodological concerns are acknowledged, but they are not given prominence in the reviewed studies. In the following section, we synthesise and critique the various approaches taken by different authors in attempting to deal with representativeness issues related to the usage of online job vacancy data for analytical and policy-making purposes.

## Existing approaches to increasing representativeness

It is not a trivial task to analyse the representativeness of findings from studies based on job advertisement data from online portals and to correct for their shortcomings. A key challenge in using online job vacancies is ascertaining whether the set of online job vacancies is a representative sample of all job vacancies in a specified economy.[5] Even if it should be

[5] Throughout the following text we assume that we can process all online job vacancies available, but this is obviously not the case for large countries. For simplicity purposes we do not consider this case however in this theoretical study.

considered finite,[6] the population of job vacancies at any given moment in time is not easily counted nor is its structure easy to determine. In only a handful of countries are employers obliged to report all job vacancies.[7]

And even if such reporting of vacancies (typically to the labour office) is mandatory, much hiring takes place internally or through informal means and networks. Jobs are reallocated in the labour market through many mechanisms, some of which do not entail a formal 'vacancy' announcement: people are reallocated internally; or tasks are split, restructured or partially switched. This is likely to affect certain aspects of the labour market more than others. For example, large international firms often first recruit from internally available candidates before announcing a vacancy in an open labour market. In smaller towns or villages where labour market participants have closer links and relations, job vacancies might first be offered to candidates known personally to the employer. Recruitment means and strategies might have sectoral and occupational specificities and can vary across countries (Teichler 2009; Keep and James 2010).

From this perspective, even if we collect all online reported job vacancies, there is a share of vacancies that are never publically advertised, and will therefore fall outside our population sample. On the other hand, vacancies tend to occur where there is either an unfulfilled demand or where employers for some reason prefer to have a selection process. Therefore, if we are interested to know which types of jobs employers find difficult to fill through internal or informal search channels, then online vacancy data can be highly useful. Due to the difficulties in identifying the structure of the *population* of vacancies, however, we argue that weighting as an adjustment technique that has been tried in improving the representativeness

[6] A finite population is one for which it is possible to count its individuals or units.

[7] Denmark is an example of a country in which the public authorities are legally obliged to report all job vacancies online.

of web-based voluntary surveys is hardly possible in the context of projects based on online job vacancy data.

The reviewed studies have – implicitly or explicitly – adopted rather different approaches to deal with the problem of (non-)representativeness of job vacancy data. Some researchers have used representative data describing the labour market structure, such as the Labour Force Survey (LFS), and judged the coverage of online vacancies based on the sectoral and occupational structure of LFS data (Štefánik 2012b; Štefánik 2012a; Jackson 2007). We find this approach problematic, however, for a number of reasons. Foremost, the Labour Force Survey is not a straightforward measure of the structure of the demand side of the labour market but rather includes supply, demand and job matches. We therefore do not see the LFS as a suitable proxy for the demand side. Furthermore, current demand is subject to seasonal trends and reflects developments in a particular sector that might not match the existing structure but rather reflect future trends in a particular labour market segment. For example, vacancies in the IT industry in many countries are due to the recent rapid expansion of this sector overrepresented among vacancies, and their share might not reflect the actual share of the industry in a national structure of employers and/or employees.

A more promising direction is the one taken by Van Ours and Ridder (1992), who tried to achieve representative results in their study of the duration of vacancies in order to determine aspects of employers' search strategy by selecting a 5% random sample from all establishments in the Netherlands. They faced attrition in the process of receiving responses from the firms they approached with a two-stage questionnaire (the first stage identified firms that were hiring, the second stage investigated aspects of the search strategy and the characteristics of the hired person). Although the authors adopted a rigorous sampling framework, they were not able to deal fully with all representativeness problems due to non-responses and other forms of data-collection problems encountered. Moreover, such a research approach is obviously also expensive and time-consuming. The availability of online vacancies could potentially simplify some steps of their research process, but this needs to be evaluated and weighed against particular research questions and objectives.

Alternative approaches have been adopted to (partially) address the representativeness issue, building on the diversification of data sources. First, a number of studies decided to focus on the segment of the labour market where the coverage bias is likely to be less of a problem.[8] Examples include focusing on graduates' CVs and vacancies targeting this labour market segment or on sectors and professions that are by definition characterised by widespread access to the Internet (IT, librarians, webmasters, etc.). Second, diversification of data sources has been used in many studies. In addition to online job ads, other types of data sources can be analysed in parallel, such as practitioner literature or administrative data. A sample of vacancies based on administrative data can be created. Eurostat (2010) gives an overview of several such examples from European countries. For instance, Martikainen (2010) uses the Finnish Business Register to create a representative sample of job vacancies by weighting the observations in an appropriate way and using a suitable estimation technique that takes potential flaws into account. Data collection was done through computer-aided interviews and through the Internet. Around 10,000 establishments out of 150,000 were drawn by stratified sampling per year. Another example of an alternative approach takes the form of complementing content analysis of job advertisements with interviews with HR managers or recruiters. Third, some scope to correct possible biases might exist if online vacancies data can be linked to firm characteristics, which can then be controlled in quantitative analysis. Fourth, market coverage and technical advancement of the online job portal(s) in a given country need to be assessed in each country. In countries where a dominant portal exists,

[8] This technique is in a way similar to stratified sampling, in which random samples are drawn from different categories that are pre-defined by the researcher. Eurostat (2010) presents this methodology in the context of job vacancies.

collected vacancies might be the best available source. These alternatives need to be weighed against the costs of collecting data by other means. Using job advertisements from an established portal and interpreting the results with caution to avoid potential biases can be a valid and acceptable choice.

### Statistical methods to address sample design problems

Another way to address the problem of representativeness of online job vacancy data *as an accurate sample* of job vacancy data is to employ statistical tools that account for sample design problems. A common type of data used for analysing employers' preferences is survey data (see Colombo 2009), and statistical tools have been developed to address estimation problems arising from the survey design.[9] We argue that some of these tools and potentially their variations can be suitably used to address problems arising from the way the respective online data has been collected.[10]

These tools can be situated in the field of statistical analysis with missing data, which is a useful field of literature to understand how to estimate population parameters in the most accurate way concerning samples that are likely to not be random. Data can be missing either because of an accidental omission and the reason for the omission is due to a variable unrelated to the variable with the missing values ('missing at random') or because of a specific reason that is related to the variable with the missing value ('not missing at random'). In our case, a job vacancy could be missing in the set of online job vacancies either because it was not posted online or because it was not advertised at all. Both reasons would point to considering our data as containing not missing at random (NMAR) values. In fact, in our case

[9] With the rise of behavioural economics, data from experiments are also gaining in importance. But since this is a new field and we are interested in established methods to study sample design issues, we will not further review data gathered through experiments.

[10] We are thankful to Mr. Nicholas Sofroniou, Expert at CEDEFOP, Thessaloniki, for this idea.

the missing values are due to the sampling procedure: online job vacancies are a sample of the population of all job vacancies in which those values are missing that have not been posted online. If the sampling procedure is under the control of the statistician, it can be addressed. However, if it is not under his or her control, the assumptions made about the mechanism leading to the missing data need to be made clear (Little and Rubin 1987). We demonstrate below how a model could be constructed that accounts for over- or underrepresentation that may arise from the selection mechanism.

Little and Rubin (1987, 2002) provide a comprehensive overview of statistical analysis with missing data and distinguish four main methodological schools in addressing missing-value issues in statistics. We will focus on a less frequently used approach:[11] the model-based approach. This method builds on the idea that the estimation of a population mean from a sample is a similar problem to the *prediction* of a population mean (Royall 1992). Several statistical works present the main features, limitations and advantages of the model-based approach (Chambers and Skinner 2003; Valliant, Dorfman, and Royall 2000; Aitkin and Aitkin 2011; Longford 2005). The idea is to predict – through an underlying probability model[12] – quantities based on data that include unobserved values. We fill in the unknown parts of a data distribution by using our knowledge of the subject matter to construct a model of how that missing data could be determined. Such a model could take the form of a density of the variable in question, conditional on *i)* a set of other variables representing information used in the survey design and *ii)* a set of parameters. We argue that it could be a useful method in the context of analysing online job vacancy data, where adjustments need to be made to allow for variables correlated with over- or under-representation of certain types of observation units.

[11] Little and Rubin (1987) also add procedures based on completely recorded units (analysing only observations with complete data), as well as design-based weighted estimation and imputation-based procedures (the missing values are filled in) to their taxonomy of methods with partially missing data.

[12] A probability model is a mathematical representation of the probability of the occurrence of an event, where an event can be the realisation of a random variable.

For the purpose of understanding employers' preferences with the help of job vacancies, the information we need from job vacancies are certain characteristics about the jobs that are advertised, such as the distribution of the need for a certain skill. It is these characteristics of occupations that are being adjusted for, such as required skills, rather than the number of each occupation *per se*. While we might expect the characteristics of a particular type of job to vary somewhat with establishment characteristics such as size, it seems reasonable to suggest that the core aspects of a given occupation are likely to remain constant across types of establishments and it is the latter conditional relationship that we are trying to compensate for.

In order to infer the social-skill needs of a nurse from a sample consisting of online vacancies – which we believe to under-represent certain parts of the population of vacancies – we would construct a conditional distribution of skill needs given a certain number of covariates. How do we choose the variables to include in the set of covariates? We might believe that only nurse vacancies in hospitals with more than 10 employees are advertised. We might know from a set of interviews with doctors or hospital directors that the skill needs for a hospital with less than 10 employees differ from those of larger hospitals: in small hospitals more social skills are needed due to higher interaction with the patients. In that case we would not be able to make correct inferences on social-skill needs of all vacancies based on a sample excluding skill needs for the smallest hospitals. We could then first assess the extent of the bias by retrieving the number of small hospitals compared to larger hospitals via external data such as an establishment panel, which contains a representative sample of companies. We might find out that the percentage of small hospitals is 15%. Our skill needs based on the online sample would consequently exclude 15% of the hospitals, in which social-skill needs are highest. We can think of the sample distribution of social-skill needs as truncated. We can now establish the trend of nurses' social-skill needs given the hospital size. In our case it would be a downward sloping curve in hospital size. To build the model that we will use to predict the social-skill needs of nurses, we would extrapolate this trend to the smallest hospital size and add a further upward sloping bit to the curve at the end of the smallest size.

Now we may ask ourselves whether hospital size is the only variable on which the online sample is underrepresented or overrepresented. To select these variables, we would need to study very carefully which vacancies we believe are placed online and which ones are not. Such variables could be region, company size, "openness of the country/ratio between tradable and non-tradable goods" (to account for whether the labour market is likely to be open for publishing vacancies on the globally available internet) or "importance of social capital in the country/occupational field" (to account for job vacancies that are not posted online because they are filled by connections). We could include all variables in our model and select the best-fitting model via model selection criteria such as the Akaike Information Criterion (AIC), which is a statistical tool to assess the quality of a statistical model.

Longford (2005) acknowledges that this approach is based on un-testable assumptions and therefore proposes to conduct a sensitivity analysis. His understanding of a sensitivity analysis is a tool to measure the impact of an assumption on the result of the analysis. He suggests in the case of not missing at random values to test sensitivity of results in one or several directions since the vast array of possible directions is not feasible to account for (Longford 2005, 79). The way to implement such a sensitivity analysis in our case would be to simulate our model for cases with more or less large companies and see how the distribution of skill needs changes.

The estimation would be conducted based on maximum likelihood or Bayesian methods. The likelihood function would be set up and it would take a certain functional form based on the assumptions about the error terms. In complex cases, the function might not be numerically tractable, in which case one would use numerical integration methods based on Bayesian or frequentist statistical methods. [13]

The limitation with this approach would be that i) we do not know how much of the missing data in the population we do not predict since we do not know the population size, and ii) we cannot predict data for which there is no information to base a model on. This problem, however, also needs to be addressed for wage assessments that do not include the black market segment.

## Conclusions, implications and suggestions

Research based on new sources of data and innovative research methods related to the spread of the Internet has been on the rise. These trends have also affected possibilities of conducting research on the labour market with respect to both its supply and demand side. This paper has critically reviewed the existing literature to summarise the various ways in which researchers have been dealing with methodological issues related to web-based sources of new data and specifically the key problem of data representativeness and generalisability of their findings. We also suggest statistical methods for dealing with missing data as a tool to estimate population parameters in the most accurate way, when the samples concerned are likely not to be random, such as the case of online job vacancies.

We find the current debate to be flawed by its failure to acknowledge that every exercise in data collection – including the census – has its limitations. Interestingly, in most studies analysing printed or online job advertisements, representativeness issues are not widely discussed and the findings and conclusions are presented in a generalising manner. We propose that rather than dismissing out of hand research efforts using online job advertisement

[13] Bayesian methods can also be used with a frequentist view; see for instance Carneiro et al. (2003).

and other types of web-based data due to weaknesses of data representativeness, a debate should be launched on how these weaknesses can be compensated for and for which types of research questions and fields they might be of a smaller concern.

We highlight that adjusting for representativeness is a particularly formidable task with respect to online job vacancy data. This is due to the fact that the population of job vacancies and its structure is practically unknown. For this reason, adjustment methods, such as weighting, that have been tested as a means of improving web-based voluntary surveys cannot be used as a technique for adjusting online job vacancy data.

Based on the review and synthesis of existing studies and broader statistical approaches to sample correction, we would also like to offer more general recommendations with respect to the usage of online job vacancies for future research. First, the representativeness and reliability of the data source used need to be evaluated at the country level. Dominant market share can be considered as an adequate source of data that can lead to reliable and transferable research results. Second, representativeness and reliability need to be assessed vis-à-vis a particular research focus. The useof a data segment or sub-sample that can be considered (more) representative can address certain aspects of coverage and sampling errors. Examples we encountered focused on professions or labour market segments that are highly exposed to the Internet (web-designers, graduate labour market), where these biases are expected to be less pronounced. Third, depending on the particular research focus, online job ads could be coupled with other sources of vacancy data or text describing analysed professions. Fourth, more sophisticated statistical methods anchored in the literature on missing data, such as the model-based approach to the correction of data not missing at random, could also be used to adjust biases stemming from the structure of online vacancy data.

The good news for this methodological debate is that Internet-based job searching is likely to become an increasingly more prominent tool for job matching, which promises to improve the coverage of the population of workers and firms that engage in it. Recent studies evaluating the quality of online job searching and matching already find a positive impact (Mang 2012; European Commission and ECORYS 2012). Compared to traditional employment channels (newspapers, friends and agencies), online job portals are able to provide a wider range of choice as well as increasingly more advanced tools to evaluate the suitability of a job or a job candidate. An especially positive impact was found for workers with interruptions in their employment history, distancing them from the labour market, such as mothers. This has important implications with respect to attempts to re-activate disadvantaged groups in various labour markets. Labour market policy can be enriched by a better understanding of what employers need, which in turn can be used to inform job counselling services,  second-chance education and training, as well as integration of disadvantaged jobseekers in the labour market (Keep and James 2010; Kureková et al. 2013). Other public policy areas could benefit from the usage of web-based data about labour markets. Examples include social policy, in particular aspects of labour market discrimination, or education and training sector which could incorporate online job vacancy information into curricula development. The motivation to pursue innovative sources of data and analytical methods and to improve the reliability of the results is not only academic, but also driven by the practical benefits they may yield.

## References

Ackers, Doede. 2012. “The Experience of EURES. Improving Access to Labour Market Information for Migrants and Employers.” High Level Conference, European Commission DG Employment Social Affairs and Inclusion Unit C4, June 11. www.labourmigration.eu/events/document/163?format=raw.

Aitkin, Murray, and Irit Aitkin. 2011. *Statistical Modeling of the National Assessment of Educational Progress*. Springer.

Askitas, Nikos, and Klaus Zimmermann. 2009. “Google Econometrics and Unemployment Forecasting.” *IZA Discussion Papers* No. 4201. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1465341.

Backhaus, Kristin B. 2004. “An Exploration of Corporate Recruitment Descriptions on Monster.com.” *Journal of Business Communication* 41 (2): 115–36. doi:10.1177/0021943603259585.

Capiluppi, Andrea, and Andres Baravalle. 2010. “Matching Demand and Offer in On-Line Provision: A Longitudinal Study of Monster.com.” In . Romania. http://roar.uel.ac.uk/995/.

Carneiro, Pedro Manuel, Flavio Cunha, and James J. Heckman. 2003. “Interpreting the Evidence of Family Influence on Child Development.” In *The Economics of Early Childhood Development: Lessons*. Minneapolis Federal Reserve Bank: The Federal Reserve Bank.

Chambers, Ray L., and Chris J. Skinner. 2003. *Analysis of Survey Data*. John Wiley & Sons. http://books.google.com/books?hl=sk&lr=&id=4pYGz69d-LkC&oi=fnd&pg=PR7&dq=Analysis+of+Survey+Data.&ots=8jgNcYGiXf&sig=mb54nLZzS-ibv8alSsBsPk-UC8E.

Colombo, Emilio. 2009. “Measuring Skill Needs through Employers’ Surveys: Problems and Methods - Colombo Presentation.pdf.” presented at the Agora conference, CEDEFOP, Thessaloniki, June 11. http://agora.cedefop.europa.eu/files/presentations/Colombo%20%20presentation.pdf.

D’Amuri, Francesco, and Juri Marcucci. 2010. *“Google It!” Forecasting the US Unemployment Rate with a Google Job Search Index*. Nota di lavoro//Fondazione Eni Enrico Mattei: Global challenges. http://www.econstor.eu/handle/10419/43536.

DeKay, Sam H. 2013. “Peering Through Glassdoor. Com What Social Media Can Tell Us About Employee Satisfaction and.” In *CONFERENCE ON CORPORATE COMMUNICATION 2013*, 45. http://www.corporatecomm.org/pdf/Abstracts_Proceedings_CCI_CCC_2013.pdf#page=60.

Dörfler, Laura, and Herman G. van de Werfhorst. 2009. “Employers’ Demand for Qualifications and Skills.” *European Societies* 11 (5): 697–721. doi:10.1080/14616690802474374.

Edelman, Benjamin. 2012. “Using Internet Data for Economic Research.” *The Journal of Economic Perspectives* 26 (2): 189–206. doi:10.2307/41495310.

European Commission, and ECORYS. 2012. *European Vacancy and Recruitment Report 2012*. Luxembourg: Publications Office of the EU.

Eurostat. 2010. *1st and 2nd International Workshops on Methodologies for Job Vacancy Statistics - Proceedings*. Methodologies and Working Papers. Luxembourg, Publications Office of the EU: European Union. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-10-027/EN/KS-RA-10-027-EN.PDF.

Fabo, Brian, and Kea Tijdens. 2014. "Using Web Data to Measure the Demand for Skills." *CELSI Discussion Paper no.21*. http://www.celsi.sk/media/discussion-papers/celsi-dp-021.pdf.

Gosling, Samuel D., Simine Vazire, Sanjay Srivastava, and Oliver P. John. 2004. "Should We Trust Web-Based Studies." *American Psychologist* 59 (2): 93–104.

Huang, Haiyan, Lynette Kvasny, K. D. Joshi, Eileen M. Trauth, and Jan Mahar. 2009. "Synthesizing IT Job Skills Identified in Academic Studies, Practitioner Publications and Job Ads." In *Proceedings of the Special Interest Group on Management Information System's 47th Annual Conference on Computer Personnel Research*, 121–28. http://dl.acm.org/citation.cfm?id=1542154.

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

Jackson, Michelle, J Goldthorpe, and C Mills. 2005. "Education, Employers and Class Mobility." *Research in Social Stratification and Mobility* 23: 3–33. doi:10.1016/S0276-5624(05)23001-9.

Keep, Ewart, and Susan James. 2010. *Recruitment and Selection – the Great Neglected Topic*. SKOPE Research Paper No. 88. Oxford: Oxford University. http://www.cf.ac.uk/socsi/research/researchcentres/skope/publications/researchpapers/SKOPEWP88.pdf.

Kennan, Mary Anne, Fletcher Cole, Patricia Willard, Concepción Wilson, and Linda Marion. 2006. "Changing Workplace Demands: What Job Ads Tell Us." *Aslib Proceedings* 58 (3): 179–96. doi:10.1108/00012530610677228.

Kennan, Mary Anne, Patricia Willard, Dubravka Cecez-Kecmanovic, and Concepción S. Wilson. 2008. "A Content Analysis of Australian IS Early Career Job Advertisements." *Australasian Journal of Information Systems* 15 (2). http://dl.acs.org.au/index.php/ajis/article/viewArticle/455.

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

Kureková, Lucia, Miroslav Beblavý, and Corina Haita. 2012. "Qualifications or Soft Skills? Studying Demand for Low-Skilled from Job Advertisements." *NEUJOBS Working Paper No. 4.3.3.*

Kureková, Lucia, Miroslav Beblavý, Corina Haita, and Anna-Elisabeth Thum. 2013. "Demand for Low-Skilled Workers across Europe: Between Formal Qualifications and Non-Cognitive Skills." *NEUJOBS Working Paper No. 4.3.3.*

Little, Roderick JA, and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons. http://www.lavoisier.fr/livre/notice.asp?id=OKLWRRAS2SXOWM.

———. 2002. *Statistical Analysis with Missing Data*. 2nd Edition. New York: John Wiley & Sons.

Longford, Nicholas T. 2005. *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician*. Springer.

Mang, Constantin. 2012. *Online Job Search and Matching Quality*. Ifo Institute for Economic Research at the University of Munich. ftp://ftp.zew.de/pub/zew-docs/veranstaltungen/ICT2012/Papers/Mang.pdf.

Martikainen, J. 2010. "Weighting and Estimation Methods: JVS Estimation in Finland by Horowitz-Thomson-Type Estimator." In , edited by Eurostat. Luxembourg, Publications Office of the EU: European Union.

http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-10-027/EN/KS-RA-10-027-EN.PDF.

Pedraza, Pablo de, Kea Tijdens, and Rafael Muñoz de Bustillo. 2007. *WP 60-Sample Bias, Weights and Efficiency of Weights in a Continuous Web Voluntary Survey*. AIAS, Amsterdam Institute for Advanced Labour Studies. http://ideas.repec.org/p/aia/aiaswp/wp60.html.

Royall, Richard M. 1992. "The Model Based (prediction) Approach to Finite Population Sampling Theory." In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, edited by Malay Ghosh and Pramod K. Pathak. Hayward, CA: Institute of Mathematical Statistics.

Sappleton, Natalie. 2013. *Advancing Research Methods with New Technologies*. Edited by Natalie Sappleton. 1st ed. IGI Global.

Shen, Kailing, and Peter Kuhn. 2013. "Do Chinese Employers Avoid Hiring Overqualified Workers? Evidence from an Internet Job Board." *Research in Labour Economics*, no. forthcoming. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2157990.

Štefánik, Miroslav. 2012a. "Internet Job Search Data as a Possible Source of Information on Skills Demand (with Results for Slovak University Graduates)." In *Building on Skills Forecasts — Comparing Methods and Applications*, edited by CEDEFOP. Luxembourg: Publications Office of the European Union. http://www.cedefop.europa.eu/EN/Files/5518_en.pdf.

———. 2012b. "Focused Information on Skills Demand Using Internet Job Search Data (with Results for Slovak University Graduates)." Institute of Economic Research, Slovak Academy of Sciences. http://doku.iab.de/fdz/events/2012/Stefanik.pdf.

Steinmetz, Stephanie, Damian Raess, Kea Tijdens, and Pablo de Pedraza. 2013. "Chapter 6: Measuring Wages Worldwide: Exploring the Potentials and Constraints of Volunteer Web Surveys." In *Advancing Research Methods with New Technologies*, edited by Natalie Sappleton, 1st ed. IGI Global.

Steinmetz, Stephanie, Kea Tijdens, and Pablo Pedraza García. 2009. *WP 76-Comparing Different Weighting Procedures for Volunteer Web Surveys*. AIAS, Amsterdam Institute for Advanced Labour Studies. http://ideas.repec.org/p/aia/aiaswp/wp76.html.

Taylor, Linnet, Ralph Schroeder, and Eric Meyer. 2014. "Emerging Practices and Perspectives on Big Data Analysis in Economics: Bigger and Better or More of the Same?" *Big Data & Society* 1 (2): 2053951714536877. doi:10.1177/2053951714536877.

Teichler, Ulrich. 2009. *Higher Education and the World of Work*. Rotterdam: Sense Publishers.

Valliant, Richard, Alan H. Dorfman, and Richard M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons. http://www.lavoisier.fr/livre/notice.asp?ouvrage=1364304.

Van Ours, J., and G. Ridder. 1992. "Vacancies and the Recruitment of New Employees." *Journal of Labor Economics*, 138–55.

Wade, Michael R., and Michael Parent. 2001. "Relationships between Job Skills and Performance: A Study of Webmasters." *Journal of Management Information Systems* 18 (3): 71–96. doi:10.2307/40398554.

Young, Kimberly S., and Carl J. Case. 2004. "Internet Abuse in the Workplace: New Trends in Risk Management." *CyberPsychology & Behavior* 7 (1): 105–11.