Webscraping for Job Vacancy Statistics

Mr. Nigel SWIER

Principal Methodologist, Office for National Statistics, United Kingdom

Abstract. Official Job vacancy statistics with the European Statistical System are produced mainly through traditional surveys but the information available is generally quite limited. The relative accessibility of on-line job advertisements provides an opportunity to produce statistics with additional variables that are more granular, timely and more frequent. This would lead to statistics that are more useful for policy-making. However, there are significant methodological challenges. There are complex relationships between job advertisements and the vacancies they relate to. For example, a job vacancy many be advertised many times on different job portals, while a single advertisement may relate to more than one vacancy. Many jobs are not advertised on-line at all, and so a key challenge is to understand the coverage of on-line job vacancies and the target population. This points to an ongoing need for surveys and for new estimation methods for integrating data from different sources.

1. Introduction

There is increasing interest from National Statistics Institutes (NSIs) in exploring how big data could be used for official statistics [1, 2]. There is a particular interest in exploring the potential of on-line job advertisements due to their accessibility and their relevance for policy-making. Securing access to relevant big data sources presents some fundamental challenges for NSIs, even just in terms of getting started in the world of big data. Therefore on-line job advertisements provide a useful departure point for NSIs in developing the skills, technology infrastructure and methods needed to incorporate big data into official statistics.

A pilot on web scraping for job vacancy statistics is being taken forward as part of a Big Data ESSNet project funded by Eurostat. This work started in February 2016 and is due to finish in May 2018. The pilot is being led by the United Kingdom with Germany, Greece, Italy, Slovenia and Sweden as partners. Additional partners (Belguim, Denmark, France and Portugal) will join in July 2017. This paper focuses on the high level challenges in using on-line job advertisements for official statistics purposes.

2. Job Vacancy Statistics

Job vacancy statistics within the ESS are subject to EC regulation No. 453/2008 [3]. This defines a job vacancy as:

"… a paid post that is newly created, unoccupied, or about to become vacant:

(a) for which the employer is taking active steps and is prepared to take further steps to find a suitable candidate from outside the enterprise concerned; and

(b) which the employer intends to fill either immediately or within a specific period of time."

EC regulation 453/2008 has several mandatory elements:

- Quarterly data that has been seasonally adjusted
- Data broken down economic activity
- Data is relevant and complete, accurate and comprehensive, timely, coherent, comparable, and readily accessible to users.

There are other elements that are optional, or subject to feasibility, including:

- Job vacancies in the agriculture, forestry and fishing sectors
- Job vacancies in public administration, defence and education
- Data on businesses with less than 10 employees
- Distinguishing between fixed term and permanent jobs.

The member states are granted considerable flexibility regarding the implementation of regulation 453/2008 in national statistical systems. Some countries use stand alone surveys, while others combine the job vacancy survey with other business surveys. Some collect the minimum information required by the regulation while others collect more. Although the regulation states that the data shall be collected using business surveys, the use of administrative data is equally permitted under the condition that the data are "appropriate in terms of quality" (according to the quality criteria of the ESS).

The regulation is focused on providing statistics for improving the functioning of labour markets at the EU level and so there is limited scope for undertaking more detailed analyses. In particular, there is a general lack of data on job types and associated skills, as well as on the location of new job vacancies. This level of detail is generally difficult to obtain using traditional business surveys, but is readily available on-line and so can be obtained (notwithstanding the legal aspects of web-scraping) without placing additional burden on respondents.

3. Job Vacancies and On-Line Data Sources

The use of the internet as a channel for advertising job vacancies is well established in many countries. A company advertising a job vacancy on-line may advertise on their own website, through a job portal website, and very often on both. In addition, a job advertisement uploaded to a job portal may be republished or indexed by another portal. This pilot uses the following typology to distinguish different types of job portals [4]:

i) *Job Boards*: Job portals that host original advertisements.

ii) *Job Search engines*: Job portals that index and republish existing job advertisements.

iii) *Hybrids*: Job portals that both host original advertisements and republish existing ones.

In addition, some portals may specialise in certain types of jobs, or in particular regions (e.g. regional on-line newssites), while other portals are more generic. In many countries, the most important on-line portal is that of the national public employment agency, but there are also many portals operated by private sector companies. In Germany alone there are an estimated 1,600 different job portals [5]. Therefore, any approach around web scraping job portals requires consideration of both the coverage of on-line job vacancies and what data is available. A complicating factor is that many jobs advertised on portals are placed by employment agencies on behalf of an employer. This makes it difficult to relate these vacancies to the correct industry sectors as required by the EU regulation.

Other possible data sources include commercial data suppliers already involved in the business of web scraping and processing on-line job advertisements for the purposes of creating data products and services (e.g. Wanted Analytics, Textkernel). In addition, the ESSNet pilot has negotiated access to data to a pilot system for analysing on-line job advertisements developed by the European Centre for Vocational Training (CEDEFOP). Given that different public agencies are interested in using the same kind of data, it makes sense to collaborate wherever possible.

This pilot is also interested in the feasibility of collecting information about job ads directly from enterprise websites. This has close links with a separate pilot within the Big Data ESSNet, which aims to develop a general framework for collecting information from enterprise websites. This involves using creating a list of enterprise website URLs linked to business registers, applying web crawling techniques to capture web pages, and text mining to extract relevant content. This approach could be used, not just for capturing information about job vacancies, but also other information (e.g. enterprise activities, use of social media, engagement in e-commerce). This has the advantage of building an explicit link with the business register, thus facilitating integration with

other data sources and avoiding problems of duplication. However, this approach has its own very significant challenges as it requires a framework where the structure of the website is not pre-defined. In contrast, a job portal website has a defined structure and so a good amount of structured data can be collected quite easily.

In conclusion, on-line job advertisements should not be considered in terms of a single source, but rather as part of a complex data eco-system involving enterprises advertising vacancies on their own websites, job portals, employment agencies, public bodies, and data analytics companies.

4. Data Access.

This pilot has involved a number of different approaches for accessing data:

3.1 Web scraping:

This involves building web scraping robots that rip data directly from the web pages of job portals. This has mainly involved the use of simple 'point and click' web scraping tools, such as Import.IO. These tools may not be suitable for large scale web scraping but have proved to be perfectly adequate for research purposes.

Work has not yet started on web scraping directly from enterprises, mainly because these approaches are still being developed elsewhere in the ESSNet. There are also important legal issues to resolve since many website prohibit web scraping as part of their terms and conditions and it is simply not viable to check the conditions of thousands of websites prior to implementing a mass web scraping approach. Proposals to tackle this issue include publishing lists of target enterprises on the NSI website, applying strict Netiquette protocols, and leaning on "helpful" aspects of statistical codes of practice (e.g. minimising respondent burden). However, the specific approaches in each country will be shaped by relevant statistical law and possibly also the risk appetite of each NSI.

3.2. Application Programming Interfaces (APIs):

This involves using APIs provided by a job portal owner to obtain data from an underlying database rather than from the webpage. This approach is easier and more robust than scraping web pages directly. It is also considered safer from a legal perspective. However, APIs tend only to be available for the very large job search engines where business models revolve around indexing and republishing job advertisements.

3.3. Direct Access:

This involves securing data access agreements with organisations that already hold job vacancy data. The main examples within this pilot are data sharing arrangement with the Swedish Government Employment Agency or "Arbetsformedlingen" and with CEDEFOP. There has also been some initial discussion around procuring data from data analytics companies but this has not been pursued further.

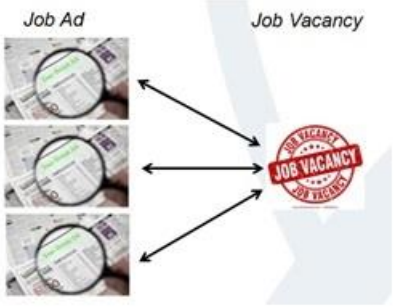
There are some important advantages with this more formal approach including avoidance of legal ambiguity and also the opportunity to gain access to time series data. There is an increasing recognition within this pilot of the benefits of this type of arrangement and there are future plans to focus more on exploring partnership models with the holders of job vacancy data.

5. Job Vacancies versus Job Advertisements

There is an important distinction between a *job vacancy* and a *job advertisement*. The formal definition of a job vacancy has already been defined in Section 2, but in simple terms can be understood as the vacant position within an enterprise that the enterprise is trying to fill. An on-line job advertisement is defined as the text on a website displaying information about a job vacancy. Thus, the job vacancy is the concept that we are aiming to measure while an online job advertisement is an artefact that indicates the existence of a job vacancy within an enterprise. However, there is very rarely a one-to-one relationship between the two concepts and this presents some formidable challenges for incorporating on-line information into job vacancy statistics.

Very often, a job vacancy will be advertised on a number of different portals (Figure 1). This may be because an advertisement uploaded on to one portal is republished by others, or because the employer uploads the advertisement multiple times. Thus, deduplication is key challenge for processing job portal data for official statistics purposes.

Figure 1: Relationship between job advertisements and a job vacancy



It is also possible for a job advertisement to contain more than one vacancy. In these case it may be possible (albeit challenging) to develop a methods for parsing job advertisements to extract this information and adjust the count of vacancies. There may even be instances where an advertisement appears to be advertising a single vacancy, but actually there is more than one. A further issue is is that of "ghost vacancies". These are job advertisements that do not advertise a genuine vacancy. It is known that employment agencies may advertise a vacancy as a means of enticing jobseekers to submit their resumes so that they can increase the number of job seekers they can offer to prospective employers to fill some future vacancy.

In summary, it is important to understand that the relationship between a job advertisement and a job vacancy is complex - producing official statistics about job vacancies is not just a question of counting the number of job advertisements.

6. Conceptual Model of Online Job Vacancy Coverage

As well as the challenge of untangling the relationship between online job advertisements and the vacancies they refer to, there is also the important question of the extent to which online job vacancies represent all current vacancies. Although many jobs are advertised online, some may be advertised in other ways, or filled through other means, such as through personal contacts. This leaves a gap between the target population and what can be measured from online advertisements. This gap may vary considerably between countries and by employment sector. For example IT jobs are typically much better covered on-line than jobs in the retail sector.

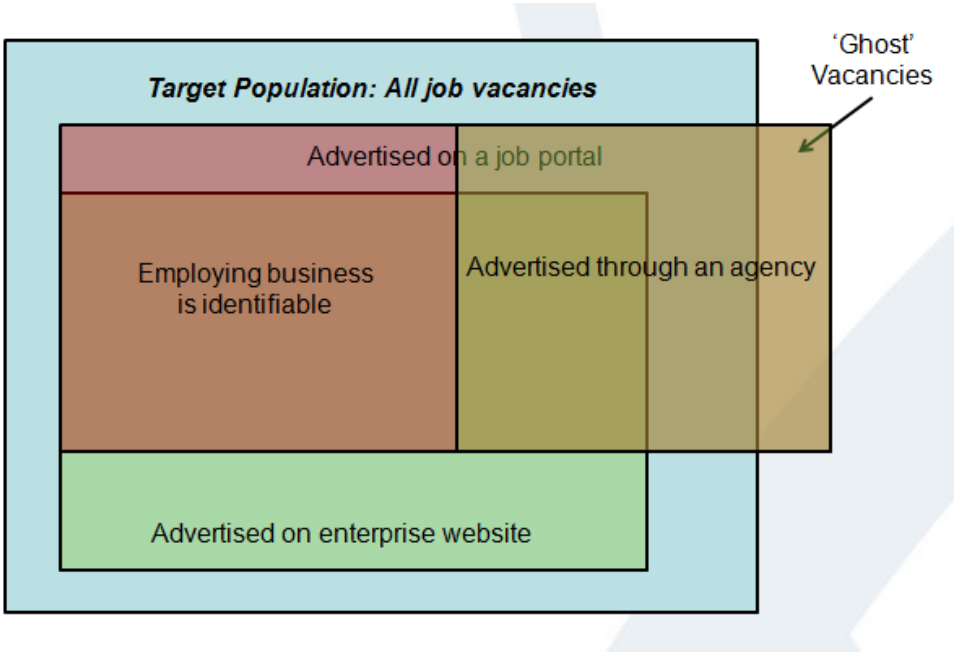
The understanding of these gaps is the key to understanding how online job advertisements might be incorporated into official statistics. It is assumed that for most (and maybe all) countries the gap between online job coverage and the target population is such that it would not feasible to produce estimates of the target population based purely on on-line job advertisements. Therefore, it is envisaged that job vacancy surveys will need to continue in some form to provide a benchmark to which data from other sources would be calibrated. Other sources would be incorporated to provide additional variables that are not captured from surveys and possibly more frequent and timely estimates. Therefore, the main methodological challenge for this pilot is around combining and integrating data from different sources.

Figure 2 proposes a conceptual model of how online job advertisements correspond to the target population. In practical terms this may be defined as all vacancies that are available to be measured by existing job vacancy surveys. This model assumes a "clean" data set with all duplicate

advertisements removed and adjustments for advertisements offering multiple vacancies. All relevant information would be reduced to a single set of variables for each on-line job vacancy. Other features of this model include:

i. The target population corresponds to what is currently estimated by job vacancy surveys.
ii. Jobs advertised through online portals include both jobs where the employer is identifiable and jobs that are advertised through an employment agency where the employer can not be identified. There are specific challenges with the latter in matching these to reporting units from the vacancy survey or business register.
iii. Jobs advertised through online job portals may also include some 'ghost' vacancies that are not within the target population. The model shows this as a subset of jobs advertised by employment agencies although a small number may also exist where the advertising business is identifiable.
iv. Jobs advertised directly on enterprise websites are a subset of the target population and will largely (although not completely) overlap with job vacancies advertised on job portals.

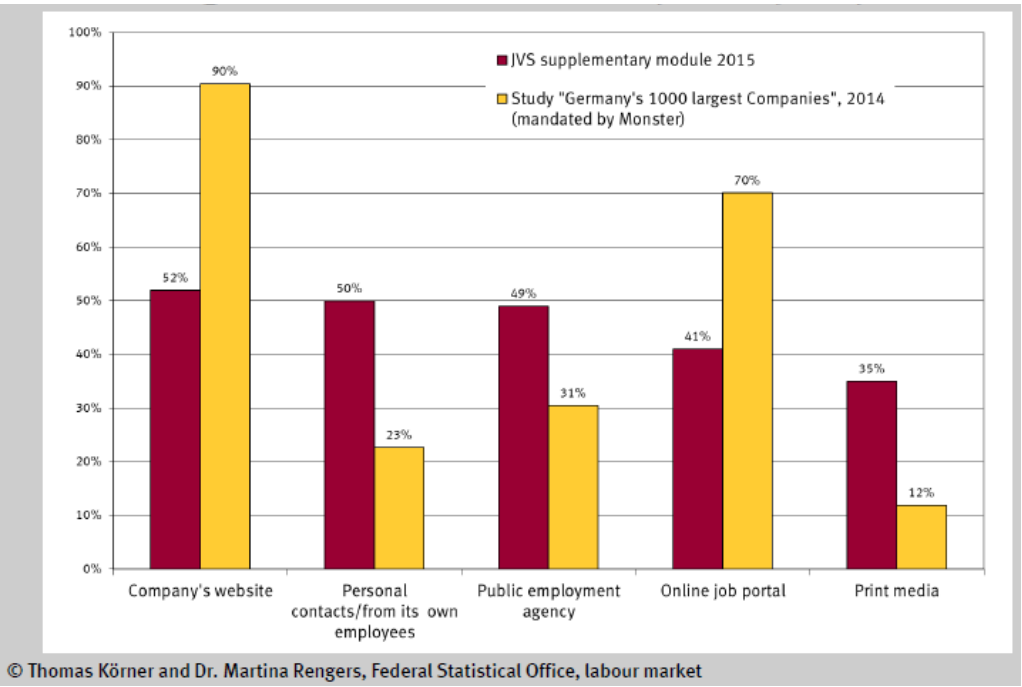
Figure 2: Conceptual model for measuring job vacancies from on-line sources.



As well as providing a conceptual framework for understanding the coverage of job vacancies from online sources and how these relate to the measurement of all job vacancies, this approach could also provide the basis for developing an estimation framework. The details of this are sketchy at present, but might involve an initial matching or reconciliation of enterprise level data between on-line sources and the job vacancy survey. This might be followed by modelling of residual data that cannot be directly matched, including jobs advertised through employment agencies.

Other information that would be useful within this kind of framework is data on the recruiting channels used by enterprises. Germany already has some of this information collected through their recent annual job vacancy survey along with equivalent information for large enterprises from a separate study (Figure 3). This shows that larger enterprises are more likely to use on-line channels whereas small businesses are more likely to use traditional channels, such as print media. This kind of data would help with understanding the differences between those enterprises that advertise on-line and those that don't, or primarily use other channels. In turn, this could be used for producing hybrid estimates that were partially based on directly on-line job advertisements and partially on model-based estimates.

Figure 3: Recruiting Channels used by German Enterprises



7. Conclusion

Data from on-line job advertisements are a very rich data source. They provide real-time information on the types of jobs and where they are located. However, the data are complex. Much of it is unstructured and it is difficult to align to established concepts - producing statistics about job vacancies is not simply a question of counting the number of job advertisements. Also, since not all jobs are advertised on-line, a framework is needed to understand their coverage in relation to all job vacancies. For this reason it is almost certain that job vacancies surveys will still be needed and the

challenge is around how to integrate data from different sources in order to produce statistics that are most useful for public policy purposes.

References

[1] United Nations (2012) "Big Data for Development, Challenges and Opportunities", Available at: http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf (accessed 14 November 2016)

[2] Eurostat (2013), "Scheveningen Memorandum", Available at: https://ec.europa.eu/eurostat/cros/content/scheveningen-memorandum_en (accessed 14 November 2016)

[3] European Union [1] Regulation (EC) No 453/2008 of the European Parliament and of the Council, Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:145:0234:0237:EN:PDF (accessed 14 November 2016)

[4] Körner. T.. M. Rengers et al.. 2016. "Inventory and qualitative assessment of job portals" Deliverable 1.1. Work Package 1 Web scraping / Job vacancies of the ESSnet on Big Data. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/File:Deliverable_1_1_draft_v5.docx (accessed 1 November 2016)

[5] Ibid