Limitations of Using Online Vacancy Data

An Overview

Cedefop Workshop on "Real-time Labour Market Information: Skills Requirements Analysis"

Karolien Lenaerts
Researcher, CEPS
December 1, 2015

Outline

- Labour market information
- Real-time labour market information
- Web-based data: advantages and limitations
- Vacancy data: advantages and limitations
- Conclusions

What is labour market information?

"All quantitative and qualitative information that relates to labour markets"
(Woods and O'Leary, 2007)

- Six components: macro labour force, labour demand, occupational supply, occupational characteristics, education and training information and classifications and crosswalks
- Initiated and made available by governments, international institutions and other organisations

Advantages:

- more accurate, better structured and more complete
- based on a randomly selected sample of the population ('representative')

Limitations:

- statistics commonly distributed with a lag, not updated regularly, or frequently revised
- small sample size and data unavailability for countries, sectors or regions with limited coverage
- new jobs and skills: not well reflected in data, occupation codes (outdated, incomplete)

Real-time labour market information

Web data for labour market analysis

- Public and private sites (job portals, intermediaries, social networks, surveys, Google Trends)
- Data can be obtained via the website or that of a partner organisation, are collected by private companies or via web crawling

Why web data?

- Rapidly advancing field, Internet as a platform and data source (Kuhn and Skuterud, 2004; Askitas & Zimmermann, 2009; 2015)
- Internet has transformed job search, matching and selection → how firms and workers search for each other, labour services are delivered and local labour demand is shaped (Autor, 2001)

Advantages of web data

Web-based data can be used to overcome issues related to traditional sources and fill the gaps where traditional sources are weak or absent

- Compile large, diverse and potentially more representative datasets in an easy, fast, flexible and relatively inexpensive way; avoid logistical issues
- Collect data in real-time (no lags or revisions)
- Collect data on phenomena that are difficult to measure with traditional sources: self-employment, on-the-job search, wage and working conditions
- Capture dynamics that are difficult to grasp otherwise (also a role for Google Trends here)

- Social networks: very large global user base comprising individuals and organisations, detailed profiles, publicly available, used for job search, selection and recruitment, on-the-job search
- Surveys: fast, flexible, easy to set up and analyse, large and diverse sample

Limitations of web data

But there are some caveats as well:

- Ethical, technical and other issues (e.g. privacy, anonymity, computer literacy, data quality, data accuracy)
- Google Trends: sample bias (only random draws when enough observations), sampling variability (data should be treated as random), no demographic information, endogeneity (Kearney and Levine, forthcoming)
- Social networks: selection, data availability
- Surveys: sample bias, measurement error, non-response, drop-out

Online job portals as a data source

Job portals are a particularly interesting source of data for labour market analysis

- Job advertisements, CVs and resumes: content analysis
  - title, description, requirements, and other information
  - over 70 'data fields' in a single job post (Carnevale et al., 2014)

- Information extracted from the portal itself: occupational structure and 'tag system'
  - occupational classification to structure database and facilitate search
  - based on tags, stored in a library called by API, or on keywords

- Wide variety of portals: public/private, general/specialised, national/international

- Increasingly developing into career communities

Research using vacancies and CVs

A growing literature, embedded in earlier work that relied on printed advertisements (Jackson et al., 2005; Jackson, 2007; Dörfler and van de Werfhorst, 2009; Barnichon, 2010)

Academic research covers a variety of topics:

- discrimination (Kuhn and Shen, 2013; Maurer-Fazio, 2012; Maurer-Fazio and Lei, 2015)
- qualifications, skills, over-qualification, mismatch, employers requirements (Kennan et al., 2008; Capiluppi and Baravalle, 2010; Kureková et al., 2012; Štefánik 2012a; 2012b; Shen and Kuhn, 2013; Kureková and Žilinčíková, 2015; Marinescu, 2015; Hershbein and Kahn, 2015)
- search behaviour, mobility (Cañibano et al., 2008; Masso et al., 2011; Kudlyak et al., 2012; Masso et al., 2013; Faberman and Kudlyak, 2014; Agrawal and Tambe, 2014)
- focus on particular sector, industry, or occupation (Wade and Parent, 2001; Huang et al., 2009)
- other topics (Martínek and Hanzlík, 2014; Marinescu and Rathelot, 2015)

Policy research on labour market, education, and social policy

Benefits of online vacancy data

- More detailed, provide more information: real job titles, job descriptions, education and skill requirements, and other information
- Easy to track the time it takes to fill particular job openings
- Vacancy data are scarce and online availability provides opportunities to access and analyse the content of job advertisements to better understand what employers require
- Support labour allocation, labour administration and development of (re-)training programs

How are these vacancy data obtained?

Online vacancies are often collected via 'web crawling' or by querying the website's API

How does this work? What are the different steps? (Carnevale et al., 2014)

- Tutorial: presentation by Lucia Kureková and Anna-Elisabeth Thum (InGRID presentation)
- Job advertisements are assembled into a database by means of a 'spider' (web bot)
- Set of advertisements is then processed:
  - extracted data from the database → parse into smaller fragments → coding
  - structure and content of the advertisement are important
  - detailed taxonomy of variables and words helpful
  - semantic analysis and text mining → synonyms, expressions, translations, …
- Careful selection of websites to be crawled (representativeness and completeness)

Obstacles in collection of vacancy data

Barriers that complicate data collection from online portals (Carnevale et al., 2014; Shapiro, 2014; Kureková et al., 2015):

- Advertisements are generally not standardised
- Advertisements are published on multiple websites or repeated over time
- Information processing and text analysis are complicated
- Portals commonly not stored information
- Portals are not standardised: heterogeneous classifications

Limitations of online vacancy data

Vacancy data are, by their very nature, incomplete:

- Not all available jobs are advertised (recruitment through internal and informal channels)
- Not all available jobs are advertised online
  - even if all jobs would be advertised online, can one capture all of them?
  - especially across regions and over time?
- Some advertisements do not correspond to real jobs or to new jobs
- Some advertisements do not list all the qualifications and skills required or lack other details
- Some advertisements refer to seasonal jobs
- Not all job seekers use the Internet to search for jobs (self-selection)
- Not all job seekers are connected to the Internet ("digital divide")
  - Autor (2001): geography and inequality: not necessarily beneficial to all groups
- Vacancies only represent a small part of the labour market and of labour demand
  - self-employment
- Vacancy data are highly volatile and may be inconsistent

Limitations of online vacancy data

Vacancy data can be biased towards specific regions, industries or applicants:

- Carnevale et al. (2014): 80% of vacancies requires at least a Bachelor degree, bias towards industries and occupations that mainly employ high-skilled, white-collar workers (STEM)

Selection issues:

- Autor (2001): adverse selection of job applicants (applying for a job is cheap and easy, so job seekers apply for many jobs, for which they could be over- or underqualified)
- Websites or online platforms could attract specific users → data representativeness and extent to which results can be generalised (Carnevale et al., 2014; Kearney and Levine, forthcoming; Kureková et al., 2015)

Data representativeness

A lot of attention is devoted to this issue by Kureková et al. (2015)

- Is the sample of online job vacancies representative for all vacancies in the economy?

- Representativeness is difficult to assess:
  - population of vacancies and its structure are unknown
  - vacancy data are not missing at random
    → results from sampling
    → vacancies that were never advertised online
  - if reporting vacancies was mandatory, then other issues still apply
    → internal selection, informal networks

- Issue not new to the literature, but only few possible solutions have been suggested
  - weighting (post-stratification weighting and propensity score adjustment, as in surveys) difficult because of unknown population of vacancies

Potential solutions

- Compare sample of vacancies with a representative dataset describing the labour market structure (LFS) and judge coverage of vacancies based on sectoral and occupational structure (Jackson, 2007; Štefanik, 2012a; 2012b)

- However, Kureková et al. (2015) argue that:
  - LFS is not a straightforward measure of the structure of labour demand (broader)
  - current labour demand may not match the existing LFS structure

- Another solution is to determine aspects of employers' search strategy by selection of a 5% random sample from all establishments in the Netherlands (Van Ours and Ridder, 1992)
  - two-stage questionnaire
  - rigorous sampling framework

- However, Kureková et al. (2015) argue that:
  - non-response and data collection issues
  - expensive and time-consuming

Potential solutions

Kureková et al. (2015) propose to use statistical models to alleviate the issue

- Draw on survey data and statistical models that have been developed to address estimation problems in survey design
  - missing data statistics: understand how to accurately estimate population parameters

- Model-based approach: estimation of a population mean from a sample is a similar to prediction of a population mean (Royall, 1992)
  - use model to determine missing values, best-fitting model is selected
  - estimated using Bayesian techniques or maximum likelihood

Model-based approach

- Model is based on the density of the variable with the missing values, and conditional on a set of variables that describe the survey design and a set of parameters

- Data on the features from advertised jobs and on the variables that determine whether or not a vacancy is posted online are necessary

- Fill in 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

  - density of the variable, conditional on a set of other variables representing information used in the survey design and a set of parameters

  - information needed are certain characteristics about the jobs that are advertised

Potential solutions

Kureková et al. (2015) make **four key points**

1. Focus on the labour market segment where the coverage bias is likely to be less problematic
2. Diversification of data sources used in many studies: other data sources analysed in parallel
3. Some scope to correct possible biases if vacancy data can be linked to firm characteristics
4. Market coverage and technical advancement of the portals need to be assessed in each country

Conclusions on using vacancy data

- Web-based data for labour market analysis: rapidly advancing research field that clearly has a bright future

- Most research is focussed on online job portals and vacancies

  - many advantages , but also limitations

  - important methodological issues, which are only addressed in a few studies

- Avenues for future research

www.ceps.eu

Thank you for your attention!

karolien.lenaerts@ceps.eu