# Extraction of occupation, competences and qualifications from online job offers for official statistics

**Maciej Beręsewicz**[1,2], **Łukasz Cywiński**[3] **& Robert Pater**[3,4]

[1]Poznań University of Economics and Business, Poland

[2]Statistical Office in Poznan, Poland

[3]University of Information Technology and Management in Rzeszów, Poland

[4]Educational Research Institute

## 1 Introduction

There is an increasing need for analyzing detailed companies' demand for workers, that is for occupation, skills or competencies and qualifications [1, 2, 3]. The *Demand for Labour* survey conducted by National Statistical Institutes does not contain information on the demand of companies for future workers' competences or qualifications. One might consider online job offers as support or alternative for surveys on the vacancy market. Due to unstructured character of these data relevant information should be extracted to conduct quality and representativeness assessment and estimation process.

Our main contribution is the proposal of an method for analyzing online job offers in the context of detailed information they contain. The method is based on gathering online job offers and screening them with text mining and machine learning tools. We present caveats behind such research and propose solutions. We apply this method for the Polish vacancy market and compare results with ongoing representative surveys to correct non-probability character of these data.

The results are especially important for economists, education sector, and labour market institutions, e.g. for shaping the OECD Skills Strategy[4]. Such detailed information may be used to adjust labour market and education policy, especially directed to reduce the structural unemployment.

## 2 Data sources

- We use country-wide general websites (25 websites with job offers or aggregators), containing job offers for all sectors and occupations.
- We disregarded small local websites because they have low coverage of job offers containing detailed information we aim to analyze, such as required skills.
- We chose websites on the basis of Google Trends volume of queries.

## 3 Data collection procedure

- We gathered the data using web-scrapping techniques.
- From November 2017 to June 2018 the data have been collected on a monthly basis, at the end of a month.
- Since July 2018 we also started to gather titles of job offers once every 5 days, while continuing gathering job offers content at the end of a month.
- Each website is stored in the original HTML format for reproducibility.
- On average 360 000 job titles and offers are collected each time.

## 4 Data cleaning and extraction of relevant information

- We lemmatized the data using Morfologik-stemming-1.9.0 library, that is, we identified the basic forms of all the words in the text.
- For some words forming the competences, this library provides inconsistent lemmatized words. To deal with such situations that we spotted during the analysis, we created an additional dictionary of exceptions.
- Because job offers can contain various words to describe the same trait (occupation, qualification, skill etc.), the algorithm should deal with the situation in which the trait is mentioned in the sentence, but with different words than in the dictionary.
- To solve this problem, we first prepared a list of occupations, qualifications, and skills. We used the European Commission ESCO, ISCO and ISCED classifications.
- Its another advantage is that it helps us detect various terms companies use to describe traits, enabling us to increase the dictionary's size. For this, we created a dictionary of synonyms based on three dictionaries.
- Moreover, we also apply several machine learning techniques to predict occupation, qualification, skills etc. based on true labelled data.

## 5 Representativeness of online job offers

- In order to assess representativeness we compare online job offers with the Demand of Labour survey.
- Figure 1 presents comparison of online data with the survey.

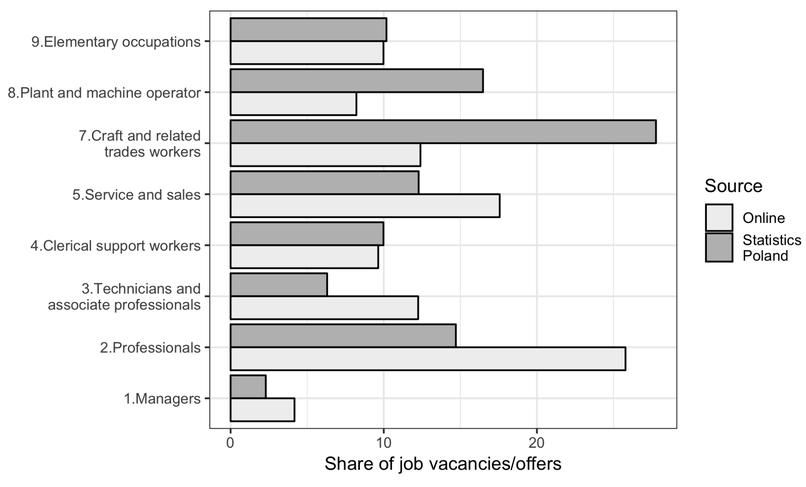


**Figure 1:** Comparison of distribution of occupations observed in online services and the Demand for Labour survey

- Due to limited number of available variables we compared distribution of occupation.
- We disregarded Armed forces and Farmers as these groups were characterized with the lowest number of vacancies and job offers.

## 6 Selected results

- Figure 2 shows the most popular transversal competences that companies require from their future workers according to Internet job offers.

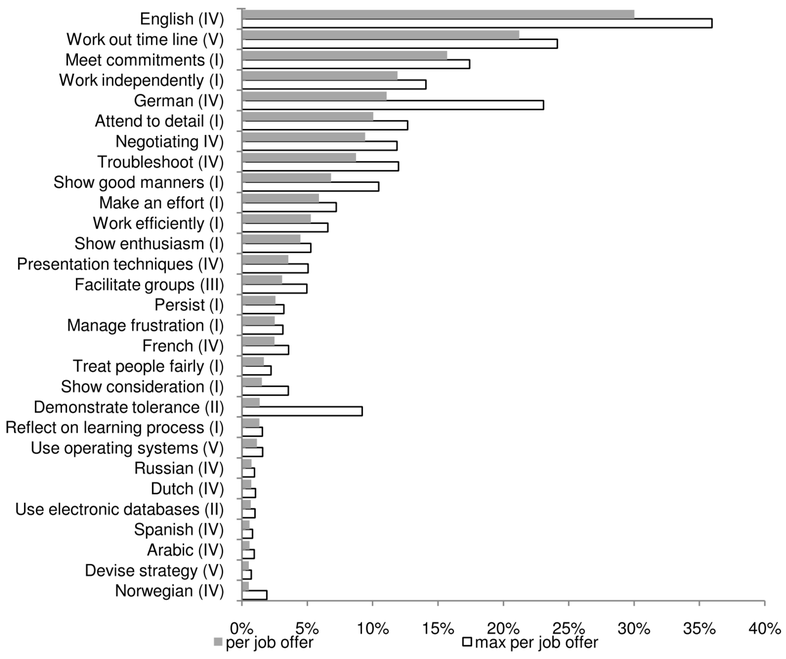


**Figure 2:** The number (per job offer) of the most frequent competences (with their ESCO categories)

- For job seekers, as well as employment and educational institutions, the up-to-date information about employers' detailed demands might be helpful.
- That is why we decided to propose a method of the analysis of demand for occupations, qualifications and competences at a macroeconomic scope by using machine learning techniques and methods to reduce bias using model-assisted estimators.
- Based on collecting and analysing Internet job offers throughout a long period of time, the method makes it possible to analyse the detailed companies demand with a relatively low cost.

## Contribution

- We propose a method for analysing online job offers in the context of demanded occupations, qualifications and skills in the labour market.
- We use official classifications, what is especially an issue in online-gathered data.
- We assess the quality, coverage and representativeness of Internet data sources for labour market statistics.
- We apply this method for the Polish vacancy market, to study demand for labour at a detailed level.
- We propose a cost-effective tool for continuous analysis of firms' demand for new workers.

## Forthcoming research

- Job offers flow analysis.
- Comparison of demand for skills with National Qualifications Framework effects of education (different terminology and approach).
- Keep representativeness in the face of a structural change.
- Merging data with other data sets.
- Checking reliability of what companies write.
- Providing continuous and comparable data on labour supply (public employment offices?).

## References

[1] D. Deming. The Growing Importance of Social Skills on the Labor Market. *Quaterly Journal of Economics*, 132:1593–1640, 2017.

[2] Kahn L. Deming, D. Skill Requirements Across Firms and Labor Markets: Evidence From Job Postings for Professionals. *NBER Working Paper*, 23328, 2017.

[3] Kahn L. Hershbein, B. Do Recessions Accelerate Routine-biased Technological Change? Evidence From Vacancy Posting. *Employment Research*, 24:1–4, 2017.

[4] OECD. Towards an oecd skills strategy. 2011.

## Acknowledgements

This study was funded by the Polish Ministry of Science and Higher Education within the Programme DIALOG (grant number DIALOG 0127/2016).