**Extraction of occupation, competences and qualifications from Internet job offers for official statistics**

**Keywords:** online data, representativeness, job offer, vacancy.

# Introduction

There is an increasing need for analysing detailed companies’ demand for workers, that is for occupation, skills or competences and qualifications ([1],[2],[3]). Current surveys conducted by Statistical Offices do not contain information on demand of companies for future workers’ competences or qualifications. One might consider online job offers as support or alternative for surveys on vacancy market. However, these data sources are unstructured and relevant information should be extracted to conduct quality and representativeness assessment and estimation process.

Our main contribution is the proposal of an efficient method for analysing Internet job offers in the context of detailed information they contain. The method is based on gathering Internet job offers and analysing them with text mining and machine learning tools. We present caveats behind such a research and propose solutions to them. 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 reducing the structural unemployment.

# Methods

We gathered job offers from several vacancy market websites in Poland. These were both country-wide general websites, containing job offers for all sectors and occupations. We chose 25 websites with job offers or aggregators of job offers. We disregarded small local websites because they have a low coverage of job offers containing detailed information we aim to analyse, such as required skills. We chose websites on the basis of Google Trends volume of queries. It is an index of the volume of Google queries by geographic location and category [5].

We gathered the data using a computer script that automated the procedure. 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.

Data collection is as follows. At the last day of each month during the studied period, all job offers scraped from the studied websites and stored in a local database in the original HTML format for reproducibility. First, we applied parsing to the data: the data stored in the standard HTML format were converted into plain text. Parsing removes such information as font size, colour, and other formatting tags.

After parsing, we lemmatized the data, that is, we identified the basic forms of all the words in the text (for example, verbs were transformed to their infinitives while nouns to their nominative cases, singular forms of the first person). For this, we developed a tool based on the Morfologik-stemming-1.9.0 library. 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. At this stage, our most important task was to find specific competences in the texts of job offers. 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. Our approach enables us to obtain data on individual traits (phrases). 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 (www.slownik-synonimow.eu, www.synonimy.pl, and sjp.pwn.pl). With a larger dictionary—with more phrases (including colloquial ones) describing competences and their groups—we will obtain more accurate representation of the competences demand in the labour market. Moreover, we also apply several machine learning techniques to predict occupation, qualification, skills etc. based on true labelled data. Knowledge of occupation is crucial as it is the only variable available in both online and Statistics Poland’s survey.

# Results

Table 1 presents a sample job offer and its treatment. Job offer is classified on the basis of a publication date, Internet portal, that it was published on, and a number. Parsed job offer contains plain text. After lemmatization this text is transformed to a basic form. Final stage is matching the lemmatized dictionary words to the text. In the example we looked for transversal competences from ESCO classification. We found six of them.

**Table 1. Sample parsed job offer, lemmatized job offer and identified transversal competences**

|  |  |
| --- | --- |
| Job offer number | 201801\_MoneyOgloszenia/105309.txt |
| Parsed job offer | **stanowisko** [**position**] programista netopis stanowiska [description] klient portalu pracapl programista net modelowanie oraz tworzenie oprogramowania z wykorzystaniem net 40 oraz sql server 2008**wymagania** [**requirements**] doświadczenie w realizacji projektów opartych o aspnet mvc entity framework znajomość programowania obiektowego opartego o net 35 lub 40 tworzenia baz danych i zapytań w języku tsql bardzo dobra znajomość języka angielskiego zdolność analitycznego myślenia otwartość terminowość kreatywność entuzjazm i samodzielność **oferujemy** [**we offer**] konkurencyjne wynagrodzenie udział w ambitnych projektach ciągły rozwój umiejętności stabilną formę współpracy liczne szkolenia pracę w wśród wiodących specjalistów na rynku forma kontaktu formularz online |
| Lemmatized job offer | stanowisko programista neta opis stanowisko klient portal programista neta opis stanowisko modelować oraz tworzyć oprogramować z wykorzystać neta oraz wymagać doświadczyć wiek realizacja projekt oprzeć ojciec znajomość programować obiektowy oprzeć ojciec neta lubić tworzyć baza dać i zapytanie wiek język bardzo dobry znajomość język angielski zdolność analityczny myśleć otwartość terminowość kreatywność entuzjazm i samodzielność oferować konkurencyjny wynagrodzić udziać wiek ambitny projekt ciągły rozwój umiejętność stabilny forma współpraca liczny szkolić praca wiek wśród wieść specjalista na rynek forma kontakt formularz online |
| Identified transversal competences | Identified phrases | Assigned to competences |
| angielskientuzjazmanalityczny myślećkreatywnośćterminowośćsamodzielność | English languageShow enthusiasmExamine evidenceGenerate new ideasMeet commitmentsWork independently |

After classifying each job offer according to phrases, we can show the aggregated results. Table 2 presents an excerpt from the classification of transversal competences. It shows competences, their groups and some descriptive statistics on the number of competences per job offer throughout analysed period. Then, estimation process consists of model-assisted estimators (e.g. calibration, LASSO).

**Table 2. An excerpt from the final classification of transversal competences**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Class | Competences | Mean | Min | Max | First | Last |
| IV | 2 |  |  | [Verbal communication](https://ec.europa.eu/esco/portal/concept?lang=en&concept=http%3A%2F%2Fdata.europa.eu%2Fesco%2Fskill%2F367189&conceptScheme=http%3A%2F%2Fdata.europa.eu%2Fesco%2FConceptScheme%2FESCO_Skills) | 0.49 | 0.39 | 0.57 | 0.40 | 0.47 |
| IV |  | 2.1 |  | [Spoken production](https://ec.europa.eu/esco/portal/concept?lang=en&concept=http%3A%2F%2Fdata.europa.eu%2Fesco%2Fskill%2F367169&conceptScheme=http%3A%2F%2Fdata.europa.eu%2Fesco%2FConceptScheme%2FESCO_Skills) | 0.22 | 0.17 | 0.27 | 0.17 | 0.21 |
| IV |  |  | 2.1.1 | [Presentation techniques](https://ec.europa.eu/esco/portal/concept?lang=en&concept=http%3A%2F%2Fdata.europa.eu%2Fesco%2Fskill%2F367347&conceptScheme=http%3A%2F%2Fdata.europa.eu%2Fesco%2FConceptScheme%2FESCO_Skills) | 0.04 | 0.03 | 0.05 | 0.03 | 0.03 |
| IV |  | 2.2 |  | [Spoken interaction](https://ec.europa.eu/esco/portal/concept?lang=en&concept=http%3A%2F%2Fdata.europa.eu%2Fesco%2Fskill%2F367171&conceptScheme=http%3A%2F%2Fdata.europa.eu%2Fesco%2FConceptScheme%2FESCO_Skills) | 0.27 | 0.22 | 0.32 | 0.23 | 0.26 |
| IV |  |  | 2.2.1 | [Debating techniques](https://ec.europa.eu/esco/portal/concept?lang=en&concept=http%3A%2F%2Fdata.europa.eu%2Fesco%2Fskill%2F367261&conceptScheme=http%3A%2F%2Fdata.europa.eu%2Fesco%2FConceptScheme%2FESCO_Skills) | 0 | 0 | 0 | 0 | 0 |
| IV |  |  | 2.2.2 | [Interrogating](https://ec.europa.eu/esco/portal/concept?lang=en&concept=http%3A%2F%2Fdata.europa.eu%2Fesco%2Fskill%2F367254&conceptScheme=http%3A%2F%2Fdata.europa.eu%2Fesco%2FConceptScheme%2FESCO_Skills) | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |
| IV |  |  | 2.2.3 | [Persuading](https://ec.europa.eu/esco/portal/concept?lang=en&concept=http%3A%2F%2Fdata.europa.eu%2Fesco%2Fskill%2F367291&conceptScheme=http%3A%2F%2Fdata.europa.eu%2Fesco%2FConceptScheme%2FESCO_Skills) | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |
| IV |  |  | 2.2.4 | [Negotiating](https://ec.europa.eu/esco/portal/concept?lang=en&concept=http%3A%2F%2Fdata.europa.eu%2Fesco%2Fskill%2F367274&conceptScheme=http%3A%2F%2Fdata.europa.eu%2Fesco%2FConceptScheme%2FESCO_Skills) | 0.09 | 0.07 | 0.12 | 0.08 | 0.09 |

Figure 1 shows the most popular transversal competences that companies require from their future workers according to Internet job offers.

****

**Figure 1. The number (per job offer) of the most frequent competences (with their categories)**

# Conclusions

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.

# References

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

[2] D. Deming, L. Kahn, Skill requirements across firms and labor markets: evidence from job postings for professionals, NBER Working Paper 23328 (2017). Available at:<http://www.nber.org/papers/w23328> (accessed 20.06.2018).

[3] B. Hershbein, L. Kahn, Do recessions accelerate routine-biased technological change? Evidence from vacancy posting, Employment Research 24 (2017), 1-4.

[4] OECD, Towards an OECD Skills Strategy, OECD (2011). Available at:<http://www.oecd.org/education/47769000.pdf> (accessed 20.06.2018).

[5] H. Choi, H. Varian, Predicting the Present with Google Trends (2009). Available at <https://www.google.com/googleblogs/pdfs/google_predicting_the_present.pdf> (accessed 31.07.2018).