Crimson Publishers
Wings to the Research

Research Article

Occupations and Skills in Demand from Web-Based Job Vacancies

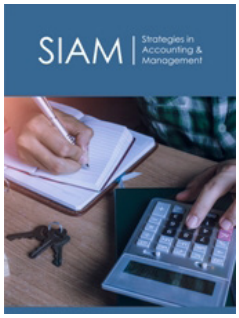
Pietro Giorgio Lovaglio*

Department of Statistics and Quantitative Methods, University of Bicocca-Milan, Via Bicocca degli Arcimboldi 8, 20126 Milan, Italy

Abstract

Online job portals collecting web vacancies have become important media for job demand and supply matching. They also represent a growing research area for the application of analytical methods to study the labour market using innovative data sources. Both the Knowledge Discovery in Databases approach and mixed supervised and unsupervised text mining approaches were typically applied to retrieve occupations associated with each web vacancy (ISCO classification up to level 4) and related skills. In the present paper we apply this method to a population of online web vacancies collected for three countries (Italy, UK and Germany) collected over a quarter in 2019, within an international project, to demonstrate the potentiality of informative power of such approach that can be considered as promising strategy providing effective support for decision making of several stakeholders such as government organizations, analysts, and recruitment agencies, as they allow for timely and fine-grained representations of complex labour market dynamics, in terms of trends, occupations, and skills. Finally, problems of representativeness that affect online vacancies are briefly discussed and possible approaches are proposed.

Keywords: Job vacancies; Scraping; Big data; Job classification



***Corresponding author:** Pietro Giorgio Lovaglio, Department of Statistics and Quantitative Methods, University of Bicocca-Milan, Via Bicocca degli Arcimboldi 8, 20126 Milan, Italy

Submission: May 5, 2021

Published: June 3, 2021

Volume 2 - Issue 4

How to cite this article: Pietro Giorgio Lovaglio. Occupations and Skills in Demand from Web-Based Job Vacancies. Strategies Account Manag. 2(4). SIAM. 000544. 2021.

Introduction

Vacancies are a crucial variable for policy analysis for assessing the degree of tightness of the labour market and its change over time is a leading indicator that underpins most monetary policy decisions, since has been demonstrated that it improves the unemployment forecasts, (see [1] for a review). Eurostat [2] publishes, in the Job Vacancy Survey (JVS), quarterly data on the number of job vacancies, defined 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.

Despite their importance vacancy measurement is generally implemented in official data through quarterly surveys that offer no detail at geographical level (only country level) and occupational level providing a limited assessment of the true underlying labour market conditions. To this end, the availability of web vacancies has prompted new research exploiting the richness and granularity of these data to provide a better understanding of local labour market conditions. In this scenario, a growing number of employers use the web to advertise job openings through web job vacancies. These usually specify a job position with a set of skills that a candidate should possess. Turning these data into knowledge can provide effective support for decision making of several stakeholders such as government organizations, analysts, and recruitment agencies. In 2015, the CRISP (The Interuniversity Research Centre on Public Services-University of Milan-Bicocca) started work on a European project supported by a grant from Cedefop (The European Center for the Development of Vocational Training). The project aims to conduct a feasibility study and create a prototype for analysing web job vacancies collected from five EU countries through extracting the requested skills from the data. The rationale behind this project was to turn data extracted from web-based job vacancies into knowledge (thus providing value) to support labour market intelligence activities.

Materials and Methods

The well-known Knowledge Discovery in Databases (KDD) process [3] was applied as a methodological framework. During this process, the quality of the data is assessed, and

cleansing activities are executed. In our context, this task deals mainly with the identification of duplicated job vacancies posted on different web source as well as job vacancies published multiple times on the same site; these tasks have been performed applying AI algorithms and details on the quality process can be found elsewhere [4-9]. In this way, the data classified according to the European classification standard ISCO-08 occupation taxonomy (which at Level 4 involves 436 occupation items) and further was enriched with information about the skills requested by the employers, thus producing a detailed portrait of the job opportunities advertised on the web.

Each title and description of the job vacancy was processed according to the following pipeline: Duplicate removal, Tokenization (splitting a sentence into its words, using a 'bag of words' approach), Stop Words removal (removing useless parts of speech), Stemming (reducing words to their base or root forms), Text Classification (selecting only a few sentences focusing on occupation descriptions useful to guess skills) and Vectorization (identifying and counting the number of n-grams located in job vacancy titles and descriptions associated with the ISCO occupation codes). Particularly, bigrams (two consecutive words) and trigrams (three consecutive words) were also considered, as suggested by successful text mining classification experiences. Furthermore, where possible, each web vacancy was classified according to a required sector of economic activity and territorial area, using site-specific codes or taxonomies from the page sections of specific web portals. This information was converted into reference/standard taxonomies, such as NUTS (Nomenclature of Territorial Units for Statistics) for territorial areas, NACE (Rev.2) for sector of economic activity. Thus, the main output of the text mining approach was a structured dataset where each line represented a job offer and the columns represented relevant information, such as:

A. Occupations: ISCO-08 classification up to level 4

B. Territorial units: Up to NUTS 3

C. Sector of economic activity: NACE classification up to level 2

D. Skill (not classified, text retrieved)

In the present paper we demonstrate the potentiality of informative power of such approach that can be considered as promising strategy providing effective support for decision making of several stakeholders such as government organizations, analysts, and recruitment agencies, as they allow for timely and fine-grained representations of complex labour market dynamics. Specifically, we analyse a population of online web vacancies collected for three countries (Italy, UK and Germany) collected over a quarter in 2019 in term of demanded occupations and related skills.

Results

In this application we analyse web job vacancies scraped from web portals of three countries between June and September 2019. Overall, after quality control and duplicate removal, the number of cleaned vacancies was reduced to 553,041 (52% UK, 28% Germany, 20% Italy). It is worth noticing that unlikely Italy, where permanent contracts cover only 45% of vacancies, in UK and Germany permanent contracts are largely dominant (92%, 71% respectively). All in all, 67% of the vacancies analysed were concentrated in the services sector, 33% in industry, manufacturing and construction. More specifically, web vacancies tend to be more concentrated in the three following activities (NACE, first level): N-Administrative and support service activities (31%UK, 23% DE, 16.4% IT), J-Information and communications (30% IT, 22% UK, 15% DE) and M-Professional, scientific and technical activities (29% DE, 23% IT, 21% UK). Figure 1 shows a complete picture over sectors and countries. For 14% of the overall vacancies it was not possible to determine the activity sector.

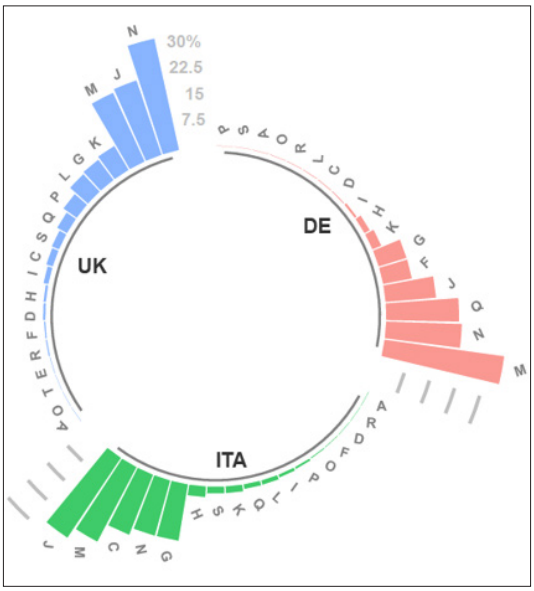


Figure 1: Most demanded jobs by economic sector (Nace Rev. 2), within countries.

Looking at demanded occupations (ISCO-08 at Level 1), web vacancies display a higher concentration of high skill occupations (48%), with the largest share by technicians and business associate professionals (35%), professionals (27%), clerical support workers (14%), crafts and related trade workers (11%), service and sales workers (10%). Moreover, demanded occupations are highly concentrated in few codes: specifically, seventeen occupations cover 66% of the entire set of demand (Table 1). To better explore country specific occupation demand, Figures 2-4 illustrates the distribution of the fifteen most required occupations at a finer level (ISCO-08 code Level 4), in UK, Italy and Germany, respectively. Accountants, accounting professionals, software developers are largely required in all three countries, whereas some difference emerges regarding education and health care professions (Germany), administrative and executive secretaries (UK) and business services agents and draughtspersons (Italy). Exploiting the richness of textual information collected in web vacancies we can assess the most relevant (recurrent) skills for each occupation and evaluate whether demanded skills may change among countries. As example, Figure 5 illustrates the word cloud of most recurrent skills for Industrial Designer in each country. Interestingly, required software for designers seems to be country specific. The presented analyses emphasized that these innovative sources presented new opportunities to collect and investigate labour market trends from a demand perspective. Examples may include the monthly stock of demanded occupations for sectors, regional variations in occupations and skill demand by industry, industrial composition of skill demand within a given area, hotspots for industry skill demand, composition of hard and soft skills for a given occupation, to name a few. The availability of such data would allow to build considerable progress and valuable that would be beneficial for research activities in the domain of labour market intelligence.

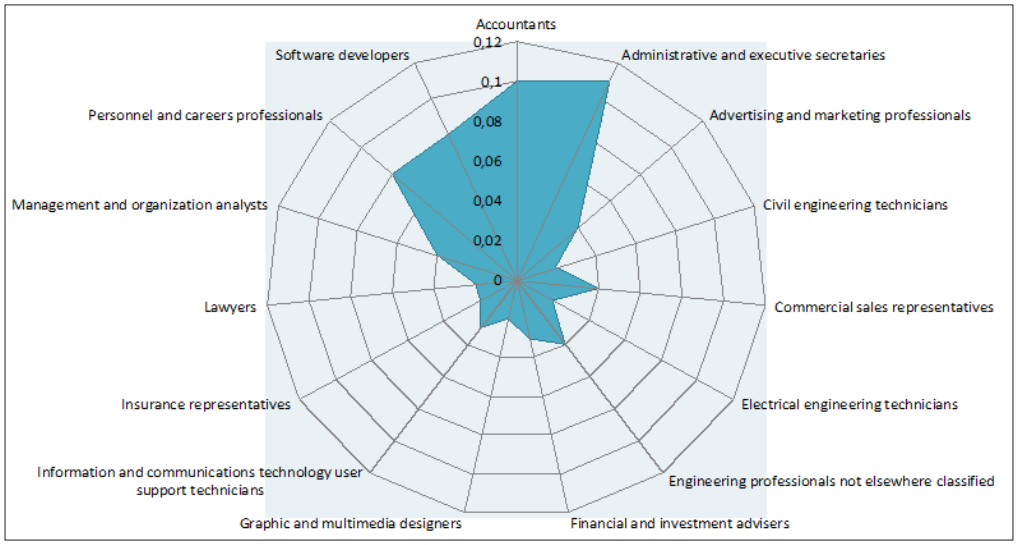


Figure 2: Most demanded jobs, by occupations in UK (ISCO 4th digit).

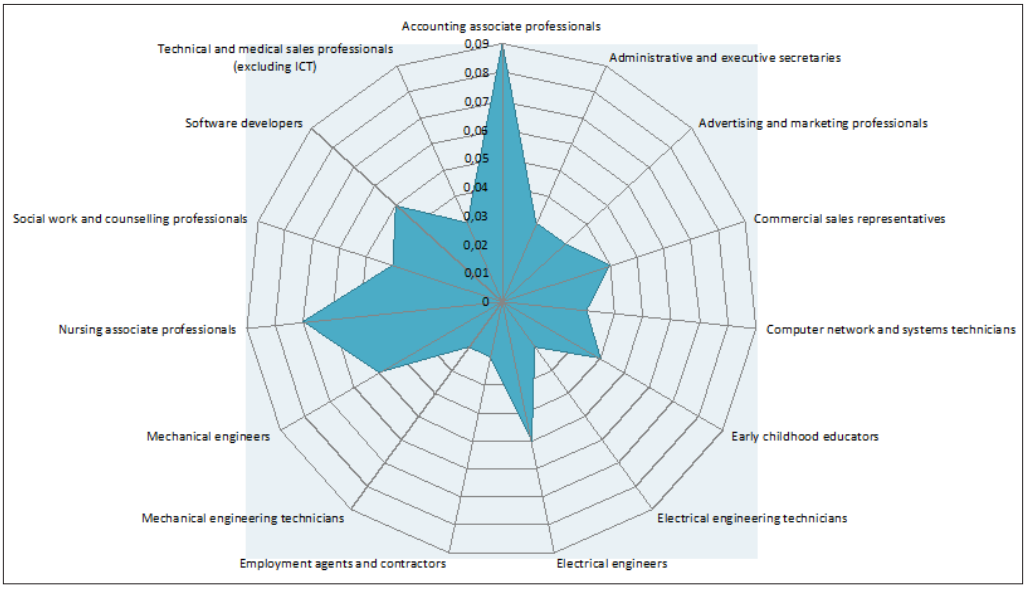


Figure 3: Most demanded jobs by occupations in Germany (ISCO 4th digit).

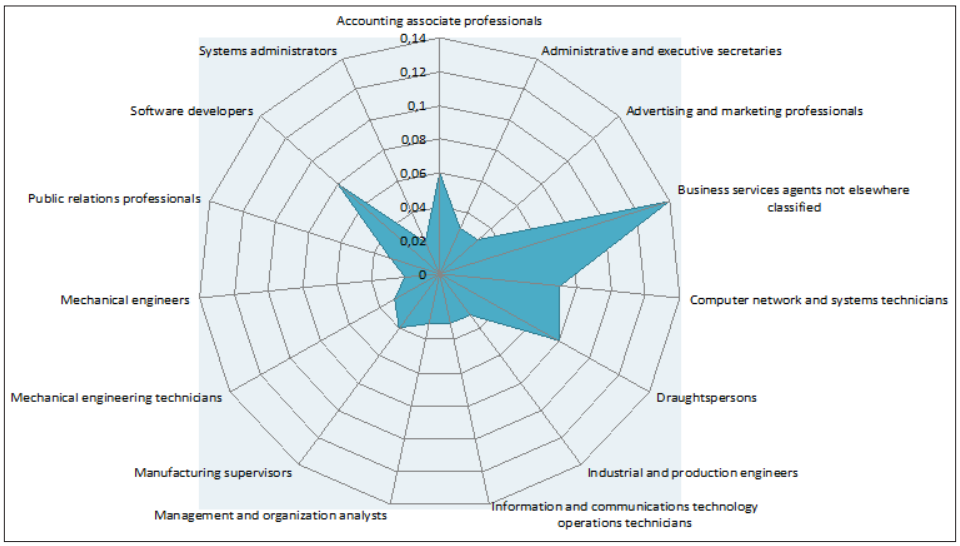


Figure 4: Most demanded jobs by occupations in Italy (ISCO 4th digit).

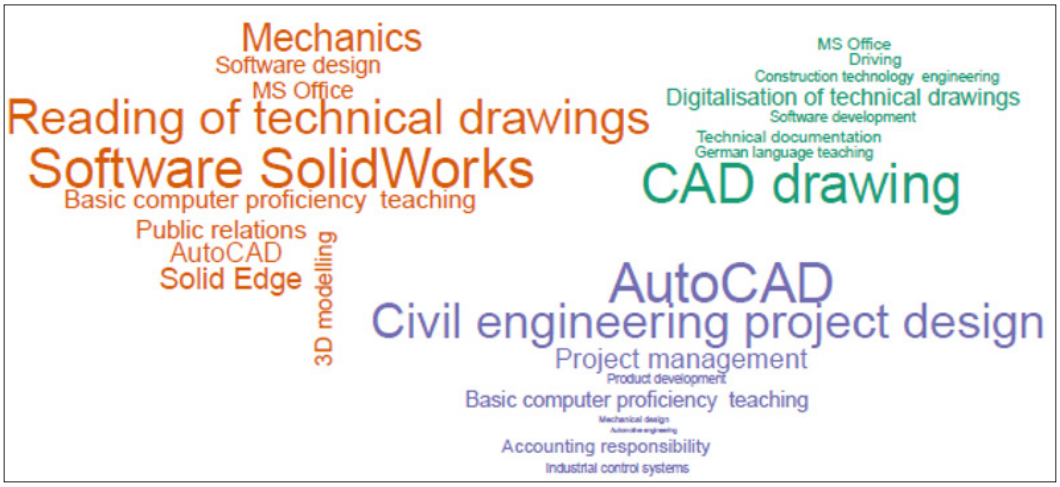


Figure 5: Most recurrent skills for industrial designer for UK (blue), Germany (green) and Italy (red).

Table 1: Most demanded occupations by ISCO (Level 4 and level 2). All three countries.

ISCO Level 4	ISCO Level 2	%
Administrative and executive secretaries	Business and administration associate professional	9.30%
Accountants	Business and administration professionals	8.20%
Software developers	ICT professionals	7.70%
Personnel and careers professionals	Business and administration professionals	6.30%
Advertising and marketing professionals	Business and administration professionals	3.90%
Commercial sales representatives	Business and administration associate professionals	3.60%
Engineering professionals not elsewhere classified	Science and engineering professionals	3.60%
Management and organization analysts	Business and administration professionals	3.50%
Accounting associate professionals	Business and administration associate professional	3.00%
ITC user support technicians	Technicians and associate professionals	2.80%
Financial and investment advisers	Business and administration professionals	2.30%
Draughtspersons	Science and engineering associate professionals	2.20%
Insurance representatives	Business and administration associate professionals	2.20%

Nursing associate professionals	Health associate professionals	2.10%
Graphic and multimedia designers	Architects, planners, surveyors and designers	2.10%
Electrical engineering technicians	Science and engineering associate professionals	2.00%
Civil engineering technicians	Science and engineering professionals	1.90%

Discussion

Despite such rich information, in term of timeliness and granularity, web data present some problems. Online vacancies data are prone to *selectivity,* a general term for self-selection error, resulting from decisions of individuals. In our context, if platforms from which data are collected are not set up for statistical purposes, the observed sample of online job ads is likely to be affected by non-random mechanism (not all online job advertisements are collected, not all websites are covered, advertisements non-conveyed through the web, sector and/or occupations which are (under)over-represented). As a result, selectivity causes coverage and non-response (or missingness) that introduce potential bias in estimates based on Online vacancies data [10,11]. Some authors [10-14] give a general overview of possible approaches to deal with non-probability samples including pseudo-randomization and the model-based approach (traditional and machine learning). A possible approach assumes that an additional 'gold standard' data source is available and adjust observed counts towards the 'gold standard' estimates, that can be a register or a survey based on a representative sample, in our case the Eurostat JVS. Most explicitly, observed vacancies are projected in a population or representative (JVS) frame using a post-stratification frame structured by known values of auxiliary variables, that should capture the selectivity process on the sample.

The greatest practical limitations to the use of full post-stratification is the need to know the proportion of the population/reference in each stratum. If we have population-level information only for certain aggregations, full poststratification is not feasible [15]. In our case, in fact, JVS data can be only used as stratification frame by two-way interactions Quarter×Nace, whereas online job vacancies data produce finer strata (for example using territory and occupation). This suggests to define as post-sampling weight balancing the quarterly stock of vacancies by industry according to online vacancies towards the quarterly stock of vacancies by industry according to the JVS: this produces a set of "post-sampling" weights for each quarter and industry, that can be assigned to each vacancy or vacancy distributions (by relevant auxiliary variables, such as NUTS, ISCO, NACE, Quarters and possible interactions). Recent works [16,17] adopt such kind of posts-stratification. Over or under-representation (for univariate or two-way or three way interactions) in online vacancies can be easily assessed by the ratio between percentage distributions of online counts and post-stratified ones: If the ratio is higher than 1, it means that a certain category (industry, occupation) is likely to be over-represented in the online job adverts dataset, whereas the opposite is true with ratio is less than 1. To conclude, data gathered from web job portals is shown to provide valuable information about job demand and is, therefore, of value to policy makers who need disaggregated real-time indicators, but, in our opinion, web data do not substitute official statistics; it rather indicates the use of official statistics as necessary benchmarks for reliable measurement of dimensions from web-based sources.

References

1. Lenaerts K, Miroslav Beblavý M, Fabo B (2016) Prospects for utilization of non-vacancy Internet data in labor market analysis-an overview. IZA Journal of Labor Economics 5(1):1-18.
2. (2018) Eurostat: Job vacancy rate.
3. Fayyad U, Piatetsky Shapiro G, Smyth P (1996) The KDD process for extracting useful knowledge from volumes of data. Commun. ACM 39(11): 27-34.
4. Hernandez MA, Stolfo SJ (1998) Real-world data is dirty: Data cleansing and the merge/purge problem. Data mining and knowledge discovery 2 (1):9-37.
5. Mezzanzanica M, Boselli R, Cesarini M, Mercorio F (2015) A model-based evaluation of data quality activities in KDD. Information Processing & Management 51(2): 144-166.
6. Boselli R, Cesarini M, Marrara S, Mercorio F, Mezzanzanica M, et al. (2017) WoLMIS: A labor market intelligence system for classifying web job vacancies. Journal of Intelligent Information Systems 51(3): 1-26.
7. Boselli R, Cesarini M, Mercorio F, Mezzanzanica M (2017) Using machine learning for labour market intelligence. In: Altun Y (edn.), Machine learning and knowledge discovery in databases. ECML PKDD 2017. Lecture Notes in Computer Science 10536.
8. Mezzanzanica M, Boselli R, Cesarini M, Mercorio F (2015) A model-based evaluation of data quality activities in KDD. Information Processing & Management 51(2): 144-166.
9. Lovaglio PG, Cesarini M, Mercorio F, Mezzanzanica M (2018) Skills in demand for ICT and statistical occupations: Evidence from web vacancies. Statistical Analysis and Data Mining 11(2): 78-91.
10. Elliott M, Valliant R (2017) Inference for non- probability samples. Statistical Science 32(2): 249-264.
11. Japec L, Kreuter F, Berg M, Biemer P, Decker P, et al. (2015) American Association for Public Opinion Research (AAPOR) Report on Big data.
12. Valliant R (2019) Comparing alternatives for estimation from nonprobability samples. Journal of Survey Statistics and Methodology 8(2): 231-263.
13. Buelens B, Burger J, Brakel J, van den (2015) Predictive inference for non-probability samples: A simulation study. Technical report.
14. Buelens B, Burger J, Brakel J, Van den A (2018) Comparing inference methods for nonprobability samples. International Statistical Review 86(2): 322-343.
15. Reilly C, Gelman A, Katz J (2001) Poststratification without population level information on the post stratifying variable with application to political polling. Journal of the American Statistical Association 96(453): 1-11.

16. Garasto S, Djumalieva J, Kanders K, Wilcock R, Sleeman C (2021) Developing experimental estimates of regional skill demand (ESCoE DP 2021-02). ESCoE Discussion Paper 2021-02.

17. Turrell A, Speigner BJ, Djumalieva J, Copple D, Thurgood J (2019) Transforming naturally occurring text data into economic statistics: The case of online job vacancy postings.