Skills in demand for ICT and statistical occupations: evidence from web based job vacancies

Pietro Giorgio Lovaglio, Mirko Cesarini, Fabio Mercorio and Mario Mezzanzanica
Dept. of Statistics and Quantitative Methods – University of Milan-Bicocca, Italy
Firstname.lastname@unimib.it

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. This paper analyses Italian web job vacancies scraped from several types of Italian web job portals between June and September 2015. After having described how the occupations associated with each web vacancy (ISCO classification up to level 4) were identified and the related skills retrieved in texts using mixed supervised and unsupervised text mining approaches, we focused on job vacancies related to ICT and statistical positions.

The principal aim of this paper is to describe these jobs in terms of the required skills that have emerged in the labour market from a demand perspective and to identify those skills that best distinguish statisticians from other ICT occupations. Hence, several machine-learning techniques were used to assess those skills that best distinguish ISCO occupations from other job groups.

After quality control and duplicate removal, the scraping collected more than 110,000 job advertisements: nearly 6,200 were classified as *ICT* or statistical positions (largely dominated by software developers). The data indicate that high-level statisticians have superior and heterogeneous professional backgrounds, linked to theoretical statistics, where analytic skills are more relevant than computing skills. Many soft and management oriented skills were also called for, which are missing among lower-level statisticians, who are restricted to more technical jobs, oriented toward general computing and informatics.

1. Introduction

The promotion of investment in the Information Communication Technology (ICT) sector, due to its key role in innovation, is a prominent feature of many national digital-economy strategies (OECD 2015). Recent trends indicate that in the European Union (EU), between 2015 and 2025, there will be an estimated 10% growth in ICT professions, translating into some 400,000 new jobs, whereas employment of science and engineering professionals (which grew by 17% between 2005 and 2015) is forecast to continue growing over the next decade driving employment up by 13% (Cedefop 2016a).

In the U.S., employment in all computer occupations is expected to increase by 12% between 2014 and 2024, adding about 500,000 new jobs, whereas employment of maths occupations (Actuaries, Mathematicians, Statisticians and Operations Research Analysts) is projected to grow 28 percent between 2014 and 2024. This should create about 43,000 new jobs (U.S. Bureau of Labor Statistics, 2015). These trends are justified by growing emphasis on new technological devices and big-data management (collection, storage and analysis), which are handled by these positions. In addition, ICT specialists (Cedefop 2016b), and particularly science, technology, engineering and mathematics (STEM) professionals top the EU list of skills shortages (Lovaglio et al 2016).

Motivation

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, as they allow for timely and fine-grained representations of complex labour market dynamics, in terms of trends, occupations, and skills.

The relevance of Web Labour Market Information

Recently, Google's Chief Economist Hal Varian stated that “the sexy job in the next ten years will be statisticians” (Varian 2012). Also, the Harvard Business Review recently called the data scientist “the sexiest job of the 21st century” (Davenport and Patil 2012). Although data science is a broad concept, it can be generally defined as a multidisciplinary blend of data inference, algorithm development, and technology used to solve complex problems analytically and to generate business value. However, although it is well recognized that STEM professionals hold high-skilled jobs, they are characterized by different task and skill sets in a wide range of economic sectors. Therefore, it is crucial to understand what these jobs actually do in businesses, especially what skills they require and what fields those skills are most readily found in. Accordingly, from a supply perspective, information was drawn from internationally recognized and defined occupation codes, which describe the trends and dynamics of these positions.

In the EU, the ISCO-08 structure defines a job as a bundle of tasks and duties performed by one person (ILO 2012; Hunter, 2014). Jobs with the same set of main tasks and duties are aggregated into 433 occupation units using a 4-digit classification code in this structure. Clearly, this classification system does not identify ‘data scientist’ per se, but it does distinguish related jobs

such as ICT, engineering and statistical positions with their associated tasks and activities. From a perspective of education, the first three major ISCO-08 (Level 1) groups, including professionals (ISCO-2), technicians and associate professionals (ISCO-3), relevant to our discussion, refer to skilled positions requiring tertiary education. According to the ISCED-97 classification system, ISCO-2 occupations require skill levels corresponding to ISCED 6 (second stage of tertiary education) and ISCED 5A (the first stage of tertiary education), whereas the ISCO-3 skill level is closer to what is taught at ISCED 5B (a more vocationally orientated education) and 5A levels.

However, apart from standardized occupational classifications, two main questions arise. First, do ICT and statistical occupations truly refer to different work activities, as assumed in standardized occupational classifications? Second, do the tasks and skills associated with these coded occupations reflect what the labour market really demands for such job profiles? The responses to these questions are crucial since current and future skill demands for these professionals will have significant influence on innovation agenda and economies in general.

Thus, online job portals that collect web vacancies have become an important medium to advertise demand and supply and for job matching used by employers and job seekers, as well as to enhance recruitment activities in new ways. Moreover, they offer fast growing areas of research with strong potential for dealing with diverse socio-economic issues related specifically to using innovative data sources and analytical methods for the study of the labour market (Bergamaschi et al. 2016; Askitas and Zimmermann 2009; D'Amuri and Marcucci 2010; Lovaglio and Vittadini 2014; Lovaglio and Verzillo 2016).

Specifically, web job vacancies can be seen as raw text posted several times on different web portals. These advertisements specify: 1) the job title, 2) the expected skills an ideal candidate should have, 3) the workplace location, 4) the economic sector and some other information, using free text descriptions. Despite the fact that these innovative sources present new opportunities to collect and analyse the labour market, especially from a demand perspective, a significant deterrent to research lies with the unstructured nature of these data, since the extraction of information from texts takes significant effort: automating profiles and skills classification with standardized and internationally recognized classification systems is no trivial task. The generation of a knowledge base derived from web portal texts requires the management of unstructured and semi-structured data (e.g., free text or precompiled text forms). More explicitly, both feature extraction and content categorization/classification are required. The first deals with recognizing a particular data structure, named Weighted Word Pairs, containing the most relevant pairs of lexical items such as single words, or chains of words, while the second requires labelling natural language texts with thematic categories from a predefined set.

Typically, in the context of job-vacancy data from web-crawled job portals, the term Weighted Word Pairs refers to the identification of particular features from texts (e.g. required skills, experience level, field of study or educational level), whereas classification requires the identification of relevant tasks into predefined categories of a target variable (such as the ISCO-08 occupations). The issues surrounding text classification have been widely studied in the database, data mining, and information retrieval communities (Aggarwal and Zhai 2012). Two major approaches have been developed in the research community (Sebastiani 2002):

1) Knowledge engineering, or explicit rules, where a set of rules to classify texts are manually defined;

2) Machine learning techniques, based on a general inductive process that automatically builds a classifier by learning from a set of pre-classified documents (labelled) by human experts.

In the first approach, the rules look for the presence of specific words or combinations of words in the text. The rule design process starts from the identification of appropriate official classifications or taxonomies or from the development of taxonomies obtained empirically through observation of the texts. Then the taxonomic entries are organized in hierarchies. The rules are then designed using the taxonomic hierarchies and entries as references, to obtain a comprehensive set of standards. The main advantage is that analysts become deeply involved in rule formulation, assessment, and tuning. The disadvantage of this method lies mainly in the huge effort required of the experts to develop all the rules.

In the second approach, the classification systems learn to predict labels (occupations/job profiles and so on) using small training samples of documents manually labelled by domain experts (gold standard). The main advantage to this approach is that less human effort is needed compared to the knowledge-engineering approach. However, a disadvantage lies in the fact that almost all categories of the target must be present and labelled in the training sample to let the classifier learn a classification model for future examples. This is a very important issue in labour market text mining, since typically the targets (occupations) present large cardinality and incomplete labels in the data.

An evaluation of the knowledge engineering and machine learning approach on classifying web-job vacancies was performed in 2015 (Amato 2015). Several authors (Lee 2005) have investigated the extraction of meaningful information from unstructured texts in support of various aspects of the recruitment process. In a supervised context, Singh and colleagues (Singh et al. 2010) proposed a system to screen candidate profiles for jobs by extracting various pieces of information from the unstructured resumes through probabilistic information extraction techniques such as Conditional Random Fields. Then (Yu et al 2005), a cascaded Information Extraction model based on SVM for

mining resumes was used, whereas Poch and colleagues (Poch et. al 2014), aimed to match appropriate candidates to a job offer, using supervised classifiers to suggest a ranked list of job offers to job seekers. In an unsupervised context, Yi and colleagues (Yi et al 2007) used structural relevance models to identify job description and resume vocabulary. Instead, a job recommender system, to dynamically update job applicant profiles by analysing their historical information and behaviours, was developed (Hong et al. 2013). Whereas, in e-recruitment within social media, (Lu 2012) focus is put on methods for deriving guidelines on user patterns and profile similarity patterns both in structured and unstructured profiles.

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. The well-known knowledge discovery in databases (KDD) process (Fayyad et al. 1996) was applied as a methodological framework. Approximately 4 million job vacancies were collected in the third quarter of 2015 over five European countries. The purpose of the data cleaning and classification task (volume, variety, and veracity) was to classify each of these vacancies according to the ISCO-08 occupation taxonomy (436 occupation items).

Contribution

In this paper, our analysis focuses on job vacancies on the web related to Italian ICT, statistical positions and skilled professions (globally six occupations, accordingly to ISCO up to level 4), were retrieved by Cedefop[1] research for Italy. Specifically, the paper describes ICT and Statistical occupations in terms of required skills, as emerged from the labour market from a demand perspective. It also aims to identify skills that best distinguish statisticians (divided into two classes of professionals: “high-level” and “low-level” technicians) from the remaining ICT occupations (merged into a single group). To achieve this, several machine learning techniques were used to assess which skills best distinguished ISCO occupations or groups of occupations. Whereas, skills were identified by looking for specific skill related n-grams (an n-gram is a set of n consecutive words) identified by domain experts using a computer assisted vocabulary identification process.

[1] Real-time Labour Market information on skill requirements: feasibility study and working prototype". Cedefop Reference number AO/RPA/VKVET-NSOFRO/Real-time LMI/010/14. Contract notice 2014/S 141-252026 of 15/07/2014 https://goo.gl/qNjmrn

The combination of occupation classification using a machine learning approach and skills identification through n-gram vocabulary was used by CRISP in the Cedefop project.

This paper is organised as follows: the next section will present the text classification methodologies and skills identification techniques used by CRISP (feature extraction and target classification) to relate Italian job vacancy descriptions to ISCO-08 occupations. Section 3 describes methodologies used to find significant skills (retrieved from texts) useful to distinguish retained occupations in general and ICT from Statistical occupations. Section 4 presents the results and Section 5 offers our conclusions.

2. Web scraping and classification

For the data collection step, a modular scraper composed of three distinct components was built. These components were ① a downloader for retrieving the web pages and storing their content into a database, ② an extractor that recognizes and extracts the main elements of a job vacancy and stores them in the database, and ③ a monitor that schedules and executes the overall scraping process periodically. This module was custom built to deal with the high heterogeneity of web sources using the Spring Framework and Talend for the orchestration activities.

Each title and description of the job vacancy was processed according to the following pipeline: **Duplicate removal (a process that removes the same job vacancy found in different web portals),** Tokenization (splitting a sentence into its words and removing punctuation), Stop Words removal (removing numbers and 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). The titles mostly focus on describing the desired occupation. Full descriptive sentences deal with several other items, such as employment and employer description, detailed job description, work location, legal disclaimers, etc. Titles were processed completely. Only those sentences focusing on the desired occupation were used for classification purposes. Sentences in full descriptions related to occupations and skills descriptions were identified using "sentinel words" e.g. "we are looking for", "the desired candidate". Sentinel words are set of consecutive words usually located in sentences focusing on a specific topic (and not in off-topic sentences) and that can be used for retrieving topic related sentences or sub-sentences.

Although the occupation classification uses a 'bag of words' approach, where a text is represented as a set of words, information about word order is not completely lost since single

words, bigrams (two consecutive words) and trigrams (three consecutive words) were also considered, as suggested by successful text mining classification experiences (Cavnar and Trenkle 1994; Cohen and Hersh 2005; Cavnar and Trenkle 1994; Liu et al. 2012).

The skill identification process first created a vocabulary of n-grams related to skills (actually single words and bigrams i.e., one or two consecutive words), which was identified by a team of domain experts using this semi-automatic process: (1) the job vacancy set was partitioned according to the ISCO code (i.e., vacancies classified on the same ISCO code were grouped together); (2) skills related sentences were identified (using sentinel words, in a similar way as the approach previously described) and extracted, each ISCO set was processed separately; (3) for each set, stop words were dropped, words were stemmed, and single word and bigram frequencies were computed. Synonyms were reduced (every element of a set of synonyms was replaced by a representative) using both lexicographic resources and expert contribution. The n-grams were ordered by frequency (separately for each ISCO code). Since the extracted sentences focused on roles and expected profile descriptions and since skills were repeated in vacancies, the skill related n-grams had high frequencies and were easily identified by domain experts (e.g. "tax knowledge" or "SAS_software). In this way, a vocabulary of skill related n-grams was built and was then used to identify skills in job vacancies.

The Text Classification task took charge of extracting features (skills and role characteristics) from job vacancy descriptions (unsupervised) using linguistic models, and specifically linguistic model modules built with custom codes using the Natural Language Processing library PyNLPl of the *SciPy* framework (http://www.scipy.org/), as well as BabelNet (a multilingual encyclopaedic dictionary, with lexicographic and encyclopaedic coverage of terms) for synonym management.

At the last step, the Vectorization process extracted features such as single words, 2-grams and 3-grams (skills) to be used to predict/classify the ISCO-08 code (up to level 4), using machine learning algorithms (supervised).

As a result, a subset of 1,007 job vacancies was randomly sampled using Italian vacancies about which the research team had extensive knowledge due to previous research conducted. The reference vacancies were classified using the ISCO-08 occupation codes by a pool of domain experts (every vacancy was reviewed by at least two experts). The dataset obtained was used as gold benchmark (also known as grand truth) for evaluating classification techniques (Amato et al. 2015). Many classifiers were used and evaluated on the reference set of job vacancies, using both Accuracy and the F1-measure. This latter criterion is a synthetic measure, which, based on a "one category-vs-all" approach in multiclass situations, balanced the average precision and recall of the the *k*-class model (p_k and r_k, respectively), where p_k is the fraction of events where we correctly

classified *k* out of all instances classified by the algorithm as *k*; whereas r_k was the fraction of events where we correctly classified *k* out of all of the cases where the true target was *k*.

The Support Vector Machine (SVM, Vapnik and Chervonenkis 1964) classifier using a linear kernel, was the best choice for both measures, giving an exceptional accuracy of 80.5%, confirming that the SVM is a very competitive classifier in text classification (see Sebastiani 2012 showing the performance of different classifiers on the Reuters-21578 dataset, commonly used as a benchmark).

Several classifiers were extensively evaluated for this task, Linear SVM, RBF Kernel SVM, fully connected Neural Networks, Convolutional Neural Networks working on Word Embedding, the interested reader can refer to (Boselli, 2017a, 2017b) for further details on the evaluated classifiers, the optimization activities, and the parameters tuned. The latter evaluation was performed on a scenario similar to the one described in this paper but focusing on English vacancies. Considering the Italian vacancies, a similar evaluation was performed (excluding Convolutional Neural Networks) and similarly to the English scenario, the Linear SVM was the best performing classifier. The classification approach is the same for both languages, the two language classification processes only differ for the stop word sets and the stemming algorithms. Here are some more details about the selected classification process. The text classification routine was developed in python using scikit-learn (Pedregosa, 2011) and NLTK (Bird, 2009). Specifically, the stop words for the Italian language provide by NLTK were used, the words were stemmed using the Snowball stemmer for the Italian language provide by NLTK, Numeric features were extracted using n-gram counts (n ranging from 1 to 4). N-grams present in less than 4 vacancies or in more than 30% of vacancies were excluded (the former are rare words, very often misspellings, the latter are very common words that are not enough discriminative for classification). The LinearSVC classifier in scikit-learn was used with the parameter C=0.01 (a numerical computing optimized implementation of the linear SVM classifier based on liblinear (Fan, 2008).

Once the Job vacancies were grouped according to the same (predicted) ISCO code, sentences related to skills for each group were identified using linguistic models, identifying words, bigrams and trigrams (skill features). These were able to identify both specific skills related occupations (e.g. “tax knowledge”, “SAS_software”) or noise such as n-grams not related to skills (e.g. “ideal_candidate”). Noise skills were processed by domain experts and BabelNet, to reduce synonyms that were replaced by a common representative.

Extracted skill features were not further grouped or classified (since there is no international standard for skill classification equivalent to ISCO for occupations) with the exception of skill features related to “experience” that were classified in the following categories: previous experience

with the position, previous experience with the working environment (field) and previous experience with the sector.

In this way, the data classified according to the ISCO-08 occupation taxonomy was enriched with information about the skills and experience requested by the employers, thus producing a detailed portrait of the job opportunities advertised on the web. Furthermore, where possible, each web vacancy was classified according to a required educational level, 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 taxonomies, such as NUTS (Nomenclature of Territorial Units for Statistics) for territorial areas, NACE for activity areas and ISCED for education levels.

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:

- Occupations: ISCO-08 classification up to level 4
- Territorial units: Up to NUTS 3
- Sector of economic activity. NACE classification up to level 2
- Educational level required (ISCED level 1)
- Skill (not classified, text retrieved)

As previously mentioned, we selected Italian web job vacancies and ISCO-08 (level 4) skilled occupations related to ICT and statistical professions, identifying six occupations, five from professionals and one for technicians from the major ISCO groups. Hence, several machine-learning classifiers were used to identify features related to skills, which best distinguish these six occupations (ISCO-08 Level 4). In a second analysis the six-level target was recoded for three levels, in order to identify the skills that distinguish two statistical occupations (ISCO-08: 2120 and 3314). The remaining occupations were merged into a single group ("Grouped Professions Level 4"). The next section briefly introduces the models that were fitted to the dataset especially concerning their features and any pros and cons for each.

3. ICT and Statistics vacancies: Classifiers

Wide data (datasets with large numbers of features, greater than observations) introduce a level of complexity that most classical models, such as generalized linear models (GLM) or algorithms, cannot manage well due to computational complexity or limitations of space or dimensions.

Specifically, when we have inputs other than observations, we cannot fit them into classical GLM using standard approaches (problems of non-convergence, non-invertible design matrices, non-uniqueness of parameters and high overfitting). Moreover, a sparse matrix (large matrices but with only a few nonzero entries, typically in dummy-coded inputs) induces problems linked to numerical approximation (either ill conditioned or quasi-singularity of design matrices) and optimization.

3.1 Logistic Regression with Ridge & Lasso penalization

Penalized regression models are tools widely used to resolve the above problems. They estimate parameters in large and sparse matrices by shrinking the maximum likelihood of estimated coefficients towards zero, where penalties are introduced to the model building process to avoid over-fitting, reduce variance of the prediction error and handle correlated predictors. The two most common penalized models are Ridge regression and Lasso-Least absolute shrinkage and selection operator (Tibshirani 1996).

Ridge regression penalizes the ℓ_2 norm of the model coefficients and provides greater numerical stability. This method keeps all the predictors in the model and shrinks them proportionally. Ridge regression reduces coefficient values simultaneously as the penalty is increased without however setting any of them to zero. Lasso regression penalizes the ℓ_1 norm of the model coefficients forcing some of the coefficient estimates to be exactly equal to zero, thus working as a variable selector (sparsity property). The two penalties also differ in the presence of correlated predictors. The ℓ_2 penalty shrinks coefficients for correlated columns towards each other, while the ℓ_1 penalty tends to select only one of them and set the other coefficients to zero.

The only tuning parameter we set using both algorithms was λ, which serves to control the relative impact of the shrinkage on the regression coefficient. A value of zero always means no shrinkage (penalized coefficient estimates coincide with maximum likelihood estimation), whereas for increasing λ the impact of the shrinkage penalty grows, and the regression coefficient estimates will approach zero.

The properties of both approaches are beneficial because they reduce the variance in the predictions and make the model more interpretable by selecting a subset of the given variables.

Therefore, the Elastic Net (EL, Zou and Hastie 2005) classifier is the weighted sum of the ℓ_1 (least absolute shrinkage and selection operator or lasso) and ℓ_2 (ridge regression) norms of the coefficients vector. The EL parameter $\alpha \in [0; 1]$ controls the penalty distribution between the ℓ_1 and ℓ_2 penalties. When $\alpha = 0$, the lasso penalty is not used and a ridge regression solution with shrunken coefficients is obtained. If $\alpha = 1$, the Lasso operator soft-thresholds the parameters by reducing all

of them by a constant factor and truncating at zero. This sets a different number of coefficients to zero depending on the α value. Moreover, while the number of predictors that can enter a Lasso model saturates at *min(n; p*) (where *n* is the number of observations and *p* is the number of variables in the model), the elastic net does not have this limitation and can fit models with a larger number of predictors.

Regarding the tuning parameters, for a specific α value, the algorithm can compute models for a single value of the tuning parameter λ or the full regularization path (α,λ) as in the *glmnet* package for R. However, whatever their virtues, Lasso and EL have been criticized since their estimates were inconsistent and biased (Fan 1997; Fan and Li 2001), respectively. Moreover, contrary to EL, Lasso does not satisfy the oracle property (an estimator both consistent in variable selection and asymptotically unbiased in parameter estimation), while the regularization parameter for model selection consistency is not optimal for prediction accuracy (Zhao and Yu 2006). Consequently, Fan and Li (2001) introduced the smoothly clipped absolute deviation (Scad, Fan and Li, 2001), which satisfies the oracle property, and other properties, such as unbiasedness, sparsity, and continuity (the estimator is continuous in the data to reduce instability in model prediction).

For the SCAD the penalty in the loss function is a folded-concave quadratic spline with a positive tuning parameter γ used to adjust the concavity of the penalty. The smaller γ is, the more concave the penalty becomes, which means finding a global minimizer is more difficult; on the other hand, the resulting estimators decrease the parameters' bias.

More recently, Wang and Leng (2007) in the context of robust regression**,** proposed the least absolute deviation Lasso (LAD-Lasso) regression estimator to carry out robust parameter estimation and variable selection simultaneously. The resulting regression estimator has oracle property, and unlike Lasso, is resistant to the outliers in the response variable (Wang and Leng 2007; Xu and Ying 2010).

These methods, however, having been developed for models with univariate responses such as continuous/binary/Poisson/time-to-event (SCAD implemented in *SIS* and *ncvreg* R-packages) or only for continuous targets (LAD-Lasso in *flare* R-package) do not cover multinomial regression, a class of models where the effect of one predictor variable is represented by several parameter/binary models. Since variable selection requires that all parameters that belong to one variable be simultaneously removed from the model and it is not clear how this complex form of selection affects the performance of estimation methods, these methods could not be applied directly to our classification problem.

3.2 K-Nearest Neighbors

K-Nearest Neighbors (KNN, Fix and Hodges 1951) is a non-parametric algorithm of machine learning used for classification and regression that does not make any assumptions on the underlying data distribution. According to KNN, given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by N_0. Then it estimates the conditional probability for class k as the fraction of points in N_0 whose response values equal k (majority vote). Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability. The KNN tuning parameter k was chosen by a grid search using cross-validation on the training dataset. Notwithstanding its simplicity, this approach typically overfits and with large dataset can be computationally expensive in the training phase.

3.3 Random Forests

Random Forests (RF, Breiman 2001) algorithm is one of the most accurate learning algorithms available. The algorithm builds a number of decision trees on bootstrapped training samples and at each split in the tree, the algorithm is not even allowed to consider a majority of the available predictors. In fact, a random sample of m predictors is chosen as split candidates from the full set p of features; this aspect differentiates RF from bagging tree (where $m = p$). The RF algorithm naturally operates as a model selector, giving, in different metrics, measures of feature importance. Its tuning parameters are the *number of trees* and the *number of features*. The first indicates the number of trees to grow, whereas the second defines the Number of features randomly sampled as candidates at each split. Typically, in the RF original formulation each tree is grown to the largest extent possible (*Max depth*) and, although pruning trees works nicely for a single decision tree because it removes noise, doing this within RF kills bagging which relies on it for having uncorrelated members during voting. However, one possible disadvantage of RF with its largest *depth* lies in the minimum size of its terminal nodes.

Hence, we specified a Random Forests algorithm adding to the grid a search the *Max depth* parameter to avoid *minimum sample size in leaves, fitting an* RF algorithm using H2O package (Aiello et al 2016), an open-source software for big-data analysis, which, with its speed and flexibility, allows users to find thousands of potential models as part of specifically discovering patterns in data.

One suitable output of RF is variable importance, measuring the discrimination power of features among target classes: *Global variable importance* is the mean decrease of accuracy over all excluded from training samples (out-of-bag) predictions, when a given variable is permuted after

training, but before prediction. In addition, *class variable importance* measures the strength and direction of each important feature toward each target class.

3.4 Gradient Boosting

Gradient Boosting is an ensemble technique, which means that prediction is done by an ensemble of base learners (typically tree), giving essentially an additive model. Unlike Bagging and Random Forests where the aggregation is made using bootstrap samples, which are independent of each other, in Gradient Boosting, when base learners are added, they are fitted on new examples sampled in relation to the base learners' accuracy in each iteration. After a base learner is added in the additive model, the data are reweighted: examples that are misclassified gain weight and examples that are classified correctly lose weight.

Being an iterative process, Gradient Boosting continues to add learner(s) until a limit is reached in the number of models or misclassified cases. Generally, boosting algorithms take longer than RF algorithms because trees are built sequentially. The most popular boosting models are AdaBoost (Freund and Schapire 1999) and Gradient Boosting (Friedman 1999, 2001).

Benchmark results have shown gradient boosting algorithms to be very competitive learners, although they are prone to overfitting. This justifies a careful tuning pre-processing phase. The main tuning parameters are the *number of trees, the maximum depth* (like RF) and the *learning rate*, a parameter bounded from *0 to 1* shrinking the contribution of each new base learner that is added in the series of the additive model. Higher learning rates occur in less stable models and lower rates result in slower convergence.

Whatever their virtues, there are still open questions about the success of Boosting algorithms and some unjustified properties when judged from a statistical point of view, such as the apparent resistance to overfitting (also when the algorithm is run for a very large number of iterations) or other counterintuitive results about regularization, complexity of base-classification trees and so on (Freidman et al. 2000; Mease and Wyner 2008).

Moreover Freidman and colleagues (Freidman et al. 2000) strongly caution the use of Boosting algorithms in multiclass cases, since such classifiers build separate two-class models for each individual class against the pooled complement classes, thereby de facto maximizing a separate Bernoulli log-likelihood for each class versus the others. Since this class pooling may produce complex decision boundaries that are difficult to approximate, the literature advocates the use of all the classes simultaneously, maximizing the log-likelihood with a classical multinomial distribution for the response.

3.5 Support Vector Machines: Linear & Non-Linear Kernel

Support Vector Machine (SVM, Vapnik and Chervonenkis 1964; Boser et al. 1992) is a supervised machine-learning algorithm, which can be used for both classification and regression challenges. The general idea in classification problems is that if data can be perfectly separated using a hyperplane, because an infinite number of hyperplanes exists, the final choice will be the *maximal margin hyperplane* (MMH), which is the separating hyperplane that is farthest from the training observations. We will then classify a test observation based on which side of the *maximal margin* hyperplane it lies.

This is a very natural way to perform classification if a separating hyperplane exists. However, if linear separability does not hold, we might consider a hyperplane that best classifies most of the training observations, fixing a maximum sustainable cost C in terms of misclassified examples. In this approach, defined by a *soft margin classifier*, the cost *C* represents the number of violations to the margin we are willing to tolerate. If $C = 0$ then no observations will be allowed to violate the margin, whereas for $C > 0$ no more than *C* observations would be misclassified.

C is essentially a tuning parameter: when *C* is small, we will have a classifier that closely fits the data, which may have low bias but high variance. On the other hand, when *C* is larger, the margin is wider and we allow more violations to it. This will lead to a classifier that is potentially more biased but may have lower variance.

The SVM is a further extension of the *soft margin classifier*, which arises from enlarging the feature space in a specific way, using kernel functions, in order to overcome the problem of non-linear class boundaries. We use both polynomials, where we have to define the correct polynomial order (*p*) and radial Kernels, depending on a free parameter γ, which essentially is the inverse of the variance of the (Gaussian) Radial Basis Kernel, as a scaling factor, weighting squared Euclidean distances between examples. Gamma is a tuning parameter, which defines the similarity between two points: a small gamma value (Gaussian Kernel with a large variance) indicates that two points can be considered similar even if they are far from each other, whereas for larger gamma values (Gaussian Kernel with a small variance) two points are considered similar only if they are close to each other. Hence, for polynomial and radial kernels, the tuning parameters are (C,γ) and (C,p), respectively.

4. Results

Generally, when scraping the selected Italian web sites between June and September 2015, the system collected 276,909 job advertisements. After quality control and duplicate removal, the number of vacancies was reduced to 137,291. A set of Italian web sites was selected by a team of

domain experts considering several types of web sites: job portals directly publishing advertisements, aggregators publishing advertisements of other sites and employment related newspaper web site sections. Nearly a quarter of these job vacancies were not used since the process was not able to extract relevant information such as Sector of economic activity or required education title (103,094 vacancies were analysed for both), or required occupation (110,950 vacancies were analysed) from the web pages due to null values.

All in all, 67% of the vacancies analysed were concentrated in the services sector, 33% in manufacturing and only 0.5% in construction (32.6%). Excelsior records (extracted in the same July-September 2015 quarter) reveal an overrepresentation of manufacturing activities (17%) and underrepresentation of services (77%) and construction (7%) in web job vacancies. More specifically, regarding industry (NACE level 1), web vacancies for services tend to be more concentrated in the following activities: wholesale and retail trade (15.1%), administrative and support service activities (13.7%), information and communications (9.8%) and professional, scientific and technical activities (9.8%), whereas for 14% of the jobs it was not possible to determine the activity sector. Moreover, only 13% of the job vacancies required a middle school education or lower (56% required secondary education and 27% a Bachelor's degree or higher).

Looking at aggregate occupations (ISCO-08 Level 1), web vacancies display a higher concentration of high skill occupations (48%), with the largest share by technicians and associate professionals (25%), professionals (20%), clerical support workers (14%), crafts and related trade workers (13%), service and sales workers (11%).

Table 1 illustrates the distribution of required occupations by ISCO-08 code Level 1 and for occupations of our interest (professionals and technicians) also Level 2 and Level 3.

Table 1: ISCO-08 occupations by Level 1 (Bold), Level 2 (Italics) and Level 3 (small caps). Counts and percentages on total (n =110,950).

A comprehensive comparison with the Excelsior records reveals that web vacancies display a higher concentration of high skill positions and an underrepresentation of service and sales positions (18% and 38%, respectively in the Excelsior database).

4.1 ICT and Statistical occupations

From the 110,950 vacancies where the ISCO-08 code was predicted, we selected those job vacancies related to ICT professions and statistics, accordingly with the codes reported in Table 2.

From the set of statistician-related vacancies we excluded 129 job offers classified as “statistical, finance and insurance clerks” (ISCO code 4312), since this occupation (contrary to the other two statistician occupations) does not belong to the skilled occupations (ISCED 6, 5A, 5B), and more concretely does not require basic statistical skills such as using standard computer software packages or applying knowledge of statistical principles and practices in the course of the work. Overall we obtained 6,222 job vacancies (representing 5.6% of the overall set in the period analysed), largely dominated by software developers.

Table 2: ICT and Statistical occupations (ISCO-08 Level 4 and grouped Professions).

Table 3 illustrates the most required skill for each occupation, divided by skill type: ICT (all skills related to software, hardware, programming, web etc.), professional (skills concerning knowledge in Mathematics, Statistics, Economy, Engineering, Law, Management, etc.) and soft skills (concerning other skills not directly related to both previous types) contained in the web vacancies and their distribution among the three types.

Table 3 Three most required skills by occupations and skill type and percentage distribution of skill type.

As expected, despite the fact that the analysed occupations referred to highly technical professions, a large amount of soft skills was also required, especially for most highly educated occupations (42.5% and 48.1% out of the total skills required for “high-level” statisticians and engineers, respectively, were soft skills). In the classification analysis aimed at identifying the most discriminant skills among the six occupations, we used all the classifiers presented in Section 3. Specifically, to determine the best combination of tuning parameters to use on the test dataset, a validation approach minimizing the cross-validated misclassification rate on training data (50% of initial data stratified by target values) was used.

Classification results were assessed using accuracy (% of testing set examples correctly classified by the classifier), error rate (mean of error for each class weighted by priors, using sample proportions of the target in the training data) and Brier score (a generalization of average square error between observed and predicted probabilities with target multiclass). Table 4 illustrates the results in terms of tuning parameters and performance measures on test data (remaining 50% of initial dataset).

Table 4: Tuning parameters and classification results on test data

As summarized in Table 4, the best model was Random Forests, both in terms of accuracy and Brier score (error rates for specific target classes ranged from 6% for ISCO-08 code 2512 to 75% for ISCO-08 code 2120), whereas the worst performance was achieved by both K-NN and Ridge regression, perhaps due to the highly imbalanced datasets. Hence, large amount of overlapping between occupations remained using available skills. Moreover, Figure 1 lists the top twenty most important skills useful to distinguish the six occupations, where text area is proportional to (scaled) the variable importance found in the RF (global variable importance).

Figure 1: Top twenty most important/discriminant skills (Random Forests), text scaled by variable importance

In the second analysis, as mentioned, we repeated the classification analysis with a three-level target to identify distinguishing skills among statistical occupations ("High level", ISCO-08 code 2120; "Low level", ISCO-08 code 3314) and other ICT occupations merged as a single group ("Others").

Table 4 reports performance results, as well as error rates for each target class. Overall, classical RF (largest trees) and Gradient Boosting were the best models leading to a suitable classification performance (accuracy higher than 95%), significantly increased from the six-level target.

Table 5: Tuning parameters and classification results on test data

As for the six-level target, we selected the most important skills that best distinguished the target classes from RF. Figures 2 and 3 show the variable importance of skills with importance not lower than 10% for the most important variable, for high-level and low-level statisticians, respectively (class variable importance).

Firstly, notice the difference between the number of variables that appear in the word clouds: for the "low-level" there are 14 variables whereas for the "high-level" there are 28, exactly double. This means that in the first case fewer skills but with deeper knowledge are required, whereas for the "high-level", a larger number of skills and competences is more relevant for choosing a good candidate for the job. The most important skill required for the high-level statisticians is data

analysis, an "acronym" probably meaning the entire process of transforming and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.

Other skills in order of importance are *MsOffice* (whose importance is 87% of the most important), *JavaProgramming* (62%), experience in the working environment/field (48%) and *BudgetSoftware* (41%). Conversely, *SAS/software* is the most important skill linked to low-level statistician with a wide gap from the second important variable in the ranking (36% of importance of the most important variable), followed by *JavaProgramming* (21%), JavaScript (16%) and *DataIntegration* (15%).

Instead, in high-level statisticians, specific statistical software does not even appear among the most important variables, whereas there were largely required programming skills linked to Java, Python and HTML. Probably, SAS is a required skill because in the lower-levels all those jobs linked were grouped more around computational oriented statistics than to data manipulation and standard analyses. For the same reason, the *DataIntegration* skill and other skills linked to computer knowledge were considered significant. Furthermore, other important skills for high-level statisticians were *ExperiencesInField*, justified because such jobs generally occupy higher positions in the business organization with more responsibilities where previous experience is highly recommended.

Figure 2: Most important skills for "High-level statisticians" (ISCO-08 code 2120), scaled by variable importance (up to 10% of the most important variable in Random Forest).

Figure 3: Most important skills for "Low-level statisticians" (ISCO-08 code 3314) scaled by variable importance (up to 10% of the most important variable in Random Forest).

Lastly, variable importance for the other occupations (not reported) showed that occupations were more oriented towards disciplines like computer science, web programming and engineering science, since the most important skills deal with specific software tools (such as AutoCAD, SapCRM, Linux), skills linked to programming languages (VBScript, Abap) and soft/managerial skills needed in management positions (CommunicationSkills, Marketing/Communications, Management and Economy/Administration).

5. Conclusions

This paper sheds light on the relevance of applying "intelligent" techniques and data engineering to manage the main issues related to big data in a real and domain-specific context. Specifically, we presented job profiles required in the third quarter of 2015 using three Italian web portals in terms of ISCO-08 occupations, while also analysing skill profiles for ICT and statistician positions.

The data showed obvious signs that different and more heterogeneous professional backgrounds are required of high-level statisticians. This level was more linked to theoretical statistics, in which analytical skills are more relevant than computing skills. Moreover, this profile involves many soft and management oriented skills, missing in the lower levels, typically restricted to more technically oriented computing and informatics jobs. These examples emphasized that these innovative sources presented new opportunities to collect and analyse labour market trends, especially from a demand perspective.

Notwithstanding, several methodological issues persist concerning the use of web vacancies which are predominantly related to the quality and representativeness of such data, and generalizability of the findings (Gosling et al. 2004; Pedraza et al. 2007; Steinmetz et al. 2009; Štefánik 2012). A key challenge in using online job vacancies was ascertaining whether the set of online job vacancies was a representative sample of all job vacancies in a specified economy. Even if the population of job vacancies can be considered finite at any given moment in time these are not easily counted nor are their structures easy to determine. Moreover, job vacancies are fundamentally voluntary, jobs and people are reallocated in the labour market and in firms, tasks are split, restructured or partially switched and recruitment strategies might have sectoral and occupational specificities (Mang 2012; Pedraza et al 2007; Gosling et al. 2004). Statistically speaking, samples of available web vacancies are prone to problems of self-selection and/or non-ignorable missing mechanism (Little and Rubin 1987). Various approaches have been taken by different authors in attempting to deal with representativeness issues related to the usage of online job vacancy data for analytical and policy-making purposes.

Among others, a number of studies decided to focus on the segment of the labour market characterized by widespread access to the Internet (e.g. ICT) in the belief that they would be relatively well represented among the job applicants due to the characteristics of Internet users and would therefore offer a representative data sample (Kennan et al. 2006; Wade and Parent 2001). Others consider online data generalizable due to the dominant market share and very high reputation of the chosen portal among employers and employees (Kureková et al. 2012).

Our research findings pave the way for future work in several directions. First, similar occupations should be automatically grouped on the basis of the skills requested by employers.

Second, collected knowledge can be represented by a graph-based model, which is a natural and convenient choice for large and highly dynamic knowledge bases including all the job vacancies (tens of millions of nodes). Third, similar evaluations could be performed in several countries and the results compared. This would represent a valuable knowledge base that would be beneficial for research activities in the domain of labour market intelligence.

The results of our previous project were validated as effective and useful by Cedefop agency. This fact stimulated an additional call-for-tenders for extending the prototype to the whole EU community. In 2017, we were awarded a grant by the Cedefop to extend the previous prototype to all the 28 EU Countries.[2]

[2] "Real-time Labour Market information on Skill Requirements: Setting up the EU system for online vacancy analysis" AO/DSL/VKVET-GRUSSO/Real-time LMI 2/009/16. Contract notice - 2016/S 134-240996 of 14/07/2016 https://goo.gl/5FZS3E

References

Aggarwal C, Zhai C (2012) Mining text data. Springer, Heidelberg.

Amato F, Boselli R, Cesarini M, et al (2015) Challenge: processing web texts for classifying job offers. In: Kankanhalli MS, Li T, Wang W (eds) Proceedings of the 2015 IEEE International Conference on Semantic Computing; *IEEE* Computer Society Press, Anaheim, CA, pp 460−463.

Askitas N, Zimmermann K (2009) Google econometrics and unemployment forecasting. IZA Discussion Papers No. 4201. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1465341. Accessed 13 June 2015

Bergamaschi S., Carlini E, Ceci M et al (2016) Big data research in Italy: a perspective. Engineering 2:163–170

U.S. Bureau of Labor Statistics (2015). Occupational Outlook Handbook. December 17, 2015. http://www.bls.gov/ooh/ Accessed 17 November 2016

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory. ACM press, Pittsburgh, pp 144–152.

Breiman L (2001) Random Forests. Mach Learn 45(1): 5–32.

Cavnar WB, Trenkle JM (1994) N-gram-based text categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Ann Arbor: Michigan, pp 161–175.

Cedefop (2016a) Information and communication technology professionals: skills opportunities and challenges. Skills panorama. http://skillspanorama.cedefop.europa.eu/en/analytical_highligths/information-and-communication-technology-professionals-skills-opportunities. Accessed 13 November 2016

Cedefop (2016b) Skill shortage and surplus occupations in Europe. Cedefop Briefing notes 9115. http://www.cedefop.europa.eu/en/publications-and-resources/publications/9115. Accessed 13 November 2016

Cohen, AM, Hersh WR (2005) A survey of current work in biomedical text mining. Brief Bioinf 6:57–71.

D'Amuri F, Marcucci J (2010) Google It! Forecasting the US unemployment rate with a Google Job search index. Nota di lavoro/Fondazione Eni Enrico Mattei: Global challenges. http://www.econstor.eu/handle/10419/43536 Accessed 22 December 2015

Davenport TH, Patil DJ (2012) Data scientist: the sexiest job of the 21st Century. HBR 90:70–76.

Fan J (1997) Comments on "Wavelets in statistics: A review" by A. Antoniadis. J Italian Statist Assoc 6:131–138.

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc 96:1348–1360.

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.

Fix E, Hodges JL (1951) Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas.

Freund Y, Schapire R (1999) A Short Introduction to Boosting. Trans Jpn Soc Artif Intell 14:771–780.

Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: A statistical view of boosting. *Ann Stat* 28:337–374.

Friedman, JH Stochastic Gradient Boosting (1999). IMS Reits lecture.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. Ann Stat 29(5):1189–1232.

Gosling SD, Simine V, Srivastava S, John OP (2004). Should we trust web-based studies. Am Psychol 59(2): 93–104

Aiello S, Eckstrand E, Fu A, Landry M, Aboyoun P (2016) Machine Learning with R and H2O. http://h2o.ai/resources/ Accessed 18 September 2016

Bird S, Ewan K, and Edward L. (2009), Natural Language Processing with Python, O'Reilly Media.

Boselli R., Cesarini M., Mercorio F., Mezzanzanica M. (2017a) Using Machine Learning for Labour Market Intelligence. In: Altun Y. et al. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2017. Lecture Notes in Computer Science, vol 10536. Springer, Cham

Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., Pasi, G., & Viviani, M. (2017b). WoLMIS: a labor market intelligence system for classifying web job vacancies. Journal of Intelligent Information Systems, 1-26.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. Journal of machine learning research, 9, 1871-1874.

Hong W, Zheng S, Wang H (2013) Dynamic user profile-based job recommender system. Proceedings of the 8th International Conference Computer Science and Education, pp 1499–1503.

Hunter D (2014) The design principles of ISCO-08: challenges for coding occupations globally. ILO Geneva. Presentation given at Amsterdam, Ingrid Workshop, February 10, 2014. https://inclusivegrowth.be/events/call6-ExpertWorkshop/programme-and-presentations Accessed 18 September 2016

ILO (2012). International standard classification of occupations ISCO-08 Volume 1: structure, group definitions and correspondence tables. International Labour Office, Geneva.

Kennan MA, Cole F, Willard P et al (2006) Changing workplace demands: what job ads tell us. Aslib Proceedings 58(3):179–196. doi:10.1108/00012530610677228.

Kureková L, Beblavý M, Haita C (2012) Qualifications or soft skills? Studying demand for low skilled from job advertisements. NEUJOBS Working Paper No. 4.3.3 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2402729 Accessed 21 January 2017

Lee I (2005) The evolution of e-recruiting: a content analysis of Fortune 100 career websites, JECO 3(3):57–68.

Little RJA, Rubin DB (1987) Statistical Analysis with Missing Data. John Wiley & Sons, New York.

Liu X, Rujia G, Liufu S (2012) Internet news headlines classification method based on the N-Gram language model. Proceedings of international conference on Computer Science and Information Processing (CSIP), pp 826–828, doi: 10.1109/CSIP.2012.6308980.

Lovaglio PG, Vacca G, Verzillo S (2016) Human Capital Estimation in Higher Education. Adv Data Anal Classif 10(4):465–489.

Lovaglio PG, Vittadini G (2014) Structural Equation Models in a Redundancy Analysis Framework with Covariates. Multiv Behav Res 49:486–501.

Lovaglio PG, Verzillo S (2016) Heterogeneous Economic Returns to Higher Education: evidence from Italy. Qual Quan 50:791–822.

Lu Y, Helou ES, Gillet D (2012) Analyzing user patterns to derive design guidelines for job seeking and recruiting website. Proceedings of 4th International Conferences on Pervasive Patterns and Applications, pp. 11–16.

Mang C (2012) Online job search and matching quality. IFO Institute for Economic Research at the University of Munich. ftp://ftp.zew.de/pub/zewdocs/veranstaltungen/ICT2012/Papers/Mang.pdf. Accessed 2 October 2016

Martikainen J (2010) Weighting and estimation methods: JVS estimation in Finland by Horowitz-Thomson-type estimator. Proceedings of 1st and 2nd International Workshops on Methodologies for Job Vacancy Statistics, Eurostat. Luxembourg, pp 32–38.

Mease D, Wyner A (2008) Evidence Contrary to the Statistical View of Boosting. J Mach Learn Res 9:131-156

Mezzanzanica M, Boselli R, Cesarini M, Mercorio F (2015) A model-based evaluation of data quality activities in KDD. Inform Process Manag 51(2):144–166.

OECD (2015) Digital Economy Outlook 2015. OECD Press, Paris.

Pedraza P, Tijdens K, Muñoz de Bustillo R (2007). WP 60-sample bias, weights and efficiency of weights in a continuous web voluntary survey. AIAS, Amsterdam Institute for Advanced Labour Studies. http://ideas.repec.org/p/aia/aiaswp/wp60.html Accessed 23 July 2015.

Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pp. 2825-2830.

Poch M, Bel N, Espeja S, Navio F (2014) Ranking job offers for candidates: learning hidden knowledge from big data. Proceedings of the Ninth international conference on Language Resources and Evaluation, pp 2076-2082.

Sebastiani F (2002) Machine learning in automated text categorization. *ACM Computing Surveys* 34(1): 1–47.

Singh A, Rose C, Visweswariah K, Chenthamarakshan V, Kambhatla N (2010) Prospect: a system for screening candidates for recruitment. Proceedings of the 19th ACM international conference on Information and knowledge management, pp 659-668.

Stefánik M (2012) Internet job search data as a possible source of information on skills demand (with results for Slovak university graduates). In: Cedefop (ed) Building on skills forecasts — comparing methods and applications, CEDEFOP, Luxembourg, pp 246-260.

Steinmetz S, Tijdens K, Pedraza P (2009) WP 76-Comparing different weighting procedures for volunteer web surveys. AIAS, Amsterdam Institute for Advanced Labour Studies. http://ideas.repec.org/p/aia/aiaswp/wp76.html. Accessed 18 September 2016

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J. R. Statist. Soc. B 58:267–288

Vapnik V, Chervonenkis A (1964) A note on one class of perceptions. Automat Remot Contr 25.

Varian H (2009) The McKinsey Quarterly, January 2009.

Wade, Michael R., and Michael Parent. 2001. “Relationships between Job Skills and Performance: A Study of Webmasters.” Journal of Management Information Systems 18 (3): 71–96. doi:10.2307/40398554.

Wang H, Leng C (2007) Unified lasso estimation by least squares approximation**.** J Amer Statist Assoc 102:1039–1048

Xu J, Ying Z (2010) Simultaneous estimation and variable selection in median regression using Lasso-type penalty. Ann Inst Stat Math 62:487–514

Yi X, Allan J, Croft WB (2007) Matching resumes and jobs based on relevance models. Proceedings of the 30th annual international ACM STGTR conference on Research and development in information retrieval, pp. 809-810.

Yu K, Guan G, Zhou M (2005) Resume information extraction with cascaded hybrid model. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics pp. 499-506.

Zhao P, Yu B (2006) On model selection consistency of Lasso. J Mach Learn Res 7:2541-2563.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J. R. Statist. Soc. B 67:301-320.