Survey Research Methods (2021)
Vol. 15, No. 2, pp. 147-167
doi:10.18148/srm/2021.v15i2.7670
European Survey Research Association

Enhancing the Demand for Labour Survey by Including Skills from Online Job Advertisements Using Model-Assisted Calibration

Maciej Beręsewicz
University of Economics and Business
Poznań, Poland, and
Poznań Statistical Office

Greta Białkowska
SKN Estymator
University of Economics and Business
Poznań, Poland

Krzysztof Marcinkowski
SKN Estymator
University of Economics and Business
Poznań, Poland

Magdalena Maślak
SKN Estymator
University of Economics and Business
Poznań, Poland

Piotr Opiela
SKN Estymator
University of Economics and Business
Poznań, Poland

Katarzyna Pawlukiewicz
SKN Estymator
University of Economics and Business
Poznań, Poland

Robert Pater
University of Information Technology and Management
Rzeszów, Poland and Educational Research Institute
Warsaw, Poland

In the article we describe an enhancement to the Demand for Labour (DL) survey conducted by Statistics Poland, which involves the inclusion of skills obtained from online job advertisements. The main goal is to provide estimates of the demand for skills (competences), which is missing in the DL survey. To achieve this, we apply a data integration approach combining traditional calibration with the LASSO-assisted approach to correct coverage and selection error in the online data. Faced with the lack of access to unit-level data from the DL survey, we use estimated population totals and propose a bootstrap approach that accounts for the uncertainty of totals reported by Statistics Poland. Our empirical results show that online data significantly overestimate interpersonal, managerial and self-organization skills while underestimating technical and physical skills. This is mainly due to the under-representation of occupations categorised as Craft and Related Trades Workers and Plant and Machine Operators and Assemblers.

Keywords: big data; non-probability samples; job vacancies; web-scrapping; data integration

1 Introduction

The process of matching job seekers with job offers, and resulting structural mismatch between labour supply and demand is one of the most challenging problems that need addressing in the labour market. It requires continuous attention of labour market and educational institutions. This mismatch can be determined by various factors. Boudarbat and Chernoff (2012) have observed that it depends more on educational characteristics than demographic and socioeconomic factors. Educational mismatch has been widely analysed from the perspective of levels and fields of education, as well as occupations (Somers, Cabus, Groot, & van den Brink, 2019).

So far, it has been possible to analyse and address such problems as under-education, over-education, occupational and educational mismatch. However, problems of skills or competences (we use both terms as synonyms) mismatch, to a large extent, remain unresolved. Hershbein and Kahn (2018) argue that by looking directly at the skill requirements in job offers rather than relying on assumptions about the skills associated with a particular occupation, it is possible to document the evolution in skill requirements for this occupation over time. Even though the demand for job-related skills can, to some extent, be evaluated by looking at the occupational and educational composition of jobs, it is impossible to make inferences about the demand for transversal skills, that is the ones not specifically related to a particular job, but transferable between different jobs.

Workers' skills have been measured from a macroeconomic perspective in a number of studies, including the

Contact information: Maciej Beręsewicz, Poznań University of Economics and Business, Al. Niepodległości 10, 61-875 Poznań, Poland (e-mail: maciej.beresewicz@ue.poznan.pl

Organisation for Economic Co-operation and Development (OECD) Survey of Adult Skills from the Program of International Assessment of Adult Competences (McGowan & Andrews, 2015). However, continuous research on the demand for skills is scarce. Surveys conducted by national statistical institutions (NSI) provide representative data on vacancies across occupational groups or economic sectors. However, they lack detailed information about measures of skills. One potential source of information on skills demand are online job offers placed by employers or entities that work on their behalf.

Big data and the Internet as a data source have become an important issue in statistics, particularly in official statistics. There are a number of multinational initiatives (e.g. European statistical system ESSnet on Big Data; American Association for Public Opinion Research, AAPOR Task Force on Big Data; European Centre for the Development of Vocational Training, Cedefop; the Makswell project, a consortium of 8 academic and national statistical institutions) that focus on the quality and suitability of estimates based on new data sources to supplement existing statistical information. For instance, Cedefop (2019a) and Cedefop (2019b) provides an overview of job vacancies and trends in European Union (EU) countries. The ESSnet on Big Data included a work package devoted to job vacancies (Workpackage 1: Webscraping—job vacancies). Its aim is to produce statistical estimates of online job vacancies using suitable techniques and specific methodologies. The intention was to explore a mix of sources including job search sites, job adverts on enterprise websites, and job vacancy data from third party sources. ESSnet on Big Data (2017) and ESSnet on Big Data (2018) was devoted to web-scraping, text mining, classification and comparison with official statistics. The latter was either at the level of statistical units (companies) or based on Statistical classification of economic activities in the European Community (NACE) sector and International Standard Classification of Occupations (ISCO) occupation variables. The project is being continued with three main tasks devoted to 1) methodological framework, 2) statistical output and 3) implementation requirements of prototypes in the relevant statistical production processes at European and national level (ESSnet on Big Data, 2019).

However, before online data can be used for official statistics, it is crucial to explore potential sources of representation errors (Reid, Zabala, & Holmberg, 2017; Zhang, 2012). In this context, Beręsewicz (2017), Daas, Puts, Buelens, and van den Hurk (2015) and Citro (2014) discussed coverage, selection and measurement errors. Japec et al. (2015), Pfeffermann (2015), and Beręsewicz, Lehtonen, Reis, Di Consiglio, and Karlberg (2018) address coverage and selection error, which can lead to a significant bias in estimates based on big data sources, in particular if it is non-ignorable.

Recently, there has been a growing interest in research on the use of non-probability samples, including big data, together with probability samples. Kim, Park, Chen, and Wu (2018), Yang and Kim (2018), Yang, Kim, and Song (2019), and Yang and Kim (2019) have developed a rigorous approach to data integration by means of mass imputation using the nearest neighbour approach and double robust estimation; Chen, Elliott, and Valliant (2018), and McConville, Breidt, Lee, and Moisen (2017) proposed using Least Angle Shrinkage and Selection Operator (LASSO) regression to conduct model-assisted estimation assuming data are missing at random and Chen, Valliant, and Elliott (2019) extended this approach assuming that only estimated totals are known. Previous papers that deal with model-assisted estimation or calibration with estimated totals include Berger, Muñoz, and Rancourt (2009), and Guandalini and Tillé (2017). Beręsewicz et al. (2018), Elliott and Valliant (2017), Valliant (2019) Buelens, Burger, and van den Brakel (2018) give a general overview of possible approaches to deal with non-probability samples including pseudo-randomization and the model-based approach, while Citro (2014), Japec et al. (2015), and Couper (2013) provide a general discussion about modern data sources for statistical purposes.

The online job market has been rapidly developing. Thus, online job offers provide interesting research possibilities. There is a growing body of economic literature on the use of online job offers, with increasing attention paid to skills (see, e.g., Kuhn and Skuterud (2004), Deming and Kahn (2018), Colombo, Mercorio, and Mezzanzanica (2019), Pater, Szkoła, and Kozak (2019), and Turell, Speigner, Djumalieva, Copple, and Thurgood (2019)). Hershbein and Kahn (2018), Marinescu and Rathelot (2018) and Turell et al. (2019) compare the distributions of online job offers they use to the results of probability-sample surveys and to the distributions of other online job offers measures. For highly correlated results they assume that the job offers non-probability sample is ”representative”. If the correlation is low, they either take a subsample of gathered job offers or weight their data. While they pay attention to obtain similar distributions of data for certain variables, they do not use new variables to reduce bias observed in online job offers.

Currently, the Demand for Labour (DL) survey conducted by Statistics Poland is used to produce estimates of vacancies by occupation, economic activity sector (NACE) or company size. The goal of our study is to enhance the DL survey by including information about skills obtained from job advertisements. We reuse online data collected for the purpose of The Study of Human Capital (HC) 2011–2014 in the module devoted to job ads. While different approaches to counteract selection bias are found in the literature, we applied pseudo-randomization (calibration) with modern assisting models (LASSO and Adaptive LASSO). In our study, we condition the ignorability of selection mechanism on the quality of auxiliary variables. We look for auxiliary variables that are highly correlated with our target variables—skills. Even though employers use various channels to seek employees with a certain set of skills, they do not make this decision based solely on any one skill. Thus accounting for other factors, with occupation being the most important, should, to some extent, decrease selection bias in online data.

We contribute to the literature in a few ways. We perform data integration of a non-probability sample with a probability sample; a task that is increasingly pursued by official statistics given big data availability. Thanks to this, we supplement a probability-sample-based Demand for Labour survey with skills. While it is difficult to include measures of skills in a job vacancy survey, this approach complements survey results. To our knowledge such an analysis has not been done before. Since companies more often seek employees with particular skills, rather than occupations, this gives us many possibilities of applying newly integrated data to labour market models. With this approach historical data may be integrated, even with other variables. Moreover, our analysis is made without having the individual level data for the probability sample, but only population totals. We propose a bootstrap approach that accounts for the uncertainty of totals reported by official statistics. Finally, we correct coverage and selection error in the online data thanks to auxiliary variables from a probability sample.

The article has the following structure. Section 2 is devoted to data sources about the Polish vacancy market, including official statistics and selected non-official sources. This section also describes data used in this study. In section 3 we describe methods of inference based on non-probability data including the bootstrap procedure to estimate variance when only limited population level data is reported. Section 4 presents empirical results for 11 skills obtained from online data. The article ends with conclusions that includes discussion on limitations of our study.

Paper is assisted with supplementary materials (RStudio Project) that contains all data and codes used in the study to reproduce results[1]. In addition, we have created a github repository with docker image that allows to run the codes online in a virtual machine[2].

2 Data

2.1 The demand for labour survey

Currently, the DL survey conducted by Statistics Poland (Statistics Poland, 2018) is the main source of information about job vacancies in Poland. It is designed to obtain information on the satisfied and unsatisfied demand, i.e. the employed (occupied jobs), the vacancies, the newly created jobs, and the liquidated jobs. In 2005 the format of the survey was changed in accordance with the Eurostat (Job Vacancy Survey) requirements in order to keep the survey content and methodology uniform across all EU Member States. Since 2007, the survey has been carried out as a sample survey and covers entities of the national economy from all NACE sections, employing at least one person.

The DL survey is carried out as a probability sample survey. The survey sample of 100,000 units is selected separately for units employing more than 9 persons (50,000), and separately for units employing up to 9 persons (50,000). With regard to large and medium-sized units, the sample is stratified by activity (19 NACE[3] sections) and by province (16 NUTS2 regions), resulting in 304 separate subpopulations. Inside each of the subpopulations, the units are sorted in a descending order according to the number of employees. The largest units in each subpopulation that meet the threshold (equalling the subpopulation size) of the number of employees are included in the survey with certainty. Then the sample of the previously determined size is selected from the remaining parts of particular subpopulations, using a stratified, proportional sampling scheme.

As regards small units, employing up to 9 persons, the main purpose of the survey is to obtain results by NACE section. Allocation is carried out between different NACE sections in order to obtain the same expected precision. Within sections, units are stratified by province and then the sample is selected using a stratified, proportional sampling scheme.

Following a significant change made in the Polish classification of occupations in 2011, data collected in the DL survey before 2011 are not comparable to those collected afterwards. Additionally, in 2018 an additional question was included in the survey questionnaire about whether the responding entity placed job offers in district employment offices (DEOs). Each NUTS4 district (Pol. *powiat*) in Poland has its own DEO.

The survey suffers from non-response, which amounted to 35.2% in 2011, 36.6% in 2012, 38.1% in 2013 and 38.2% in 2014. Correction for this error involves multiplying sampling weights by the inversion of response rates within particular strata and calibration to meet the known population totals (number of units in 19 NACE sections and 16 NUTS 2 levels).

The survey defines the following terms for the measurement of labour demand:

Vacancies are positions or jobs unoccupied owing to labour turnover or newly created that simultaneously meet the

[1]Unfortunately, unit level data from The Study of Human Capital is anymore available online and thus our materials is the only place where it may be found.

[2]To access the repository please follow the link: https://github.com/BERENZ/bkl-paper-srm

[3]The Statistical classification of economic activities in the European Community, abbreviated as NACE, is the classification of economic activities in the European Union; Various NACE versions have been developed since 1970.

Table 1
Estimated total number of vacancies at the end of Q1 based on the DL survey

2011	2012	2013	2014
71,775	52,424	42,889	52,725

following three conditions: (1) were actually unoccupied on the survey day, (2) the employer had made efforts to find people willing to take up the job, (3) if adequate candidates were found to occupy the vacancies, the employer would readily take them in.

Newly created jobs are jobs created in the course of organizational changes, expansion or change of business activity, as well as all jobs available in newly established companies.

The DL survey provides quarterly estimates about the number of job vacancies by 1) occupation (9 major groups and sub-groups of more detailed occupations denoted by 2-digit codes; with the exclusion of the 10th major group—occupations in the armed forces), 2) NACE, 3) company size (1–9, 10–49 and 50+ employees), 4) ownership type (public, private) and 5) province (16 units). In addition to marginal distributions, estimates of joint distributions of job offers that are published are limited to two-way interactions. Information about precision, measured by relative standard error, is published on an yearly basis only for marginal distributions of auxiliary variables and varies from 2% to 20%.

In the study we examined occupation (2-digit codes), NACE and province as potential auxiliary variables to reduce the selection bias in skills described in job offers. We used only estimated totals reported by Statistics Poland as we did not have access to micro-data from the survey. We decided to disregard *Skilled agricultural, forestry and fishery workers* as this group accounts for less than 1% of all job vacancies. Table 1 contains information about estimated vacancies for the first quarters of 2011–2014 based on the DL survey.

Unfortunately, the DL survey does not cover skills required by employers but this information may be find in description of online job advertisements. To complement DL survey we used online job ads that were collected during the Study of Human Capital survey that is introduced in next section.

2.2 Online job advertisements

The Study of Human Capital in Poland. The Study of Human Capital in Poland (HC), a cross-sectional survey to monitor the labour market, was carried out between 2010–2015. The survey was resumed in 2017 but in a narrower scope. The survey was conducted by the Polish Agency for Enterprise Development (PAED) and the Centre for Evaluation and Analysis of Public Policies at the Jagiellonian University (CEAPP). The survey consisted of four modules: (1) survey of employers, (2) survey of job offers, (3) working-age population survey and (4) representatives of training institutions. The aim was to keep track of the situation in the Polish labour market, monitor the supply and demand for skills as well as the system of education and professional training in Poland in the period 2010–2015. The data from the current the surveys are freely available from the HC survey website The HC survey (2020), data from previous waves are available only at request from the authors, the methodology of the study is described in Czarnik et al. (2011) and the data collection procedure is described in the report of Polish Agency for Enterprise Development (2011). The focus of the survey is limited to the working-age population.

The goal of the HC survey of job offers was to provide characteristics of skills and occupations included in job offers, not produce estimates of these characteristics for all job offers in Poland.

The statistical unit defined in the job ads module was *a unique job offer for a single position, published on a given day, excluding internships for students and pupils and jobs in foreign countries*. The survey did not distinguish between seasonal, part-time or full-time job offers. This definition differs from the one used in official statistics because it is related to the job description rather than the vacancy. However, we assumed that information included in the job ad can be taken to reflect the job vacancy description.

A mixed mode of data collection was used. Job offers were obtained from a random sample of 160 public employment offices (DEOs; stratified by 16 provinces) in 2010 and the job brokerage portal www.Careerjet.pl. We will focus on the description of the latter data source, because we do not use data from public employment offices to avoid the under-representation of skills in job offers (see explanation in section 2.2). According to PAED and CEAPP (Czarnik et al., 2011, p. 41) the reasons for the selection of this online job board are mainly twofold: 1) Careerjet.pl contains job advertisements that are directly published by companies on this portal, as well as ones from other online job boards; so it is not limited to only one source of data, and 2) Careerjet.pl does not promote any specific job board, and also accounts for local job advertisements, what ensures coverage of labour demand of small firms, not only large ones, which are known to dominate job offers on certain online job boards.

Job offers had to meet specific requirements for the day of the survey, which was the 4th Monday of March of every year (except for 2010, when the data collection was conducted in September). Therefore, the survey was designed to be comparable between successive years.

Data from the Careerjet.pl website were collected in a semi-automated manner. Interviewers took a screenshot of a displayed job offer or saved the page as an html file, and then entered the data using the copy-and-paste method according to a preset format (table in MS Word). Each offer was a separate text file with a corresponding identifier. Then specially prepared software transformed the dataset for the coding process.

The job offers were coded according to a categorization key containing a list of skills, occupations and other features. Each offer was coded independently by two coders. Table B1 in the Appendix contains information about the coding precision for occupations and NACE sections indicated by the number of digits in a given code; the higher the number of digits, the more detailed the occupation specification is. Each year, the coding reliability index was calculated based on a sample of 100 job offers, which represents the total number of codes used and coding consistency of coders. These ratios are presented in Table 2.

Verification of offers consisted in removing duplicate ads and those that did not meet the adopted selection criteria. The database did not include offers of low-quality data (where it was impossible to determine the place of work and the recruitment area, as well as offers with insufficient information). The uniqueness of offers from the second survey edition in 2011 was verified at the level of the database, not at the stage of obtaining job offers. Duplicates were distinguished by comparing (1) publication date, (2) source, (3) city, (4) province, (5) job offer reference number (not the ad's ID), (5) company name and (6) occupation. Without access to the raw data, we assumed that publicly available databases contained unique job offers.

First the job offers registered on the day of data collection were downloaded and coded. The target sample size for each year was set at 20,000, which included all job offers collected from from DEOs plus as many ads from Careerjet.pl as necessary to reach the target. Table 2 presents the initial sample size (before deduplication) and final sample size for each survey year, including the collection date. Ads from Careerjet.pl accounted for about 60% of all job offers collected in the survey.

The coding precision was lowest in 2010, which is not surprising as this was the first year of the study. The index increases in the subsequent years. Unfortunately, more detailed information regarding the coding procedure was not available. One should keep in mind that mis-classification error may cause bias in estimated proportions (see for instance van Delden, Scholtus, & Burger, 2016) and models parameters (see for instance Dlugosz, Mammen, & Wilke, 2017).

Results reported in the article are based on the survey of job offers conducted between 2011 and 2013–2014 (three waves). For 2012, the publicly available dataset contained only 1-digit occupations (9 groups, see Table B2) and it was not possible to obtain the full dataset from the survey administrator. Therefore, we decided to take the following steps regarding the final dataset:

- disregard occupations with single digit code (143 records),
- disregard the 6th occupation category (i.e. skilled agricultural, forestry and fishery workers) because of the small number of job vacancies reported in the DL survey,
- disregard the following NACE sections: A (Agriculture, Forestry And Fishing), B (Mining And Quarrying), D (Electricity, Gas, Steam And Air Conditioning Supply), E (Water Supply; Sewerage, Waste Management And Remediation Activities), L (Real Estate Activities) for lack of population totals with estimated standard errors.

As a result, the final dataset for the waves in 2011, 2013 and 2014 consisted of a total of 38,100 observations. There were 34 two-digit occupation codes, 16 provinces and 16 NACE [4].

Skills measured in the study. The HC survey proposes a classification of skills for the analysis of the vacancy market. It was prepared after reviewing various skills classifications used by different international institutions, including: institutions dealing with statistical data (e.g. the Australian Bureau of Statistics), those that develop skills standards (e.g. National Classification of Professional Standards), and enterprises responsible for the development of professional skills (e.g. O*NET. The Occupational Information Network). For more details see Czarnik et al. (2011, chap. 2) attached in the Online Supplementary Materials. The survey distinguished the following skills:

1. Artistic—artistic and creative skills,
2. Availability—availability to work for the employer,
3. Cognitive—finding and analyzing information, drawing conclusions,
4. Computer—working with computers and using the Internet,
5. Interpersonal—contacts with others,
6. Managerial—managerial skills and organization of work,
7. Mathematical—performing calculations,
8. Office—organization of and conducting office tasks,
9. Physical—physical fitness,
10. Self-organization—self-organisation, initiative, punctuality,
11. Technical—handling, assembling and repairing equipment.

A detailed description of the skills categories is presented in Table A1. During the coding process 1 was used if a given skill was included in the job description, 0 otherwise. There were almost no missing data in variables denoting skills as the lack of a given skill was indicated by 0.

[4] We collapsed underrepresented NACE sections and occupation codes for job ads and did the same for the DL survey data. See supplementary materials for the whole data processing report.

Table 2
Sample sizes and the coding precision in the DL survey in the online job ads module

Year Day	2010 10th Sep	2011 28th Mar	2012 26th Mar	2013 25th Mar	2014 28th Mar
Initial Sample	21,195	22,243	23,366	22,795	23,452
Final sample	20,009	20,634	21,594	20,081	21,456
DEOs	8,198	7,018	7,253	5,614	8,542
the Internet (Careerjet.pl)	11,811	13,618	14,342	14,467	12,914
Coding quality	0.72	0.89	0.96	0.96	0.96

Source: Polish Agency for Enterprise Development (2011, 2012, 2013, 2014, 2015).

Table 3
Share of skills included in job offers by data source based on pooled data for 2011, 2013 and 2014

Skill	Careerjet.pl	DEOs	Carrerjet.pl & DEOs
Artistic	15.8	2.2	11.2
Availability	21.0	2.9	14.8
Cognitive	20.8	1.5	14.3
Computer	33.2	8.9	25.0
Interpersonal	55.9	6.9	39.3
Managerial	29.2	2.0	20.0
Mathematical	0.3	0.1	0.2
Office	3.8	1.5	3.0
Physical	6.0	2.0	4.7
Self-organization	59.1	7.6	41.6
Technical	4.3	5.1	4.6

Table 4
Cramer's V between skills and occupation, NACE section and province, based on the HC survey pooled data for 2011, 2013 and 2014

Skill	Occupation (2 digits)	NACE	Province
Artistic	0.22	0.11	0.05
Availability	0.15	0.14	0.05
Cognitive	0.21	0.06	0.06
Computer	0.45	0.23	0.10
Interpersonal	0.42	0.23	0.06
Managerial	0.34	0.15	0.04
Mathematical	0.05	0.02	0.03
Office	0.11	0.06	0.03
Physical	0.17	0.09	0.04
Self-organization	0.34	0.19	0.04
Technical	0.31	0.11	0.07

Table 3 presents the share of given competences included in job offers according to the data source—Careerjet.pl (the Internet), DEOs, and combined data from both data sources. For example, self-organisation was included in 59.1% of all job offers on the Internet while only in 7.6% of offers in DEOs. Spearman correlation coefficient between shares of competences measured in online offers and DEOs was equal to 0.74.

Data from the Internet were much richer in terms of published content in job offers, compared to DEO data. Employers placing job offers online tended to prepare much more detailed descriptions and therefore managed to better specify their requirements. As regards DEOs, the content (and form) of ads was limited by the input format, which allows the employer to enter the sought-after occupation and any preferences regarding education or knowledge of a foreign language.

The design of the HC survey did not include imputation of missing data; for instance 254 records had missing values in the occupation and 268 in the province. The highest number of missing values was recorded for NACE (over 22,000), which is mainly due to the lack of information about the company in the ads. The share of missing data varied between the survey waves as presented in Table B3. Therefore, we imputed missing data in occupation, NACE and province based on one nearest neighbour with Gower distance and weights assigned to columns that are based on variable importance from random forest. This approach is implemented in the VIM package (Kowarik & Templ, 2016) and was applied to the original dataset.

Correlation with auxiliary variables. Given limited access to totals estimated from the DL survey, the correlation with skills derived from online job ads was assessed only for the following auxiliary variables: occupation, NACE section and province. Cramer's V correlation coefficients are presented in Table 4. The most correlated variable is occupation and the least correlated is province. This is reasonable because skills are specified for occupation rather than for the place of work or the company's type of activity.

The highest correlations are observed for interpersonal, computer and managerial skills, which suggests that the use

Table 5
Basic idea of data integration when variables are available at unit-level or domain-level

Data source	$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{d}$	$\boldsymbol{T}^X$	$\widehat{\boldsymbol{T}}^X$
Population data	✓	–	–	✓	–
Online data (A)	✓	✓	–	–	–
Sample survey (B)	✓	–	✓	–	✓

of auxiliary variables could reduce selection bias in the case of these skills. The weakest relationship is observed for office, physical and mathematical skills. Since we focus on model-assisted estimators under generalized linear model with main effects, correction for selection bias based on these variables may be limited due to lack of interactions (as in classification trees).

3 Methods

3.1 Data integration approach

Enhancing probability survey with online data (i.e. non-probability sample) may be achieved by data integration. Table 5 presents the case when population, sample survey and online data are considered. First three columns denote variables available at unit-level data. $\boldsymbol{X}$ denote matrix of auxiliary variables such as occupation or NACE sector, target variable(s) denoted by vector $\boldsymbol{Y}$ and $\boldsymbol{d} = \boldsymbol{\pi}^{-1}$ is vector of probability weights used for inference based on sample survey. Last two columns contain either known totals $\boldsymbol{T}^{\boldsymbol{X}}$ or estimated totals $\widehat{\boldsymbol{T}}^{\boldsymbol{X}}$ for auxiliary variables $\boldsymbol{X}$. Note that we assume $\boldsymbol{X}$ are available in all sources, while $\boldsymbol{Y}$ only for online data. For simplicity, we assume that weights available in sample survey are already corrected for coverage and non-response errors. That is often the case when National Statistical Institutions provide unit-level data with only one set of weights.

The goal of data integration is to estimate some quantity (e.g. mean, total) of target variables $\boldsymbol{Y}$ present only in online data. Elliott and Valliant (2017) summarised possible approaches that consider pseudo-randomization (i.e calibration) or model-based approach. In addition, Kim and Wang, 2018 consider mass imputation and double robust estimation that take into account propensity score weighting.

In the paper we consider a pseudo-randomization approach in which pseudo-weights from non-probability sample are calibrated to estimated totals $\widehat{\boldsymbol{T}}^X$ or to the population sum of $\boldsymbol{Y}$ predictions based on approach introduced by Wu and Sitter (2001) and further developed for non-probability samples by Chen (2016). Detailed description is presented in the sections below.

3.2 Traditional calibration

Calibration was proposed by Deville and Särndal (1992) and is a method of searching for so called calibrated weights by minimizing the distance measure between the sampling weights and the new weights, which satisfy certain calibration constraints. As a consequence, when the new weights are applied to the auxiliary variables in the sample, they reproduce the known population totals of the auxiliary variables exactly. It is also important that the new weights should be as close as possible to sampling weights in the sense of the selected distance measure (Särndal & Lundström, 2005).

Following the notation in Chen et al. (2019), let us define the online (non-probability) sample from population of job vacancies as $s_{A,t}$ of size $n_{A,t}$ where $t = 1, \ldots, T$ denotes the wave. For simplicity, we drop subscript t. This sample contains variables of interest Y_k, where $k = 1, \ldots, K$. Further, let $\boldsymbol{d}^A_{n_A\times 1}$ be a vector of pseudo-weights that are typically set to $\frac{N}{n_A}$ for all units $i \in s_A$ and for all K variables, where N is the size of the target population. In this approach we assume simple random sampling design for sample s_A due to lack of knowledge of the data generation process or selection mechanism. If N is unknown then we may use the estimated size $\widehat{N} = \sum_{i\in s_B} d_i^B$ based on the probability sample B. As traditional calibration is done for the whole dataset, we drop subscript k denoting k-th target variable.

Let $\boldsymbol{D}^A$ be a diagonal matrix of pseudo-design weights $\boldsymbol{d}^A$ and $\boldsymbol{w}_{n_A\times 1}$ be calibrated weights that minimize an expected distance measure with respect to the design of $\mathcal{A}$

$$E_{\mathcal{A}}\left[\frac{\sum_{i\in s_A} g(w_i, d_i^A)}{q_i}\right], \tag{1}$$

under the constraint:

$$\sum_{i\in s_A} w_i \boldsymbol{x}_i^T = \boldsymbol{T}^{\boldsymbol{X}}, \tag{2}$$

where $\boldsymbol{T}^{\boldsymbol{X}}$ is a row vector of population totals of sample calibration variables $\boldsymbol{X}$ and $g(w_i, d_i^A)$ is a differentiable function with respect to w_i, strictly convex on an interval containing d_i^A, $g(d_i^A, d_i^A) = 0$ and $\boldsymbol{q}$ is a vector of weights, unrelated to $\boldsymbol{d}$, that accounts for heteroscedasticity (in our case $\boldsymbol{q} = \boldsymbol{1}$). The commonly used generalized regression (GREG) estimator uses the χ^2 distance $g(w_i, d_i^A) = \frac{(w_i - d_i^A)^2}{d_i^A}$. For this distance measure:

$$\boldsymbol{w}^{\text{GREG}} = \boldsymbol{d}^A + \boldsymbol{D}^A\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{D}^A\boldsymbol{X})^{-1}(\boldsymbol{T}^{\boldsymbol{X}} - (\boldsymbol{d}^A)^T\boldsymbol{X})^T, \tag{3}$$

where $\boldsymbol{D} = \operatorname{diag}(\boldsymbol{d}^A)$. The estimate of the population mean of outcome $\boldsymbol{y}_k$ assuming that we have K target variables is based on calibrated weights:

$$\widehat{\overline{T}}_{y_k}^{\text{GREG}} = \frac{\sum_{i\in s_A} w_i^{\text{GREG}} y_{ki}}{\sum_{i\in s_A} w_i^{\text{GREG}}}. \tag{4}$$

The calibrated weights defined in (4) do not rely on any outcome variable and can be applied for all variables on the non-probability sample s_A.

In the case when only estimates of totals $\widehat{\boldsymbol{T}}^X$ based on probability sample B are known, Dever and Valliant (2010) introduced estimated control calibration. In this framework, we replace $\boldsymbol{T}^X$ in (3) with $\widehat{\boldsymbol{T}}^X$, which results in

$$\boldsymbol{w}^{\text{ECGREG}} = \boldsymbol{d}^A + \boldsymbol{D}^A\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{D}^A\boldsymbol{X})^{-1}(\widehat{\boldsymbol{T}}^X - (\boldsymbol{d}^A)^T\boldsymbol{X})^T, \quad (5)$$

and thus the estimated mean is given by

$$\widehat{\overline{T}}_{y_k}^{\text{ECGREG}} = \frac{\sum_{i\in s_A} w_i^{\text{ECGREG}} y_{ki}}{\sum_{i\in s_A} w_i^{\text{ECGREG}}}. \quad (6)$$

Following Chen et al. (2019) we denote this estimator as ECGREG (Estimated control GREG) to distinguish it from GREG with known population totals.

3.3 Model-assisted calibration

Following results obtained by Chen (2016), and Chen et al. (2018), we consider a model-assisted calibration approach using a plausible model. Model-assited calibration was proposed by Wu and Sitter (2001) and further extended by the above mentioned authors. The basic idea of model-assisted calibration is as follows. We build K separate models for each target variable Y_k using the same set of covariates denoted by $\boldsymbol{X}_k$:

$$E_\xi(y_{ki}|\boldsymbol{x}_{ki}) = \mu(\boldsymbol{x}_{ki},\boldsymbol{\beta}_k),\ V_\xi(y_{ki}|\boldsymbol{x}_{ki}) = v_{ki}^2\sigma_k^2, \quad (7)$$

where $\boldsymbol{\beta}_k = (\beta_{k1},\ldots,\beta_{kp})^T$ and σ_k are unknown superpopulation parameters. $\mu(\boldsymbol{x}_{ki},\boldsymbol{\beta}_k)$ is a known function of $\boldsymbol{x}_{ki}$ and $\boldsymbol{\beta}_k$, and v_{ki} is a known function of $\boldsymbol{x}_{ki}$ or $\mu(\boldsymbol{x}_{ki},\boldsymbol{\beta}_k)$. E_ξ and V_ξ are expectation and variance with respect to the model ξ.

Let $\boldsymbol{B}_k$ be the finite population (or census) estimate of $\boldsymbol{\beta}_k$ and $\hat{\mu}_{ik} = \mu(\boldsymbol{x}_{ki},\hat{\boldsymbol{B}}_k)$, where $\hat{\boldsymbol{B}}_k$ is the sample estimate of $\boldsymbol{B}_k$. For each Y_k we need to obtain set of pseudo-weights so we define $\boldsymbol{d}^A$ for each k variable separately. As we assume simple random sampling then pseudo-weights are the same for each target variable ($\boldsymbol{d}_k^A = \boldsymbol{d}^A$). Then, for each k, the model-assisted calibrated weights $\boldsymbol{w}_k$ minimize a distance measure $E_{\mathcal{A}}\left[\frac{\sum_{i\in s_A} g(w_{ik},d_i^A)}{q_i}\right]$ under constraints $\sum_{i=1}^n w_{ik} = N$ and $\sum_{i=1}^n w_{ik}\hat{\mu}_{ik} = \sum_{i=1}^N \hat{\mu}_{ik}$. Under χ^2 distance measure with $q_i = 1$, the model-assisted calibrated weights are:

$$\boldsymbol{w}_k^{MC} = \boldsymbol{d}^A + \boldsymbol{D}^A\boldsymbol{M}_k(\boldsymbol{M}_k^T\boldsymbol{D}^A\boldsymbol{M}_k)^{-1}(\boldsymbol{T}_k^M - (\boldsymbol{d}^A)^T\boldsymbol{M}_k)^T, \quad (8)$$

where superscript MC denotes model-assisted calibration, $\boldsymbol{D}^A = \text{diag}(\boldsymbol{d}^A)$, $\boldsymbol{T}_k^M = (N, \sum_i^N \hat{\mu}_{ik})$ and $\boldsymbol{M}_k = (\boldsymbol{1}^A, (\hat{\mu}_k^A)_{i\in s_A})$. Note that in this approach we obtain K sets of weights for each $\boldsymbol{Y}_k$ variable separately. In this setting the population mean is given by

$$\widehat{\overline{T}}_{y_k}^{\text{MC}} = \frac{\sum_{i\in s_A} w_{ik}^{\text{MC}} y_{ik}}{\sum_{i\in s_A} w_{ik}^{\text{MC}}}. \quad (9)$$

If the totals are estimated from the reference, independent probability sample of size n_B, then constraints are $\sum_{i\in s_A} w_{ik} = \sum_{i\in s_B} d_i^B = \widehat{N}$ and $\sum_{i\in s_A} w_i\hat{\mu}_{ik} = \sum_{i\in s_B} d_i^B\hat{\mu}_{ik}$, where $\boldsymbol{d}^B$ are weights from probability sample B drawn from population of vacancies. Similarly, as in the case of GREG, we replace $\boldsymbol{w}^{MC}$ with the $\boldsymbol{w}^{ECMC}$ obtained from the estimated totals and get

$$\widehat{\overline{T}}_{y_k}^{\text{ECMC}} = \frac{\sum_{i\in s_A} w_{ki}^{\text{ECMC}} y_{ki}}{\sum_{i\in s_A} w_{ki}^{\text{ECMC}}}. \quad (10)$$

Further, we assume that $\mu(\cdot)$ is defined as a generalized linear model (i.e. logistic regression), LASSO and adaptive LASSO regression are described in the following section.

3.4 Model-assisted calibration using adaptive LASSO

Least Angle Shrinkage and Selection Operator (LASSO) is a regularized regression that can perform both variable selection and parameter estimation (Tibshirani, 1996); it gained popularity because it prevents model over-fitting by selecting more accurate and parsimonious models. In the paper we focus on skills defined as binary variable that are modelled with logistic regression. Thus, for the classical LASSO the objective function for the penalized logistic regression uses the negative binomial log-likelihood, and is

$$\widehat{\boldsymbol{\beta}}_k = \underset{\boldsymbol{\beta}_k}{\text{argmin}}\left(\sum_{i\in s_A}\Big[-y_{ik}(\boldsymbol{x}_{ki}\boldsymbol{\beta}_k) + \log(1+\exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}_k))\Big] + \lambda_{n^A k}\sum_{j=1}^p\Big|\beta_{kj}\Big|\right), \quad (11)$$

where the first part is loss function and the second is the penalty term. $\lambda_{n^A k}$ is a penalty parameter that optimizes a model-fitness measure (e.g. AIC, BIC) for k-th target variable and depends on the sample size n^A. $\lambda_{n^A k}$ is derived in cross-validation procedure. In (11) we use absolute values of model parameters $|\beta_{kj}|$ which in the literature is named as L-1 regularization. If we use β_{kj}^2 then it is L-2 penalty and then the model becomes the ridge regression.

Traditional LASSO does not meet the oracle property (Chen, 2016), i.e. does not correctly select outs zero and non-zero parameters. To overcome this issue, an adaptive LASSO was proposed by Zou (2006). The main difference with classical LASSO is an extra parameters α that prevents LASSO from selecting covariates with large effect sizes in favor of lowering prediction error when the sample size is small. In the case of logistic regression assuming k target variables (for $k = 1$, see Chen et al., 2019, p. 5, eq. (13)), is given as

$$\widehat{\boldsymbol{\beta}}_k = \underset{\boldsymbol{\beta}_k}{\operatorname{argmin}} \left(\sum_{i \in s_A} \left[-y_{ik}(\boldsymbol{x}_{ik}\boldsymbol{\beta}_k) + \log(1 + \exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}_k)) \right] + \lambda_{n^A k} \sum_{j=1}^{p} \alpha_{kj}^{\gamma_k} |\beta_{kj}| \right), \quad (12)$$

where $\alpha_{kj}^{\gamma_k}$ is an adjustable weight and γ_{n^A} is a penalty used to optimize a model fit measure, while other parameters remain as defined previously. Therefore, in adaptive LASSO there are additional penalty parameters that makes the prediction better but requires extra optimization effort. Given $\lambda_{n^A k}$ and γ_k, one can estimate $\widehat{\boldsymbol{\beta}}_k$ through iterative procedures. Common choice for α_{kj} is $\frac{1}{|\widehat{\beta}_{kj}^{\text{MLE}}|}$ where $|\widehat{\beta}_{kj}^{\text{MLE}}|$ is the maximum likelihood estimate (MLE) of β_{kj} or $\frac{1}{|\widehat{\beta}_{kj}^{\text{RIDGE}}|}$ obtained from ridge regression (RIDGE). If $\alpha_{kj}^{\gamma_k} = 1$ then we get standard LASSO model. The power of the weight parameter, γ_k, is a constant greater than 0 that interacts with kj to control LASSO from selecting or excluding parameters. LASSO can be estimated using the glmnet package (Simon, Friedman, Hastie, & Tibshirani, 2011).

Then, to obtain the population mean we need to replace $\boldsymbol{w}_k^{MC}$ with corresponding $\boldsymbol{w}_k^{\text{ECLASSO}}$ from the standard LASSO model or $\boldsymbol{w}_k^{\text{ECALASSO}}$ obtained under the adaptive LASSO model. To obtain $\widehat{\boldsymbol{\beta}}_k$ we followed the approach proposed by Chen et al. (2018) and used the cross-validation procedure. For more details refer to Chen et al. (2019).

3.5 Estimators used in the paper

The outcome variable of interest is whether the description of a job offer ($i = 1, \ldots, n_{A,t}$) contained a given skill. Let us define the binary indicator for the outcome variable $\boldsymbol{Y}_{kt}$ for each $k = 1, \ldots, 11$-th skill and for each $t = \{2011, 2013, 2014\}$

$$y_{ikt} = \begin{cases} 1 & \text{if } i\text{-th job offer contains } k\text{-th skill in year } t \\ 0 & \textit{otherwise}. \end{cases} \quad (13)$$

For each variable k we calculate the following estimators:

$$\widehat{T}_{y_{kt}}^{\text{HTSRS}} = \frac{\sum_{i \in s_{A,t}} \frac{\widehat{N}_{A,t}}{n_{A,t}} y_{ikt}}{\sum_{i \in s_{A,t}} \frac{\widehat{N}_{A,t}}{n_{A,t}}},$$

where $\widehat{N}_{A,t} = \sum_{i \in s_B} d_i^B$ was estimated based on the DL survey, which is Horvitz-Thompson estimator using pseudo-weights calculated for each period separately.

$$\widehat{T}_{y_{kt}}^{\text{ECGREG}} = \frac{\sum_{i \in s_{A,t}} w_{it}^{\text{ECGREG}} y_{ikt}}{\sum_{i \in s_{A,t}} w_{it}^{\text{ECGREG}}},$$

where we use estimated totals for occupation (2-digit code; 34 levels). See Table B6. This estimator was calculated for each period separately.

$$\widehat{T}_{y_{kt}}^{\text{ECMC}} = \frac{\sum_{i \in s_{A,t}} w_{ikt}^{\text{ECMC}} y_{ikt}}{\sum_{i \in s_{A,t}} w_{ikt}^{\text{ECMC}}},$$

where we use a logistic regression model for each y_k separately based on pooled data from all periods and one auxiliary variable denoting occupation (2-digit code; 34 levels).

$$\widehat{T}_{y_{kt}}^{\text{ECLASSO1}} = \frac{\sum_{i \in s_{A,t}} w_{ikt}^{\text{ECLASSO1}} y_{ikt}}{\sum_{i \in s_{A,t}} w_{ikt}^{\text{ECLASSO1}}},$$

where we use LASSO regression for each y_k separately based on pooled data from all periods and one auxiliary variable denoting occupation (2-digit code; 34 levels).

$$\widehat{T}_{y_{kt}}^{\text{ECLASSO2}} = \frac{\sum_{i \in s_{A,t}} w_{ikt}^{\text{ECLASSO2}} y_{ikt}}{\sum_{i \in s_{A,t}} w_{ikt}^{\text{ECLASSO2}}},$$

where we use LASSO regression for each y_k separately based on pooled data from all periods and two auxiliary variable denoting occupation (2-digit code; 34 levels) and NACE (14 levels).

$$\widehat{T}_{y_{kt}}^{\text{ECALASSO1}} = \frac{\sum_{i \in s_{A,t}} w_{ikt}^{\text{ECALASSO1}} y_{ikt}}{\sum_{i \in s_{A,t}} w_{ikt}^{\text{ECALASSO1}}},$$

where we use adaptive LASSO regression with the same settings as ECLASSO1.

3.6 Variance estimation

Chen et al. (2018) and Chen et al. (2019) proposed analytical formulas for the asymptotic design variance which consists of two parts: 1) variance with respect to non-probability sample A, and 2) variance with respect to probability sample B. However, this approach requires access to unit-level data from the s_B sample, which is not always the case. For example, these data cannot be obtained owing to the risk of disclosure or the cost of purchasing these data is very high.

Moreover, the estimated totals and their uncertainties can only be published in a limited form. For instance, the DL survey reports standard errors for the estimated totals of vacancies by size, type of company and NACE section separately. In addition, estimated errors are only published for the last quarter of each year in the annual report. There are no estimates of uncertainty measures for vacancies by occupation; fortunately, there are cross-classification estimates of vacancies for occupation by NACE. Table B4 in Appendix presents estimated relative standard errors reported by Statistics Poland for the DL survey for 2011, 2013 and 2014. The precision varies between domains defined by NACE section but, in almost all cases, is lower than 20%. The highest standard errors are for Accommodation and Catering, and Administrative and Support Service Activities, while the lowest—for Manufacturing and Public Administration and Defence. Also, the estimates and relative standard errors for vacancies within NACE sections are stable over time.

In view of the limitations of the reporting procedure in the DL survey, we made the following assumption: *standard errors are similar in a given year and we can approximate standard errors from the 1st quarter based on information from the 4th quarter*. Without access to unit-level data, we could not verify the validity of this assumption but as the estimates of vacancies by NACE (and also by occupation) are stable over time this assumption is likely to be valid.

We used the bootstrap method to account for uncertainty in estimating the model based on $s_{A,t}$ and estimated totals from $s_{B,t}$, which is described in Algorithm 1 below.

There are two main issues with reference to non-probability samples. First, we assume simple random sampling with is certainly an simplification and may underestimate the true variance. We decided to use this approach due to lack of knowledge about the data generation process (i.e. selection mechanism). Second, variance of non-probability sample is conditional on the data gathered online during this study, which does not represent all online job advertisements for the days of data collection.

The following steps were taken to calculate variance in the bootstrap procedure. Let

$$\hat{\theta}^{(r)}_{y_k} = \widehat{\overline{T}}^{\text{HTSRS(r)}}_{y_k}, \widehat{\overline{T}}^{\text{ECGREG(r)}}_{y_k}, \widehat{\overline{T}}^{\text{ECMC(r)}}_{y_k}, \widehat{\overline{T}}^{\text{ECLASSO1(r)}}_{y_k}, \widehat{\overline{T}}^{\text{ECLASSO2(r)}}_{y_k}, \widehat{\overline{T}}^{\text{ECALASSO1(r)}}_{y_k}$$

which are then used to derive variance and relative standard errors (CV) given by the following equations:

- Variance

$$\text{var}(\hat{\theta}_{y_k}) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\theta}^{(r)}_{y_{ki}} - \bar{\hat{\theta}}^{(r)}_{y_k}\right)^2, \quad \bar{\hat{\theta}}^{(r)}_{y_k} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}^{(r)}_{y_{ki}} \tag{14}$$

- Relative Standard Error (CV)

$$\text{CV}(\hat{\theta}_{y_k}) = \frac{\sqrt{\text{var}(\hat{\theta}_{y_k})}}{\bar{\hat{\theta}}^{(r)}_{y_k}} \times 100\%. \tag{15}$$

To estimate variances for the aforementioned estimators we used bootstrap with $R = 500$ replicates. We compared the results in terms of relative standard errors. All calculations were done in R statistical software (R Core Team, 2018) using codes written by the authors and LASSO procedure provided in Chen et al. (2019). Data and R scripts to reproduce all calculations (including estimated models), tables and figures are available in the supplementary files and on the github repository[5].

4 Estimation of the demand for skills

Table 6 presents point estimates produced by means of the estimators presented in section 3.5. Column HTSRS is used for comparison to verify whether the models corrected the bias resulting from the specificity of online data. All bias-corrected estimates show similar demand for skills. The biggest differences between the bias-uncorrected (HTSRS) and corrected estimates are visible for the skills with high Cramer's V correlation presented in Table 4, i.e. interpersonal, managerial or computer skills. For almost all categories, online job ads overestimate the share of skills required by employers.

The biggest difference (almost 54% vs 35%) between all estimators can be observed for interpersonal competences. Other groups where there is a high difference after adjusting for known population totals are managerial skills (almost 10 p.p.) and computer skills (over 10 p.p.). There are two groups that online job advertisements underestimate: technical and physical competences. This is mainly due to underrepresentation of two categories of occupations: (7) Craft and related trades workers and (8) Plant and machine operators and assemblers. See Table B6 in the supplementary materials.

The actual demand for most of the skills is lower than raw results would suggest. It is especially the case for most wanted skills. This means that the skills mismatch in the Polish labour market may be lower than initially expected. These results also show that the studies not taking into account the extent of selection bias across skill requirements in online job postings, may overvalue or undervalue some skills. For example, Deming and Kahn (2018) show higher relative demand than our results do, especially for cognitive and interpersonal (social) skills, and lower demand only for managerial skills in the US economy. Similarly, Hershbein and Kahn (2018) show higher than ours share of online postings containing cognitive and computer skills requirements, also for the US economy. However, these differences to a high extent may result from large differences between the US economy and Polish economy.

As can be seen, the estimates based on ECGREG, ECMC, ECLASSO1,2 and ECALASSO1 are similar. This suggests that the variables used for the estimation provide comparable information despite the underlying model. Table C1 provides information about Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) for each skill, ECLASSO1, ECLASSO2 and ECALASSO1 model. Based on this table, it can be concluded that the inclusion of two

[5] https://github.com/BERENZ/job-offers-bkl

Algorithm 1: Algorithm to obtain bootstrapped population totals from probability sample B and replications of non-probability sample s_A

for *For r in 1:R* **do**

  if *Reference sample data* s_B, t **then**

    For each period t generate independently one finite population bootstrap based on the procedure;

    1. Generate
    $$\widehat{\boldsymbol{T}}_t^{\text{NACE(r)}} \sim N\left(\widehat{\boldsymbol{T}}_t^{\text{NACE}}, \text{SD}\left(\widehat{\boldsymbol{T}}_t^{\text{NACE}}\right)\right).$$
    2. Obtain
    $$\widehat{\boldsymbol{T}}_t^{\text{NACE, OCCUP(r)}} = \widehat{\boldsymbol{T}}_t^{\text{NACE(r)}} \times \frac{\widehat{\boldsymbol{T}}_t^{\text{NACE,OCCUP}}}{\widehat{\boldsymbol{T}}_t^{\text{NACE}}}.$$
    3. Use
    $$\widehat{\boldsymbol{T}}_t^{\text{NACE, OCCUP(r)}} \text{ and } \widehat{\boldsymbol{T}}_t^{\text{OCCUP(r)}}$$
    as estimated population totals for the estimation procedure.

  end

  if *Non-probability sample* $s_{A,t}$ **then**

    For each period t generate independently sample $s_{A(r),t}$ from $s_{A,t}$ using simple random sampling with replacement

  end

  Based on $s_{A(r),t}$, $\widehat{\boldsymbol{T}}_t^{\text{NACE, OCCUP(r)}}$ and $\widehat{\boldsymbol{T}}_t^{\text{OCCUP(r)}}$ calculate

  1. $\widehat{\overline{T}}_{y_{k,t}}^{\text{HTSRS(r)}}$ based on $s_{A(r),t}$, for each t independently,
  2. $\widehat{\overline{T}}_{y_{k,t}}^{\text{ECGREG(r)}}$ using new calibration weights for each $s_{A(r),t}$ using $\widehat{\boldsymbol{T}}_t^{\text{NACE, OCCUP(r)}}$ and $\widehat{\boldsymbol{T}}^{\text{OCCUP(r)}}$, for each period t independently,
  3. $\widehat{\overline{T}}_{y_{k,t}}^{\text{ECMC(r)}}, \widehat{\overline{T}}_{y_{k,t}}^{\text{ECLASSO1(r)}}, \widehat{\overline{T}}_{y_{k,t}}^{\text{ECLASSO2(r)}}, \widehat{\overline{T}}_{y_{k,t}}^{\text{ECALASSO1(r)}}$ using new calibration weights that meet new totals obtained from models that are fitted on pooled samples $s_{A(r),t}$.

end

SD() denotes standard errors derived from Table B4, $N()$ is normal distribution, $\widehat{\boldsymbol{T}}^{\text{NACE, OCCUP}}$ denotes estimated totals for cross-classification of NACE and Occupation (2-digit codes). Note that the part $\widehat{\boldsymbol{T}}^{\text{NACE(r)}} \times \frac{\widehat{\boldsymbol{T}}^{\text{NACE,OCCUP}}}{\widehat{\boldsymbol{T}}^{\text{NACE}}}$ assumes that we can split $\widehat{\boldsymbol{T}}^{\text{NACE(r)}}$ according to the estimate share of vacancies by occupation in a given NACE section. Table B5 in Appendix presents information about relative standard errors of the estimated $\widehat{\boldsymbol{T}}^{\text{OCCUP(r)}}$ in the bootstrap procedure. Detailed results are presented in Table B6 in Appendix. Uncertainty varies from 2% to 20%, which is inline with errors for NACE or other variables reported in the DL survey.

Table 6
Point estimates of the fraction of skills for the pooled sample for 2011, 2013 and 2014 (in %)

SKILLS	HTSRS	ECGREG	ECMC	ECLASSO1	ECLASSO2	CALASSO1
Artistic	15.8	12.3	12.4	12.5	13.0	12.5
Availability	20.9	19.8	19.7	19.6	21.5	19.5
Cognitive	20.9	14.3	14.3	14.6	14.0	14.6
Computer	33.0	22.2	22.0	22.3	23.0	22.6
Interpersonal	53.8	34.5	34.5	35.1	35.0	34.9
Managerial	26.2	16.7	16.5	16.8	17.7	16.8
Mathematical	0.4	0.4	0.4	0.4	0.4	0.4
Office	3.9	3.1	3.1	3.2	3.4	3.2
Physical	5.4	7.4	7.6	7.5	8.2	7.6
Self-organization	58.6	43.8	43.5	43.9	46.2	43.8
Technical	4.3	7.5	7.7	7.7	8.3	7.7

variables—occupation and NACE—results in a better model for each skill. The AUC varies from 0.644 for cognitive skills to 0.829 for technical competences, which indicates that the standard LASSO model is better than the adaptive one. Also, there are almost no differences between ECLASSO1 and ECALASSO1, which suggests that despite additional penalty the estimated parameters are close. Figure C1 provides a more detailed comparison of the estimated share of skills over the reference period.

Table 7 provides information about estimated relative standard errors for skills estimates for 2011, 2013 and 2014. ECMC and ECLASSO estimators are more efficient than ECGREG and ECMC is less efficient than estimators with LASSO. This is because ECGREG assumes a linear model and auxiliary variables are high dimensional and LASSO produce predictions with less variation by construction. Note that despite higher AUC for ECLASSO2, it provides less precise estimates mainly due to the high number of dimensions of the auxiliary variables and variability in totals from the DL survey. Moreover, there are almost no differences between adaptive and non-adaptive LASSO, which suggests that the estimated parameters are probably correctly specified. Finally, model-assisted calibration estimators (ECMC, ECLASSO1,2, ECALASSO1) were calculated based on models fitted to pooled data from three years which may be also reason for lower variability of totals used in the calibration procedure. Based on this result, we can choose estimates based ECLASSO1 as the final ones.

5 Conclusion

In the article we described our attempt to enhance the Demand for Labour survey conducted by Statistics Poland by including information about skills listed in online job advertisements. We considered online data as non-probability sample and apply methods that are developed for purpose of integration of probability and non-probability sample. In particular, we applied model-assisted estimators including generalized linear, LASSO and Adaptive LASSO models. Based on these results we conclude that the application of these methods reduced bias in online data for several skills but not for all. This can be explained mainly by the small correlation with the auxiliary variables used.

To our knowledge this is the first attempt to extend labour market surveys conducted by National Statistical Agencies by data from the Internet. Previous applications were devoted to non-probability samples based on web surveys or opt-in panels. Our approach shows that methods developed for non-probability samples may be applied for modern data sources such as big data. The latter is currently discussed in terms of auxiliary variables for small area estimation, nowcasting of selected indicators or creating new official statistics. However, there are some issues that should be discussed in detail.

The main limitation involved in the use of online data and combining them with existing surveys is the lack of auxiliary variables. For example, occupations or NACE information need to be extracted from the ad description or may not be even provided by employers. On the other hand, official statistics about the demand for labour are based on probability samples with restricted access to unit-level data (which limit possible approaches) or estimated totals for a certain level or cross-classification (often without uncertainty measures).

More generally, research on non-probability samples shows that using these data for statistics requires availability of good independent data sources. The main sources are either probabilistic samples or administrative records. Not always official statistics collects data that is required for the data integration purpose.

Another issue is measurement and unit error. In our study, we associated a job advertisement with a job vacancy, but this may not always be the case. This aspect also limits possibilities to verify whether selection mechanism is ignorable

Table 7
Average estimates of relative standard errors for skills over 2011, 2013 and 2014 (in %)

Skills	ECGREG	ECMC	ECLASSO1	ECLASSO2	ECALASSO1
Artistic	11.1	3.5	3.4	3.4	3.4
Availability	22.0	1.0	0.9	1.5	1.0
Cognitive	25.4	8.5	8.1	9.3	8.2
Computer	24.9	12.9	12.4	12.7	12.4
Interpersonal	17.6	6.6	6.3	6.6	6.4
Managerial	15.3	5.6	5.3	5.5	5.4
Mathematical	15.6	4.1	4.0	3.2	4.1
Office	33.5	4.7	4.4	4.4	4.6
Physical	32.6	4.1	4.2	4.7	4.3
Self-organization	16.7	3.8	3.6	3.5	3.6
Technical	25.1	5.3	5.2	7.8	5.2

or not. But taking into account our results, non-ignorable selection in job ads would suggest even lower demand for skills measured in this study. We plan to verify this in the future, when access to unit-level data from the DL survey will be possible. Further, we also assumed that description included on job ads may be related to job vacancy occupations reported by Statistics Poland. This should be verified in the future by investigating job descriptions reported by entities in the DL survey.

In our study, we used online data for Poland to enhance Polish Demand for Labour survey with skills. We applied skills classification proposed in the Study of Human Capital in Poland. Since the DL survey is conducted in all EU countries (as Job Vacancy Survey), our method can be reproduced for any EU country, for which appropriate information on job advertisements is collected. In applying this method, some country differences in the DL survey methodology should be considered[6]. For example, while Polish DL survey includes all NACE sections, some countries exclude certain sections from the survey. There are also differences between countries in: the source of vacancy data (survey or administration), statistical unit (local or enterprise), statistical population (companies with 1+ or 10+ employees), and reference dates (last day of a quarter or other).

Finally, in our study we used online data from 2011–2014 that was already coded and did not require text mining extract occupation or skills. These data may be actually used for preparing training data for machine learning. This is because original descriptions of job advertisements are associated with labels suited for machine learning purposes. However, one should keep in mind that data from the past not necessarily may hold for future job advertisements.

Despite these problems we conclude that online data combined with official statistics can provide a better picture of competences, education and other requirements made by employers and can be used to monitor changes by interested entities. In the time of decreasing response rates and budget cuts using data that is already "out there on the Internet" is tempting but requires a attention to its quality and selection of appropriate methods of inference.

Acknowledgments

Robert Pater and Maciej Beręsewicz's contributions were partially financed by the Polish Ministry of Science and Higher Education under the "Dialog" Program, as part of the project: Horizontal educational mismatch—a new method of measurement with application to Poland (grant no. 0127/2016).

References

Beręsewicz, M. (2017). A two-step procedure to measure representativeness of internet data sources. *International Statistical Review*, *85*(3), 473–493. doi:https://doi.org/10.1111/insr.12217

Beręsewicz, M., Lehtonen, R., Reis, F., Di Consiglio, L., & Karlberg, M. (2018). *An overview of methods for treating selectivity in big data sources*. Eurostat Statistical Working Papers.

Berger, Y. G., Muñoz, J. F., & Rancourt, E. (2009). Variance estimation of survey estimates calibrated on estimated control totals. An application to the extended regression estimator and the regression composite estimator. *Computational statistics & data analysis*, *53*(7), 2596–2604.

Boudarbat, B., & Chernoff, V. (2012). Education–job match among recent Canadian university graduates. *Applied Economics Letters*, *19*(18), 1923–1926.

Buelens, B., Burger, J., & van den Brakel, J. A. (2018). Comparing inference methods for non-probability samples. *International Statistical Review*, *86*(2), 322–343.

[6]Detailed description of the Labour Force Survey methodology can be found on the Eurostat website: https://ec.europa.eu/eurostat/cache/metadata/en/jvs_esms.htm

Cedefop. (2019a). Online job vacancies and skills analysis. A cedefop pan-European approach. doi:10.2801/097022

Cedefop. (2019b). The online job vacancy market in the EU. Driving forces and emerging trends. doi:10.2801/16675

Chen, J. K. T. (2016). Using LASSO to calibrate non-probability samples using probability samples. Retrieved from https://deepblue.lib.umich.edu/handle/2027.42/120734

Chen, J. K. T., Elliott, M. R., & Valliant, R. (2018). Inference for nonprobability samples. *Survey Methodology*, *44*(1), 117–144.

Chen, J. K. T., Valliant, R. L., & Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(3), 657–681.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, *40*(2), 137–61.

Colombo, E., Mercorio, F., & Mezzanzanica, M. (2019). AI meets labor market: Exploring the link between automation and skills. *Information Economics and Policy*, *47*, 27–37.

Couper, M. P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, *7*(3), 145–156.

Czarnik, S., Dobrzyńska, M., Górniak, J., Jelonek, M., Keler, K., Kocór, M., . . . Worek, B. (2011). Report concluding the 1st round of the study conducted in 2010. Study of Human Capital in Poland. Retrieved from https://www.parp.gov.pl/storage/publications/pdf/2011_study_hr_en.pdf

Daas, P. J., Puts, M. J., Buelens, B., & van den Hurk, P. A. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, *31*(2), 249–262. doi:https://doi.org/10.1515/jos-2015-0016

Deming, D., & Kahn, L. B. [Lisa B]. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, *36*(S1), S337–S369.

Dever, J. A., & Valliant, R. (2010). A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology*, *36*(1), 45–56.

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382.

Dlugosz, S., Mammen, E., & Wilke, R. A. (2017). Generalized partially linear regression with misclassified data and an application to labour market transitions. *Computational Statistics & Data Analysis*, *110*, 145–159.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249–264.

ESSnet on Big Data. (2017). ESSnet big data. Final technical report, work package 1—web scraping / job vacancies—deliverable 1.3. Retrieved from https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5d/WP1_Deliverable_1.3_Final_technical_report.pdf

ESSnet on Big Data. (2018). ESSnet big data. Final technical report, work package 1—web scraping / job vacancies—deliverable 1.3. Retrieved from https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5e/SGA2_WP1_Deliverable_2_2_main_report_with_annexes_final.pdf

ESSnet on Big Data. (2019). Work package B overview. Retrieved from https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPB_Overview

Guandalini, A., & Tillé, Y. (2017). Design-based estimators calibrated on estimated totals from multiple surveys. *International Statistical Review*, *85*(2), 250–269.

Hershbein, B., & Kahn, L. B. [Lisa B.]. (2018). Do recessions accelerate routine-biased technological change? Evidence from vacancy postings. *American Economic Review*, *108*(7), 1737–1772. doi:DOI:10.1257/aer.20161570

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., . . . Usher, A. (2015). Big data in survey research. AAPOR task force report. *Public Opinion Quarterly*, *79*(4), 839–880. doi:http://dx.doi.org/10.1093/poq/nfv039

Kim, J. K., Park, S., Chen, Y., & Wu, C. (2018). *Combining non-probability and probability survey samples through mass imputation*. arXiv preprint arXiv:1812.10694.

Kim, J. K., & Wang, Z. (2018). Sampling techniques for big data analysis. *International Statistical Review*, *87*(S1), S177–S191. doi:10.1111/insr.12290

Kowarik, A., & Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*, *74*(7), 1–16. doi:10.18637/jss.v074.i07

Kuhn, P., & Skuterud, M. (2004). Internet job search and unemployment durations. *American Economic Review*, *94*(1), 218–232.

Marinescu, I., & Rathelot, R. (2018). Mismatch unemployment and the geography of job search. *American Economic Journal: Macroeconomics*, *10*(3), 42–70. doi:10.1257/mac.20160312

McConville, K. S., Breidt, F. J., Lee, T. C., & Moisen, G. G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, *5*(2), 131–158.

McGowan, M. A., & Andrews, D. (2015). Skill mismatch and public policy in OECD countries. OECD Economics Department Working Papers. Retrieved from http://dx.doi.org/10.1787/5js1pzw9lnwk-en

Pater, R., Szkoła, J., & Kozak, M. (2019). A method for measuring detailed demand for workers' competences. *Economics: The Open-Access, Open-Assessment E-Journal*, *13*(2019-27), 1–29. doi:10.5018/economics-ejournal.ja.2019-27

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics: 24th annual Morris Hansen lecture. *Journal of Survey Statistics and Methodology*, *3*(4), 425–483. doi:http://dx.doi.org/10.1093/jssam/smv035

Polish Agency for Enterprise Development. (2011). Methodology of the Study of Human Capital in Poland (1st round; in Polish). Retrieved from https://www.parp.gov.pl/component/publications/publication/raport-metodologiczny-z-badan-bkl-z-2011-r

Polish Agency for Enterprise Development. (2012). *Methodology of the Study of Human Capital in Poland (2nd round; in Polish)*. Available at request.

Polish Agency for Enterprise Development. (2013). *Methodology of the Study of Human Capital in Poland (3rd round; in Polish)*. Available at request.

Polish Agency for Enterprise Development. (2014). *Methodology of the Study of Human Capital in Poland (4th round; in Polish)*. Available at request.

Polish Agency for Enterprise Development. (2015). *Methodology of the Study of Human Capital in Poland (5th round; in Polish)*. Available at request.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Reid, G., Zabala, F., & Holmberg, A. (2017). Extending TSE to administrative data: A quality framework and case studies from Stats NZ. *Journal of Official Statistics*, *33*(2), 477–511. doi:https://doi.org/10.1515/jos-2017-0023

Särndal, C.-E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*(5), 1–13. Retrieved from http://www.jstatsoft.org/v39/i05/

Somers, M. A., Cabus, S. J., Groot, W., & van den Brink, H. M. (2019). Horizontal mismatch between employment and field of education: Evidence from a systematic literature review. *Journal of Economic Surveys*, *33*(2), 567–603. doi:10.1111/joes.12271. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joes.12271

Statistics Poland. (2018). The demand for labour in 2017. Retrieved from https://stat.gov.pl/en/topics/labour-market/demand-for-labor/the-demand-for-labour-in-2017,1,12.html

The HC survey. (2020). Data files 2017-2018 from the Study of Human Capital in Poland, Polish Agency for Enterprise Development (in Polish). Retrieved from https://www.parp.gov.pl/component/site/site/bilans-kapitalu-ludzkiego#pobierzdane

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

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

Valliant, R. (2019). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*. doi:10.1093/jssam/smz003. eprint: http://oup.prod.sis.lan/jssam/advance-article-pdf/doi/10.1093/jssam/smz003/28324553/smz003.pdf

van Delden, A., Scholtus, S., & Burger, J. (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics*, *32*(3), 619–642.

Wu, C., & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, *96*(453), 185–193.

Yang, S., & Kim, J. K. (2018). *Integration of survey data and big observational data for finite population inference using mass imputation*. arXiv preprint arXiv:1807.02817.

Yang, S., & Kim, J. K. (2019). Nearest neighbor imputation for general parameter estimation in survey sampling. In K. Huynh, D. T. Jacho, & T. Gautham (Eds.), *The econometrics of complex survey data: Theory and applications* (pp. 209–234). Emerald Publishing Limited.

Yang, S., Kim, J. K., & Song, R. (2019). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63. doi:https://doi.org/10.1111/j.1467-9574.2011.00508.x

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Appendix A
Skills measured in the online data

Table A1
Eleven general skills categories used in the Study of Human Capital in Poland

Skill	Behavior dimension	Behavior sub-dimension
Artistic	artistic and creative skills	-
Availability	availability	readiness to travel frequently; flexible working hours (no fixed slots)
Cognitive	seeking an analysis of information, and drawing conclusions	quick summarising of large volumes of text; logical thinking, analysis of facts; continuous learning of new things
Computer	working with computers and using the Internet	basic knowledge of MS Office-type package; knowledge of specialist software, ability to write applications and author websites; using the Internet: browsing of websites, handling e-mail
Interpersonal	contacts with other people (with colleagues, clients, people in the care)	cooperation within the group; ease in establishing contacts with colleagues and/or clients; being communicative and sharing ideas clearly; solving conflicts between people
Managerial	managerial skills and organisation of work	assigning tasks to other members of staff; coordination of work of other staff; disciplining other staff—taking them to task;
Mathematical	performing calculations	performing simple calculations; performing advanced mathematical computations
Office	organisation and conducting office works	-
Physical	physical fitness	-
Self-organization	self-organisation of work and showing initiative (planning and timely execution of tasks at work, efficiency in pursuing a goal)	independent making of decisions; entrepreneurship and showing initiative; creativity (being innovative, inventing new solutions); resilience to stress; timely completion of planned actions
Technical	technical imagination and handling technical devices	handling technical devices; repairing technical devices

Appendix B
Details about the online data

Table B1
Coding precision, measured by the number of job offers with codes of differing accuracy (different number of digits) based on pooled data from 2011, 2013 and 2014 for occupation

Number of digits	Job offers
6 digits	33 966
5 digits	2 663
4 digits	715
3 digits	614
2 digits	142
1 digits	138

Table B2
Information available in published datasets from the Study of Human Capital in Poland despite its quality

Variable	2011	2012	2013	2014
Occupation (up to 6 digits)	X	-	X	X
Occupation (only 1 digit)	X	X	X	X
NACE (up to 3 digits)	X	X	X	X
Industry	X	X	X	X
Province	X	X	X	X
Subregion	X	X	x	X
Education	X	X	X	X
Foreign languages	X	X	X	X
Work experience	X	X	X	X

Table B3
Percentage of missing data in selected variables in each wave of the Human Capital in Poland survey

Variables	2011	2013	2014
Occupation	0.33	0.40	0.49
NACE	6.04	56.86	41.98
Voivodeship	1.06	0.01	0.21

Table B4
Relative standard errors of estimators [in %] for vacancies in the Demand for Labour survey for Q4 of 2011, 2013 and 2014

Section	2011	2013	2014
Total	3.40	4.01	3.98
C - Manufacturing	5.50	5.27	5.64
F - Construction	13.86	19.21	15.12
G - Trade; repair of motor vehicles	13.69	15.75	16.33
H - Transportation and storage	8.07	9.93	9.17
I - Accommodation and catering	15.99	20.78	18.26
J - Information and communication	6.30	7.04	11.50
K - Financial and insurance activities	7.00	8.36	7.43
M - Professional, scientific and technical activities	8.12	8.71	12.01
N - Administrative and support service activities	23.09	12.89	17.76
O - Public administration and defence; compulsory social security	3.19	3.50	2.56
P - Education	8.85	10.65	12.06
Q - Human health and social work activities	5.53	6.88	6.00
R - Arts, entertainment and recreation	7.08	8.68	9.28
S - Other service activities	18.09	21.29	20.77

Table B5
Relative standard errors of estimators [in %] for vacancies by occupation (2-digit code) in Q4 of 2011, 2013 and 2014 based on the proposed bootstrap procedure

Year	Min	Q1	Median	Mean	Q3	Maximum
2011	2.38	4.78	6.42	7.84	10.20	20.05
2013	3.49	4.94	6.50	8.18	10.12	20.40
2014	2.32	5.50	6.55	8.26	10.70	17.56

Table B6

Distribution of Occupation (ISCO-08 2-digit codes) in Population and HC data (average over 2011, 2013 and 2014)

Occupation	Population	Online data	CV
11 - Chief executives, senior officials and legislators	0.49	1.68	4.50
12 - Administrative and commercial managers	1.70	2.22	4.53
13 - Production and specialized services managers	1.27	2.02	9.56
14 - Hospitality, retail and other services managers	0.29	2.78	11.84
21 - Science and engineering professionals	4.45	4.09	4.64
22 - Health professionals	3.33	1.47	6.45
23 - Teaching professional	0.91	2.00	10.75
24 - Business and administration professionals	6.73	14.65	3.74
25 - Information and communications technology professionals	3.94	8.17	8.10
26 - Legal, social and cultural professionals	0.71	0.91	5.54
31 - Science and engineering associate professionals	1.51	1.50	6.03
32 - Health associate professionals	0.79	0.58	6.31
33 - Business and administration associate professional	4.37	19.33	3.67
34 - Legal, social cultural and related associate professionals	0.97	0.53	5.66
35 - Information and communications technicians	1.73	0.78	6.73
41 - General and Keyboard Clerks	1.41	1.82	2.90
42 - Customer Services Clerks	5.20	2.53	6.70
43 - Numerical and Material Recording Clerks	1.28	1.46	6.63
44 - Other Clerical Support Workers	2.52	0.51	4.09
51 - Personal Services Workers	2.41	3.10	16.14
52 - Sales Workers	8.79	16.49	15.35
54 - Protective Services Workers	1.28	1.16	13.77
71 - Building and Related Trades Workers (excluding Electricians)	9.53	1.71	17.09
72 - Metal, Machinery and Related Trades Workers	7.09	2.36	5.27
73 - Handicraft and Printing workers	0.72	0.25	5.76
74 - Electrical and Electronics Trades Workers	2.09	1.23	13.99
75 - Food Processing, Woodworking, Garment and Other Craft and Related Trades Workers	5.64	1.18	5.35
81 - Stationary Plant and Machine Operators	2.71	0.35	5.19
82 - Assemblers	2.72	0.21	6.27
83 - Drivers and Mobile Plant Operators	7.99	1.67	9.23
91 - Cleaners and Helpers	1.35	0.19	10.09
93 - Labourers in Mining, Construction, Manufacturing and Transport	2.20	0.49	8.52
94 - Food Preparation Assistants	1.26	0.26	19.34
96 - Refuse Workers and Other Elementary Workers	0.60	0.31	5.41

Appendix C
Estimation process and results

Table C1

Quality of the model measured by Area Under the ROC Curve (AUC; average over 500 bootstrap replicated)

SKILLS	$ECMC$	$ECLASSO1$	$ECLASSO2$	$ECALASSO1$
Technical	0.817	0.829	0.846	0.829
Mathematical	0.753	0.784	0.818	0.784
Artistic	0.657	0.665	0.672	0.665
Computer	0.742	0.748	0.755	0.748
Cognitive	0.636	0.644	0.654	0.644
Managerial	0.715	0.722	0.731	0.722
Interpersonal	0.724	0.731	0.750	0.731
Self-organization	0.689	0.695	0.708	0.695
Physical	0.674	0.687	0.713	0.687
Availability	0.597	0.605	0.635	0.604
Office	0.655	0.671	0.681	0.670

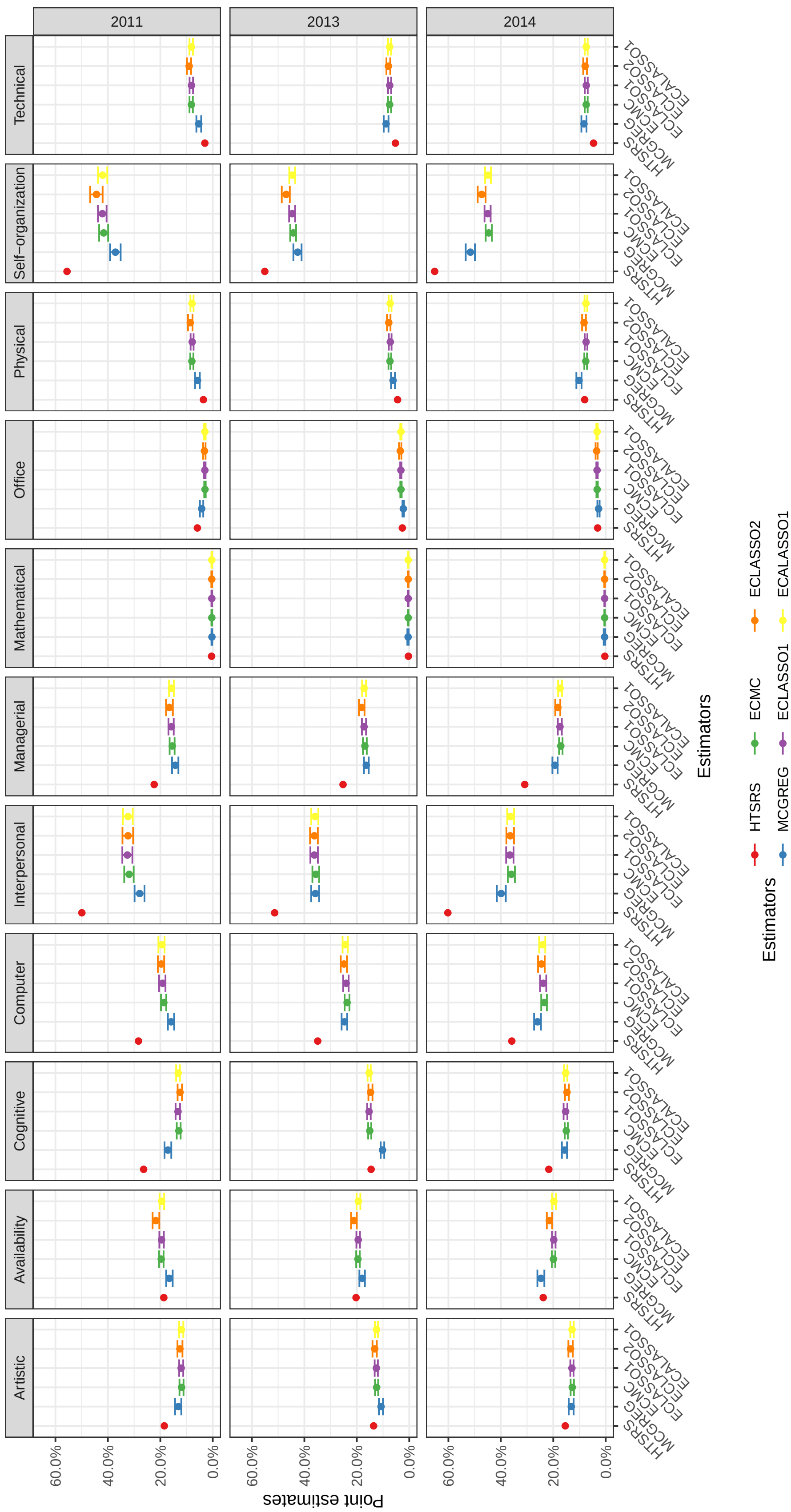

Figure C1. Point estimates for the estimators and HTSRS estimator for each skill for 2011, 2013 and 2014 separately