UNIWERSYTET EKONOMICZNY W POZNANIU

Instytut Informatyki
i Ekonomii Ilościowej

Paulina Reus i Alicja Smaruj

Estymacja wielkości błędu pokrycia w Centralnej Bazie Ofert Pracy

Estimation of the over-coverage error in Central Job Offers Database

Praca licencjacka

Promotor: dr Maciej Beręsewicz, prof. UEP

Pracę przyjęto dnia:

Podpis promotora

Kierunek: Informatyka i Ekonometria

Poznań 2022

Contents

Introduction

The analysis of labour demand allows an ongoing assessment of the use of jobs and at the same time allows a broader characterisation of the actors of the national economy. Official statistics mainly focuses on using sample surveys (i.e. job vacancy surveys) to estimate the demand. These surveys are usually carried out on a quarterly or less frequent basis, which does not allow for dynamic changes in employment, which results in the fact that, unfortunately, they do not always follow the current trends or are limited in scope.

In this case, non-statistical sources, which are result of registration of job vacancies, seem to be attractive. These are mainly online services such as job advertisement portals or administrative sources with vacancies submitted to public employment offices. Although theoretically the non-statistical sources should only include entities looking for employees, there are entities that do not have vacancies or have fewer/more vacancies, which results in coverage, selection and measurement errors and therefore cannot be used directly to study population characteristics. In this thesis we chose to investigate the magnitude of one of these errors, namely the coverage error, one of the non-statistical sources.

Non-statistical sources, in the case of research on labour demand, are advertising platforms. In Poland, we can distinguish platforms where employers place their own job advertisements, such as olx.pl, pracuj.pl, jaob.pl and many others. Data coming from the first type of service are not always properly structured. In the case of the second type of services, the data are more structured, which is conditioned by the form in which the offer is submitted to the employment office. A more structured structure of the second type of database will make it much easier to carry out the analysis on these resources.

In our research we used Central Job Offer Database (CBOP), due to its structured character and easy way of accessing data from the database. CBOP is created as a result of job offers registration at public employment offices and held by the Ministry of Family and Social Policy.

Estimation of coverage error in the aforementioned source will be obtained by verifying whether there are entities in the database which should not be there, i.e. entities which do not have vacancies despite the advertised vacancies. In order to do this, a random sample should be examined in terms of the timeliness of posted offers and the results should be used in an appropriate model.

Due to the dichotomous nature of the results of the study of whether an offer is current, the best model for the estimation of these data is logistic regression. This method allows to assess the impact of many different factors (explanatory variables) on the probability of the occurrence of a given event. By selecting appropriate explanatory variables, the model is able to assess whether a given offer is or is not valid. This thesis focuses on estimating the coverage error in the CBOP using logistic regression.

The main objective of this work is to estimate the coverage error in the Central Job Database, which consisted of smaller sub-objectives that made it possible to achieve. The sub-objectives consisted of:

1. Cleaning the data extracted from the Central Job Database in order to assess the errors resulting from the presence in the database of units outside the scope and duplicate offers
2. Designing and carrying out an audit sample to assess the over-coverage resulting from outdated vacancies.
3. Based on the results, appropriately apply logistic regression to predict which ads are current.
4. Evaluate the different sources of over-coverage over time, for each sub-population.

The dissertation is structured as follows. The first chapter contains an introduction to the terminology used in labour demand research, a description of the sources of data, as well as the ways of obtaining them and the various errors that occur in statistical research. *Paulina Reus* was responsible for the theoretical introduction to the study. Chapter two introduces the idea of the logistic regression model and explains the reasons for choosing this type of model. The theoretical explanation of the logistic regression model was prepared by *Paulina Reus*, while the description of the use of regression in the following study was done by *Alicja Smaruj*. The third chapter written by *Alicja Smaruj* focuses mainly on data analysis and estimation. It contains a description of the data reliability verification performed, the analysis of its results and their use to create a logistic regression model applied to the cleaned CBOP data. The verification of

the drawn sample of CBOP offers was carried out jointly by *Alicja Smaruj* and *Paulina Reus*, and the analysis of its results was prepared by *Paulina Reus*. A thorough cleaning of the data from undesirable values was carried out by *Alicja Smaruj*, as well as the design and use of the logistic regression model.It is summarised with conclusions from the results.

Chapter 1

Measuring the demand for labour

1.1 Labour market

Labor market data is crucial for policy evaluation. According to the definition of Statistics Poland, the demand for labor is defined as *the number of vacancies that can be offered by the economy of a given country under certain socio-economic conditions*. They are often used in policy making or setting employment targets, and in the calculation of labour market indicators, including international ones comparisons (International Labour Organization 2015). Employment shapes conditions and relationships of work and life of the population of a given region. Together with capital, they constitute a given market. It is also worth mentioning that, the conducted research confirms the existence of a positive relationship between the development of the labor market (in terms of the demand for labor) and the level of economic development of the region (Nagel 2015).

The statistics of labour demand themselves form the basis of many European Union policies and are used, among others, to monitor the European Employment Strategy (EES), the European Pillar of Social Rights and the Agenda for Sustainable Development 2030, which is why many institutions, both national and international (EURES), conduct regular research and analysis of labour demand.

Regarding the evolution of labour demand in the European Union, based on data collected by Eurostat, the vacancy rate in Q4 2021 (Eurostat 2022) was 2.6%, with the highest number of job vacancies in Q4 2021 in the Czech Republic (4.9%), Belgium (4.7%) and the Netherlands (4.2%), while the lowest rates were recorded in Bulgaria, Greece and Spain (0.7% in all three countries). In Poland, the vacancy rate is 1.2%. However, it should be remembered that the

economy of the European Union is recovering after two years of the pandemic, when the demand for labor was initially suppressed by restrictions related to the high number of cases. The data used to compile these statistics comes exclusively from statistical sources.

Each country has statistical agencies that collect data through labor force surveys and national censuses. As it is summarised in the study Gathering and analysing labour market data (International Labour Organization 2015) these data are often limited or very general, which is often due to limited resources Resource limitations. Unions do not always provide reliable data on their establishments, which affects the effectiveness of statistics. However, it seems interesting to examine this issue based solely on non-statistical sources to broaden the overall picture of the problem.

1.2 Basic definitions

A vacancy is a paid position that is newly created, vacant, or to be vacated. Statistics Poland (2022a) defines vacancies as positions or jobs vacated as a result of employee turnover or newly created that simultaneously meet the following three conditions:

1. positions and jobs were actually unoccupied on the reference day,
2. employer 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.

Statistics Poland (2022b) defines ”job offer” which is at least one vacancy declared by an employer to public employment offices (PEOs). This definition is used solely to compile statistics based on administrative data that is created based on PEOs registration. This is a broader definition as job offer at PEOs may refer to the two main types of contracts: mandate contracts and employment contracts. In both cases, the employee undertakes to perform an activity for the employer as specified in the contract. However, in the case of an employment contract, the employee is entitled to employee privileges, as it is a contract concluded on the basis of the Labour Code, while a mandate contract is only a formal civil law contract.

1.3 Data sources

In general we may distinguish two types of data sources: statistical and non-statistical. First group contains mainly of censuses, sample surveys and statistical registers. The latter cover sources that were created for non-statistical purposes, for instance administrative data or commercial databases. In Poland we may distinguish the following sources regarding measuring job demand that are described below.

1.3.1 The Demand for Labour Survey

The Labour Demand Survey is carried out by the Statistical Office. The subject scope of the survey covers economic entities with the number of employees equal to or greater than 1. The subject scope of the survey covers changes in the number of employed and not yet filled jobs, including newly created jobs. The survey of demand for work is carried out quarterly by a representative method. The sample in the survey is selected by stratifying the general population, which means that sub-populations are determined from the population, in which independent samples are drawn. Also during the sampling process units with up to 9 employees and units with more than 9 employees are distinguished. Data accuracy is estimated on the basis of random and non-random error analysis. The data themselves are obtained through the use of an electronic form, via the Statistics Poland's reporting portal. The form includes, interãlia, the total number of employed persons, the number of employed women, the number of newly created places since the beginning of the year, the number of closed places since the beginning of the year.

The basic indicator which is calculated on the basis of the survey results is the Job Vacancy Rate (JVR), which defines the share of vacancies in the total number of occupied and vacant jobs filled and vacant jobs. Data related to occupation groups are compiled on the basis of the Classification of Occupations classification of occupations and specialities 2014, while data on the type of activity of national economy entities are presented according to the Polish Classification of Activities 2007 (Statistics Poland 2022c).

The survey is obligatory under the *Ustawa z dnia 29 czerwca 1995 r.o statystyce publicznej* and is conducted quarterly. Since Poland's accession to the European Union, data from this survey are the basis for developing the job vacancy rate, which is important for comparisons at

the international, national and regional level. The survey is primarily used for monitoring and evaluating the effectiveness of measures taken in the field of labour market policy.

1.3.2 Administrative data – Central Job Offers Database

The Central Job Offer Database (CBOP) is one of the databases where official job offers are stored. It is a database shared by labour offices, which contains information on current job offers and internship proposals from a given office. Job offers are submitted directly by employers using an e-form downloaded from www.praca.gov.pl. The database is managed and made available by the Ministry of Family, Labour and Social Policy. CBOP operates on the basis of the *Ustawa z dnia 20 kwietnia 2004 r. o promocji zatrudnienia i instytucjach rynku pracy* on employment promotion and labour market institutions. The main purpose of the database is to help unemployed people find a job matching their education, experience and competences. The database is regularly updated and expanded. Employers can post their offers using an e-form. The database contains approximately 27,000 job offers registered with employment offices.

In order to group the offers, CBOP uses the Classification of Occupations and Specialities, which is a list of occupations identified on the labour market. It is developed on the basis of International Standard Classification of Occupations ISCO-08. The classification distinguishes 9 main groups, which are:

- 1 – Personal representatives authorities, public authorities, senior officials and managers;
- 2 – Professionals ;
- 3 – Technicians and associate professionals intermediate personnel;
- 4 – Employees office workers;
- 5 – Workers, service workers and sales workers;
- 6 – Farmers, gardeners, foresters and fishermen;
- 7 – Workers, industrial workers, Industrial workers and craftsmen;
- 8 – Machine operators and assemblers of machines machine and plant operators;
- 9 – Workers who perform simple work.

The above groups are subdivided into subgroups that further specify the occupational group.

The CBOP allows external entities to download in an automated way the information on job offers made available by the CBOP, but the entity requesting the possibility of downloading the data must complete the application presented in Figure 1.1.

1.3.3 Commercial database – Pracuj.pl

With the growing popularity of online recruitment portals, more and more employers are choosing to post their offers on websites instead of employment offices, which results in the availability of data on job offers from these websites. However, the offers posted on recruitment portals are not verified in any way, so the reliability of these sources is not always high. According to the ”Megapanel PBI/Gemius” study, conducted by Polskie Badania Internetu (PBI) and the Gemius research company, which is the standard for measuring website audience in Poland, one of the most popular job portals in 2021 in Poland was ”Pracuj.pl” with 3.39 million users and about 12 percent reach (MegaPanel 2021). As far as user satisfaction is concerned, pracuj.pl has been rated very high, both from the perspective of employers and job seekers (Wiącek et al. 2011).

Pracuj.pl is a private company and was registered in 2000. In addition to the possibility of remote recruitment, the platform also provides individual salary reports and analyses of labour market trends.The portal contains a job search engine, as well as a rich database of articles and numerous tools, including a test of professional predispositions, an interview simulator, and a salary calculator. The Internet as a job search tool is also a source of advice, professional and interpersonal skills tests, presentations of employers and their internships, as well as industry sections containing detailed information on job offers and salaries in a given sector (Wieczorek 2011). In order to place an ad on pracuj.pl you need to pay a fee.

To place an ad on pracuj.pl, you must pay a fee. The cost for placing an ad for 30 days ranges from 499 PLN to 999 PLN. After the expiry. After the expiration date, the offer changes its status to expired and is archived.

ZGŁOSZENIE KRAJOWEJ OFERTY PRACY

I. Informacje dotyczące pracodawcy krajowego

1. Nazwa pracodawcy / agencji zatrudnienia:……

……

2. Adres pracodawcy – Kod pocztowy …….. Miejscowość ……

Ulica …… Gmina ……

3. Numer telefonu …… e-mail ……,

Osoba wskazana przez pracodawcę do kontaktów……

4. REGON ……..……..…… NIP ……..

5. Czy pracodawca jest agencją zatrudnienia zgłaszającą ofertę pracy tymczasowej?
NIE ☐ TAK ☐ nr wpisu do rejestru …….

6. Liczba zatrudnionych pracowników .....

7. Czy pracodawca w okresie 365 dni przed dniem zgłoszenia oferty pracy został ukarany lub skazany prawomocnym wyrokiem za naruszenie przepisów prawa pracy albo jest objęty postępowaniem dotyczącym naruszenia przepisów prawa pracy? TAK ☐ NIE ☐

8. Czy oferta jest w tym czasie zgłoszona do innego PUP na terenie kraju? TAK ☐ NIE ☐

II. Informacje dotyczące zgłaszanego miejsca pracy

1. Nazwa stanowiska ……..

2. Liczba wolnych miejsc pracy …… w tym tylko dla osób niepełnosprawnych ……

3. Wnioskowana liczba kandydatów ……

4. Miejsce /adres/ wykonywania pracy ……

5. Data rozpoczęcia pracy ……..

6. Rodzaj umowy: ☐ umowa o pracę na czas nieokreślony; ☐ umowa o pracę na czas określony;
☐ umowa o pracę na okres próbny; ☐ umowa o pracę na zastępstwo;

☐ umowa zlecenie /uzasadnienie/ ……..……..;

☐ inne/uzasadnienie/ ……..

7. Okres zatrudnienia/okres wykonywania umowy OD ……..…….. DO ……..……..…….

8. Czy oferta pracy jest ofertą pracy tymczasowej? TAK ☐ NIE ☐

9. System i rozkład czasu pracy: ☐ jednozmianowa; ☐ dwie zmiany; ☐ trzy zmiany; ☐ inne

10. Wymiar czasu pracy ……..……; praca w godzinach: ……..……..

Niepełny wymiar czasu pracy /uzasadnienie/ …… ……..

11. Wysokość proponowanego wynagrodzenia brutto ……..

Premie/dodatki……..

12. System wynagradzania: ☐ miesięczny; ☐ godzinowy

13. Ogólny zakres obowiązków ……..…

……..……

……

Urząd Pracy m.st Warszawy 28.03.2022 r. – Załącznik nr 1 do procedury nr ewid.: P-7.11.

Figure 1.1. Application form for submitting an job offer to Public Employment Offices

1.4 Local and international experiences in measuring job vacancies

In addition to the methods mentioned above for capturing job vacancies, there are many other methods.

An interesting measure of job vacancies is the Conference Board's Help-Wanted Advertising Index, which expresses the change in the number of jobs by examining the change in the number of ads in 51 major U.S. newspapers. It was used from 1951 to 2008. (Abraham i Wachter 1987)

The Job Openings and Labor Turnover Survey (JOLTS) is also based on surveys, but unlike European Statistics, which is conducted quarterly, it is reported monthly. JOLTS is a report from the Bureau of Labor Statistics of the U.S. Department of Labor that counts job vacancies and separations, including the number of workers who voluntarily leave employment. (Bossler et al. 2020)

A revolution in job search is WoLMIS, a system that collects and automatically classifies multilingual job offers on the Internet with reference to a standard classification of occupations. The system was developed for the European agency Cedefop, which supports the development of European VET policy and contributes to its implementation. In particular, WoLMIS enables labour market analysts and specialists to understand labour market dynamics and trends in several European countries. (Boselli et al. 2018)

1.5 Non-sampling errors

Every statistical study is subject to errors, of course one should try to maximize the representativeness of the results by minimizing these errors. Among errors we can distinguish random errors and non-random errors.

- *Random errors* this is the category of error that is directly related to the mathematical model used in sampling and inference. The error arises from the random nature of the sample as well as from the fact that random units are included in the sample. The lack of an appropriate model, which is the most frequent cause of non-random error, makes proper analytical description difficult (Szreder 2015). In order to eliminate them, the whole sample should be examined (Bethlehem i Biffignandi 2021).

- *Non-random errors* are all errors that arise as a result of incorrect preparation of the sampling frame, taking of measurements, or occurring data deficiencies. We can distinguish among them:
  - Coverage error which occurs when the sampling frame does not include some units from the study population (under-coverage) or includes units outside of the study population (over-coverage).
    - Omission error (part of the under-coverage error) the selected sampling frame / list / database does not include all units of the population under study (for example: employer in the application form, reports one job although he/she is looking for more people)
    - Inclusion error (part of the over-coverage error) is the error when the selected sampling frame includes units that do not belong to the population under study.These are cases where the database contains units that have not issued advertisements
    - Repetition / duplicate error (part of the over-coverage error) occurs when a sampling frame contains more than one occurrence of a given unit.
    - The contact error occurs when there are entities in the survey operation that cannot be contacted (for example, institutions that have provided incorrect contact details, making it impossible to verify the information).
  - Measurement errors are related to incorrect collection and recording of information. They are usually caused by incorrect recording of answers by the interviewer or misunderstanding of the question by the respondent (Bethlehem i Biffignandi 2021).
  - Non-response error occurs when a survey fails to participate in the study (unit non-response) or answer one or more questions (item non-response).
  - Auto-selection error is a type of error that results from the general nature of the study on which the data are based. The study units are not randomly selected, they decide for themselves to participate in the study.

1.6 Conclusion

In our study, we decided to use administrative source – CBOP – to broaden the general picture of the labour demand in Poland. In order to verify the information provided by the CBOP, an audit sample by means of the telephone survey was conducted. During the survey, respondents answered questions related to the existence of a given offer, its validity period and the number of unfilled positions offered by employers.

Chapter 2

Methods to assess over-coverage error

2.1 Decomposing over-coverage error

The goal of the thesis is study the over-coverage error in Central Job Offers Database. Over-coverage error needs to be decomposed in order to understand sources of errors. To do so we assume that we should make this administrative data coherent with the Demand of Labour (DfL) survey conducted by Statistics Poland.

Having that, we may distinguish two sources: referring to the target population (i.e. statistical units, companies with at least 1 employed) and to definition of job vacancy. The latter consider errors due to not meeting the definition of job vacancy (i.e. contract vs internship or contract of mandate), duplication and expired job vacancies.

$$\text{over-coverage error} = n_1 + n_2 + n_3, \tag{2.1}$$

where

- n_1 – the number of units that are out-of-scope regarding definition used in the DfL survey.
- n_2 – the number of units that are duplicated.
- n_3 – the number of units that are outdated i.e. despite of visibility on the website vacancy is already occupied.

Then, we can calculate over-coverage rates as specified below

$$\text{Out-of-scope units rate} = M_i = \frac{n_i}{n}, \tag{2.2}$$

where i denotes type of over-coverage and n denotes the initial number of observations/vacancies.

The last number, i.e n_3 is hard to verify using only observed data and thus need to be estimated based on some external source, such as sample survey. Then, this external sample is used to build a model that allow to predict which vacancy (ad) is already taken and should be removed. In our study we decided to use logistic regression that is presented below.

2.2 Logistic regression

2.2.1 Basic information

The logistic regression model is a special case of the generalised linear models (GLMs). The distinguishing feature of logistic regression is that the dependent variable is dichotomous, meaning that it takes the value of 0 or 1 (i.e. it assumes Bernoulli or binomial distribution). The goal of logistic regression is to model conditional probability of the target variable given set of dependent variables.

The main advantage of logistic regression is its usefulness with any type of variable. For example, variables can be continuous or discrete, or any combination of both types, and referring to Lee (2005) there is no need for normal distributions. By the nature of logistic regression, the analysis is not affected by qualitative, quantitative or categorical dependent and independent variables. According to Agresti (2018), logistic regression is beneficial for the study of binary interpretation and categorical data.

2.2.2 Bernoulli and binomial distribution

Let n denote sample size and $i = 1, ..., n$ denote index of a unit in this sample. Consider a binary variable y_i taking the values 0, indicating failure, and 1, indicating success, and π_i denotes probability of success. Then we can write probability density function of Bernoulli distribution (Alvarez-Santiago et al. 1996) using the following formula :

$$\Pr(Y_i = y_i) = (\pi_i)^{y_i}(1 - \pi_i)^{1-y_i}. \tag{2.3}$$

Having that, the expected value and variance are given by the following equations:

$$E(Y_i) = \mu_i = \pi_i$$
$$Var(Y_i) = \pi_i(1 - \pi_i). \tag{2.4}$$

Assume that the units under study can be divided into I groups according to the factors of interest such that all units in the group have identical values of all variables. Let n_i denote the number of observations in group i, and let Y_i denote the number of individuals who have the characteristic of interest in group i. Y_i takes values from 0 to n_i. Assuming that the observations n_i are independent in each group and that their probabilities of having the given characteristic πi are equal, the distribution Y_i is binomial and can be written as follows

$$Y_i \sim B(n_i, \pi_i), \tag{2.5}$$

and its probability function will take the form

$$Pr(Y_i = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{(n_i - y_i)}, \tag{2.6}$$

where y_i is a realisation of Y_i. The mean and variance of Y_i can be shown as

$$E(Y_i) = \mu_i = \pi_i n_i, \tag{2.7}$$

$$var(Y_i) = (\sigma_i)^2 = n_i \pi_i (1 - \pi_i), \tag{2.8}$$

Y_i is a Bernoulli random variable with mean and variance dependent on probability, from which it follows that a factor affecting this probability will affect the variance of the observation and the expected value. We can write the number of successes Y_i in group i as the sum of the individual indicator variables, so $Y_i = \sum_j Y_{ij}$, which suggests that the mean Y_i is the sum of the individual means.

2.2.3 Logit transformation

The next step in building the model is to capture the probability π_i from the vector of observed variables $\boldsymbol{x}_i$. The simplest solution would be to assume that π_i is a linear function of the variables. A linear probability model can be used for this

$$\pi_i = x_i' \beta, \tag{2.9}$$

where β is a vector of regression coefficients (Denk i Finkel 1992).

Unfortunately, in this model the probability π_i on the left must be between zero and one, but the linear predictor x' and β on the right can take any real value, which means that the predicted values may not be in the right range. This can be done by removing the range constraints from the likelihood function and modelling this transformation as a non-linear function of the variables.

First, the probability must be transformed into odds, which will be defined as the ratio of probability to its complement, the ratio of success to failure

$$\text{odds}_i = \frac{\pi_i}{1-\pi_i}. \tag{2.10}$$

The above function should then be logarithmised by calculating logit or log-odds, resulting in the removal of the lower bound as given below

$$\text{logit}(\pi_i) = log\frac{\pi_i}{1-\pi_i} \tag{2.11}$$

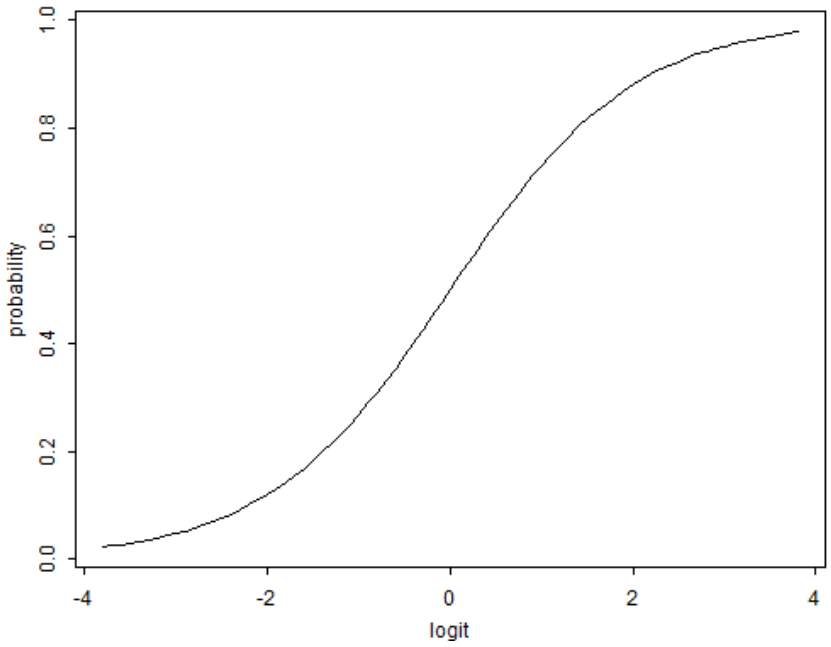

Figure 2.1. The Logit Transformation

It is worth noting that as the probability drops to zero, the odds approach zero and the logit approaches $-\infty$. Conversely, when the probability approaches unity, the odds approach $+\infty$ and so does the logit. Thus, the logit map the linear predictor in the interval (0;1) to the

whole real line. Note that if the probability is 1/2, then the chances are equal and the logit is zero. Negative logit represent probabilities less than half, and positive logits correspond to probabilities greater than half.

Solving for π_i in equation

$$\text{logit}(\pi_i) = log\frac{\pi_i}{1-\pi_i}, \tag{2.12}$$

we obtain

$$\pi_i = \text{logit}^{-1}(\eta_i) = \frac{\exp\{\eta_i\}}{1+\exp\{\eta_i\}}, \tag{2.13}$$

where $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ is the linear predictor.

Assuming that the logit probability π_i, is consistent with the linear model, one can proceed to define the logistic regression model.

The above expression should be defined as a multiplicative probability model, in which $\exp(\beta_j)$ is the odds ratio. Then solving for the probability π_i in the logit model, we obtain the model, in which the left-hand side of the equation is on a known probability scale, while the right-hand side is a non-linear function of the predictors, which makes it very difficult to express with it the effect of a one-unit increase in the predictor, while leaving the others unchanged, on the probability. If the predictors are continuous, derivatives with respect to x_j should be performed.

$$\frac{d\pi_i}{dx_{ij}} = \beta_j\pi_i(1-\pi_i) \tag{2.14}$$

The equation above shows that the effect of the j-th predictor on the probability π_i depends on the coefficient β_j and the probability value.

2.3 Using logistic regression to estimate over-coverage error

The section above has explained the theory behind the logistic regression. Knowing the nature of the calculation method, we chose to apply it to our data to examine whether job offer is valid or invalid. It was perfect to our case because of its dichotomy.

At first, it was necessary to select the right explanatory variables and based on them built a model, which would allow us to extract the average probability of success $\hat{\pi}_i$, which would be indicative of validity of the job offer.

Then, we would be able to compute the value of over-coverage error n_3 according to the equation:

$$n_3 = \sum_{i=1}^{n} v_i - \sum_{i=1}^{n} \hat{\pi}_i \times v_i \tag{2.15}$$

where

- n_3 – the number of free vacancies that are invalid despite of visibility on the website
- $\hat{\pi}_i$ – the average probability of success attesting to the validity in an i-th job offer
- v_i – the number of free vacancies in an i-th job offer
- n – the number of reported job offers

This application enables the better estimation of the labour demand in Poland and simultaneously the better estimation of over-coverage error n_3. On the basis that we have an ability to count the approximate number of free vacancies by multiplying the number of free vacancies and the average probability of success and then subtract it from the theoretical number of vacancies, as a results we obtain the volume of the over-coverage error n_3.

2.4 Conclusion

In order to appropriately estimate the volume of over-coverage error we decomposed it into three types of errors. The error resulting in the lack of job offers' validity could be handled by the presented concept of logistic regression. Having regard to the available methods, we are ready to work on the real data and estimate the over-coverage error by deduplication and creating the model presenting the predicted labour demand.

Chapter 3

Estimation of the over-coverage error in Central Job Offers Database

3.1 Preparation of admin data

3.1.1 Aligning the definitions and population of job vacancies with the Demand for Labour survey

Data for the study was obtained through Central Job Offers Database (CBOP) API that allow to download full database at single query. In order to access the database one needs to formally inquire for a token and specify the IP of the computer which will be connecting to the database.

Being aware of the good quality of the data used, we were able to proceed with in-depth work on the data. To begin with processing all the collected job offers from 29 September 2020 to 17 February 2022. They were transformed from `json` format to `rds` (R Core Team 2021) using `RcppSimdJson` package (Eddelbuettel i Knapp 2021) and properly cleaned using several R packages, namely: `tidyverse` (Wickham, Averick, et al. 2019), `data.table` (Dowle i Srinivasan 2021) and `lubridate` (Grolemund i Wickham 2011).

In order to assess the coverage error in CBOP we need to make the original data coherent with definitions used on the Demand for Labour survey (JVS). This required the following steps:

1. aligning definitions regarding the type of contract (e.g. employment contract vs contract of mandate),
2. selecting place of work i.e. it should be work in Poland,

3. deduplication of job ads as multiple files were processed (e.g. we get data for each day of a month and one ad may have different details at the beginning and end of this period),
4. being active in given month.

The definition of contract type used in the CBOP allows for aligning administrative population with the one defined in JVS. In particular, to meet the criteria we selected the following types of contracts:

- Employment contract for a definite period (pol. umowa o pracę na czas opreślony),
- Employment contract for a trial period (pol. umowa o pracę na okres próbny),
- Employment contract for an indefinite period (pol. umowa na czas nieokreślony),
- Contract for the time of performance of specific work (pol. umowa o pracę na czas wykonania określonego zadania),
- Substitute employment contract (pol. umowa o pracę w zastępstwie).

One should be noted. This definition is accurate and without measurement error because given employer need to specify type of contract when submitting a job vacancy to public employment offices (PEOs). Potential candidates for this vacancy can be only employed on this contract, it cannot be changed as it is verified by PEOs staff.

Furthermore, population of JVS cover only positions in Poland thus we need to narrow down it to places in Poland. Please note that this requirement does not mean that the job vacancy is for Polish citizen, it can be offered to any person.

Finally, determining whether job vacancy is active is the most problematic. On the one hand, when downloading the job ads on a given day we get only these ads that were placed in the CBOP online service. On the other, from various studies we know that over-coverage error is present due to out-of-data ads that were not taken down from the service or PEOs staff were not informed about the change in the status. According to interviews with PEOs' staff this process looks as follows:

- employers are required to inform given PEO that job vacancy is occupied, and PEOs' staff are required to change this in the dataset,
- it may happen that employers are reporting with delay. There may be also a delay in reporting by PEOs' staff i.e. changing the information in the IT system,
- to make this period as short as possible PEOss staff are contacting once two weeks or a month to verify if the vacancy is available and how many places are occupied.

To assess this error we undertaken the following steps: 1) we used information about whether the ad is still available (column `czyWazne` in CBOP database) and 2) we conducted an audit sample survey to directly contact employers and verified availability and the correct number of unoccupied positions.

3.1.2 Data deduplication

Due to the fact that job offers were downloaded every day, a lot of them were duplicated. However, that was not the only reason of the duplication. We chose to sort them by date of download and group by year. Our main column was hash value as it could not be repeated. Thanks to the package `janitor` created by Firke (2021) we were able to distinguish differences and embrace an appropriate way to cope with them.

The table 3.1 presents the changing number of job offers after successive steps of taken methods of deduplication. The first row, ”start value”, represents the number of job offers at the beginning before any action. Then we removed all the job offers, which differ just by the date of download and nothing else, so they were ”identical” to each other. This step was the most crucial, because the number of job offers dropped drastically by approximately 94%. Another deduplication problem was generated by human fault and it was ”spelling mistakes”, which was correctable by replacing polish letters to Latin using `stringi` (Gagolewski 2021) and constitutes only 0.21% of whole deduplication excluding difference in download date which was counted applying `scales` (Wickham i Seidel 2020).

Table 3.1. Number of job offers after different stages of deduplication

	Deduplication stage	2020	2021	2022	all
1	Start value	1,219,863	7,208,013	786,707	9,214,583
2	Identical	74,416	352,960	53,965	470,402
3	Spelling mistakes	74,391	352,806	53,947	470,205
4	Actualised - different validity date	64,416	293,702	49,460	392,996
5	Actualised - different start date	63,895	292,270	49,335	390,886
6	Actualised - different end date	63,803	291,734	49,283	390,201
7	Actualised - different salary	63,623	290,005	48548	386,845
8	Actualised - combined	62,816	285,092	48,118	380,379
9	Different contact data	62,294	281,560	47,624	375,819
10	Different number of vacancies	62,162	280,642	47,516	374,572
	Sum of duplicates	1,157,701	6,927,371	739,191	8,840,011

Further issues are connected with the work place. The employers repeatedly struggle with finding the right person for a free vacancy and there is additionally ”the phenomenon of continuous job offers”, which caused that the announces on the website were often prolonged and ”actualised”. Therefore the validity date, the start date and the end date were postponed many times. That gets us to make a decision of remaining only those job offers which had the latest possible date. From all of the events the mostly common was delaying the validity date that ranks with 80.57%. The companies sometimes changed the level of salary from unknown reasons, they might increase it to encourage candidates to submit an application or because of inflation. This value is not relevant for a subsequent survey, so it did not matter which job offer did we kept. Next step was to combined recent method and eliminate as many duplicates as we could.

The final two approaches were the most difficult as they required not only removing job offer, but also summing free vacancies from it with regard to the distinction between free vacant and free vacant for disabled people. In situation of appearing two ”different contact data”, as phone number and recruiter's name, we assumed that job offers could come from the same company, but from at least two independent departments. That would means the job offers are equally important. In consequence, we took action to sum the number of vacancies within the identical hash, and reduce unnecessary, single offers. Similarly in a case of ”different number of vacancies”, knowing the dates were without change, these job offers had to be for distinct departments. That is why we acted analogous to previous situation.

We ended deduplication at this point, as we were unable to find other ways to intelligently remove replicas and decided that our data was good enough to work on it.

3.1.3 Difficulties encountered

The aim of our data cleaning was to prepare the data for further analysis. We knew that even with our intensified effort the data would not be perfect. There were some missing information, unclear phrases and inconsistent data. Part of them was unaccounted for us, so we were forced to make several assumptions. During the deduplication itself, we faced a lot of complications which hindered complete harmonisation of data.

The table 3.2 presents information about the number of job offers before taking any steps, number of job offers after the deduplication methods earlier and our theoretically target number of job offers, which was generated based on the number of unique hash values. As ex-

pected, we did not achieve our target value. The theoretically surplus of job offers makes up 9% of current number of job offer. Nonetheless this value is enough, especially knowing that practically in surplus there are a great deal of valid job offers such as the case with different contact data showed.

Table 3.2. Comparison of number of job offers after deduplication and theoretically target

	2020	2021	2022	all
After deduplication value	62.162	280.642	47.516	374.572
Unique number of hash ids	58.449	257.462	43.725	342.511

We were able to distinguish some individual and more popular problems, which was too difficult for us to resolve. For example, one job offer with the identical hash value was represented by two records, both with work place in town named Chełm. However, one job offer took place in Chełm in voivodeship ”lubelskie”, and another one in voivodeship ”dolnośląskie”. Furthermore, it was not the only one job offer with incoherence in location. Under one hash value, there were offers with a precisely defined work place and offers that only display the town name.

Cleaning CBOP required Natural Language Processing as the occupation names were sometimes slightly different, for example, because of distinction in person conjugation (one offer: ”główna/y księgowa/y”, another: ”główny księgowy”). Here and there, the company's name was written as abbreviation or in full (one offer: ”Top Design Studio sp. z o.o.”, another: ”Top Design Studio spółka z ograniczoną odpowiedzialnością”). Similar examples were noticeable in columns related to salary, job description or responsibilities.

The most surprising inaccuracy concern cases, where to one unique hash value is assigned several job offers and additionally, each of them has a distinct job title or occupational code. It was quite common incident. For instance, we have job offer for pedagogue, psychologist for 20h and psychologist for 11h, and all of them share the same hash value. We do not know the real reason of this situation, however it could be induced by human mistake, actualisation or system fault.

Analysing the whole data cleaning, it is visible that it was essential to prepare the data for further work. Without deduplication there would still be more than 8 mln redundant job offers. The data could include instances that were unprocessed and we did not want to consider. Moreover, we had a chance to distinguish some anomalies found in the data.

3.2 Audit sample results

3.2.1 Design of the sample

In order to have any possibility to work on the data and draw conclusions, the quality of the data needed to be examined using `rcran` (R Core Team 2021), `tidyverse` (Wickham, Averick, et al. 2019), `lubridate` (Grolemund i Wickham 2011), `readxl` (Wickham i Bryan 2019). Therefore the samples, which consist of 510 job offers from CBOP, have been drawn by lot and saved in xlsx format.

The survey conducted on data from CBOP began on 13 August 2021, which was the day when the sample was drawn. Data collection process, i.e. interviews lasted 15 days. Considering the fact that it only could be held on working days, it took 11 days to collect all the information from employers and .

The study used contact information provided in the advertisements. In the beginning, we introduced ourselves as students of the University of Economics and Business in Poznań, who need some information for the study and then we moved on to questions. The conversation was held according to Figure 3.1. The purpose was to check whether the job offer is valid. In the case of an invalid job offer, the key knowledge was since when its status has changed and why. However, if the job offer had been valid, in case of the study conducted on the data from CBOP the current number of free vacancies would have been compared to the one appearing on the website.

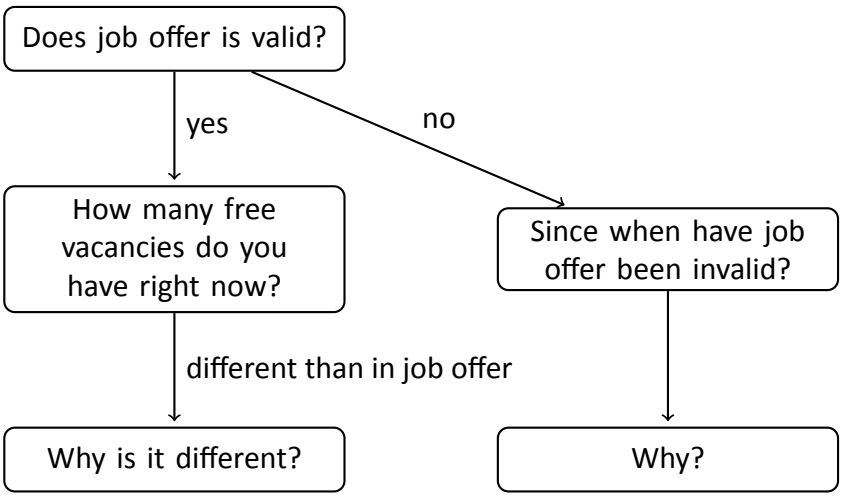

Figure 3.1. Diagram of conducted conversations

3.2.2 Results

Examining data from CBOP we managed to reach 465 employers and despite many trials the remaining 45 did not answer. The reasons for unanswered calls were diverse, for example, there was no signal, the number was incorrect, the answering machine and voicemail responded, or the number was unavailable. Because of bias towards the study, some of the recruiters had unfriendly attitudes:

- they were reluctant to give any information,
- the angst of disclosing confidential data and mistrust were visible,
- the fear of punishment for not updating information on website was critical point,
- they showed a lack of time and willingness to help,
- the cause of the study made no sense for them, although it was explained to them multiple times.

These interviewees required extra time and devotion. Therefore 5 of the phone call ended without receiving any data and 2 more with receiving only partial data. Fortunately, a considerable group of employers were very helpful. They shared more than necessary information, which resulted in creating a broader picture of the case.

After the analysis of the received data, the validity of the job offer could be determined. It turned out that 388 job offers (84.35%) were valid. We consider a valid job offer as an offer that is still published on a website, and until now, there is a demand for a new employee in the workplace. Another 65 job offers (14.13%) were found to be invalid. The median time of becoming inactive to the phone call where the inadequacy was detected is 5 days. To our surprise, 7 job offers (1.52%) came out as incorrect. The interviewees said they did not need anyone for this position, and they did not recognise the given offer because they had not reported it.

Taking into account solely the valid job offers, we wanted to compare the number of free vacancies. Based on the job offers on the website, this number makes a total of 1799 vacancies, but in reality, according to conveyed information, it amounts to 1165 vacancies. Hence, the difference is 634 free vacancies. The lack of knowledge of the number of vacancies from the remaining 130 job offers needs to be taken into consideration while drawing conclusions.

During the conversations, we identified 320 job ads to represent the correct number of vacancies and 60 job ads had different number of vacancies (still grater than 0). The recruiters explained that they had already hired some workers, but they still needed ones. In other cases,

some employees did not prove themselves, so they did not work anymore, and the demand for workers raised. We could not collect information about the remaining 8 offers because of a lack of cooperation or knowledge about factual data in the workplace.

Due to friendly interviewees from the phone conversations related to CBOP, we could distinguish "the phenomenon of continuous job offers". It only involves companies where the demand for new employees changes at a dynamic rate. There occurs continuous employee turnover, once they need 3 workers, another time as many as 15. Therefore, their job offer is repeatedly prolonged. They decided to constantly collect job application forms in case of a sudden rise in demand. In consequence, it is almost impossible for them to declare the correct number and modify it au courant. However, this situation adds to the difficulty of making a correct estimate of the level of vacancies.

Table 3.3 illustrates how many job offers are how many free vacancies there were at the time of the interview and according to CBOP. The most common number is one free job vacancy, to which 201 job offers are allocated in real terms and to which 304 job offers are assigned based on website. There is one offer, which declared to need 300 employees, but does not need so many people anymore.

Further, this data was used to model probability that the ad is inactive using logistic regression. The process is described in the next section.

It was crucial to maintain the unity by simply changing the voivodeships' names from polish letters to Latin. Being aware of the importance of some information, we created new variables, such as the occupational group and the time interval of the visibility of the job offer.

Moreover, we realised that having knowledge about employers could impact on a validity also. That's why we download the data containing among others the National Business Registry Number, the class of the company, the Gross Domestic Product sector and the actualisation date. We have ensured that the specific company's data is as up-to-date and got rid of possible older duplicates.

At the end, the sample was connected to the file with employer's information and all job offers with missing values were eliminated.

Table 3.3. Number of job advertisements by the number of vacancies and source of information (admin or sample survey)

# vacancies	CBOP	Sample survey	Proportion
1	304	201	0.66
2	74	68	0.92
3	28	27	0.96
4	15	18	1.20
5	29	20	0.69
6	6	3	0.50
7	0	2	–
8	2	2	1.00
9	1	0	0.00
10	30	24	0.80
12	3	3	1.00
15	6	7	1.17
16	1	1	1.00
20	5	2	0.40
25	1	1	1.00
30	3	0	0.00
40	1	1	1.00
300	1	0	0.00
Sum	510	380	

3.3 Logistic regression results

3.3.1 Structure of the model

Our explanatory variable is `valid_boolean` variable that testifies about job offer's validity. Because of its dichotomous, we chose to apply Binomial distribution modelled via logistic regression. We selected several independent variables, namely:

- the occupational group – consists of the first number of six-digit occupational code, it corresponds sequentially:

  1 – "Public officials, officers, managers",

  2 – "Professionals – high positions",

  3 – "Technicians and associate professionals",

  4 – "Office workers",

  5 – "Service & sales workers",

6 – ”Farmers, gardeners, fishermem”,

7 – ”Craft and related trade workers”,

8 – ”Plant and machine operators and assemblers”,

9 – ”Employees performing simple work”.

- voivodeship (pol. województwo) – where the place of work is located. Due to not big enough sample, we could not be more detailed and use cities or towns.
- the logarithmised number of vacancies available – we used log transformation because this variable is highly skewed.
- the time interval – the range of visibility of the job offer on the website expressed in days. We calculated that as differential between the date of acceptance the entry form and the date of expiry of the job offer. We isolated five intervals, that is:

  < 25 days,

  $< 25; 50)$ days,

  $< 50; 75)$ days,

  $< 75; 100)$ days,

  $>= 100$ days.

- size of the company - represents the size of the company, where:

  M – the enterprise hiring maximally 9 people,

  S – the enterprise hiring between 10 and 49 people,

  D – the enterprise hiring at least 50 people.

- the sector - determines what type of workplace the enterprise is:

  1 – public,

  2 – private.

To generate a model, we used `glm()` function (R Core Team 2021). After preliminary analysis, the odd value in standard error of the occupational group number 6 was noticed, which totals more than a thousand. In our sample only 3 job offers concern this occupational group, which resulted in anomaly. Therefore we decided to completely remove job offers connected to this group. Then we could proceed to further estimators analysis.

3.3.2 Estimated model parameters

The generated model needed to be assess to gain confidence in its right fit to our case. We began our evaluation by studying variety of factors and results.

The table 3.4 presents the estimated parameters from logistic regression along with their standard errors. The reference group is occupational Group1, `visibilityTimeInterval>25`, `voivodeship dolnoslaskie`, `class D` and `sek 1`. 60% of the variables cause a decrease in the size of the probability of a valid job offer and another 40% cause an increase. Unfortunately, not all variables are statistically significant according to the p-value 0.05 rule.

The usefulness of our model could be determined by analysing the null and residual deviance. There are responsible for estimation of the degree of fit of the response variable to the model with only an intercept term (null deviance) or with n endogenous variables (residual deviance). Based on theirs volume we can conducted the chi-square test and find out that our model is statistically significant with the p-value lower than 0.0005. From which we can conclude that our model predicts the explanatory variable really well .

We wanted to examine the influence of the variables on the validity of the job offer, so we used `Anova()` function from `car` package (Fox i Weisberg 2019). Based on the table 3.5 it looks like the location, in our case voivodeship, has the biggest impact on the probability of success, next ones are an occupational group and the kind of sector, to which employer belongs. All three variables are statistically significant, as well as logarithm of available vacancies. The size of the company (variable class) has the lowest influence.

The steps taken in building a model and the profound analysis made was a reason of the optimally good fit of the model to our data. The model describe the explanatory variable in a reliable way, which means we can continue the analysis.

Afterwards, there was a need for creating a new variable, the visibility time interval, based on the existing data. Additionally, we were forced to extract only job offers with no missing values and excluding the occupational group number 6, that represents the agricultural labour sector.

By using function `predict()` from `car` package (Fox i Weisberg 2019), we were able to calculate the probability of a validity for each job offer. Then we multiplied it with the number of vacancy to gain the estimated number of vacancy, which we could use to make the analysis and draw conclusions.

Table 3.4. The summary of the created model

Variable	Estimate	Std. Error
occupational Group 2	1.306	(0.871)
occupational Group 3	2.161**	(1.002)
occupational Group 4	0.722	(0.901)
occupational Group 5	2.289**	(0.965)
occupational Group 7	2.091**	(0.922)
occupational Group 8	1.241	(0.913)
occupational Group 9	0.852	(0.882)
visibilityTimeInterval <25; 50)	−0.284	(0.521)
visibilityTimeInterval <50; 75)	−0.463	(0.631)
visibilityTimeInterval <75; 100)	1.220	(0.931)
visibilityTimeInterval >=100	−0.425	(0.734)
voivodeship kujawsko-pomorskie	−2.104***	(0.788)
voivodeship lodzkie	−1.031	(0.848)
voivodeship lubelskie	−2.004**	(0.897)
voivodeship lubuskie	−0.185	(0.916)
voivodeship malopolskie	−1.064	(0.940)
voivodeship mazowieckie	−0.745	(0.826)
voivodeship opolskie	−1.126	(0.938)
voivodeship podkarpackie	−1.892**	(0.835)
voivodeship podlaskie	14.128	(650.890)
voivodeship pomorskie	−0.798	(0.845)
voivodeship slaskie	−0.247	(0.843)
voivodeship swietokrzyskie	−0.762	(1.306)
voivodeship warminsko-mazurskie	−1.263	(0.870)
voivodeship wielkopolskie	−1.422*	(0.858)
voivodeship zachodniopomorskie	1.648	(1.239)
class M	−0.865	(0.755)
class S	−0.418	(0.434)
sek 2	1.443****	(0.423)
log(vacanciesAvailable)	0.478*	(0.249)
Constant	0.342	(1.125)
Observations	450	
Log Likelihood	−152.140	
Akaike Inf. Crit.	366.281	
Note:	****p<0.001 ***p<0.01 **p<0.05 *p<0.1	

3.4 Over-coverage error

3.4.1 Tabular results

One might think that raw data collected from CBOP would be enough to draw reasonable conclusions. However, as early analysis has shown, we need to face a problem with over-

Table 3.5. Anova type 2 results – a glimpse of variable importance

Response: valid_boolean			
	LR Chisq	Df	Pr(>Chisq)
occupationalGroup	15.541	7	0.0296554 *
visibilityTimeInterval	5.596	4	0.2314189
voivodeship	33.552	15	0.0039338 **
class	1.903	2	0.3861739
sek	12.006	1	0.0005303 ***
log(vacanciesAvailable)	4.128	1	0.0421696 *
Note:	***p<0.001 **p<0.01 *p<0.05		

coverage errors in many views, for example duplication, invalidity etc. The table 3.6 presents the number of free vacancies per year at the turn of the end of the 2020 and the beginning of the 2022 based on the notation from equation 2.1.

The data deduplication, which ranks on first place, was the biggest challenge, because of elimination difficulties. Nonetheless, the issue with invalidity was the least expected one and it occupies the third position. Other coverage error also caused difficulties resulting in distorted results.

Table 3.6. The comparison of volume of the decomposed over-coverage errors expressed in the number of free vacancies

	2020	2021	2022	all
n_1	1,479,494	9,105,277	1,200,295	11,785,066
n_2	2,731,198	15,728,399	1,748,722	20,245,037
n_3	28,879	137,880	26,277	185,555
sum	4,239,571	24,971,556	2,975,294	32,215,658

3.4.2 Detailed results regarding over-coverage due to out-date ads

Generally, the validity survey is not conducted and people analyse only appropriately cleaned data without applying models. In our case, we went a step further and we addressed this topic. Our aim was to compare the demand for labour observed in CBOP with correction based on logistic regression. In order to visualise results, we used diverse packages, such as `tidyverse` (Wickham, Averick, et al. 2019), `lubridate` (Grolemund i Wickham 2011) and function `melt()` from `reshape2` (Wickham 2007).

At the beginning, we focused on the whole labour demand in Poland and the comparison of the observed and estimated value is represented on the figure 3.2. As it's visible, the esti-

mated volume is significantly lower. It makes up approximately 90% of the total. As the year comes to an end the number of vacancies decreases and at the very start of the year this number increases. It reaches its peak equals to more than 125 thousands free vacancies in the half of the year, which is during summer months.

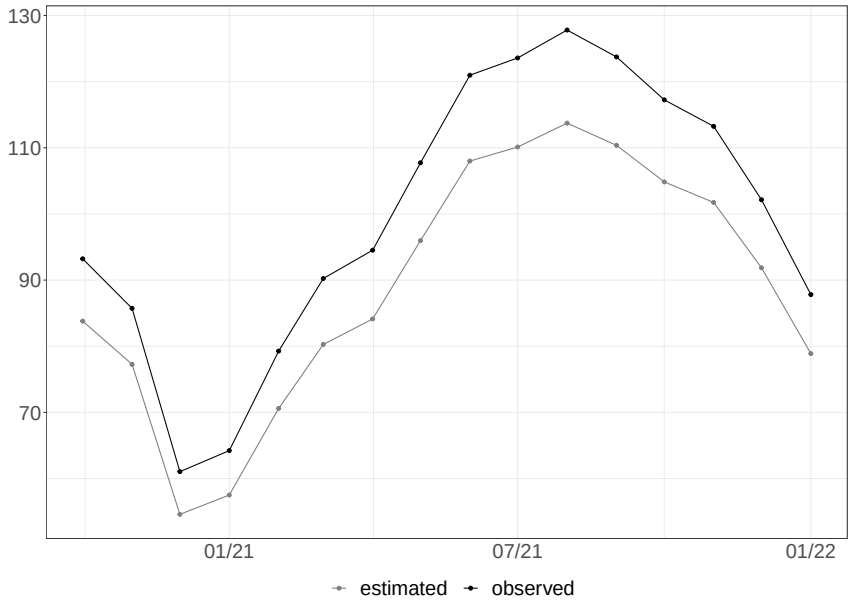

Figure 3.2. Number of vacancies in thousands per month over the period 01/10/2020-31/01/2022

To each job offer is assigned an occupational group to indicate which work sector it is targeted at. We wanted to examined how the structure of labour demand looks like divided into occupational groups, so we created the figure 3.3. The biggest difference in estimated and observed values is manifested by simple work employees and the smallest ones by farmer, public sector's representatives and technicians. However, it is caused because of small amount of reported job offers related to the occupational group no. 1 and no. 6. Moreover, there is no visible correlation between the level of competence and the timeliness of offers.

An important factor of taking up work is the workplace. Therefore, we grouped our data according to the voivodeship in which the work takes place. On the figure 3.4 there are observed and estimated number of vacancies in distinction between all 16 voivodeships. Base on the study conducted by Statistics Poland (2021), we could draw conclusions that the voivodeships with the lowest level of population tend to have a observed labour demand almost equal to the estimated one, these are voivodeship dolnośląskie, lubuskie, opolskie, podlaskie, świętokrzyskie and zachodnipormorskie. The interesting fact is that most of these voivodeship

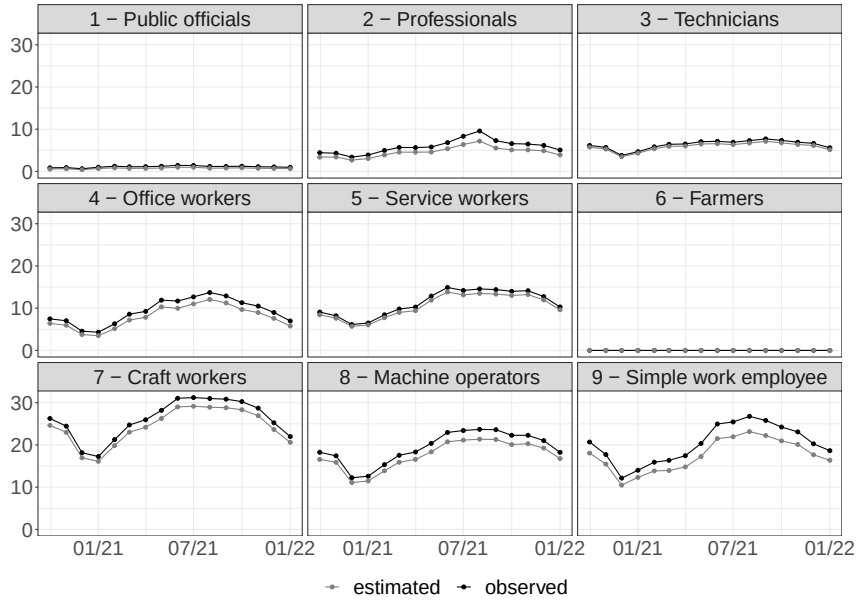

Figure 3.3. Number of vacancies in thousands by specific occupational groups over the period 01/10/2020 - 31/01/2022

are located at the west border. More than 20% of difference declare voivodeship lubelskie, kujawsko-pomorskie and podkarpackie.

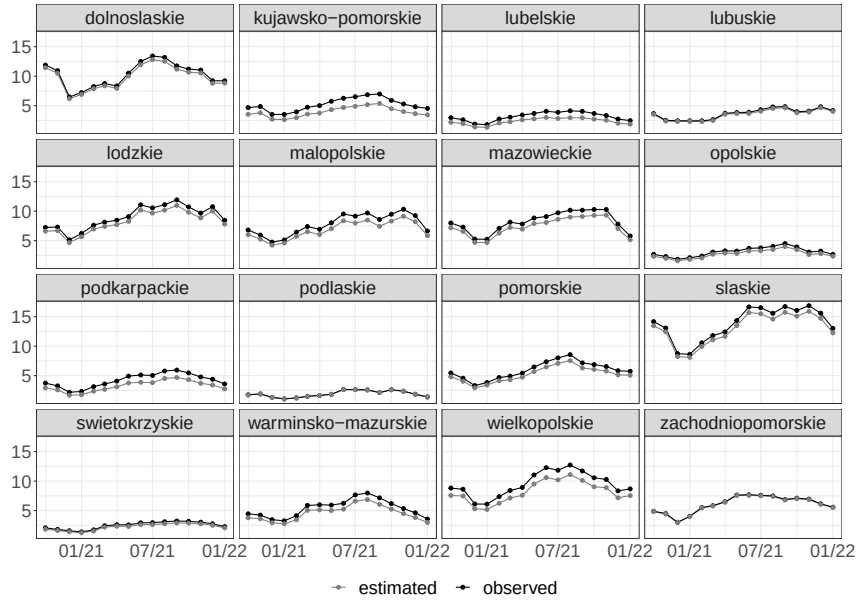

Figure 3.4. Number of vacancies in thousands by voivodeships over the period 01/10/2020 - 31/01/2022

The size of the gap between estimated and observed values is influenced not only by the conditions related to the job offer itself, but also by the employer. The figure 3.5 focuses on the class of the hiring company which determine its scope. It is visible that similar dependencies

are found in small and average enterprises, which means that all the companies engaging to 49 people display averagely 16% difference. However, the corporation employing more than 49 people seems to have a significantly lower volume of difference between the estimated and observed value of labour demand even though there are much more job offers on the website from them. The average size of the gap makes up 8% of the observed number of vacancies, which is twice less than in the previously mentioned classes.

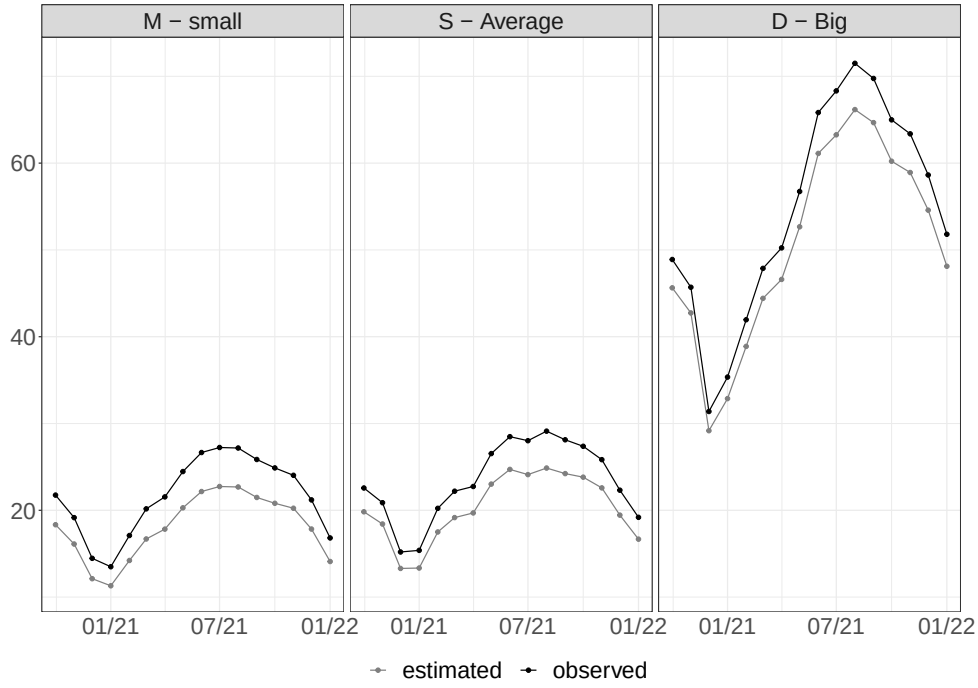

Figure 3.5. Number of vacancies in thousands by class type over the period 01/10/2020 - 31/01/2022

As we have elaborated, the influence of over-coverage error n_3 could be spread into some categories, as a workplace, an occupational group or a size of a company. Unfortunately, it must be reckoned that not all of the job offers visible on the website are valid. The figures have shown that the difference between estimated and observed volumes is significant in a many cases.

3.5 Summary

The aim of our analysis was to estimate the over-coverage error occurring in non-statistical source, such as CBOP. Therefore, we needed to appropriately prepare the given data, which revealed the problem of the data duplication. We examine the data quality by studying the

draw by lot sample. Then we built a model which allows us to estimate the estimated level of labour demand and compare it to be observed one. Studying the results we can see that the difference between the estimated and observed value is significant.

Conclusions

Many units to appropriately measure the level of labour demand decide to use statistical data sources, such as census conducted by the Statistical Office. Another way is to operate on non-statistical data sources, for example the Central Job Offer Database (CBOP). However, that is strictly connected with the risk of occurring non-random errors. Therefore, we wanted to assess one of the non-random error, which is over-coverage error, appearing in data from CBOP. This may allow correct inferences in the future about labour demand in Poland.

In our dataset we distinguished three types of over-coverage errors, these are errors made by inconsistency with definition, data duplication and job offer's invalidity despite of its visibility on the website. To manage the first two errors our actions were to minimise the number of wrong job offers by filtering data, profoundly analysing it and then removing the questionable ones. The over-coverage error including invalidity required preparing the logistic regression model and estimating the predicted number of free vacancies.

The whole analysis considering multiple aspects of survey were conducted by using R programming language for statistical computing and graphics and its code is available on GitHub.

Beginning our analysis with working on admin data, we were able to precisely specify the units we want to consider as meeting the definitions. The accepted offers needed to be an active job offer held in Poland and relate to determined contract type. After first selection, we noticed how significant was the first over-coverage error.

Next challenge involve removing all duplicates from data. As a reason of daily downloading data, we had a lot of unnecessary data, which made up 94% of the total. Unfortunately, in the remaining data, the duplicates could still be distinguished. It was caused by for example spelling mistakes, constant actualisation or differences in contact data. The most common one was delaying the validity date. The data after deduplication constitutes approximately 4% of the base value.

The estimation of the last over-coverage error connected to job offers' invalidity needed to be preceded by quality survey on the CBOP sample. Therefore, the information about validity was gathered through phone calls conversations. It turned out that 84.35% job offers were valid, 14.13% were invalid and 1.52% were incorrect. The interviews enable also verification of the posted number of free vacancies. Subsequently we build a model that estimated the probability of job offer's validity. We depended it from the occupational group, the province, the size and the sector of the company, the time of offer's visibility and the logarithmised number of vacancies available. The increase in value of probability is caused by 60% of used variables. According to the null and residual deviance the generated model is useful and Anova attests about the high variables importance. In consequence, we counted the estimated number of free vacancies.

In general the over-coverage error could be counted in millions. The most input has a data duplication error, because it equals to 2/3 of whole. Nonetheless, the detailed definition and the consideration of the validity of offers also allowed a more accurate estimation of labour demand. The over-coverage error itself would add improperly more than 30 million free vacancies to our reflections, which would have substantial impact on inferences.

The differences between the estimated and observed number of free vacancies were easily visible on presented figures. The whole labour demand might be overestimated by approximately 10%. The job offers considering the disposition of hiring a simple work employee tend to have disturbance in the theoretical and observed value more often than other occupational groups. The location of the workplace has the biggest influence on the job offers' validity. It turned out, that the level of population in the voivodeship is positive correlated with the level of difference between the estimated and observed value. The companies hiring more than 49 people also exhibit lower volume of difference.

According to our multilevel survey we can deduce that over-coverage errors are a real problem we needed to take into account during estimation of the labour demand. Meeting the determined criteria, data uniqueness and validity are crucial to appropriately count the number of free vacancies. Therefore, we hope our conducted survey would have an actual impact on the future estimation of labour demand.

Bibliography

Abraham, K. G., & Wachter, M. (1987). Help-wanted advertising, job vacancies, and unemployment. *Brookings papers on economic activity*, *1987*(1), 207–248.

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Alvarez-Santiago, S. A., García-Oliva, F., & Varela, L. (1996). Analysis of vesicular-arbuscular mycorrhizal colonization data with a logistic regression model. *Mycorrhiza*, *6*(3), 197–200.

Bethlehem, J., & Biffignandi, S. (2021). *Handbook of web surveys*. John Wiley & Sons.

Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., Pasi, G., & Viviani, M. (2018). WoLMIS: a labor market intelligence system for classifying web job vacancies. *Journal of Intelligent Information Systems*, *51*(3), 477–502.

Bossler, M., Gürtzgen, N., Kubis, A., Küfner, B., & Lochner, B. (2020). The IAB Job Vacancy Survey: design and research potential. *Journal for Labour Market Research*, *54*(1), 1–12.

Denk, C. E., & Finkel, S. E. (1992). The aggregate impact of explanatory variables in logit and linear probability models. *American Journal of Political Science*, 785–804.

Dowle, M., & Srinivasan, A. (2021). *data.table: Extension of 'data.frame'* [R package version 1.14.0]. https://CRAN.R-project.org/package=data.table

Eddelbuettel, D., & Knapp, B. (2021). *RcppSimdJson: 'Rcpp' Bindings for the 'simdjson' Header-Only Library for 'JSON' Parsing* [R package version 0.1.5]. https://CRAN.R-project.org/package=RcppSimdJson

Eurostat. (2022). *Job vacancies between 2019 and 2022*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Job_vacancy_statistics&oldid=558537#Job_vacancies_between_2019_and_2022

Firke, S. (2021). *janitor: Simple Tools for Examining and Cleaning Dirty Data* [R package version 2.1.0]. https://CRAN.R-project.org/package=janitor

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gagolewski, M. (2021). *stringi: Fast and portable character string processing in R* [R package version 1.6.2]. https://stringi.gagolewski.com/

Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. https://www.jstatsoft.org/v40/i03/

International Labour Organization. (2015). National Employment Policies: A Guide for Workers' Organisations.

Lee, S. (2005). Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data. *International Journal of remote sensing*, *26*(7), 1477–1491.

MegaPanel, P. (2021). *Gowork.pl przed pracą na OLX i Pracuj.pl, najmniej zyskał Linkedin*. Retrieved March 1, 2021, from https://www.wirtualnemedia.pl/artykul/gowork-pl-przed-praca-na-olx-i-pracuj-pl-najmniej-zyskal-linkedin-top10

Nagel, K. (2015). Relationships between unemployment and economic growth–the review (results) of the theoretical and empirical research. *Journal of Economics & Management*, *20*, 64–79.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

SEJM. (1995). Ustawa z dnia 29 czerwca 1995 r.o statystyce publicznej [art.30 ust.1 pkt.3].

SEJM. (2004). Ustawa z dnia 20 kwietnia 2004 r. o promocji zatrudnienia i instytucjach rynku pracy [https://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU20040991001/T/D20041001L.pdf].

Statistics Poland. (2021). *Powierzchnia i ludność w przekroju terytorialnym w 2021 roku*. https://stat.gov.pl/obszary-tematyczne/ludnosc/ludnosc/powierzchnia-i-ludnosc-w-przekroju-terytorialnym-w-2021-roku,7,18.html

Statistics Poland. (2022a). *Demand for labour*. https://stat.gov.pl/en/metainformation/glossary/terms-used-in-official-statistics/3011,term.html

Statistics Poland. (2022b). *Job offer*. https://stat.gov.pl/en/metainformation/glossary/terms-used-in-official-statistics/2960,term.html

Statistics Poland. (2022c). *Methodological report The demand for labour*. Retrieved June 21, 2022, from https://stat.gov.pl/obszary-tematyczne/rynek-pracy/popyt-na-prace/zeszyt-metodologiczny-popyt-na-prace,3,1.html

Szreder, M. (2015). Wiadomości Statystyczne Nr 1 - Styczeń 2015 r. [https://stat.gov.pl/files/gfx/portalinformacyjny/pl/defaultaktualnosci/5982/7/8/1/wiadomosci_statystyczne_01_2015.pdf].

Wiącek, A. et al. (2011). Ocena porównawcza wybranych polskich portali rekrutacyjnych z perspektywy kandydata, pracodawcy i kierownika. *Informatyka Ekonomiczna*, (22), 337–345.

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, *21*(12), 1–20. http://www.jstatsoft.org/v21/i12/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J. et al. (2019). Welcome to the Tidyverse. *Journal of open source software*, *4*(43), 1686.

Wickham, H., & Bryan, J. (2019). *readxl: Read Excel Files* [R package version 1.3.1]. https://CRAN.R-project.org/package=readxl

Wickham, H., & Seidel, D. (2020). *scales: Scale Functions for Visualization* [R package version 1.1.1]. https://CRAN.R-project.org/package=scales

Wieczorek, G. (2011). Internet jako narzędzie poszukiwania i doboru personelu. *Prace Naukowe Akademii im. Jana Długosza w Częstochowie. Edukacja Techniczna i Informatyczna*, *6*, 147–162.

List of Tables

List of Figures