COMPLEMENTING DATA GAPS ON WAGES IN THE LABOUR FORCE SURVEY DATA SET: EVIDENCE FROM POLAND

Wojciech Grabowski[1], Karol Korczak[2]

[1] *University of Łódź, Faculty of Economics and Sociology, Department of Econometric Models and Forecasts (Institute of Econometrics), Poland, ORCID: 0000-0002-6707-3736, wojciech.grabowski@uni.lodz.pl;*

[2] *University of Łódź, Faculty of Economics and Sociology, Department of Computer Science in Economics (Institute of Applied Economics and Informatics), Poland, ORCID: 0000-0003-1936-1423, karol.korczak@uni.lodz.pl.*

Abstract: *Due to the low level of quality of the Labour Force Survey (LFS) data set, studies devoted to matching the LFS data with data from alternative sources are frequent. In this paper, we propose a novel method of complementing data gaps on wages in the Labour Force Survey data set. The method is based on estimataing the parameters of the multilevel model explaining wages on the basis of the Structure of Earnings Survey (SES) data set. In such a way, we identify the impact of individual characteristics and enterprise-level features on wages. We also find evidence of random differences between the wages of workers from different professional groups. The relative importance of consecutive groups of variables is evaluated on the basis of the estimates of the parameters of the full model and reduced models. The results of the estimation of the parameters are in line with expectations. The estimates of parameters and predictions of random effects are used in order to calculate the theoretical wages of individuals who do not report wages in the Labour Force Survey. When the predicted wages are compared with the observed ones, some discrepancies are observed. Rationales for these discrepancies are provided. Therefore, the use of a correction factor is proposed. Correction factors are provided for different features of workers and different features of enterprises. The use of the microeconometric multilevel model, as well as the correction factor, leads to reasonable wage estimates of workers not reporting them in the Labour Force Survey. The proposed method may be used in order to complement data gaps on wages for other EU countries.*

Keywords: *LFS, SES, microeconometrics, mixed-effects model, data gaps.*

JEL Classification: *C50, J01, C21.*

APA Style Citation: Grabowski, W., & Korczak, K. (2020). Complementing Data Gaps on Wages in the Labour Force Survey Data Set: Evidence from Poland. *E&M Economics and Management, 23*(3), 4–22. https://doi.org/10.15240/tul/001/2020-3-001

Introduction

The European Union Labour Force Survey (EU LFS) is a widely used source of information on the participation in the labour force of citizens from the countries of the European Union. The LFS data set contains quarterly collected, anonymized data on individuals representing various industries and occupations. The data are collected using common classifications, concepts and definitions. In each country, the same set of characteristics is collected.

Despite common standards for data collection, the use of the unified LFS data set may sometimes encounter various difficulties. First of all, from the very beginning of the LFS, there have been numerous methodological changes in sampling (Kerr & Wittenberg, 2015), definitions and classifications (European Commission, 2018). These changes make it difficult to compare data with previous years. Secondly, different cross-national classification rules may produce various problems. Cross-national

differences in coding data (regarding parental leave beneficiaries) can be a source of bias in international comparisons (Mikucka & Valentova, 2013). Another problem which limits the use of the LFS data is related to the data gaps. Research studies devoted to the analysis of the LFS data point out the problem of non-response (see, e.g., Pastore & Socha, 2006; ADB, 2012). The problem of non-response may introduce sample-selection bias when characteristics for the population are analysed. In some countries, in the presence of a non-response, the designated households that cannot be interviewed are replaced with other households in the vicinity. However, such a practice may lead to the disadvantages far outweighing the benefits (Vehovar, 1999).

The low quality of the data is reflected when it is compared with data from other sources (ADB, 2012). The lack of coherence of data from various sources is likely to surprise the user when faced with different figures referring to similar concepts. Therefore, studies devoted to matching the LFS data with data from alternative sources are commonly used (see, e.g., Ormerod & Ritchie, 2007; O'Mahony & Timmer, 2009; Orche Galindo & Bueno Maroto, 2011). Moreover, researchers also propose alternative measurement methods, which prove to be more accurate compared with those based on LFS estimates (see, e.g., Skinner et al., 2002).

In the LFS data set for Poland, significant data gaps occur, especially in relation to wages. Our analysis shows that in the LFS data set for people employed in Poland in the years 2010–2016, data gaps regarding wages varied from 47 to 71%. This is due to the fact that the purpose of the LFS is to describe the situation on the labour market in Poland (economic activity of the population, characteristic of employed and unemployed persons, economically inactive population), and all other data, including data on wages, are treated as additional (see, e.g., Central Statistical Office, 2018). In general, citizens are not willing to provide information concerning their wages. They willingly provide information concerning their level of education, type of education, marital status, etc, while fields with information about wages are often not filled in. Therefore, in Poland (as in other countries), the LFS data set is not treated as a reliable source of data regarding wages (see, e.g., Ormerod & Ritchie, 2007; Pastore & Socha, 2006; ADB, 2012).

The main goal of our research study is to propose a novel method for complementing the LFS data on wages. It is based on the use of estimates of parameters of a microeconometric model, estimated on the basis of the European Union Structure of Earnings Survey (SES) data set. The SES data set is a source of complete information on individuals' wages. For all individuals, reliable information about wages is provided. The reliability of this source of data is due to the fact that all information concerning wages is provided by employers. Therefore, the problem of individuals being ashamed of their wages is avoided. Moreover, the paper contributes to the economic literature in several other ways. A secondary aim of the study is to analyse the factors determining wage levels. We propose the original specification of the econometric model, which takes into account random differences among the wages of workers from different occupational groups. Since workers are classified according to 1-digit, 2-digit, 3-digit and 4-digit codes, the use of the multilevel model is justified. Therefore, our specification extends specifications from other studies devoted to the determinants of wages (see, e.g., Skinner et al., 2002; Majchrowska & Strawiński, 2018). The use of data covering the period 2010–2016 allows us to point out the variability of the results of the estimation for different years. Finally, we compare the characteristics of empirical wages of individuals who declare them with the theoretical wages for workers who do not. We learn about the reasons for not declaring, overestimating and underestimating wages.

Despite the fact that the SES data set is complete, there are still good reasons to complement the LFS data set. The SES data set is narrower as it includes only data on wages, professions and basic characteristics of employment. The LFS data set, in turn, gathers additional information concerning, participation in training courses, doing extra work, commuting to work etc. Therefore, completing the LFS data set appears to be useful for purposes of conducting more in-depth analyses of labour market.

The paper has the following structure. In the first section, a literature review on the microeconometric determinants of wages is provided. In the second section, the data preparation and research method are described. The results and their discussion are

provided in section 3. The final section offers concluding remarks.

1. Literature Review on Microeconometric Determinants of Wages

In this paper, we propose a novel method for complementing data on wages when the Labour Force Survey does not provide such information. Since data from the Structure of Earnings Survey are used, the microeconometric model explaining wages should be specified. Therefore, the parameters of the econometric model are estimated in this paper. This specification is based on similar studies devoted to the determinants of wages.

Firstly, the analysis of individual wages should be based on the classical theoretical framework proposed by Mincer (1974) and Mincer and Polachek (1974). These authors argued that individual wages grow as the number of years of schooling and the level of experience increases. As a result, variables associated with experience (general experience, experience in an enterprise a worker had worked in before the poll was conducted) are included as explanatory variables. Since the rate of return of different types of education may not be constant, binary variables associated with the level of education are taken into account (Psacharopoulos & Ng, 1994).

In order to take into account the problem of the gender wage gap, a variable associated with gender should be included (see, e.g., Blau & Kahn, 2007; Peter & Drobnič, 2013; Ahn & Sanchez-Macros, 2017; Gallen et al., 2018; Hara, 2018; Artz & Taengnoi, 2018). Moreover, inter-industry disparities in wages make the inclusion of sectoral dummies reasonable (Gannon et al., 2007). Wages in the private and public sectors should differ (Melly, 2005), the size of an enterprise should have a positive impact on wages (Zimermanová, 2010), and workers employed for an indefinite period are expected to earn more than employees with other types of contracts. Therefore appropriate dummy variables are included in the specification.

According to the New Economic Geography approach, there is a correlation between wages and the economic potential of the region (agglomeration patterns, market access, migration processes) (see, e.g., Cieślik & Rokicki, 2016). The results of other studies indicate that spatial differentiation of wages is an especially important problem in emerging and transition economies (Majchrowska & Strawiński, 2016). Therefore, dummy variables associated with regions are taken into account. Though variables associated with marital status and town size are also included as regressors in the microeconometric model of wages (see, e.g., Ahituv & Lerman, 2007; Grabowski, 2019), unfortunately, the SES by Occupations data does not enable identification of these categories. Therefore these variables are not included in the specification.

Labour markets in developed but also in developing countries have been experiencing significant multidimensional changes in recent years, reflected in the occupational-wages structure affecting the distribution of income between various groups of employees. These processes have been examined within the framework of the theory of Skill-Biased Technical Change (SBTC) (see, e.g., Freeman & Katz, 1994; Katz & Autor, 1999) and the concept of Routinization-Biased Technical Change (RBTC), leading to labour market polarization (Autor et al., 2003). According to this first concept, changes in demand for workers with different skill levels result in an increase in the gap between the wages of skilled and unskilled employees. The RBTC hypothesis is rooted in the model proposed by Autor et al. (2003), which assumes that different types of tasks are defined, and the wages of workers from the same task group depend on whether performed activities are routine or non-routine.

All in all, both theories (RBTC and SBTC) assume that wages strongly depend on occupations. Using data from the SES by Occupations, theoretically, workers belonging to the same 4-digit code occupational group should have similar wages, ceteris paribus. If large differences between wages of workers from the same group exist, workers with lower wages may quit and start working for enterprises offering higher wages. Workers belonging to the same 3-digit code occupational group should have similar wages, but within group variation, they may be larger than in the case of workers from the same 4-digit code occupational group. Changes of enterprises for workers from the same 3-digit code occupational group and different 4-digit code occupational groups are easy but more difficult than in the case of workers from the same 4-digit code occupational

group. Analogously, changes of enterprises for workers from the same 2-digit code occupational group and different 3-digit code occupational groups seem to be more difficult than in the case of workers from the same 3-digit code occupational group. Changes of enterprises for workers from the same 1-digit code occupational group are easier than in the case of workers from different 1-digit code occupational groups but more difficult than in the case of workers from the same 2-digit code occupational group. Therefore, the largest differences among wages should be observed for employees from different 1-digit code occupational groups.

2. Data and Method

2.1 Data Description

We use data on Polish individuals from the Labour Force Survey and the Structure of Earnings Survey. The LFS as well as the SES are conducted in the 28 Member States of the European Union, candidate countries and countries of the European Free Trade Association (EFTA). The LFS data set contains data reported from employees while the SES data set contains data reported from employers. The LFS data are published annually (quarterly data sets) while data from the SES are published at two-year intervals (annual data sets). The SES data set contains a complete set of data on individuals' wages. We use annual data from the SES and data (only for employed people) from the fourth quarter from the LFS. It should be mentioned that in the SES, wages from October are reported. Therefore, we take wages from the fourth quarter in the case of the LFS. Finally, our analysis covers the years 2010, 2012, 2014 and 2016. Tab. 1 presents the number of observations in our research process, including the number of missing data on wages in the LFS.

Tab. 1: Number of individual observations

Data set	2010	2012	2014	2016
The LFS	19,215 (9,097)	28,471 (20,197)	24,785 (14,408)	22,345 (13,824)
The SES	688,383	725,239	730,498	795,879

Source: own

Note: Numbers in brackets show missing data on wages.

A novel method of complementing the LFS data on wages has been developed on the basis of the advantage of the SES. Since there is no gap in data concerning wages in the case of the SES, this data set is used in order to identify the relationship between wages and the features of the workers and the enterprises they represent. Since data has a hierarchical structure, the parameters of the hierarchical model should be estimated. Among the variables affecting the level of wages, the following groups are distinguished: features of workers (gender, experience, education), locations of enterprises (all 16 regions in Poland), NACE sections (statistical classification of economic activities in the European Community) represented by enterprises, other features of enterprises (size, sector). All variables representing the mentioned groups are listed in Tab. 2.

2.2 Method

The procedure of complementing the individual LFS data on wages can be implemented in a few steps. In the first step, we define vector **SES**, which consists of all variables listed in Tab. 2. Apart from this vector, we define vector **LFS**, which consists of the same variables, but their values concern individuals analysed in the Labour Force Survey. For example, the variable *LFS_GENDER* has the same definition as the variable *SES_GENDER*, but its values are observed for individuals from the Labour Force Survey. It should be mentioned that for some variables, there were differences in the LFS and SES data structure. For example, in the SES, gross wage is reported by the employer, while in the LFS, net wage is reported by the employee. In addition, the education and size of the enterprise were divided by some other

Tab. 2: Independent variables used in the model

Variable	Definition of variable
SES_GENDER	Dummy variable taking a value of 1 for females and 0 for males
SES_EXP_GEN	Experience of worker (number of years) in all enterprises he/she had worked before the poll was conducted
SES_EDU[1–5]	Dummy variables taking a value of 1 for workers with: a basic level of education (*SES_EDU1*); vocational education (*SES_EDU2*); secondary general education (*SES_EDU3*); secondary technical education (*SES_EDU4*); tertiary education (*SES_EDU5*)
SES_IND	Dummy variable taking a value of 1 if a respondent is employed for an indefinite period and 0 otherwise
SES_LOC_[DLN-ZACH]	Dummy variables taking a value of 1 for workers from the region: Dolnośląskie (*SES_LOC_DLN*); Kujawsko-Pomorskie (*SES_LOC_KP*); Lubelskie (*SES_LOC_LUBE*); Lubuskie (*SES_LOC_LUBU*); Łódzkie (*SES_LOC_LODZ*); Małopolskie (*SES_LOC_MAL*); Mazowieckie (*SES_LOC_MAZ*); Opolskie (*SES_LOC_OPOL*); Podkarpackie (*SES_LOC_PODK*); Podlaskie (*SES_LOC_PODL*); Pomorskie (*SES_LOC_POM*); Śląskie (*SES_LOC_SL*); Świętokrzyskie (*SES_LOC_SW*); Warmińsko-Mazurskie (*SES_LOC_WM*); Wielkopolskie (*SES_LOC_WIEL*); Zachodniopomorskie (*SES_LOC_ZACH*)
SES_SECTION_[A-S]	Dummy variables taking a value of 1 in the case of enterprises from section: Agriculture, forestry and fishing (*SES_SECTION_A*); Mining and quarrying (*SES_SECTION_B*); Manufacturing (*SES_SECTION_C*); Electricity, gas, steam and air conditioning supply (*SES_SECTION_D*); Water supply; sewerage, waste management and remediation activities (*SES_SECTION_E*); Construction (*SES_SECTION_F*); Trade; repair of motor vehicles (*SES_SECTION_G*); Transportation and storage (*SES_SECTION_H*); Accommodation and catering (*SES_SECTION_I*); Information and communication (*SES_SECTION_J*); Financial and insurance activities (SES_SECTION_K); Real estate activities (*SES_SECTION_L*); Professional, scientific and technical activities (*SES_SECTION_M*); Administrative and support service activities (*SES_SECTION_N*); Public administration and defense; compulsory social security (*SES_SECTION_O*); Education (*SES_SECTION_P*); Human health and social work activities (*SES_SECTION_Q*); Arts, entertainment and recreation (*SES_SECTION_R*); Other service activities (*SES_SECTION_S*)
SES_SIZE_[1–4]	Dummy variables taking a value of 1 for individuals employed in enterprises with: 1–9 workers (microenterprises) (*SES_SIZE1*); 10–49 workers (*SES_SIZE2*); 50–249 workers (*SES_SIZE3*); more than 250 workers (*SES_SIZE4*)
SES_SECTOR	Dummy variable taking on a value of 1 for individuals employed in enterprises from the public sector and 0 for enterprises from the private sector

Source: own

categories. In order to compare the results, we use appropriate methods of transforming their values (see, e.g., Strawiński, 2015).

In the second step, we estimate the parameters of the following multilevel model (McCulloch, 1997) explaining individuals' wages:

$$log(wage_i) = \boldsymbol{SES}_i\boldsymbol{\beta}^{SES} + \sum_j z1_i^j\, u1^j + \sum_k z2_i^k\, u2^k + \sum_l z3_i^l\, u3^l + \sum_m z4_i^m\, u4^m + \varepsilon_i \quad (1)$$

where $log(wage_i)$ is the logarithm of wage of an *i*-th individual. $\boldsymbol{SES}_i$ is the vector of explanatory

variables which are listed in Tab. 2. $\boldsymbol{\beta}^{SES}$ consists of consecutive parameters for explanatory variables. ε_i denotes error term following normal distribution with average equal to zero and constant variance. $u1^j$, $u2^k$, $u3^l$ and $u4^m$ denote random effects associated with the 1-digit code, 2-digit-code, 3-digit code and 4-digit code occupational groups respectively. The letters *j, k, l*, and *m* index 1-, 2-, 3- and 4-digit occupational groups. It is also assumed that random effects follow normal distribution with means $E(u1^j)$, $E(u2^k)$, $E(u3^l)$ and $E(u4^m)$ and standard deviations $\sqrt{Var(u1^j)}$, $\sqrt{Var(u2^k)}$, $\sqrt{Var(u3^l)}$, $\sqrt{Var(u4^m)}$. The variable $z1^j_i$ equals unity, if the *i*-th worker belongs to the *j*-th 1-digit code occupational group and 0 if he/she belongs to another group. Variables $z2^k_i$, $z3^l_i$ and $z4^m_i$ are defined analogously. The proposed approach assumes that each worker is assigned to one of 4-digit code occupational groups, each 4-digit code group belongs to one of the 3-digit code groups, each 3-digit code occupational group is included in one of the 2-digit code groups etc. Therefore, the data set has a hierarchical structure and the use of a hierarchical model is justified. Arendt and Grabowski (2019) suggest that individuals assigned to the same 4-digit code occupational group should have similar wage levels, ceteris paribus. If large differences between workers' wages from the same 4-digit code occupational group exist, workers with lower wages may quit and start working for enterprises offering higher wages. The wages of workers from the same 3-digit code occupational group may differ slightly more, but we expect that between-group variation exceeds the within-group variation etc. Therefore, a multilevel model seems to be appropriate. We were unable to take into consideration the problem of endogeneity of education level due to missing data on the workers' family members in the SES data set.

After estimating the parameters of model (Formula 1), we receive estimates of $\boldsymbol{\beta}^{SES}$ parameters, the expected values of random effects ($E(u1^j)$, $E(u2^k)$, $E(u3^l)$, $E(u4^m)$) with their standard deviations ($\sqrt{Var(u1^j)}$, $\sqrt{Var(u2^k)}$, $\sqrt{Var(u3^l)}$, $\sqrt{Var(u4^m)}$). These results are used further in the third step of the analysis in which we complement the LFS data on earnings on the basis of estimates of structural parameters, means and standard deviations for predicted random effects. We use the following Formula for the wage of an individual from the LFS data set:

$$W\_theor\_LFS_i = \begin{cases} W\_LFS_i & if\ \ AV\_LFS_i = 1, \\ exp(\hat{f}_i) & if\ AV\_LFS_i = 0, \end{cases} \quad (2)$$

where AV_LFS_i is a dummy variable taking on the value 1 if data concerning workers' wages (W_LFS_i) is available in the LFS data set, and $\hat{f}_i$ is defined as follows:

$$\hat{f}_i = \boldsymbol{LFS}_i\widehat{\boldsymbol{\beta}}^{SES} + \sum_j z1_i^j\, w1^j + \sum_k z2_i^k\, w2^k + \\ + \sum_l z3_i^l\, w3^l + \sum_m z4_i^m\, w4^m\ , \quad (3)$$

and variables $w1^j$, $w2^k$, $w3^l$ and $w4^m$ are defined as follows:

$w1^j$ – random number generated from the normal distribution with mean $E(u1^j)$ and standard deviation $\sqrt{Var(u1^j)}$,

$w2^k$ – random number generated from the normal distribution with mean $E(u2^k)$ and standard deviation $\sqrt{Var(u2^k)}$,

$w3^l$ – random number generated from the normal distribution with mean $E(u3^l)$ and standard deviation $\sqrt{Var(u3^l)}$,

$w4^m$ – random number generated from the normal distribution with mean $E(u4^m)$ and standard deviation $\sqrt{Var(u4^m)}$.

The algorithm of complementing the individual data on wages in the LFS source file is presented in the Appendix (Fig. 1).

3. Results and Discussion

Firstly, we identified factors which turned out to be statistically significant in the equation explaining the logarithm of wages (Formula 1). Tab. 3–5 present the results of the estimation of the parameters of the multilevel model for the analysed years (2010–2016).

Results of the estimation indicate that the wages of females in the Polish economy are significantly lower than the wages of males. This result is in line with the results of investigations conducted by, among others, Majchrowska and Strawiński (2016; 2018), who analysed this phenomenon in the Polish labour market. Moreover, the positive relationship between the level of education and wages is in line with the Mincerian framework and confirms the results of other studies for the Polish economy (see, among others, Florczak & Grabowski, 2018). However, the decreasing role of tertiary

Tab. 3: Estimates of parameters for variables associated with features of workers and features of enterprises

Variable	2010	2012	2014	2016
SES_GENDER	-0.106***	-0.112***	-0.110***	-0.106***
SES_EDU5	0.194***	0.184***	0.188***	0.185***
SES_EDU4	0.009***	0.012***	0.011***	-0.005***
SES_EDU2	-0.059***	-0.052***	-0.045***	-0.065***
SES_EDU1	-0.069***	-0.061***	-0.045***	-0.048***
SES_IND	0.159***	0.171***	0.169***	0.132***
SES_EXP_GEN	0.007***	0.007***	0.006***	0.004***
SES_SIZE2	0.079***	0.058***	0.088***	0.074***
SES_SIZE3	0.176***	0.158***	0.184***	0.186***
SES_SIZE4	0.318***	0.277***	0.304***	0.304***
cons	7.598***	7.748***	7.782***	7.909***

Source: own

Note: *, **, *** denote significant at the 10%, 5% and 1% level respectively.
Reference categories:
– workers with secondary general education (*SES_EDU3*);
– microenterprises with 1 to 9 workers (*SES_SIZE1*).

Tab. 4: Estimates of parameters for variables associated with the location of an enterprise

Variable	2010	2012	2014	2016
SES_LOC_KP	-0.067***	-0.066***	-0.084***	-0.098***
SES_LOC_LUBE	-0.090***	-0.103***	-0.089***	-0.117***
SES_LOC_LUBU	-0.048***	-0.044***	-0.045***	-0.031***
SES_LOC_LODZ	-0.039***	-0.048***	-0.045***	-0.067***
SES_LOC_MAL	-0.031***	-0.035***	-0.031***	-0.047***
SES_LOC_MAZ	0.111***	0.085***	0.069***	0.059***
SES_LOC_OPOL	-0.035***	-0.048***	-0.054***	-0.066***
SES_LOC_PODK	-0.105***	-0.111***	-0.109***	-0.122***
SES_LOC_PODL	-0.037***	-0.100***	-0.068***	-0.087***
SES_LOC_POM	0.012***	0.010***	-0.008***	-0.024***
SES_LOC_SL	-0.014***	-0.009***	-0.025***	-0.044***
SES_LOC_SW	-0.066***	-0.089***	-0.078***	-0.108***
SES_LOC_WM	-0.061***	-0.052***	-0.076***	-0.063***
SES_LOC_WIEL	-0.030***	-0.032***	-0.025***	-0.036***
SES_LOC_ZACH	-0.034***	-0.009***	-0.026***	-0.024***

Source: own

Note: *, **, *** denote significant at the 10%, 5% and 1% level respectively. Dolnośląskie region (SES_LOC_DLN) is treated as reference category.

education attainment and the increasing role of vocational education attainment is in line with the results obtained by Strawiński et al. (2018) as well as Parteka (2018). Some changes in patterns are also observable. For example, in the years 2010–2014, workers with secondary technical education had an advantage over workers with secondary general education, but in 2016, the latter earned more. The findings of these papers confirm that due to the very low supply of workers with vocational education and the high demand for labour-intensive jobs in the Polish economy, the relative wages of physical workers are increasing compared to the wages of cognitive workers. The phenomenon of the positive relationship between experience and wages was positively verified. In order to avoid the problem of multicollinearity, only one variable associated with experience was taken into account (*SES_EXP_GEN*). In order to check the stability of the results, a variant with the second variable *SES_EXP_ENTERP* was also considered. This variable is defined as the number of years a worker had worked for an enterprise (in the year of the poll). The results of the estimation were very similar.

Analysis of wage differences in the regional dimension points to the dominant position of the Mazowieckie region (Tab. 4). This result confirms the leading role of the capital city, cumulating the social and economic potential of the country. Regions such as Pomorskie and Dolnośląskie also offer higher than median wages. The lowest level of wages is observed in such regions as Lubelskie, Podkarpackie, Podlaskie and Świętokrzyskie, which are classified as the least developed Polish regions with low innovation potential, thus creating relatively low-wage jobs. The obtained results are in line with the new economic geography approach, arguing that there exists a strong correlation between wages and the economic potential of the region (agglomeration patterns, market access, migration processes) (see,

Tab. 5: Estimates of parameters for variables associated with the NACE section of an enterprise

Variable	2010	2012	2014	2016
SES_SECTION_A	0.034***	0.098***	0.062***	0.105***
SES_SECTION_B	0.317***	0.283***	0.220***	0.163***
SES_SECTION_D	0.178***	0.170***	0.172***	0.185***
SES_SECTION_E	0.047***	0.002	0.002	-0.005
SES_SECTION_F	-0.039***	-0.057***	-0.068***	-0.057***
SES_SECTION_G	-0.018***	-0.052***	-0.032***	-0.031***
SES_SECTION_H	-0.034***	-0.058***	-0.044***	-0.053***
SES_SECTION_I	-0.036***	-0.112***	-0.100***	-0.091***
SES_SECTION_J	0.148***	0.107***	0.081***	0.094***
SES_SECTION_K	0.194***	0.125***	0.084***	0.136***
SES_SECTION_L	0.025***	-0.039***	-0.021***	-0.027***
SES_SECTION_M	0.045***	0.054***	0.045***	0.043***
SES_SECTION_N	-0.216***	-0.173***	-0.147***	-0.140***
SES_SECTION_O	-0.024***	-0.097***	-0.124***	-0.135***
SES_SECTION_P	-0.116***	-0.187***	-0.148***	-0.163***
SES_SECTION_Q	-0.093***	-0.160***	-0.169***	-0.188***
SES_SECTION_R	-0.036***	-0.133***	-0.119***	-0.103***
SES_SECTION_S	-0.157***	-0.048***	-0.041***	-0.061***

Source: own

Note: *, **, *** denote significant at the 10%, 5% and 1% level respectively. Section C (Manufacturing) (SES_SECTION_C) is treated as the reference category.

Tab. 6: Predictions of random effects for selected 4-digit professional groups

	2010	2012	2014	2016
Professions with very low predicted random effects (lower than -0.4)	2642, 6223, 7113, 7536, 7537, 9334, 9510	2642, 5242, 6114, 6221, 6222, 9334, 9510	2642, 2656, 4213, 6129, 6221, 6222, 7113, 9123, 9334	2642, 2656, 4213, 6222
Professions with very high predicted random effects (higher than 0.8)	1111, 1112, 1120, 2330, 2341, 2352, 2612, 3153, 3154	1111, 1112, 1120, 2333, 2341, 2352, 2612	1111, 1112, 1120, 2612	1111, 1120, 2352, 2612

Source: own

e.g., Cieślik & Rokicki, 2016). However, it should be stressed that information concerning regions is very limited. Regions in Poland are characterized by a high variance of the intra-regional level of development. For example, in the Mazowieckie region, although Warsaw is rich and the enterprises located there offer high wages, the peripheral counties of the region are much poorer.

The results from Tab. 5 reveal significant inter-industry differences in wages. The highest wages were recorded in Mining and Quarrying (section B), Professional, Scientific and Technical Activities (section J), as well as the Information and Communication industries (section K). The high level of wages in the IT and Scientific industries is due to the very high demand for jobs related to these NACE sections in the labour market. The strong historical position of trade unions in the Mining and Quarrying industry determined the very high level of wages in this NACE section. However, some changes in patterns are observable. Differences between wages in this section and wages in other industries decreased substantially from 2010 to 2016. As a consequence of the trade unions' strategy in the wage bargaining process, the observed levels of wages do not take into account market conditions and may be above productivity levels (see Jonek-Kowalska, 2014). On the other hand, substantial increases in wages between 2010 and 2016 can also be observed (e.g. in section N – *Administrative and support service activities).* Estimates of the parameters for the remaining variables are in line with expectations and the results of other similar studies. Workers employed in larger enterprises, private enterprises (the positive relationship between membership of an enterprise in the public sector and wages for emerging countries was found by Seshan, 2013, among others), or on permanent contracts (the impact of the type of contract on wages was verified by, among others, Dias da Silva and Turini (2015)) earn more.

Tab. 6 provides predictions of random effects for selected 4-digit professional groups. Groups with predicted random effects below -0.4 and above 0.8 are taken into account. The lower (in absolute value) negative threshold is due to the fact that wages are generally right-skewed.

The predictions of random effects indicate that, for some professions, wages are higher than expected (on the basis of a microeconometric model without random effects) in the whole analysed period (2010–2016). In particular, this concerns members of public authorities (1111), senior public administration officials (1112), general and executive directors, and judges (2612). High (in absolute value) and negative predictions of random effects are found in the case of professions requiring a lower quality of worker and are characterized by a decreasing demand for them (for example workers stacking shelves (9334), employees of pawnshops and loan institutions (4213), or inland fishermen (6222)). The results from Tab. 6 are in line with expectations and indicate that it is not only the level of education but also the task content of a job that affects wages (see Hardy et al., 2018).

After the estimation of the parameters of multilevel models, the validity of the assumptions concerning random effects should be verified. The presence of random effect for all 4-digit code occupational groups is verified on the basis of testing the validity of the following hypothesis:

$$H_0: \overset{\forall}{m} u4^m = 0, \quad H_1: \sim H_0. \tag{4}$$

If H_0 hypothesis is not rejected, the presence of random effects for 3-digit code occupational groups should be verified on the basis of testing the validity of the following hypothesis:

$$H_0: \overset{\forall}{l} u3^l = 0, \quad H_1: \sim H_0. \tag{5}$$

If random effects are not present for 4-digit nor for 3-digit code occupational groups, the presence of random effects for 2-digit code occupational groups should be verified on the basis of testing validity of the following hypothesis:

$$H_0: \overset{\forall}{k} u2^k = 0, \quad H_1: \sim H_0. \tag{6}$$

If H_0 hypothesis is not rejected in all three cases (Formulas 4, 5 and 6), the presence of random effects for 1-digit code occupational groups should be verified on the basis of testing the validity of the following hypothesis:

$$H_0: \overset{\forall}{j} u1^j = 0, \quad H_1: \sim H_0. \tag{7}$$

Tab. 7 summarises the results of the verification of the hypothesis concerning the presence of random effects for 4-digit code occupational groups for all analyzed years.

The results from Tab. 7 indicate that in all analysed years random effects were present for

Tab. 7: The verification of the hypothesis concerning presence of random effects for 4-digit code occupational groups

Year	2008	2010	2012	2014	2016
P-value	0.000	0.000	0.000	0.000	0.000

Source: own

Tab. 8: Maximum values of the Cramer's V statistic for pairs of binary explanatory variables

	SES_GENDER	*SES_EDU5*	*SES_EDU4*	*SES_EDU2*	*SES_EDU1*	*SES_IND*	*SES_SIZE2*	*SES_SIZE3*	*SES_SIZE4*
SES_GENDER	–	0.18	0.05	0.17	0.04	0.01	0.07	0.02	0.08
SES_EDU5		–	–	–	–	0.10	0.03	0.04	0.06
SES_EDU4			–	–	–	0.00	0.02	0.04	0.05
SES_EDU2				–	–	0.05	0.02	0.01	0.02
SES_EDU1					–	0.07	0.02	0.00	0.03
SES_IND						–	0.01	0.01	0.01
SES_SIZE2							–	–	–
SES_SIZE3								–	–
SES_SIZE4									–

Source: own

4-digit code occupational groups. It indicates that the use of the multilevel mode was justified.

In order to measure the scale of the dependence of binary variables, the Cramer's V statistic was calculated for all pairs and every time periods. Tab. 8 presents maximal values of the Cramer's V statistic for all (excluding regional and sectional dummy variables) pairs of two dychotomic variables. To save space, the Cramer's V statistic for pairs including regional and sectional dummies were not enclosed, however it should be pointed here that these values remained at a low level.

The results from Tab. 8 indicate that the problem of strong dependence among explanatory variables does not exist (as indicated by low values of the Cramer's V statistic).

In order to evaluate the relative importance of consecutive groups of variables, we compared the sum of squared residuals for the full model with the sum of squared residuals for models without specific variables or groups of variables. We considered reduced models without the following variables/groups of variables, associated with: gender; experience; employment for an indefinite period; regions; NACE sections; size of enterprise; level of education. For each variant of the reduced model we calculated the following quantity:

$$S_j = \frac{e_j^T e_j - e^T e}{e^T e}, \tag{8}$$

where $\boldsymbol{e^T e}$ denotes the sum of squared residuals for the full model, while $\boldsymbol{e_j^T e_j}$ is the sum of squared residuals for a model without the *j*-th group of variables.

Next, we calculated the relative importance of each group of variables on the basis of the following Formula:

$$RI_j = \frac{s_j}{\sum_k s_k}. \tag{9}$$

Tab. 9 presents the relative importance of consecutive groups of variables for all years.

The results of the estimation of the parameters of the full and reduced models indicate that the unexplained part of the variation of the dependent variable increases substantially if the variables associated with the size of an enterprise are excluded from the basic specification. It means that the relative importance of the size of an enterprise contributes the most in explaining workers' wages. Though the importance of some variables which reflect individuals' features in explaining wages turned out to be large (experience, level of education), the role of the variable reflecting type of employment contract as well as most of the company-related characteristics (region, NACE section, size) turned out to be important as well. This result is in line with Ryczkowski and Maksim (2018), who have evidenced that the impact of company-related characteristics is not weaker than the impact of personal characteristics. The relative importance of groups of variables seems to be stable across the whole period 2010–2016. However, it should be stressed that the role of gender in explaining wages decreased in 2016 compared with 2010–2014. In fact, the gender pay gap has decreased in Poland in recent years (see OECD, 2017).

Tab. 9: Relative importance of consecutive groups of variables

Variable/group of variables	2010	2012	2014	2016
Gender	0.055	0.063	0.068	0.017
Experience	0.151	0.157	0.151	0.137
Regions	0.129	0.105	0.098	0.113
NACE sections	0.161	0.174	0.143	0.156
Size	0.244	0.213	0.232	0.276
Level of education	0.130	0.133	0.143	0.151
Indefinite employment period	0.129	0.153	0.163	0.146
Public sector	0.000	0.002	0.002	0.003

Source: own

Tab. 10: Actual and predicted average wages for workers from the Labour Force Survey (in PLN)

Group created on the basis of features of workers and features of enterprises	2010		2012		2014		2016	
	Pred. (1)	True (2)	Pred. (3)	True (4)	Pred. (5)	True (6)	Pred. (7)	True (8)
All workers	2,912	2,646	3,189	2,872	3,549	3,091	3,877	3,471
Males	3,124	2,747	3,278	3,099	3,612	3,252	3,954	3,721
Females	2,586	2,533	3,078	2,703	3,476	2,919	3,793	3,213
Private sector	2,560	2,434	2,750	2,532	3,170	2,903	3,479	3,307
Public sector	3,931	3,056	4,138	3,278	4,364	3,485	4,767	3,808
1–9 workers	1,894	2,138	2,135	2,301	2,448	2,492	2,726	2,704
10–49 workers	2,654	2,558	2,987	2,783	3,409	2,955	3,700	3,310
50–249 workers	3,143	2,777	3,489	3,021	3,840	3,273	4,274	3,649
More than 250 workers	3,960	3,108	4,012	3,378	4,266	3,536	4,625	4,076
Basic education	2,071	1,883	2,123	2,011	2,317	2,188	2,569	2,369
Vocational education	2,319	2,219	2,456	2,389	2,609	2,588	2,870	2,801
Secondary general education	2,379	2,294	2,503	2,489	2,676	2,685	2,946	3,116
Secondary technical education	2,697	2,514	2,878	2,756	3,087	2,923	3,328	3,195
Tertiary education	4,121	3,838	4,457	3,975	4,895	4,161	5,306	4,509

Source: own

In order to evaluate the quality of our novel method for complementing data on the wages of workers who did not provide any information about their wages in the LFS, descriptive statistics both for workers who gave information about their wages as well as for workers who did not provide any information have been calculated. The mean values for all workers, as well as for members of consecutive groups, are presented in Tab. 10.

According to the results from Tab. 10, the higher predicted wages correspond to actual wage levels. This tendency is optimistic, since it seems that the microeconometric model correctly predicts the distribution of wages. For workers who have completed tertiary education and who are employed in large enterprises or the public sector, predicted wages are higher, which is in line with reality. On the other hand, the theoretical wages of women with basic or vocational education in microenterprises belonging to the private sector are low, which corresponds with reality. However, the differences between the predicted wages and the observed ones in some groups are not negligible. Therefore, an in-depth analysis of these differences should be conducted.

Analysis of the ratios of predicted wages to observed ones indicates that they are not equal in all groups. The analysed ratio is much higher in the case of workers from public sector enterprises than for employees representing private enterprises. Moreover, the ratio strongly depends on the size of the enterprise (the larger a company, the higher the analysed ratio) and the level of the workers' education (positive correlation). The results are in line with expectations and common knowledge concerning discrepancies between wages obtained from the LFS and the SES data sets. As Strawiński (2015) found, wages from the two data sets are almost equal in the case of employees with below median wages. When wages exceed the median, the difference between SES wages and LFS wages is an increasing quantile function of the wage distribution.

Large differences between predicted and observed wages in the public sector and the largest enterprises may be due to the fact that

employees do not take into account perks or other additional benefits when they report their net wages. A similar tendency may be observed in the case of employees with higher education. Additional benefits and perks are more often provided in larger enterprises, the public sector and departments with a high ratio of well-educated workers. Moreover, in large and public enterprises, a quarterly or yearly bonus is very often provided to workers. This bonus is taken into account by Human Resources (HR) departments when they provide information about the wages of employees, but may not be included by individual employees when they reveal their net wage. It should be stressed that high earning workers may be less aware of their wages than low earning workers. Suppose that there is one worker with a constant wage of 2,247 PLN and a second worker with a wage of 6,247 PLN. The second one may report 6,200 since 47 is not a large part of his/her wage. For the first worker, the difference between 2,247 and 2,200 is large, so this employee will probably give his/her exact wage. A positive correlation between the quantile of the wage distribution and the discrepancy between the predicted and observed wages may be due to changes in wages as well as polarization and the SBTC phenomenon (see, e.g., Goos & Manning, 2007). The increasing gap between high and low earners results from the fact that the wages of professionals change more often. When workers report the level of their wages in the Labour Force Survey, they may mention wages earned a few months before. In the case of highly-qualified employees, the difference between wages observed a few months previously and the current level may be much larger than in the case of low earning workers.

In the case of workers from some groups (low level of education, the private sector, microenterprises), the predicted wages are higher than the observed ones. This tendency may be due to the envelope wage phenomenon, which is popular in transition economies (see, e.g., Williams, 2015; Kukk & Staehr, 2014; 2017; Williams & Hordonic, 2015). Employees with a low level of education, working for poor enterprises in transition countries, very often agree to receive a so-called envelope wage. Employers decide to provide an envelope wage in order to reduce costs. Officially, they are given very low wages, which is reported by the HR departments of the enterprises. In reality, they are given this minimum low wage plus an envelope wage. They report their true income in the Labour Force Survey. This may result in higher actual wages than predicted ones for particular groups of workers.

All in all, discrepancies between the predicted wages and the actual ones indicate that additional steps are required when the predicted data on wages are combined with the empirical data. The results of the ANOVA analysis (available upon request) indicate that four factors explain more than 95% of the variation of the predicted to observed ratio. They are: sector, employee's level of education, employee's gender, and size of the enterprise. Therefore, for each year (2010, 2012, 2014, 2016) we define the following variables:

$pre^{t}_{se,e,g,si}$ – average predicted wage of an employee from sector *se*, employed in an enterprise from size class *si* with the level of education *e* and gender *g* in year *t*,

$obs^{t}_{se,e,g,si}$ – the average observed wage of an employee from sector *se*, employed in an enterprise from size class *si* with the level of education *e* and gender *g* in year *t*.

On the basis of these variables, the correction factor is calculated according to the following Formula:

$$corr^{t}_{se,e,g,si} = \frac{obs^{t}_{se,e,g,si}}{pre^{t}_{se,e,g,si}}. \tag{10}$$

As a result, a modified Formula for calculating wage for the *i*-th individual from the LFS who works in an enterprise belonging to sector *se* and size class *si* is as follows:

$$W\_theor\_LFS\_CORR_i = \begin{cases} W\_LFS_i & if \quad AV\_LFS_i = 1, \\ W\_theor\_LFS_i * corr^{t}_{se,e,g,si} & if\ AV_{LFS_i} = 0. \end{cases} \tag{11}$$

Formula 11 indicates that when the wage of an individual in the LFS is given, its prediction is redundant. When it is not given, a theoretical value should be calculated on the basis of the SES and microeconometric model estimates. Moreover, after comparing the wages of individuals' reporting their wages with workers not reporting them, a correction factor should be calculated for all groups. This correction factor should next be used to calculate corrected theoretical wages. After using Formula 11, we obtain the improved LFS data set with wages for all individuals (even those not reporting them).

The correction factors turned out to be similar in all analysed years (2010, 2012,

Tab. 11: Geometric mean of a correction factor calculated according to the Formula 10

Size of an enterprise and educational attainment of workers	Females, Public sector	Females, Private sector	Males, Public sector	Males, Private sector
Microenterprises, Basic education	–	1.07	–	1.08
10–49 workers, Basic education	0.87	0.94	1.03	0.99
50–249 workers, Basic education	0.80	0.92	0.93	0.93
At least 250 workers, Basic education	0.73	0.90	0.75	0.89
Microenterprises, Vocational education	–	1.07	–	1.12
10–49 workers, Vocational education	0.85	1.05	0.99	1.03
50–249 workers, Vocational education	0.87	0.99	0.93	0.94
At least 250 workers, Vocational education	0.83	0.96	0.83	0.87
Microenterprises, Secondary general education	–	1.06	–	1.15
10–49 workers, Secondary general education	0.98	1.06	0.90	1.06
50–249 workers, Secondary general education	0.86	1.09	0.86	0.95
At least 250 workers, Secondary general education	0.77	1.00	0.95	0.89
Microenterprises, Secondary technical education	–	1.08	–	1.13
10–49 workers, Secondary technical education	0.96	1.00	0.93	1.00
50–249 workers, Secondary technical education	0.93	0.93	0.90	0.90
At least 250 workers, Secondary technical education	0.79	0.91	0.85	0.85
Microenterprises, Tertiary education	–	0.96	–	1.01
10–49 workers, Tertiary education	0.79	0.97	0.85	0.95
50–249 workers, Tertiary education	0.85	0.91	0.82	1.00
At least 250 workers, Tertiary education	0.84	0.94	0.84	0.87

Source: own

Note: – denotes microenterprises that are very rare in the public sector. Therefore means for these groups are omitted.

2014, 2016). To save space in the article, only the geometric means of the ratio for different sectors, size groups, genders and levels of educational attainments are reported. These geometric means are presented in Tab. 11.

The results from Tab. 11 comply with the results from Tab. 10. After predicting the wages (according to Formula 2) of employees with tertiary education working in larger public enterprises, they are multiplied by a correction factor lower than 1. After predicting the wages (according to Formula 2) of employees with a lower level of educational attainment, working for smaller, private enterprises, a correction factor higher than 1 is used. After using the correction factor according to Formula 11, the predicted and actual data can be combined, and a researcher may use a larger data set in order to analyse the level of wages.

Conclusions

To sum up, the presented study highlights the problem of the large data gaps in wages in the Polish Labour Force Survey data set. However, this problem is universal and also appears in data sets from other EU countries. We propose a novel method to complement the LFS data on wages on the basis of the SES data set and microeconometric model estimates. The proposed method can be adapted and implemented on the LFS data sets of other countries. The use of this method makes it possible to get a more complete profile of respondents from the LFS. Moreover, the completed data on wages can be used in further LFS data analyses without having to skip a large number of incomplete observations.

As a result of using our method of complementing missing data, we found some discrepancies between the predicted and observed wages. However, these discrepancies turned out to depend mainly on two features of the enterprises (size class and sector) and two features of the workers (level of education and gender). Therefore, on the basis of these features, we created correction factors which are used in order to adjust wages. The analysis of correction factors provided information about the mechanism of not declaring, underestimating and overestimating wages. Well-educated employees from large and public enterprises may not include additional benefits when providing information about wages. Workers with a low level of educational attainment who are employed in small, private enterprises may agree to receive a so-called envelope wage and report higher earnings in the LFS than is reported in the SES.

The results of the estimation of the parameters of the multilevel model show the impact of features of employees on wages. Generally, they are in line with expectations. However, some changes in patterns are observable. For example, workers with secondary technical education had an advantage over workers with secondary general education in the years 2010–2014. However, in 2016, the latter earned more, ceteris paribus. Differences between wages in the mining and quarrying industry and wages in other industries decreased substantially from 2010 to 2016. Predictions of random effects indicate that there are professions with very high (e.g. members of public authorities) and very low (e.g. workers stacking shelves) wages in the whole analysed period. When random effects are not included, the explanatory power of the econometric model decreases substantially.

We should also emphasize that our research could have some limitations. Not all variables affecting wages are included in the SES data set. For example, variables associated with the class of town/village and marital status are not available. Therefore, some variables that are traditionally used in models explaining wages cannot be included in the specification. Moreover, information concerning the location of an enterprise is very limited. Regions may be characterized by a high-variance of the intra-regional level of development. Therefore, information that an enterprise is located in a specific region does not reflect the full picture of conditions in the local labour market.

In the future, we will use completed data on wages in further LFS data analyses, related to, for example, the polarization hypothesis, the skill-biased technical change phenomenon, the gender wage gap, and inter-regional and section differences in wages. On the basis of newer LFS and SES questionnaires, we will develop the proposed method in order to obtain more accurate estimates. Moreover, we will try to conduct similar estimations for other EU countries. After that, we will compare the differences between Poland and other countries.

Acknowledgement: *This paper was prepared within the framework of the research project*

entitled "The polarization of the Polish Labour Market in the context of technical change," financed by the National Science Centre, Poland (contract number 2016/23/B/HS4/00334).

We are grateful to Iwona Kukulak-Dolata, Łukasz Arendt, Leszek Kucharski, Paweł Baranowski and anonymous reviewers for their helpful comments, which enabled us to improve the initial version of this paper.

References

ADB. (2012). *Labour Force Data Analysis: Guidelines with African Specificities*. Retrieved March 28, 2019, from www.afdb.org/fileadmin/uploads/afdb/Documents/Publications/Labour%20Force%20Data%20Analysis_WEB.pdf

Ahituv, A., & Lerman, R. I. (2007). How do marital status, work effort, and wage rates interact? *Demography, 44*(3), 623–647. https://doi.org/10.1353/dem.2007.0021

Ahn, N., & Sanchez-Macros, V. (2017). Emancipation under the great recession in Spain. *Review of Economics of the Household, 15*(2), 477–495. https://doi.org/ 10.1007/s11150-015-9316-7

Arendt, Ł., & Grabowski, W. (2019). Technical change and wage premium shifts among task-content groups in Poland. *Economic Research – Ekonomska Istraživanja, 32*(1), 3392–3410. https://doi.org/10.1080/1331677X.2019.1661788

Artz, B., & Taengnoi, S. (2019). The Gender Gap in Raise Magnitudes of Hourly and Salary Workers. *Journal of Labor Research, 40*(1), 84–105. https://doi.org/10.1007/s12122-018-9277-8

Autor, D. H., Levy, F., & Murane, R. J. (2003). The skill content of recent technological change. An empirical exploration. *The Quarterly Journal of Economics, 118*(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Blau, F. D., & Kahn, L. M. (2007). The gender pay gap: Have women gone as far as they can? *Academy of Management Perspectives, 21*(1), 7–23. https://doi.org/10.5465/amp.2007.24286161

Central Statistical Office. (2018). *Labour Force Survey in Poland. II quarter 2018.* Retrieved from http://stat.gov.pl/download/gfx/portalinformacyjny/en/defaultaktualnosci/3293/2/30/1/labour_force_survey_in_poland_2nd_quarter_2018.pdf

Cieślik, A., & Rokicki, B. (2016). Individual wages and regional market potential: evidence from the Polish Labour Force Survey. *Economics of Transition, 24*(4), 661–682. https://doi.org/10.1111/ecot.12102

Dias da Silva, A., & Turrini, A. (2015). *Precarious and less well-paid? Wage differences between permanent and fixed-term contracts across the EU countries* (Economic Papers 544). Brussels: European Commission.

European Commission. (2018). *EU Labour Force Survey Database User Guide.* Retrieved from http://ec.europa.eu/eurostat/documents/1978984/6037342/EULFS-Database-UserGuide.pdf

Freeman, R., & Katz, L. (1994). Rising wage inequality: the U.S. versus other advanced countries. In R. Freeman (Ed.), *Working Under Different Rules* (pp. 29–62). New York, NY: Russel Sage Foundation.

Gallen, Y., Lesner, R. V., & Vejlin, R. (2019). The labor market gender gap in Denmark: Sorting out the past 30 years. *Labour Economics, 56,* 58–67. https://doi.org/10.1016/j.labeco.2018.11.003

Gannon, B., Plasman, R., Tojerow, I., & Rycx, F. (2007). Inter-industry wage differentials and the gender wage gap: Evidence from European countries. *Economic and Social Review, 38*(1), 135–155. Retrieved from http://hdl.handle.net/2262/60127

Goos, M., & Manning, A. (2007). Lousy and lovely jobs: The rising polarization of work in Britain. *Review of Economics and Statistics, 89*(1), 118–133. https://doi.org/10.1162/rest.89.1.118

Grabowski, W. (2019). Does the use of professional legal assistance bring measurable benefits? *Applied Economics Letters, 26*(17), 1444–1447. https://doi.org/10.1080/13504851.2019.1578850

Hara, H. (2018). The gender wage gap across the wage distribution in Japan: Within- and between-establishment effects. *Labour Economics, 53,* 213–229. https://doi.org/10.1016/j.labeco.2018.04.007

Hardy, W., Keister, R., & Lewandowski, P. (2018). Educational upgrading, structural change and the task composition of jobs in Europe. *Economics of Transition, 26*(2), 201–231. https://doi.org/10.1111/ecot.12145

Jonek-Kowalska, I. (2015). Employment and Remuneration Trends in Polish Hard Coal Mines in the Context of the Relations Between Boards and Trade Unions. *International Journal of Synergy and Research, 3,* 27–43. http://dx.doi.org/10.17951/ijsr.2014.3.0.27

Katz, L. F., & Autor, D. H. (1999). Changes in the wage structure and earnings inequality.

In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. 3, pp. 1463–1555). Amsterdam: Elsevier.

Kerr, A., & Wittenberg, M. (2015). Sampling methodology and fieldwork changes in the October Household Surveys and Labour Force Surveys. *Development Southern Africa, 32*(5), 603–612. http://dx.doi.org/10.1080/0376835X.2015.1044079

Kukk, M., & Staehr, K. (2014). Income underreporting by households with business income: Evidence from Estonia. *Post-Communist Economies, 26*(2), 257–276. https://doi.org/10.1080/14631377.2014.904110

Kukk, M., & Staehr, K. (2017). Identification of households prone to income underreporting. Employment status or reported business income? *Public Finance Review, 45*(5), 599–627. https://doi.org/10.1177/1091142115616182

Majchrowska, A., & Strawiński, P. (2016). Regional differences in gender wage gaps in Poland: New estimates based on harmonized data for wages. *Central European Journal of Economic Modelling and Econometrics, 8*(2), 115–141. Retrieved from http://www.cejeme.org/publishedarticles/2016-09-30-636028709931093750-5582.pdf

Majchrowska, A., & Strawiński, P. (2018). Impact of minimum wage increase on gender wage gap: Case of Poland. *Economic Modelling, 70,* 174–185. https://doi.org/10.1016/j.econmod.2017.10.021

McCulloch, C. E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association, 92*(437), 162–170. https://doi.org/10.1080/01621459.1997.10473613

Melly, B. (2005). Public-private sector wage differentials in Germany: Evidence from quantile regression. *Empirical Economics, 30*(2), 505–520. https://doi.org/10.1007/s00181-005-0251-y

Mikucka, M., & Valentova, M. (2013). Employed or inactive? Cross-national differences in coding parental leave beneficiaries in European Labour Force Survey data. *Survey Research Methods, 7*(3), 169–179. http://dx.doi.org/10.18148/srm/2013.v7i3.5308

Mincer, J. (1974). *Schooling, Experience, and Earnings. (Human Behavior & Social Institutions No. 2).* Cambridge, MA: National Bureau of Economic Research.

Mincer, J., & Polachek, S. (1974). Family Investments in Human Capital: Earnings of Women. *Journal of Political Economy, 82*(2), S76–S108.

OECD. (2017). *Closing the Gender Gap – Poland*. Paris: OECD. Retrieved March 31, 2019, from www.oecd.org/gender/Closing%20the%20Gender%20Gap%20-%20Poland%20FINAL.pdf

O'Mahony, M., & Timmer, M. P. (2009). Output, input and productivity measures at the industry level: the EU KLEMS database. *The Economic Journal, 119*(538), F374–F403. http://dx.doi.org/10.1111/j.1468-0297.2009.02280.x

Orche Galindo, E. J., & Bueno Maroto, H. (2011). Obtaining Statistical Information in Sampling Surveys from Administrative Sources: Case Study of Spanish LFS 'Wages from the Main Job'. In *ESSnet Data Integration Workshop, Madrid, Spain*.

Ormerod, C., & Ritchie, F. (2007). Linking ASHE and LFS: can the main earnings sources be reconciled? *Economic & Labour Market Review*, *1*(3), 24–31. https://doi.org/10.1057/palgrave.elmr.1410041

Parteka, A. (2018). Import Intensity of Production, Tasks and Wages: Micro-Level Evidence for Poland. *Entrepreneurial Business and Economics Review, 6*(2), 71–89. https://doi.org/10.15678/EBER.2018.060204

Pastore, F., & Socha, M. (2006). The Polish LFS: A Rotating Panel with Attrition. *Ekonomia, 15*(3), 3–24.

Peter, S., & Drobnič, S. (2013). Women and their memberships: Gender gap in relational dimension of social inequality. *Research in Social Stratification and Mobility, 31,* 32–48. https://doi.org/10.1016/j.rssm.2012.09.001

Psacharopoulos, G., & Ng, Y. C. (1994). Earnings and Education in Latin America. *Education Economics, 2*(2), 187–207. https://doi.org/10.1080/09645299400000016

Ryczkowski, M., & Maksim, M. (2018). Low wages – Coincidence or a result? Evidence from Poland. *Acta Oeconomica, 68*(4), 549–572. https://doi.org/10.1556/032.2018.68.4.4

Seshan, G. K. (2013). Public-private-sector employment decisions and wage differentials in peninsular Malaysia. *Emerging Markets Finance and Trade, 49*(S5), 163–179. https://doi.org/10.2753/REE1540-496X4905S510

Skinner, C., Stuttard, N., Beissel-Durrant, G., & Jenkins, J. (2002). The measurement of low pay in the UK Labour Force Survey.

Oxford Bulletin of Economics and Statistics, *64*(S1), 653–676. https://doi.org/10.1111/1468-0084.64.s.5

Strawiński, P. (2015). Krzyżowe porównanie danych o wynagrodzeniach z polskich badań przekrojowych. *Bank i Kredyt, 46*(5), 433–462. Retrieved from http://bankikredyt.nbp.pl/homen.aspx?f=/content/2015/05/bik_05_2015_en.html

Strawiński, P., Majchrowska A., & Broniatowska, P. (2018). Wage Returns to Different Education Levels. Evidence from Poland. *Ekonomista, 2018*(1), 25–49.

Williams, C. C. (2015). Evaluating cross-national variations in envelope wage payments in East-Central Europe. *Economic and Industrial Democracy, 36*(2), 283–303. https://doi.org/10.1177/0143831X13505120

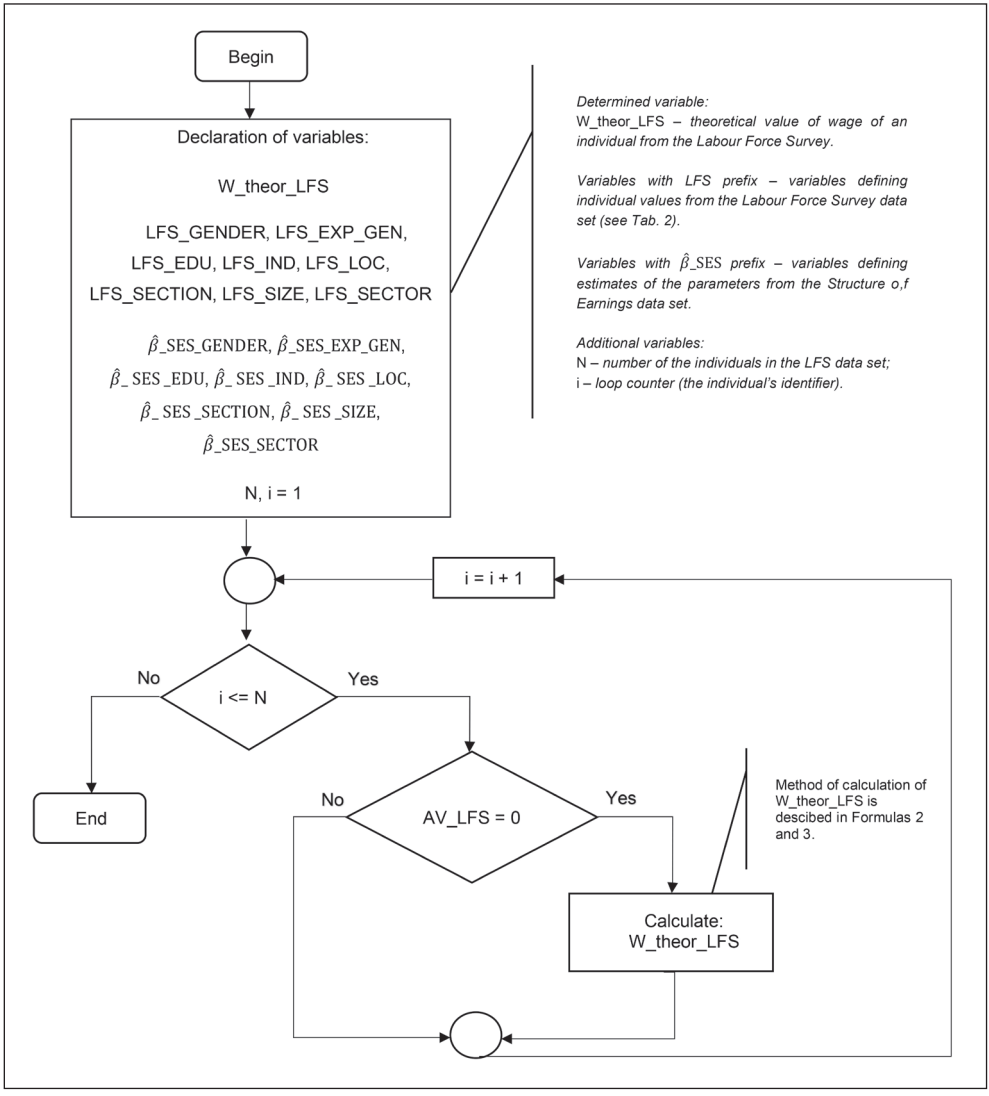
Williams, C. C., & Hordonic, I. A. (2015). Evaluating the prevalence of the undeclared economy in Central and Eastern Europe: An institutional asymmetry perspective. *European Journal of Industrial Relations, 21*(4), 389–406. https://doi.org/10.1177/0143831X14568835

Vehovar, V. (1999). Field substitution and unit nonresponse. *Journal of official Statistics*, *15*(2), 335–350.

Zimermanová, K. (2010). Selected actual aspects of employees remuneration in small and medium-sized Companies. *E&M Economics and Management*, *13*(3), 33–44.

Appendix

Fig. 1: Simplified algorithm of complementing data on wages in the LFS source file



Source: own

Note: Fig. 1 presents a simplified algorithm of complementing data on wages in the LFS source file. Implementing the Formula on W_theor_LFS requires the use of conditional statements for each dummy variable (see Tab. 2). In a complete algorithm, there are 44 conditional statements for variables of this type (available upon request).