STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1 pp. 137–157, DOI 10.21307/stattrans-2020-008
Submitted – 10.01.2019; accepted for publishing –13.12.2019

Robust estimation of wages in small enterprises: the application to Poland's districts

Grażyna Dehnel[1], **Łukasz Wawrowski**[2]

ABSTRACT

The paper presents an empirical study designed to test a small area estimation method. The aim of the study is to apply a robust version of the Fay-Herriot model to the estimation of average wages in the small business sector. Unlike the classical Fay-Herriot model, its robust version makes it possible to meet the assumption of normality of random effects under the presence of outliers. Moreover, the use of this version of the Fay-Herriot model helps to improve the precision of estimates, especially in domains where samples are of small sizes. These alternative models are supplied with auxiliary variables. The study seeks to present the characteristics of and differences among small business units cross-classified by selected NACE sections and district units of the provinces of Mazowieckie and Wielkopolskie. It was carried out on the basis of data from a survey conducted by the Statistical Office in Poznan and from administrative registers. It is the first study which attempts to produce estimates of average wages for this sector of the national economy.

Key words: small area estimation, indirect estimation, robust Fay-Herriot model, administrative registers, enterprise statistics

JEL Classification: C13, C51, M20

1. Introduction and motivation

Nowadays, it is widely known that small and medium-sized enterprises (SME) are a strong pillar of the economy. They play a crucial role in the economic and social sphere, not only for the country as a whole, but, even more importantly, at the regional level. This is because there is a strong correlation between the development of the SME sector and the regional development. The growth of the SME sector helps to eliminate regional differences, contributes to the improvement of living conditions of local communities and fosters creation of new jobs; in other words, it has a positive impact

[1] Poznan University of Economics and Business, Department of Statistics, Poznań. E-mail: grazyna.dehnel@ue.poznan.pl. ORCID: https://orcid.org/0000-0002-0072-9681.
[2] Poznan University of Economics and Business, Department of Statistics, Poznań. E-mail: lukasz.wawrowski@ue.poznan.pl. ORCID: https://orcid.org/0000-0002-1201-5344.

on the region's economic growth. In addition, entrepreneurs tend to locate their capital close to their places of residence, which enables them to rely on local resources and the local market (Strużycki 2004). The scope and intensity of investment depend on entrepreneurs' assessment of the degree of regional development, which in turn creates demand for regular publishing of economic information, such as the level of entrepreneurship, the labour market situation and investment activity.

In the context of regional development, it is small companies which play a most important role in this process. Despite that, they are usually treated as a mere component of the SME sector, and thus are rarely the subject of separate studies or analyses. The scope of available data on small businesses is limited, especially at low levels of aggregation. Such data are most often collected by sample surveys conducted by Statistics Poland. The study presented in this article is the attempt to partly fill this gap. The aim of the study is to estimate one indicator of entrepreneurship, namely average monthly wages in small enterprises at the level of districts. The analysis was limited to four NACE sections (manufacturing, construction, trade, transportation) and two provinces representing the highest level of entrepreneurial activity in Poland – Wielkopolskie and Mazowieckie. The province of Mazowieckie comprises 42 districts (including 5 cities) and the province of Wielkopolskie 35 districts (including 4 cities).

In the case of small companies, information on monthly financial results by NACE section is available only at the country and province levels. However, using the methods of indirect estimation offered by small area estimation (which are resistant to outliers), the authors managed to obtain estimates for a more detailed domain of interest, by cross-classifying NACE section with the territorial division into districts.

The article consists of five parts. The first part is devoted to the difficulties encountered while estimating average wages on the basis of Polish survey data. The second part describes data sources used for the estimation and provides additional details about the empirical study. The methodological considerations of the analysis are presented in the third part, the summary of the results and their interpretation in the fourth part, and the conclusions and suggestions for further work in the fifth part. The study presented in this paper is the continuation of the authors' previous research (Dehnel and Wawrowski 2018).

That study involved small enterprises employing from 10 to 49 employees. Owing to data availability, the analysis is limited to the year 2011, which was the time when the labour market was going through a downturn. The rate of registered unemployment grew to 12.5% compared to the previous year's 9.5% (according to the Labour Force Survey). The only positive trend that could be observed at that time was a rise in the average employment in the enterprise sector (a-7.6% growth in the construction section). The year 2011 also saw a slight increase in wages, but the rate of real wage growth was insignificant in view of the relatively large rise of consumer prices. The

average monthly wages was PLN 3481, which indicated the annual growth of 5.5%. However, the average monthly wages varied considerably across the groups of enterprises, depending on their size and type of activity. In small companies, the average gross wages totalled PLN 2583, ranging from PLN 1818 in companies involved in other service activities to PLN 4737 in information and communication companies; in medium-sized enterprises the average gross salary stood at PLN 3568, while in large enterprises at PLN 4255 (GUS 2013) (Figure1). In companies representing the four selected NACE sections, the average monthly salary reached similar levels, ranging from PLN 2150 in transportation enterprises to PLN 2460 in manufacturing companies.

There was also a considerable variation in average monthly salaries across provinces. The highest values were obtained for Mazowieckie province, which was an outlier. The average monthly wages in this province is considerably higher than the average monthly wages in the remaining provinces, regardless of the company size. Relatively high average monthly wages were also observed in Pomorskie, Dolnośląskie, Śląskie and Wielkopolskie.

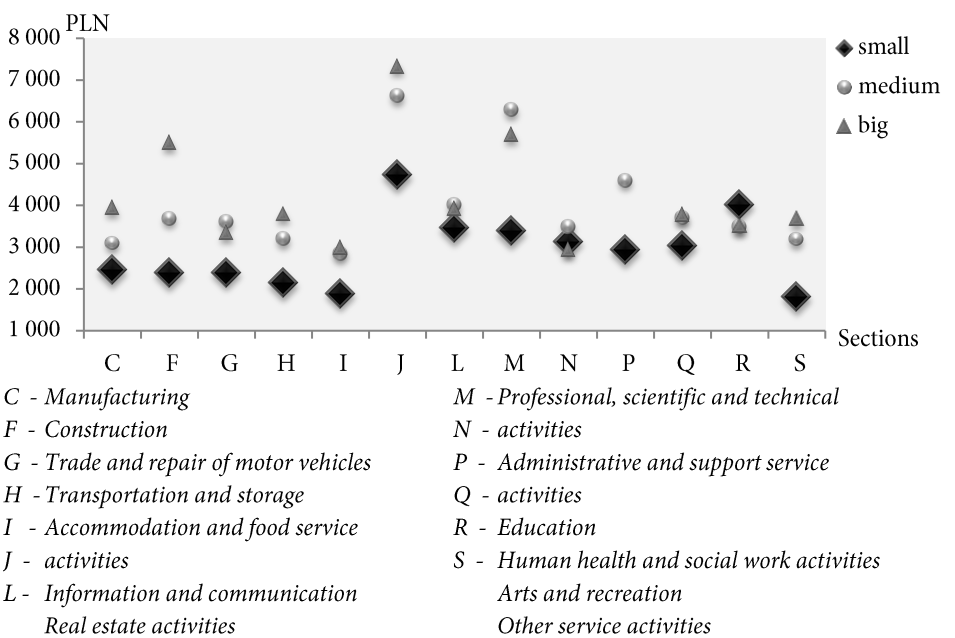

Figure 1. Average monthly wage per 1 employee according to the enterprise's principal activity of size class in 2011

Source: Based on data by Statistics Poland (GUS 2013).

Figure 2 shows the average monthly wages for the whole subregion (NUTS 2), but does not account for the internal variation, which is explained, for example, by the classic theory of growth poles. In view of this fact, and also trying to satisfy the demand for detailed statistical information, an attempt was made to estimate the average monthly wages at a lower level of spatial aggregation, i.e. across districts.

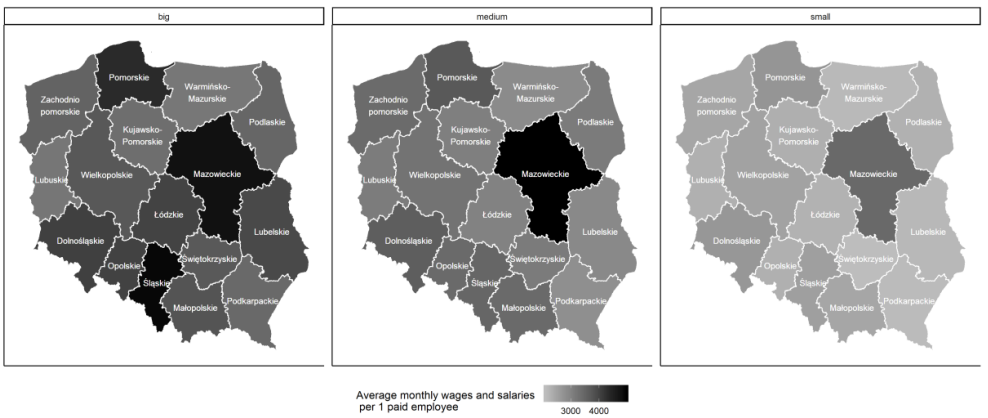

Figure 2. Average monthly wage per 1 employee across provinces by the enterprise's size class in 2011
Source: Based on data by Statistics Poland (GUS 2013).

2. Estimation methods

2.1. Direct estimation

The traditional approach in survey methodology to estimate population means and totals involves the use of the direct Horvitz-Thompson (1952) estimator. Let U denote a population consisting of N units divided into D domains $U_1, \ldots, U_D$ with the sample size denoted by N_d, where $d = 1, \ldots, D$. The sample denoted by s and $s \in U$ can also be divided into $s_1, \ldots, s_D$ with the sample size n_d for each domain.
Let y_{di} denote the value of the target variable for i-th unit in domain d. Each unit has a sampling weight w_{di}. The population mean in area d is denoted by θ_d. Thus, the Horvitz-Thompson estimator is expressed by the following formula:

$$\hat{\theta}_d^{HT} = \sum_{i=1}^{n_d} y_{di} w_{di}. \quad (1)$$

The Horvitz-Thompson (HT) estimator is unbiased and effective for large sample sizes n_d. However, when estimating detailed domains, sample sizes tend to be very small or even zero, which causes a big variance of the HT estimator and makes it impossible to obtain direct estimates.

2.2. Indirect estimation

To estimate population means in domains characterized by a small sample size n_d, it is necessary to use indirect estimators. This approach is also known as small area estimation, where the term *area* does not necessarily refer to a geographical unit. In this type of estimation, auxiliary variables from sources other than the survey are utilised. These variables should not contain sampling errors, so they should be taken from censuses or administrative registers.

The most common indirect approach is a model-based estimation, in which the target variable is the dependent variable in the linear mixed model. The methods within this approach can be divided into area-level and unit-level models. In an area-level model the dependent variable and auxiliary variables are aggregated at the target level of districts. Values of the dependent variables are often estimated by the Horvitz-Thompson (1952) estimator on the basis of survey data, while aggregated values of covariates come from the source that is not measured with error, such as the census or administrative registers (Guadarrama et al. 2016). The most popular example of an area-level model is the Fay-Herriot model (Fay and Herriot, 1979) and its numerous variants, e.g. spatial or robust (Rao and Molina 2015). On the other hand, there are unit-level models in which the dependent variable is taken directly from the survey and auxiliary variables also come from the census or administrative registers, but in a raw form. The Empirical Bayes method (Molina and Rao 2010) utilizes a nested error linear regression model and the Monte Carlo approximation to estimate the variable of interest in target domains, while the M-Quantile approach (Chambers and Tzavidis 2006) uses quantile regression to ensure robustness of that method. In both approaches, the random effect is usually assigned to the geographical area. Because access to unit-level data is limited, area-level models are used more frequently in practice.

Fay and Herriot used their model to estimate income in small geographical units in the USA (Fay and Herriot, 1979). However, the model's application is wider than that – e.g., it can be used for poverty estimation (Wawrowski 2016). Let us assume that the direct estimate of the population mean is the sum of true values of the parameter and random error, which is expressed by the formula:

$$\hat{\theta}_d^{HT} = \theta_d + e_d, \tag{2}$$

where $e_d \overset{ind}{\sim} N(0, \sigma_{ed}^2)$. In practice, variance σ_{ed}^2 is unknown and has been estimated on the basis of survey data.

Then, the true value of the parameter can be described by a linear model with area random effect:

$$\theta_d = x_d^T \beta + u_d, \tag{3}$$

where x_d is a vector of auxiliary information for area d, β is a vector of regression parameters and u_d is area random effect with distribution $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$.

By combining the two equations above, we obtain the Fay-Herriot model:

$$\hat{\theta}_d^{HT} = x_d'\beta + u_d + e_d. \tag{4}$$

In order to obtain Empirical Best Linear Unbiased Predictor (EBLUP) of the Fay-Herriot model, it is necessary to estimate area random effect variance (σ_u^2). It can be done by various methods, e.g. Fay-Herriot method, Prasad-Rao method, REML or ML. Then, EBLUP is expressed by:

$$\hat{\theta}_d^{FH} = x_d^T\hat{\beta} + \hat{u}_d = \hat{\gamma}_d\hat{\theta_d} + (1-\hat{\gamma}_d)x_d^T\hat{\beta}, \; d = 1, \ldots, D \tag{5}$$

where

$$\hat{\beta} = \left(\sum_{d=1}^{D}\hat{\gamma}_d x_d x_d^T\right)^{-1}\sum_{d=1}^{D}\hat{\gamma}_d x_d\hat{\theta}_d$$

and $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\hat{\sigma}_{ed}^2}$.

It is worth remembering that EBLUP is a weighted average of direct and regression model estimates. The weight $\hat{\gamma}_d$ is a share of area random effect variance in the total variance and measures the uncertainty of the fitted model. For small values of sample variance ($\hat{\sigma}_{ed}^2$), a larger part of the final estimate will be contributed by the direct estimate, which will decrease as the sample variance increases (Rao and Molina, 2015).

The Fay-Herriot model is an example of a *shrinkage* estimator. Its drawback is the fact it cannot deal with outliers. For this reason, it is necessary to use robust small area estimation methods. In practice, this can be achieved by replacing $\hat{\beta}$ and $\hat{u}$ in the Fay-Herriot model with their outlier-resistant alternatives (Chambers et al., 2014).

Let us replace values of $\hat{\sigma}_{ed}^2$ and $\hat{\sigma}_u^2$ variances with covariance matrices Σ_e and Σ_u and let $V = \Sigma_e + \Sigma_u$. Then, the vector of fixed effects β is expressed by:

$$\beta = (X^TV^{-1}X)^{-1}XV^{-1}y \tag{6}$$

and random effects vector u is:

$$u = \Sigma_u Z^T V^{-1}(y - X\beta). \tag{7}$$

It can be noted that equations (6) and (7) could be transformed into:

$$X^TV^{-1}(y - X\beta) = 0 \tag{8}$$

and

$$\Sigma_u Z^T V - 1(y - X\beta) - u = 0. \tag{9}$$

Sinha and Rao (2009) proposed a robust version of equations (8) and (9):

$$X^T V^{-1} U^{1/2} \psi(U^{1/2}(y - X\beta)) = 0 \tag{10}$$

where $U = diag(V)$. A robust random effects vector is defined by:

$$\psi((y - X\beta)^T U^{\frac{1}{2}}) U^{\frac{1}{2}} V^{-1} (\partial V / \partial \theta) V^{-1} U^{\frac{1}{2}} \psi(U^{\frac{1}{2}}(y - X\beta)) = \mathrm{tr}(\mathrm{D}^{\Psi}(\partial \mathrm{V} / \partial \theta)), \tag{11}$$

where $\partial V / \partial \theta$ is the first order partial derivative of V with respect to the variance component θ and for $Z \sim N(0,1)$, $D^{\psi} = E(\psi^2(Z))V^{-1}$.

Moreover, Warnholz (2016) proposed a modification of the above equation in which only diagonal elements of the V matrix are used to standardise the residuals. In a robust Fay-Herriot model, this matrix is diagonal, but the transformation can be useful in models with correlated random effects, e.g. SAR(1) and AR(1) models, where calculations might be time-consuming.

The final equation of the robust Fay-Herriot model can be written as:

$$\hat{\theta}_d^{RFH} = x_d^T \hat{\beta}^{\psi} + \hat{u}_d^{\psi} \quad d = 1, \ldots, D. \tag{12}$$

The Fay-Herriot model and its robust version can also be used for estimation in non-sampled domains. In such cases, the estimated population mean is obtained from a regression model.

The mean square error (MSE) of the population means obtained with the help of the methods described above can be estimated using bootstrap methods. For the Horvitz-Thompson estimator, one can employ the standard replication weights procedure, whereas for the Fay-Herriot and robust Fay-Herriot models, the recommended option is parametric bootstrap proposed by González-Manteiga et al. (2008). Different estimates can be compared in terms of relative root mean square error (RRMSE), calculated as a square root of MSE divided by the estimate.

All methods described in this section are implemented in R language (R Core Team, 2018) in *survey* (Lumley, 2004), *sae* (Molina and Marhuenda, 2015) and *saeRobust* (Warnholz, 2018) packages.

3. Description of the DG1 dataset

The study is based on the data from the DG1 survey, which is the main source of information about Polish entrepreneurs. In the case of small companies, the survey uses a-10% sample of enterprises employing between 10 and 49 persons. These selected companies are then asked to complete a questionnaire about the basic company characteristics (Dehnel 2016).

The sampling design of the DG1 survey enables direct estimation while using the HT estimator, in order to obtain precise estimates at the level of province or for NACE sections. The structure of small companies in selected provinces and sections is presented in Table 1.

Table 1. Number of small enterprises in Mazowieckie and Wielkopolskie provinces in 2011

Province / NACE section	Wielkopolskie	Mazowieckie	Total	Mazowieckie (%)	Wielkopolskie (%)
Manufacturing	2782	2693	21583	12.9	12.5
Construction	1896	1423	12736	14.9	11.2
Trade	4038	2473	22677	17.8	10.9
Transportation	930	617	5000	18.6	12.3

Source: Based on data from the DG1 survey.

The number of manufacturing companies in both provinces is similar – 2782 in Mazowieckie and 2693 in Wielkopolskie, which accounts for about 13% of all enterprises in this section. Construction companies in Mazowieckie province account for 14.9% of the total number of units in the section, while in Wielkopolskie province for 11.2%. The biggest absolute and relative differences could be observed within the trade section – 4038 (17.8%) enterprises in Mazowieckie and 2473 (10.9%) in Wielkopolskie. Transportation is the smallest section, containing only 5000 companies from the whole country. Almost a fifth of them, 930 enterprises, are located in Mazowieckie province, whereas 617 in Wielkopolskie.

A detailed analysis of the number of companies at the level of district shows that there are no transportation enterprises in two districts (Lipski and Żuromiński) of Mazowieckie. These districts were excluded from the estimation process. Figure 3 shows the spatial distribution of population size at the level of district in each province.

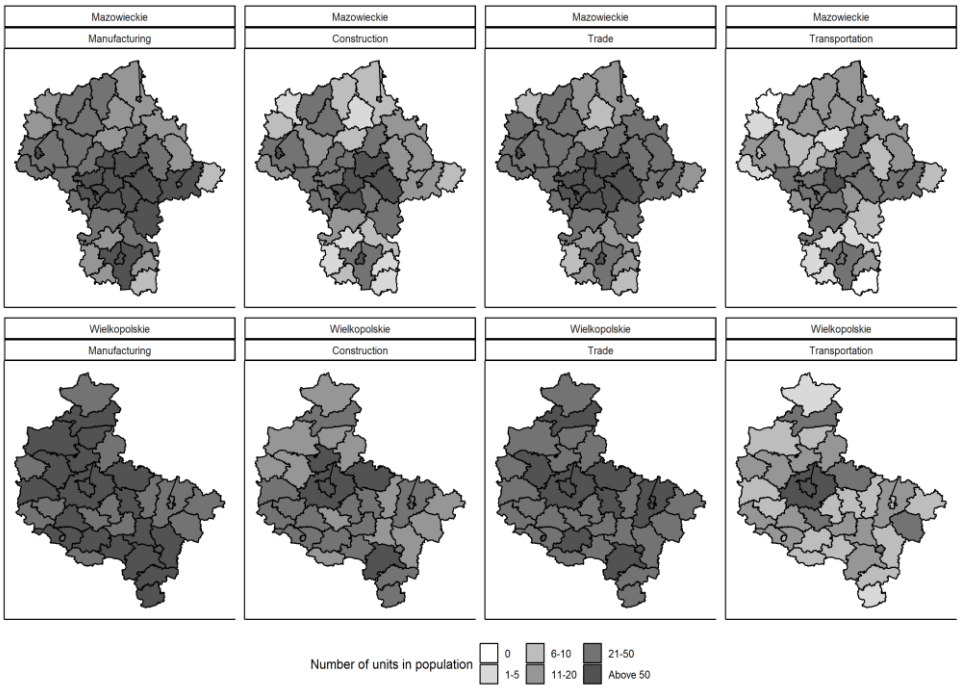

Figure 3. Number of enterprises in the population across districts in Mazowieckie and Wielkopolskie province

Source: Based on data from the DG1 survey.

Unit counts across districts indicate spatial variation as well as differentiation among and within the four NACE sections. The largest numbers of companies operate within the manufacturing and trade sectors, whereas the smallest in the construction and transportation. As already mentioned, the DG1 sample should include at least 10% of the population, and this condition is met. Table 2 shows the number of districts for different intervals of the sample size.

Table 2. Sample size at district level and NACE section

Province / NACE section	Number of units in the sample						
	Non-sampled	1	(1,5]	(5,10]	(10,20]	(20,50]	Above 50
Mazowieckie							
Construction	10	13	14	3	1	0	1
Manufacturing	3	3	15	15	3	2	1
Trade	4	5	21	5	5	1	1
Transportation	15	8	16	0	0	1	0

Table 2. Sample size at district level and NACE section (cont.)

Province / NACE section	Number of units in the sample						
	Non-sampled	1	(1,5]	(5,10]	(10,20]	(20,50]	Above 50
Wielkopolskie							
Construction	4	8	17	4	0	2	0
Manufacturing	0	0	6	14	12	2	1
Trade	0	1	13	14	5	1	1
Transportation	9	12	11	1	2	0	0

Source: Based on data from the DG1 survey.

As could be observed, the sample size in most districts lies within two intervals: (1-5] and (5-10]. There are also many unsampled domains. Interestingly, the number of unsampled districts is much higher in Mazowieckie province than in Wielkopolskie. The biggest sample sizes (above 20 enterprises) occurred in Warsaw, the capital of Poland, for the whole Mazowieckie province, for Poznań, the capital city of Wielkopolskie province, and for Poznański district that belongs to the areas surrounding Poznań (Figure 4). To facilitate the comparison, Figures 3 and 4 have the same legend. White areas denote non-sampled districts. The largest number of non-sampled districts can be observed while analysing the transportation section. This is mostly due to the small population sizes from this section.

In conclusion, the sample size across districts is usually small or even zero (Table 2). In this case, the use of direct estimation is either impossible or is likely to entail big values of the mean square error. One possible solution is to switch from the currently used methodology of estimation to new techniques of indirect estimation offered by small area estimation (SAE). The indirect estimation method was selected for this study mainly because of the characteristics of the enterprise population, i.e. its high degree of differentiation and the presence of outliers.

Figure 4. Number of units sampled across districts in Mazowieckie and Wielkopolskie province
Source: Based on data from the DG1 survey.

4. Estimation of average monthly wages at district level

The wage estimation process for selected domains consists of 4 stages: (1) direct estimation, (2) model fitting, (3) indirect estimation of population means for sampled and non-sampled domains and (4) MSE estimation. At the beginning, Horvitz-Thompson estimates were obtained for all sampled domains, together with their mean square errors. These estimates could be calculated for 201 of the 306 domains. For the remaining domains the sample was either zero or contained only one enterprise, which is not sufficient to estimate mean square error. The resulting direct estimates of the average wages ranged from PLN 2102 (section F, Mazowieckie province) to PLN 6494 (section G, Wielkopolskie). Using the rule of Tukey's fences (Hoaglin et al. 1986) for outlier identification, 11 outliers were detected, of which 8 in Wielkopolskie province. The outlying values represent estimated wages in the capitals of provinces, but also in districts where direct estimation was based on a very small sample, e.g. in Śremski district. The Horvitz-Thompson estimate, calculated there on the basis of just two units, amounted to PLN 6494. Consequently, the relative root mean square error was relatively high, at 40%. The above demonstrates that direct estimates calculated on the basis of small sample sizes are not reliable and cannot be analysed apart from their root mean squared errors.

There are no general precision thresholds for small sample domains. Eurostat (2013) recommends that they should be survey-specific, purpose-specific and should be determined taking users' needs into consideration. According to the guidelines by Statistics Poland, RRMSE of the estimates should not exceed 10%, and those above 20% should not be published (GUS 2013a), which is the usual publication practice (Tzavidis 2018). Table 3 presents domain frequencies for the specific RRMSE intervals.

Table 3. Number of districts with RRMSE of direct estimates within a given interval

Province / NACE section	Direct estimates RRMSE range		
	(0,10]	(10,20]	Above 20
Mazowieckie			
Construction	5	8	5
Manufacturing	16	16	5
Trade	11	19	4
Transportation	8	6	1
Wielkopolskie			
Construction	7	11	5
Manufacturing	23	8	3
Trade	17	13	4
Transportation	4	2	4

Source: Based on data from the DG1 survey.

For each NACE section, there is at least one district in each province where RRMSE of direct estimates exceeds 20%.

To improve the precision of direct estimates, the model-based approach was used. More than 20 variables measured at the level of district were examined as potential auxiliary variables for the model describing wages. The final model contains 4 variables: NACE section (X_1), income of enterprises drawn from the Ministry of Finance's registers (X_2), number of enterprises per 10 thousand of people (X_3) drawn from the Polish National Business Register (REGON), and the average monthly gross wages (X_4). Variable X_1 is categorical and the other three are continuous. Table 4 presents fixed effects of the Fay-Herriot and robust Fay-Herriot model and their standard errors (in brackets).

Table 4. Beta coefficients and their standard errors in the Fay-Herriot and the robust Fay-Herriot model

Variable	*Beta coefficients*	
	Fay-Herriot	Robust Fay-Herriot
Intercept	96.2421 (207.2802)	1011.2566 (231.1063)***
X_1 Construction	270.4663 (88.0113)**	1002.5664 (93.5293)***
X_1 Trade	207.8377 (208.0932)	208.7219 (75.2006)**
X_1 Transportation	-275.6738 (94.4806)**	-424.0788 (98.4990)***
X_2	0.0068 (0.0020)***	0.0071 (0.0021)***
X_3	1.1833 (0.1352)***	1.4082 (0.1607)***
X_4	0.3071 (0.0851)***	-0.0986 (0.1011)
P value significance codes: < 0.001 - ***; < 0.01 - **; < 0.05 - *		

Source: Based on data from the DG1 survey.

Compared to manufacturing, monthly wages in the construction sector were on average PLN 270 higher when estimated using the FH model, and PLN 1002 higher when using the RFH model. Compared to the trade section, they were higher by PLN 208 and PLN 209, respectively. A significant difference between the results yielded by the two models occurred also while estimating average wages in the transportation section. The average wages in that section was estimated at PLN 275 less than in the manufacturing section when using the FH model, and at PLN 424 less than in the manufacturing section when using the RFH model. Higher values of revenue and the number of enterprises per 10 thousand people in districts are correlated with higher wages. In the case of Fay-Herriot model, the average monthly gross wages can be interpreted in the same way. This covariate is insignificant in the robust Fay-Herriot model.

The beta coefficients presented in Table 4 were then used to estimate the average wages for non-sampled domains. Figure 5 presents the distribution of the estimation results.

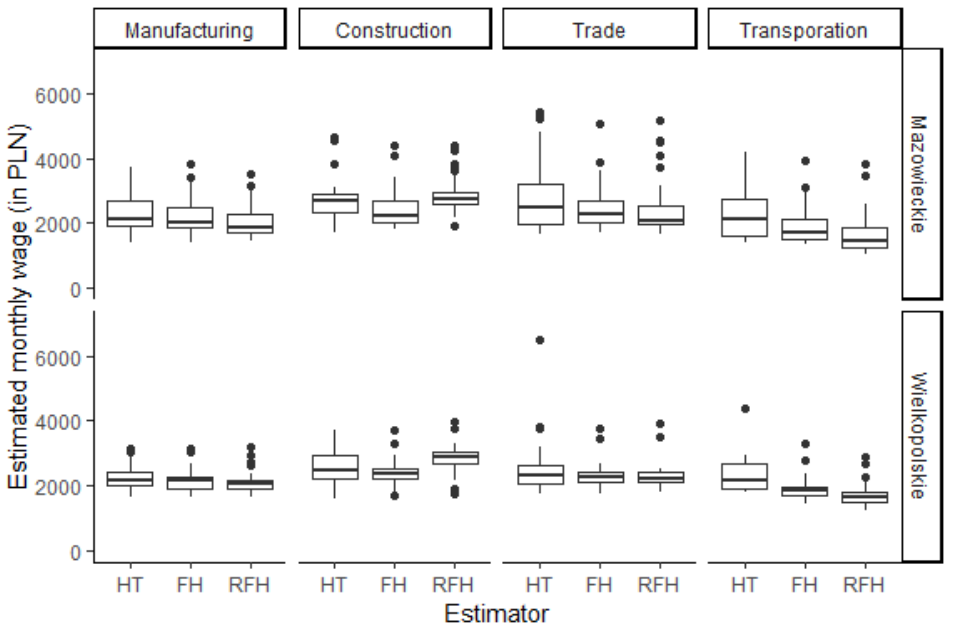

Figure 5. Distribution of estimates of average monthly wages obtained by using three approaches: Horvitz-Thompson estimator (HT), Fay-Herriot model (FH) and robust Fay-Herriot model (RFH)

Source: Based on data from the DG1 survey.

Unlike direct estimation, small area estimation methods make it possible to obtain estimates of an average monthly wages for all target domains (districts). For almost all domains, estimates based on the robust Fay-Herriot model are characterized by the lowest coefficient of variation, except for the trade and transportation sections in Mazowieckie province. The reduced impact of outliers is especially evident in the trade and transportation sections in Wielkopolskie province. The lowest wages are estimated for the transportation section, whereas the highest – over PLN 5000 – for the trade section in Warsaw. The comparison of average wages between the provinces indicates small differences. For example, the estimated average monthly wages in manufacturing companies in Mazowieckie province equals PLN 2060, while in Wielkopolskie it is PLN 2106. Bigger differences could be observed in the maximum wages – in the transportation section in Wielkopolskie it is PLN 2879, while in Mazowieckie PLN 3832.

Another aspect worth analysing is the precision of estimates measured by relative root mean square error. Distributions of these values are presented in Figure 6.

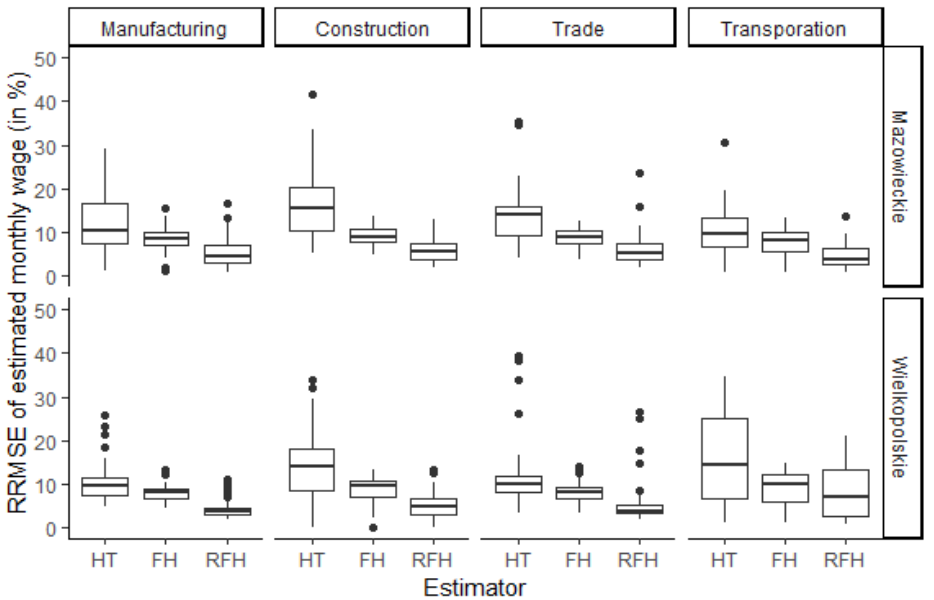

Figure 6. Distribution of RRMSE of the estimates of average monthly wage obtained through the three approaches - Horvitz-Thompson estimator (HT), Fay-Herriot model (FH) and robust Fay-Herriot model (RFH)

Source: Based on data from the DG1 survey.

It should be noted that both the Fay-Herriot model and its robust version are characterized by smaller values of RRMSE than the Horvitz-Thompson estimator. Because of many outliers in the trade section, RRMSE values for the RFH model in a few districts are higher than those for the FH model. However, descriptive statistics indicate that, on the whole, this approach yields more precise estimates than direct estimation. The median value of RRMSE for the robust Fay-Herriot model is at least half the value obtained for the Horvitz-Thompson, in all considered domains. The maximum value of RRMSE exceeds the 20% threshold only for 4 districts. Figure 7 shows a comparison of direct and indirect estimates using scatterplots.

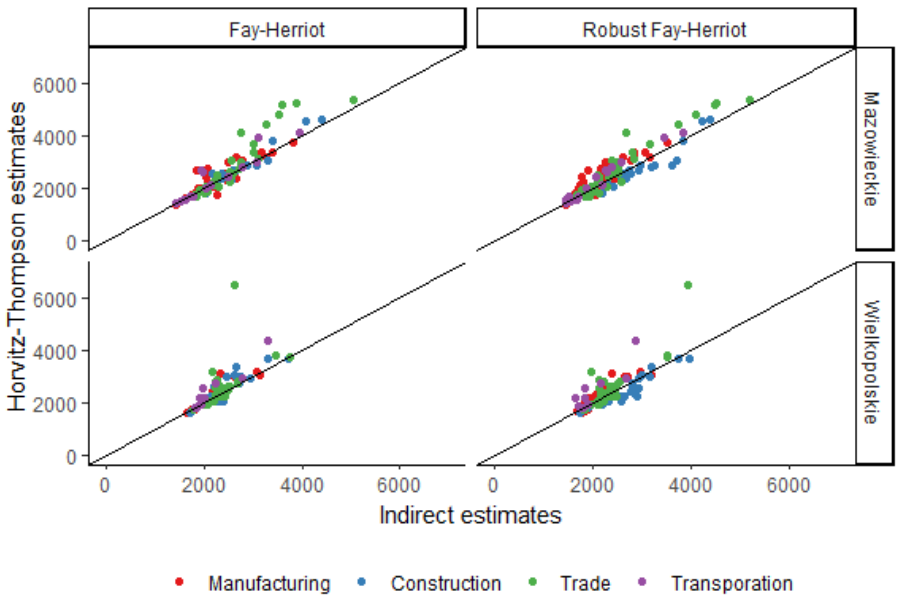

Figure 7. Comparison of estimates of average monthly wage at district level by NACE section
Source: Based on data from the DG1 survey.

The direct and indirect estimates are similar in most cases. The similarity of estimates measured by Pearson's coefficient of linear correlation is the highest for Mazowieckie and equals $r = 0.923$ for the Fay-Herriot model and $r = 0.926$ for the robust Fay-Herriot model. In the case of Wielkopolskie, the values are smaller, mainly due to the impact of the outlier (Śremski district). Pearson's coefficient of linear correlation for direct estimates and the FH model equals $r = 0.758$ and $r = 0.801$ for the robust model.

Figure 8 shows the correlation between RRMSE values obtained using different approaches. There is a difference in the shape of scatterplots for the two indirect approaches. In the case of the Fay-Herriot model, the precision of direct and indirect estimates is similar up to the level of 10%. For higher values of RRMSE, however, the FH model outperforms the direct estimator in terms of precision; this is the result of the γ weight, which accounts for the precision of the direct estimator. The use of the robust FH model improves the precision of most estimates but because values of fixed and random effects differ from those in the FH model, the resulting RRMSE values are relatively higher.

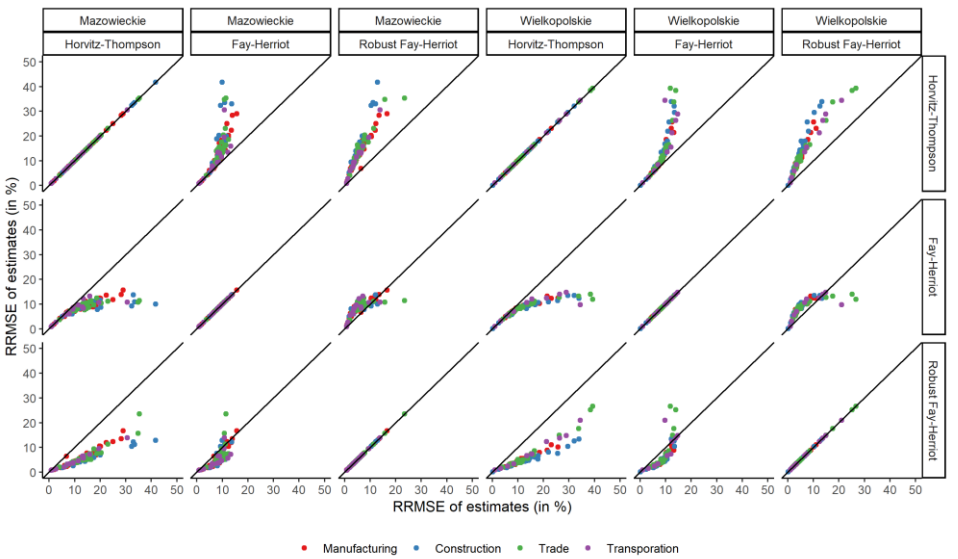

Figure 8. Comparison of RRMSE of estimates of average monthly wage at district level by NACE section

Source: Based on data from the DG1 survey.

As is demonstrated above, the results obtained by means of the Fay-Herriot model and its robust version are, in most cases, more precise in terms of RRMSE than direct estimates. The robust FH model is characterised by the highest average precision. Figure 9 shows mean estimates obtained using this model across districts.

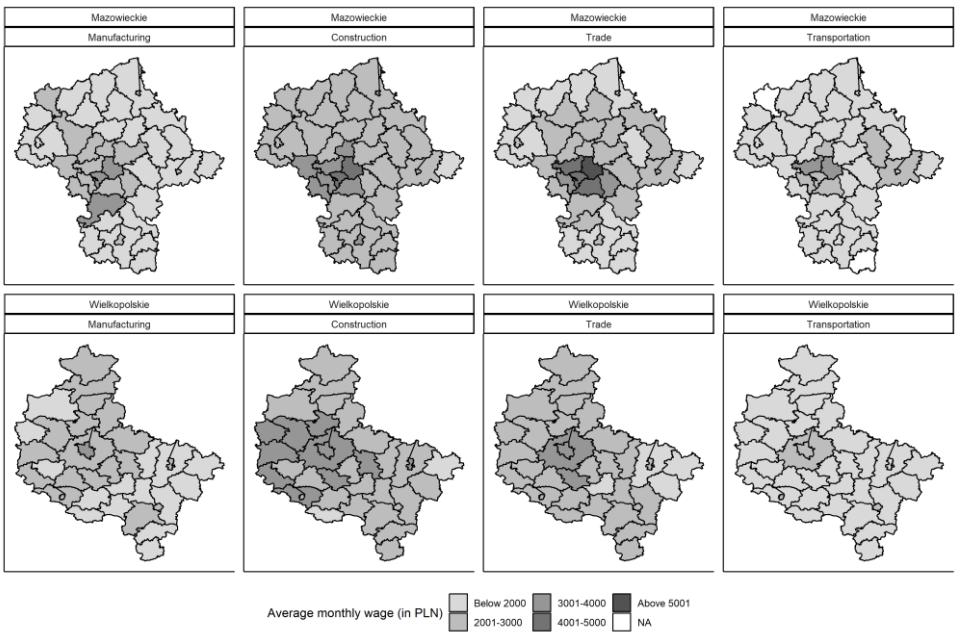

Figure 9. Spatial variation in average monthly wages at district level
Source: Based on data from the DG1 survey.

The highest estimates of average monthly wages were obtained for large cities and the neighbouring territorial units. In Mazowieckie province, the level of average wages was higher than in Wielkopolskie province.

The graphic presentation facilitate the identification of groups of similar districts characterised by similar values of the target variable and those that represent different level of average wages. The biggest difference between the units of Mazowieckie and Wielkopolskie provinces could be observed for province capitals and the neighbouring districts. Using the section of industry as the differentiating criterion, the highest average wages was observed in the group of construction companies, while the lowest in the group of transportation companies.

5. Conclusions

The article describes a study whose aim is to estimate the average monthly wages in the districts of Wielkopolskie and Mazowieckie provinces, for companies representing four major industry sections: manufacturing, construction, trade and transportation. Two methods of indirect estimation were applied: the FH model and its robust version. The analysis focused on districts for which no official estimates had been published before. A potential problem with performing analysis at such a low

level of aggregation is that sample sizes in districts are relatively small. However, thanks to the use of small area estimation methods, it was possible to obtain relatively precise estimates, i.e. with lower values of RRMSE compared to the results obtained using direct estimates. Robust estimation affected outlier values of monthly wages and decreased the range of estimates. Moreover, the use of auxiliary variables for indirect estimation made it possible to obtain estimates of the average wages in non-sampled domains. The study boasts a degree of novelty, because it made it possible to estimate monthly wages in small enterprises according to NACE sections, at the level of districts, for the very first time. The highest estimates of average wages were obtained for large cities and their neighbouring districts in all the four main NACE sections. The results of the study moreover indicate that both Warsaw and Poznań serve as poles of growth for the neighbouring districts.

Further work into the subject will focus on the application of robust area-level models with spatial autocorrelation (Warnholz 2016), and, depending on data availability, unit-level models (Chambers and Tzavidis 2006).

Acknowledgements

The project is financed by the Polish National Science Centre DEC-2015/17/B/HS4/00905.

REFERENCES

CHAMBERS, R., TZAVIDIS, N., (2006). M-quantile models for small area estimation, Biometrika, 93(2), pp. 255–268.

CHAMBERS, R., CHANDRA, H., SALVATI, N., TZAVIDIS, N., (2014). Outlier robust small area estimation, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1), pp.47–69.

DEHNEL, G., (2016). M-estimators in business statistics. Statistics in Transition, New Series, 2016, Vol. 17, No. 4, pp. 749–762, ISSN 1234-7655, http://dx.doi.org/10.21307/stattrans-2016-050.

DEHNEL, G., WAWROWSKI. Ł., (2018). Robust Estimation Of Revenues Of Polish Small Companies By NACE Section And Province, In: M. Papież and S. Śmiech (Eds.), The 12th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena, Conference Proceedings, Foundation of the Cracow University of Economics, Cracow, pp. 110–119.
URL http://dx.doi.org/10.14659/SEMF.2018.01.11.

EUROSTAT, (2013). Handbook on precision requirements and variance estimation for ESS households surveys, European Commission, Belgium, DOI:10.2785/13579.

FAY III, R. E., HERRIOT, R. A., (1979). Estimates of income for small places: an application of James-Stein procedures to census data, Journal of the American Statistical Association, 74(366a), pp. 269–277.

GONZÁLEZ-MANTEIGA, W., LOMBARDIA, M., MOLINA, I., MORALES, D., SANTAMARIA, L., (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. Computational Statistics and Data Analysis, 52(12), pp. 5242–5252, http://EconPapers.repec.org/ RePEc:eee:csdana:v:52:y:2008:i:12:p:5242-5252.

GUADARRAMA, M., MOLINA, I., RAO, J. N. K., (2016). A comparison of small area estimation methods for poverty mapping, Statistics in Transition new series, 1(17), pp. 41–66.

GUS, (2013). Działalność przedsiębiorstw niefinansowych w 2011 r., Zakład Wydawnictw Statystycznych, Warszawa, (Activity of Non-financial Enterprises in 2011), Central Statistical Office of Poland, Warszawa. URL, http://stat.gov.pl/obszary-tematyczne/podmioty-gospodarcze-wyniki-finansowe/przedsiebiorstwa-niefinansowe/dzialalnosc-przedsiebiorstw-niefinansowych-w-2016-r-,2,12.html [Accessed 14 November 2018].

GUS, (2013a). Narodowy Spis Powszechny Ludności i Mieszkań. Ludność. Stan i struktura demograficzno-społeczna, Zakład Wydawnictw Statystycznych, Warszawa, URL http://stat.gov.pl/spisy-powszechne/nsp-2011/nsp-2011-wyniki/ludnosc-stan-i-struktura-demograficzno-spoleczna-nsp-2011,16,1.html [Accessed 14 November 2018].

HOAGLIN, D. C., IGLEWICZ, B., TUKEY, J. W., (1986). Performance of some resistant rules for outlier labeling, Journal of the American Statistical Association, 81(396), pp. 991–999.

HORVITZ, D. G., THOMPSON, D. J., (1952). A generalization of sampling without replacement from a finite universe, Journal of the American statistical Association, 47(260), pp.663–685.

LUMLEY, T., (2004). Analysis of complex survey samples, Journal of Statistical Software 9(1): pp.1–19

MOLINA, I., MARHUENDA, Y., (2015). "sae: An R Package for Small Area Estimation", The R Journal, 7(1), pp. 81–98. https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf

MOLINA, I., RAO, J. N. K., (2010). Small area estimation of poverty indicators. Canadian Journal of Statistics, 38(3), pp. 369–385.

R Core Team, (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RAO, J. N. K., MOLINA, I., (2015). Small area estimation, John Wiley & Sons.

SINHA, S. K., RAO, J. N. K., (2009). Robust small area estimation, Canadian Journal of Statistics, 37(3), pp.381–399.

STRUŻYCKI, M., (2004). Małe i średnie przedsiębiorstwa w gospodarce regionu, PWE, Warszawa.

TZAVIDIS, N., ZHANG, L. C., LUNA, A., SCHMID, T., ROJAS-PERILLA, N., (2018). From start to finish: a framework for the production of small area official statistics, Journal of the Royal Statistical Society: Series A (Statistics in Society), 181(4), pp. 927–979.

WARNHOLZ, S., (2016). Small Area Estimation Using Robust Extensions to Area Level Models, Doctoral dissertation, Freie Universität Berlin.

WARNHOLZ, S., (2018). saeRobust: Robust Small Area Estimation. R package version 0.2.0, https://CRAN.R-project.org/package=saeRobust.

WAWROWSKI, Ł., (2016). The Spatial Fay-Herriot Model in Poverty Estimation, Folia Oeconomica Stetinensia, 16(2), pp. 191–202.