Central European Journal of Economic Modelling and Econometrics

Wage Estimation for Small Business Statistics in Poland Using Area-Level Models with Register and Survey Data

Grażyna Dehnel*, Łukasz Wawrowski†

Submitted: 14.11.2023, Accepted: 17.12.2024

Abstract

In this article, we propose an application of extended Fay-Herriot models with the Generalized Variance Function and the use of benchmarking for estimating short-term business characteristics at a low level of aggregation. This method of estimation, based on the area-level approach, does not require access to unit-level data, which is a clear advantage from the perspective of data protection requirements. The purpose of the study was to estimate the average monthly wage in small companies (employing between 10 and 49 people) in Poland in 2011 for domains defined by the interaction of subregion and type of economic activities. The properties of small area estimators are examined using real data from a short-term business sample survey and administrative registers. Results indicate that the proposed models increase the precision of estimates, which can be used for detailed regional economic studies.

Keywords: Fay-Herriot model, business statistics, robust estimation, small area estimation

JEL Classification: C13, C15, C51, M20

*Department of Statistics, Poznań University of Economics and Business, Poznań, Poland; e-mail: grazyna.dehnel@ue.poznan.pl; ORCID: 0000-0002-0072-9681

†Computer Science Research Center, Łukasiewicz Research Network – Institute of Innovative Technologies EMAG, Katowice, Poland; e-mail: lukasz.wawrowski@emag.lukasiewicz.gov.pl; ORCID: 0000-0002-1201-5344

1 Introduction

Nowadays it is expected that information about economic indicators will be analyzed at a low level of aggregation, in a timely and systematic manner. For this to happen, one must have access to the latest, detailed, and precise estimates about businesses for disaggregated domains. Existing business surveys are often not designed to meet the growing demand for small domain characteristics. Designing a new statistical survey is usually expensive and time-consuming. It is therefore necessary to turn to new, more sophisticated estimation methods to provide such estimates.

One possible solution is to use small area estimation (SAE) methods. The main idea of these methods is to supplement existing survey data with auxiliary variables, e.g. from a census or administrative registers, in order to improve estimation accuracy (Harmening et al., 2020). The term *small area (domain)* refers to any cross-classifications or geographical areas which a given survey was not designed for, which means that the actual sample sizes are not sufficient to obtain precise estimates (Rao and Molina, 2015). In the case of business statistics, the estimation of economic indicators proves particularly challenging owing to problems, such as the uneven distribution of economic variables in space and across sectors defined by the type of economic activity, which is the result of a non-uniform concentration of economic activity. Furthermore, distributions of data about businesses tend to be strongly asymmetrical, highly variable and concentrated, which is usually caused by the presence of outliers. As a result, it is difficult to retain the properties of classical estimators used in sample surveys (Riviére, 2002).

The growing demand for detailed business characteristics is driving the development of SAE methods, which can be classified as robust methods. Recently, there have been many publications presenting results of studies in this area (Militino et al., 2015; Schmid et al., 2016; Ferrante and Pacei, 2017; Dehnel and Wawrowski, 2020; Fabrizi et al., 2018; Luzi et al., 2018). This trend is a response to the growing availability of register data, which can be used in estimation as a source of auxiliary variables.

Most SAE techniques rely heavily on parametric statistical models. When used in sample surveys, they usually have to be adapted to specific sample designs. Some aspects of model adjustment to a design-based approach in business surveys are considered by Hidiroglou and Smith (2005). Falorsi and Righi (2015) provide a general framework for sample design in multivariate and multi-domain surveys. Burgard et al. (2014) analyzed the effects of different sampling designs used in business surveys on the effectiveness of SAE methods.

Business survey variables may contain a large proportion of zero-valued observations (Chandra and Chambers, 2011; Karlberg et al., 2015) and many of them are right-skewed. The problem of positively skewed variables can also be solved by log-transforming data for area-level models. In this way, normal linear mixed models can be used used (Fay and Herriot, 1979; Kreutzmann et al., 2019). However, one should take into account an additional source of bias from back-transforming the

final estimates on the original scale (Slud and Maiti, 2006; Sugasawa and Kubokawa, 2017). Karlberg (2000) proposed a transformation-retransformation estimator for highly skewed populations in the presence of zeroes. Fabrizi et al. (2018) presented area level modeling on the log scale in a business survey considering a Bayesian approach and effectively dealing with back-transformation. The above-mentioned publications rely on examples from Italy, Poland, and the United Kingdom. There are many spatio-temporal studies based on Italian data (Martini and Giannini, 2020), while papers from the UK are rather theoretical. It shows that access to small area business statistics for scientific purposes is rather difficult.

This article describes a study whose goal was to develop a new method for handling data from a Polish short-term business survey and presents the possibility of applying small domain approaches in order to estimate business characteristics at the level of aggregation for which official statistics are not published. To deal with specific properties of business data three different area-level approaches were used, all of which are based on the Fay-Herriot (FH) model: robust, spatial, and robust-spatial. In addition, the study procedure involved the Generalized Variance Function, which was used to decrease the variance of direct estimates resulting from small sample sizes. Moreover, the authors used benchmarking to adjust estimates to their corresponding values at a higher level of aggregation, for which official statistics are available. This is the first attempt in Polish short-term business statistics to adjust this procedure with the goal of producing estimates of the monthly wage in small companies for the domain defined as the interaction of subregion (NUTS 3 level of the Nomenclature of territorial units for statistics) and economic activity of companies defined according to the NACE classification (NACE is fully compliant with the Polish classification of activities PKD-2007 (Eurostat, 2008)). The study is a continuation of research efforts to develop a method of estimating characteristics of small businesses for domains for which official statistics are not available (Dehnel and Wawrowski, 2019b, a, 2020). One of the main criteria for selecting area-level modeling was the fact that this approach does not require access to unit-level data and is less subject to restrictions associated with preserving statistical confidentiality (Moura et al., 2017). This is an evident advantage from the perspective of data protection requirement. The use of area-level methods for business data is also advisable since they enable the construction of more suitable and effective models for skewed data (Ferrante and Pacei, 2017). It can be noted that the area-level approach can produce estimates even for areas with no sample information and can provide more reliable estimates by exploiting information from contextual or area-level effects in the small area distribution of the target variable. Ignoring these effects in unit-level models can lead to biased estimates (Namazi-Rad and Steel, 2011; Luzi et al., 2018).

The empirical study was based on official statistics from a business survey known as DG1, data from the business register, and administrative registers maintained by the Ministry of Finance and the Social Insurance Institution (ZUS). Owing to data availability, the analysis was conducted only for the year 2011. The DG1 survey,

officially called "Activity of non-financial small enterprises", is the largest survey in Polish short-term business statistics. It is used to collect data from businesses employing over 9 people. Each month about 30 thousand companies participate in the survey. Administrative registers are maintained mainly for administrative purposes, which is why their exploitation for statistical purposes can present certain challenges. In the study, registers were used as the source of auxiliary variables.
The purpose of the article is to examine the applicability of extended Fay-Herriot (FH) models with the Generalized Variance Function to estimate the monthly wage of employees in small enterprises in Poland at a level of aggregation for which official statistics are not available, namely for domains defined by the interaction of subregion and economic activity. The estimation was conducted independently for four largest types of economic activity (NACE sections) – Manufacturing, Construction, Trade, and Transportation – at the level of subregions in an attempt to meet users' needs for statistics at lower levels of aggregation. Information about small companies across subregions can be used to improve the monitoring of the economic, demographic, and social situation, as well as the efficiency of managing resources, such as the workforce, raw materials, and infrastructure. Additionally, in EU countries, estimates at this level serve as the basis for allocating structural and investment funds. The study is based on real data from the DG1 survey. The tax register maintained by the Ministry of Finance and a social insurance register maintained by ZUS are used as sources of auxiliary variables. The novelty of the study consists in using register data and one of the small area estimation methods in order to estimate of business characteristics at the level of subregions. We designed a general framework for the production of small area statistics dedicated to short-term business official statistics. By using real data and a simulation study we demonstrated that the proposed approach delivers more precise and less biased estimates compared to the basic model.
The rest of the article consists of four parts. The following section outlines methodological considerations of the analysis and is followed by a description of the data sources used for the estimation. The fourth section provides details about the empirical study and contains a summary of the results and their interpretation. The article ends with conclusions and suggestions for further work.

2 Methods

The methods describes in this section were used to estimate the average monthly wage for domains defined as the interaction of NACE section and NUTS 3 unit. We start with the Horvitz-Thompson (direct) estimator, which is the standard approach used in official statistics. It does not require model assumptions but can be very inefficient for areas with very small sample sizes, leading to very imprecise estimates. We expected that direct estimates of the monthly wage for the target domains would also be affected by this problem. One possible way to overcome it is to use linear-mixed models, especially the Fay-Herriot (FH) model (Fay and Herriot, 1979) and

its extensions. In its basic form, the FH model is a weighted combination of the direct and the regression-synthetic estimators. It relies on aggregated data, so there are no confidentiality issues as in the case of unit-level data. The FH model can be easily extended by accounting for spatial, temporal, or robust components. Our goal was to test these additional effects for wage estimation, as this variable tends to be characterized by spatial relationships and outliers.

2.1 Horvitz-Thompson estimator

The direct estimator (Horvitz and Thompson, 1952) is the traditional approach used in survey methodology for estimating population means and totals. Let U denote a population, which consists of N units divided into D domains $U_1, ..., U_D$ with the population size denoted by N_d, where $d = 1, ..., D$. The sample, denoted by s, such that $s \in U$, can be also divided into $s_1, ..., s_D$ with sample size n_d for each domain. Let y_{di} denote the value of the target variable for the i-th unit in domain d. Moreover, 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 given by the formula:

$$\hat{\theta}_d^{HT} = \frac{1}{n_d} \sum_{i=1}^{n_d} y_{di} w_{di}. \tag{1}$$

The Horvitz-Thompson (HT) estimator is design-unbiased and efficient for large sample sizes n_d (Pfeffermann, 2013). For a sampling design without second-order inclusion probabilities, the variance estimator is given by:

$$\hat{var}(\hat{\theta}_d^{HT}) = N_d^{-2} \sum_{i \in s_d} w_{di}(w_{di} - 1) y_{di}^2. \tag{2}$$

In applications, the variance is approximated by applying linearization or bootstrap methods (Molina et al., 2022). For some domains, the sample size can be very small or even zero, which leads to big variances of the HT estimator and situations in which it is impossible to obtain direct estimates.

2.2 Fay-Herriot model

Fay and Herriot proposed a model that made it possible to estimate income in small areas in the US. The FH model is defined in two stages. Firstly, let's assume that the true value of the parameter can be described by a linear model with an 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 an random area effect with the distribution $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$.

The model in (3) is called the *linking model*. Its parameters cannot be estimated, since the true values of θ_d are not observed. Nevertheless, using direct estimates based on survey unit-level data it can be assumed that the direct estimator of the population mean is the sum of the true value of the parameter and the random (survey) error. The model defined in this way is called the *sampling model* and it is given by the formula:

$$\hat{\theta}_d^{HT} = \theta_d + e_d, \tag{4}$$

where $e_d \overset{ind}{\sim} N(0, \sigma_{ed}^2)$. Error variances σ_{ed}^2 need to be assumed as known because otherwise, the number of unknown parameters would be greater than the number of observations used to fit the FH model. In practice, these variances are estimated based on survey data (Molina et al., 2022).
Then, the combination of equations (3) and (4) gives the Fay-Herriot model:

$$\hat{\theta}_d^{HT} = x_d'\beta + u_d + e_d. \tag{5}$$

The random area effect variance (σ_u^2) needs to be estimated in order to obtain the Empirical Best Linear Unbiased Predictor (EBLUP) of the Fay-Herriot model. This can be achieved by various methods e.g. Fay-Herriot method, Prasad-Rao method, REML, or ML (Chambers et al., 2014; Rao and Molina, 2015; Dehnel and Wawrowski, 2020).

2.3 Spatial Fay-Herriot model

The spatial Fay-Herriot model is the classical FH model with a spatial autocorrelation coefficient and a proximity matrix (Pratesi and Salvati, 2008). In a generalized form, the model is given by

$$\hat{\theta}_d^{HT} = x_d^T\beta + (I - \rho W)^{-1}u_d + e_d, d = 1, ..., D, \tag{6}$$

where: $\hat{\theta}_d^{HT}$ - the estimated value of the mean for area d, x_d^T – a $p \times 1$ vector of explanatory variables for area d, ρ – a spatial autocorrelation coefficient, W – a proximity matrix, u_d – random area effect with $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$, e_d – random error with $e_d \overset{ind}{\sim} N(0, \psi_d)$ with known variance ψ_d.
The proximity matrix W represents the neighborhood relationship between analyzed areas, while ρ is a measure of spatial correlation between random effects of neighboring areas. In the original proximity matrix W^0, the diagonal elements are equal to 0, while the remaining ones are equal to 1 if two areas are contiguous, and 0 otherwise. This is a first-order contingency matrix (common border matrix), also known as the queen matrix. The matrix W is derived from the matrix W^0 by dividing each row element by the row total. In this way, the matrix is expressed in row-standardized form, where row elements add up to 1.
The Spatial Empirical Best Linear Unbiased Predictor of this model is given by:

$$\hat{\theta}_d^{SFH} = x_d^T\tilde{\beta} + b_d^T\left\{\hat{\sigma}_u^2\left[(I - \hat{\rho}W)(I - \hat{\rho}W^T)\right]^{-1}\right\} \times$$

$$\times \left\{ diag(\psi_d) + \hat{\sigma}_u^2 \left[(I - \hat{\rho}W)(I - \hat{\rho}W^T) \right]^{-1} \right\}^{-1} (\hat{\theta}_d^{HT} - x_d^T \tilde{\beta}), \tag{7}$$

where b_d^T is the vector $1 \times D$ with value 1 in the d-th position and $\hat{\rho}$ is the estimated spatial correlation coefficient.

2.4 Robust Fay-Herriot model and spatial robust Fay-Herriot model

The Fay-Herriot model is an example of a shrinkage estimator, but 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 β and σ_u^2 in the Fay-Herriot model with their outlier robust alternatives. Detailed equations are presented in (Sinha and Rao, 2009; Chambers et al., 2014) and in this paper, we present only the last stage of these derivatives. A robustified equation for u_i in the Fay-Herriot model is given by:

$$\psi_d^{-1/2} \psi_{b_1} [\psi_d^{1/2} (\hat{\theta}_d - x_d^T \beta - u_i)] - \sigma_u^{-1} \psi_{b_2} (\sigma_u^{-1} u_d) = 0, \ d = 1, ..., D. \tag{8}$$

Robustified ML equations for β and σ_u^2 are following:

$$\beta : \sum_{d=1}^{D} x_d (\sigma_u^2 + \psi_d)^{-1/2} \psi_b(r_d) = 0$$

$$\sigma_u^2 : \sum_{d=1}^{D} (\sigma_u^2 + \psi_d)[\psi_b(r_d) - c]$$

where $r_d = (\sigma_u^2 + \psi_d)^{-1/2} (\hat{\theta}_d - x_d^T \beta)$ and $c = E[\psi_b^2(u)]$ with $u \sim N(0, 1)$. Finally, a robust estimator (REBLUP) is given by:

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

The methods presented above handle symmetric outliers in u_d and e_d. The robust spatial Fay-Herriot model is a special case of the model in (5).

2.5 Mean square error estimation

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

2.6 Generalized Variance Function

Generally, estimates of $v\hat{a}r(\hat{\theta}_d^{Dir})$ are not precise since the sample sizes at this level are generally very small. An alternative estimation procedure uses auxiliary data and applies a log-linear model for the variance estimates of direct estimators. Schall (1991) proposed the general GVF method that fits the model:

$$log(v\hat{a}r(\hat{\theta}_d^{HT})) = x_d^v \beta^v + e_d^v. \tag{9}$$

Many applications can be found in the literature, one of which (Fuquene et al., 2019) utilizes the following formula:

$$log(v\hat{a}r(\hat{\theta}_d^{HT})) = \beta_1 \hat{\theta}_d^{HT} + \beta_2 \sqrt{n_d} + \beta_0, \tag{10}$$

where $\hat{\theta}_d^{HT}$ denotes the direct estimator and n_d – the sample size.

2.7 Model diagnostics methods

In order to use the Fay-Herriot model, a number of assumptions, mostly relating to normality, must be satisfied, in particular the normal distribution of random effects and random errors. The best method to verify these assumptions is a quantile-quantile plot. Apart from displaying the shape of the distribution, it also helps to identify outliers.

The quality of the resulting estimates is generally assessed in terms of the mean squared error. However, it is worth noting that a more complete description of the estimation results requires a disjoint evaluation of each component of the MSE. In other words, both efficiency and bias values should be considered independently. Brown et al. (2001) proposed a test that can be used to verify the hypothesis that differences between direct estimates and those obtained from area-level models are not statistically significant. The Wald statistic takes the following form:

$$W(\hat{\theta}_d^{FH}) = \sum_{d=1}^{D} \frac{(\hat{\theta}_d^{HT} - \hat{\theta}_d^{FH})^2}{v\hat{a}r(\hat{\theta}_d^{HT}) + \hat{MSE}(\hat{\theta}_d^{FH})}. \tag{11}$$

The value of the test statistic is compared with the quantile of the distribution χ^2 with the number of degrees of freedom equal to the number of units considered in the analysis. The null hypothesis represents the equality of the expected value of the direct estimate and the indirect estimate.

2.8 Benchmark

The idea of benchmarking is that the aggregated FH estimates should sum up to estimates obtained at a higher level τ:

$$\sum_{d=1}^{D} \xi_d \hat{\theta}_d^{FH,bench} = \tau, \tag{12}$$

where ξ_d denotes the share of the population size of each area in the total population size (N_d/N).
Datta et al. (2011) proposed a general estimator:

$$\hat{\theta}_d^{FH,bench} = \hat{\theta}_d^{FH} + \left(\sum_{d=1}^{D} \frac{\xi_d^2}{\phi_d} \right)^{-1} \left(\tau - \sum_{d=1}^{D} \xi_d \hat{\theta}_d^{FH} \right) \frac{\xi_d}{\phi_d}. \tag{13}$$

Depending on the weight ϕ_d there are 3 benchmarking options:

1. $\phi_d = \xi_d$,
2. $\phi_d = \xi_d / \hat{\theta}_d^{FH}$,
3. $\phi_d = \xi_d / \hat{MSE}(\hat{\theta}_d^{FH})$.

In the first case, all FH estimates are adjusted by the same factor, in the second and third options, the adjustment factor depends on FH estimates or the MSE of FH estimates. In general, larger adjustments will be made for domains with larger FH/MSE estimates (Kreutzmann et al., 2019).

3 Data

The empirical study was based on official statistics from the DG1 business survey, officially called "Activity of non-financial small enterprises", data from the business register, and administrative registers maintained by the Ministry of Finance and the Social Insurance Institution (ZUS). Owing to data availability, the analysis was conducted only for the year 2011.

3.1 Business survey

The DG1 business survey is the largest survey in Polish short-term business statistics. It is used to collect data from businesses employing over 9 people. The DG1 survey is carried out monthly and collects essential data about each business unit, its activity, products, including information about basic measures of economic activity, such as sales revenue (from goods and services), the number of employees, gross

wages, wholesale and retail sales, excise tax and product specific subsidies (Dehnel and Wawrowski, 2020). The survey includes a 10% stratified sample of businesses employing between 10 and 49 persons. The sample is allocated to make sure that shares of individual NACE divisions are consistent with the structure of the business population in a given province. Strata are also created for different forms of ownership (private and state-owned companies).
The sampling frame consists of about 98 thousand units. 19 thousand of them are large and medium-sized companies (with more than 49 employees), while the remaining 80 thousand are small companies (with 10–49 employees). Each month about 30 thousand units participate in the survey. The sampling design of the DG1 survey for small businesses enables direct estimation by means of the HT estimator to obtain precise estimates at the province level or for NACE sections.

3.2 Administrative registers

Administrative registers (AR) are the basis of many organizational activities in the country. These information systems are maintained mainly for administrative purposes, which is why their exploitation for statistical purposes can present certain challenges. In order to estimate the average monthly wage, the authors used information maintained by the Ministry of Finance and the Social Insurance Institution. These administrative data are regularly provided to Statistics Poland to support the construction of the business register and survey administration. They contain reliable information about such variables as revenue, costs, income, and the number of employees for nearly the entire business population. In the study, these registers were used as the source of auxiliary variables.
Using data stored in administrative registers the authors were able to check the quality of statistical data by comparing values collected by official statistics with those registered by tax offices. The comparison was performed for the revenue variable and revealed that revenue values from the DG1 are highly correlated with those registered by tax offices. While about 55% of companies report virtually identical values, over 30% of businesses under-report their revenue in the DG1 in comparison with their tax statements.

4 Results

As mentioned earlier, the goal of the study was to investigate the applicability of of extended Fay-Herriot (FH) models with the GVF for estimating the monthly wage in small companies in Poland. In this section, we present results obtained in the simulation study in which properties of the Generalized Variance Function were investigated. In the second part, real data from the DG1 survey and administrative registers were used to estimate monthly wage.

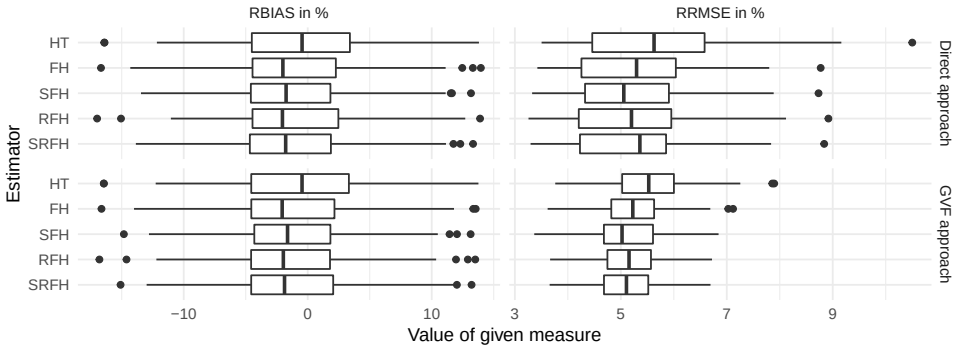
4.1 Design-based simulation study

To assess the properties of estimators presented in Section 2 we conducted a design-based simulation study. The population data was generated using data from the DG1 survey – each row was repeated according to the sample weight. In this way, we obtained the true value of the average wage, which was used as the reference value in the comparison of results.
The goal of the simulation was to analyse the bias and MSE of the studied estimators. The simulation consisted of the following steps:

1. Use the synthetic population to draw $I = 100$ 10% samples $(i = 1, ..., I)$ using the same sampling design as the one used in the DG1 survey.
2. Use each of the estimators $\in \{HT, FH, SFH, RFH, SRFH\}$ and variance estimates $\in \{direct, GVF\}$ to produce estimates of the average wage for each sample I
3. Calculate:
$$BIAS(\hat{\theta}) = \frac{1}{I}\sum_{i=1}^{I}(\hat{\theta}^{(i)} - \theta), \tag{14}$$
$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(\hat{\theta}^{(i)} - \theta)^2}. \tag{15}$$
4. Calculate another performance measure expressed in % — relative bias (RBIAS) and relative root mean square error (RRMSE):
$$RBIAS(\hat{\theta}) = \frac{BIAS(\hat{\theta})}{\theta} \cdot 100, \tag{16}$$
$$RRMSE(\hat{\theta}) = \frac{RMSE(\hat{\theta})}{\theta} \cdot 100. \tag{17}$$

Figure 1 presents the performance of each estimator under the two approaches to the estimation of variance. Each point on the plot represents one of the NUTS 3 units for which a given measure was calculated. The median value of relative bias closest to zero can be observed for the Horvitz-Thompson estimator: -0.46% regardless of the variance estimation approach. The best results were obtained using the spatial Fay-Herriot model. In this case, the median value of relative bias is slightly lower for the GVF approach (-1.63%) than for the direct variance estimation approach (-1.74%).
Compared to the GVF approach, the values of relative root mean square error have a much larger interquartile range when the direct variance estimator is used. Differences between median values of RRMSE are not so evident. The lowest RRMSE is observed

Figure 1: Performance of domain predictions – relative bias (RBIAS) and relative root mean square error (RRMSE)

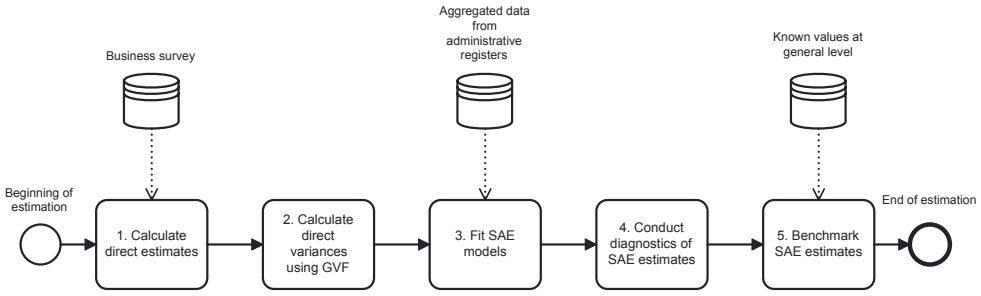


for the spatial Fay-Herriot model, with 5.06% in the case of the direct approach, compared to 5.02% when the GVF approach is used.
The design-based simulation study shows that the model-based estimators could improve the RRMSE of the target variable without increasing bias compared with the Horvitz-Thompson estimator. The results indicate that the use of the generalized variance function in the estimation process helps to decrease the variance of RRMSE estimates, though the improvement was not found to be statistically significant.

4.2 Application to real data

This section presents results obtained by fitting the Fay-Herriot model and its extended versions in order to estimate the average monthly wage of employees working in small companies in Poland for domains defined as the interaction of the NUTS 3 level and NACE section. Figure 2 presents the 5 stages of the estimation procedure.

Figure 2: Estimation procedure for the area-level model



First (stage 1), we calculated direct estimates of the average wage for the target domains using data from the DG1 survey, which is the main source of wage information. Unfortunately, owing to the small sample sizes in some domains (NUTS 3 x NACE), these estimates are not reliable. Therefore, in the next stage, we tried to improve the direct estimates first by smoothing the variance of direct estimates using the Generalized Variance Function, and then, by using data from administrative registers. Administrative registers contain information about units in the entire population but they do not include information about wages. However, they do contain covariates which are correlated with our variable of interest. By combining data from these two sources in SAE models we can improve the quality of average wage estimates for the target domains. After fitting the models, we conducted some diagnostics and moved on to the final step of the process, in which we benchmarked the new estimates to ensure they were consistent with the average wage calculated at the country level. All calculations were conducted in the statistical software environment R (R Core Team, 2023), mainly using the emdi package (Kreutzmann et al., 2019).
In our study, we analysed wages for four NACE sections: Manufacturing (C), Construction (F), Trade (G), and Transportation (H). Table 1 contains descriptive statistics of the sample for each NACE section at NUTS 3 level.

Table 1: Descriptive statistics of the sample size by NACE section and the number of NUTS 3 units

NACE	No. of NUTS 3 units	Min	Mean	Std. dev.	Median	Max
C: manufacturing	73	12	49	24	41	131
F: construction	73	4	18	13	15	71
G: trade	73	14	50	38	39	298
H: transportation	68	2	8	6	6	37

In Poland, there 73 NUTS 3 units, called subregions. As can be seen, the sample contained companies from all subregions for three NACE sections, while in the case of the Transportation section, five subregions were not represented in the sample. The section with the largest number of companies is Trade, where the minimum number of companies selected from a single NUTS 3 unit is equal to 14. The number of companies in the Transportation section was much smaller, compared to the other three – the median sample size is equal to 6. Given the high degree of variability in sample size across the four NACE sections, we were able to assess the performance of SAE methods under different scenarios.
The data from the sample were first used to estimate the average monthly wage using the direct Horvitz-Thompson estimator. These estimates were used as a dependent variable in the models during the process of indirect estimation and were treated as reference values for the evaluation of the results. There were 15 outliers across all 287 domains, especially in bigger cities, which were not found in other statistical

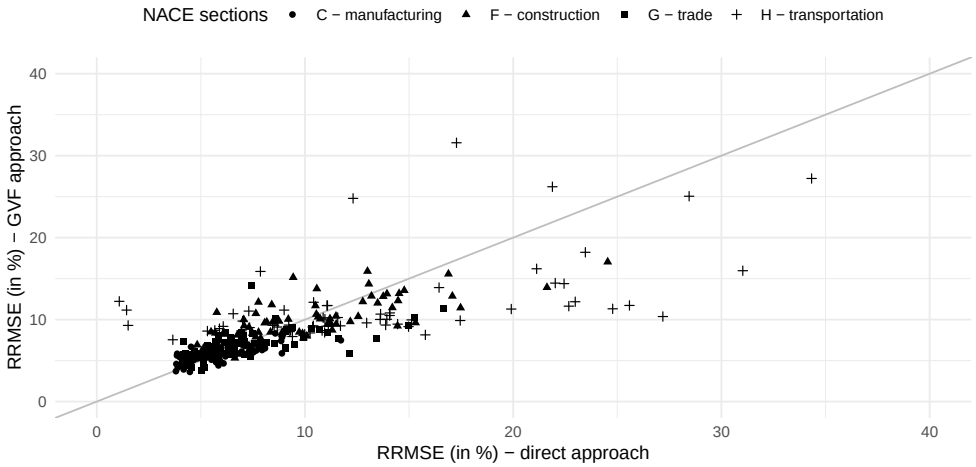
sources containing wage data, such as Statistical Year-books for selected subregions. At this stage of the analysis, the outliers were left unchanged in the dataset. They are partially plotted in Figure 4.
The maximum precision of estimates, measured by RRMSE, ranged from 12% (for Manufacturing) to 34% (for Transportation). Given the small sample size and imprecise estimates of variance, the second stage involved the use of the generalized variance function (GVF). After obtaining direct estimates, the calculation of pooling variances using GVF was the next stage of the estimation procedure, which was intended to improve the quality of estimates of the average monthly wage at the NUTS 3 level. Pooled variance values for each NACE section were estimated using formula (10). Beta coefficients of the models with GVF are presented in Table 2.

Table 2: Beta coefficients with standard errors (in brackets) of the generalized variance function by NACE section

NACE	Intercept	$\hat{\theta}_d^{HT}$	$\sqrt{n_d}$
C: manufacturing	8.6161 (0.2867)	0.0011 (0.0001)	-0.2225 (0.0287)
F: construction	8.8226 (0.2900)	0.0015 (0.0001)	-0.4904 (0.0491)
G: trade	8.6819 (0.2567)	0.0010 (0.0001)	-0.2978 (0.0360)
H: transportation	7.6706 (0.5794)	0.0017 (0.0002)	-0.3164 (0.1939)

The model's regression coefficients were statistically significant at $\alpha = 0.05$ (for Transportation – at significance level $\alpha = 0.1$). For all sections, β_1 coefficients were positive, while β_2 coefficients were negative. This is consistent with the expected direction of the relationship between the estimator variance and the number of units in the sample and the value of the estimate, as determined by the formula for estimator variance. Higher estimates are associated with higher variance values, while higher sample sizes result in lower variance (Rao and Molina, 2015). Coefficients of variance calculated using the GVF were strongly or moderately correlated with those obtained for direct estimates (see Figure 3). Pearson's correlation coefficient was equal to 0.67 for Manufacturing, 0.75 for Construction, 0.63 for Trade, and 0.55 for Transportation. By applying the GVF, it was possible to decrease the variability of variance estimates, especially for Transportation, where the number of sampled companies was the smallest. In extreme cases, the sample size was 2, while the population included 62 units (rzeszowski subregion). This resulted in unreliable and unstable direct estimates of variance. The inclusion of the GVF made it possible to deal with near zero values of direct estimates of variance and decreased the degree of their variability. In line with the proposed procedure (see Figure 2), pooled variance values calculated with the GVF are used in SAE models instead of direct variance values.
The analysis accounted for the possibility of spatial relationships. The level of variables for domains of interest is often determined by geographical conditions and

Figure 3: Comparison of RRMSE for direct estimates and those obtained using GVF by NACE sections and NUTS 3 units



is affected by the values of these variables observed in neighboring domains. In order to verify the existence of spatial autocorrelation, the authors used direct estimates of the average monthly wage to calculate values of Moran's I and Geary's C (see Table 3).

Table 3: Spatial autocorrelation coefficients by NACE section with p-values given in brackets

NACE	Moran's I	Geary's C
C: manufacturing	0.3684 (0.0000)	0.4857 (0.0000)
F: construction	0.2420 (0.0005)	0.6138 (0.0000)
G: trade	0.3935 (0.0000)	0.4669 (0.0000)
H: transportation	0.1729 (0.0127)	0.6653 (0.0005)

All spatial autocorrelation coefficients are statistically significant. The lowest value of Moran's I is observed for Transportation (0.1729), and the highest – for Trade (0.3925). As regards Geary's C, the highest values were obtained for Transportation (0.6653) and the lowest – for Trade (0.4669). These results indicate the existence of a weak positive spatial correlation.

The third stage consisted in estimating the average monthly wage by applying four models: the Fay-Herriot model (FH), the robust Fay-Herriot model (RFH), the spatial Fay-Herriot model (SFH) and the spatial robust Fay-Herriot model (SRFH).

Estimation precision for these FH models was assessed using the MSE estimated through the bootstrap method.
Access to information from administrative registers was limited owing to legal restrictions and only a small set of variables was available. Auxiliary information used in the models was selected based on data completeness and correlation with the dependent variable. We looked for possible associations between these variables and wages. The finally selected auxiliary variables are listed below:

i) logarithm of average revenue (from the Ministry of Finance),

ii) the number of companies per 100 thousand population (from the National Register of Business Entities),

iii) the average number of employees (from the Social Insurance Institution),

iv) an indicator variable connected with population size (takes the value of 1 if the number of people living in a given NUTS 3 unit is higher than the third quantile of the distribution of population size values for all NUTS 3 units, and 0 otherwise).

Register data could be used after being linked with records in the DG1 survey using the REGON identifier, which is unique for each entrepreneur in Poland. The integrated dataset, consisting of administrative and DG1 survey data, made it possible to verify values of such variables as revenue and the number of employees, which were available in at least two sources. In some cases, it was necessary to impute the number of employees in the register based on the value from the DG1 survey. After data editing was complete, values of each variable were aggregated at the level of domains defined as the interaction of NACE section and NUTS 3 unit to enable estimation by means of area-level models. The models were estimated for each NACE section independently. Descriptive statistics of the dependent variable (from the DG1 survey) and continuous auxiliary variables (from the administrative registers) are presented in Table 4.
Estimates of the average monthly wage obtained by applying the methods described in Section 2 with random effect at NUTS 3 level, along with their precision measured by relative root mean square error, are shown in Figure 4.
Estimates obtained by applying the indirect estimators are similar. In 2011, the average monthly wage in Manufacturing ranged from PLN 1724 to 3527, with a mean of PLN 2367. In Construction, the range was bigger: over PLN 2200 (minimum – PLN 1747, maximum – PLN 3993). The average wage was also higher and amounted to PLN 2553. A similar average wage was recorded for companies in the Trade section – PLN 2548, which is also characterized by the highest variability in wages (from PLN 1847 to 4888). In transportation companies, the average wage was the lowest of the four sections – PLN 2237, with minimum and maximum values equal to PLN 1638 and PLN 3044, respectively.

Table 4: Descriptive statistics of the dependent and continuous auxiliary variables

Source	Variable name	Minimum	Median	Mean	Std. dev.	Maximum
DG1	average wage	1519.47	2396.86	2517.90	575.12	5417.21
AR	logarithm of average revenue	7.62	8.54	8.63	0.50	11.12
AR	the number of enterprises per 100 thousand population	0.48	3.80	3.87	2.16	12.26
AR	the average number of employees	16.54	19.71	19.80	1.28	23.58

The application of indirect estimation improved estimation precision for all four sections. In the Manufacturing section, the biggest gain in precision, compared to direct estimates, was observed for the spatial Fay-Herriot model (SFH): the maximum value of the relative root mean square error (RRMSE) for direct estimation (HT) – 8.32% – was reduced to 7.39%. Mean RRMSE values were equal to 5.84% for HT and 5.18% for SFH. When it comes to Construction, FH models helped to reduce the maximum RRMSE value from 17.04% to 11.00%. In the case of Trade, the biggest decrease in mean RRMSE values was obtained for RFH model (from 6.90% for HT to 5.61% for RFH). In the least numerous Transportation section, the biggest gain in precision was achieved by applying the SFH model. The maximum RRMSE value dropped from 31.57% to 12.50%, while the mean value – from 11.64% to 8.23%.

4.3 Small area model diagnostics

The forth stage of the study involved performing model diagnostics. First, the authors tested the assumed normality of random effects and residuals and concluded that they were approximately normally distributed in all four sections. Figure 5 presents quantile-quantile plots for the above-mentioned values.

In the next stage of the analysis, the authors conducted tests proposed by (Brown et al., 2001), which can be used to verify the hypothesis that differences between direct estimates and those obtained from area-level models are not statistically significant. The hypothesis is verified by the Wald test. The test results clearly indicate that there are no statistically significant differences between direct estimates and those obtained by applying the four different variants of the FH model. In all cases, the p-value was close to 1, which indicates that model-based estimates might be design-based unbiased (see Figure 6). Values of the correlation coefficient between

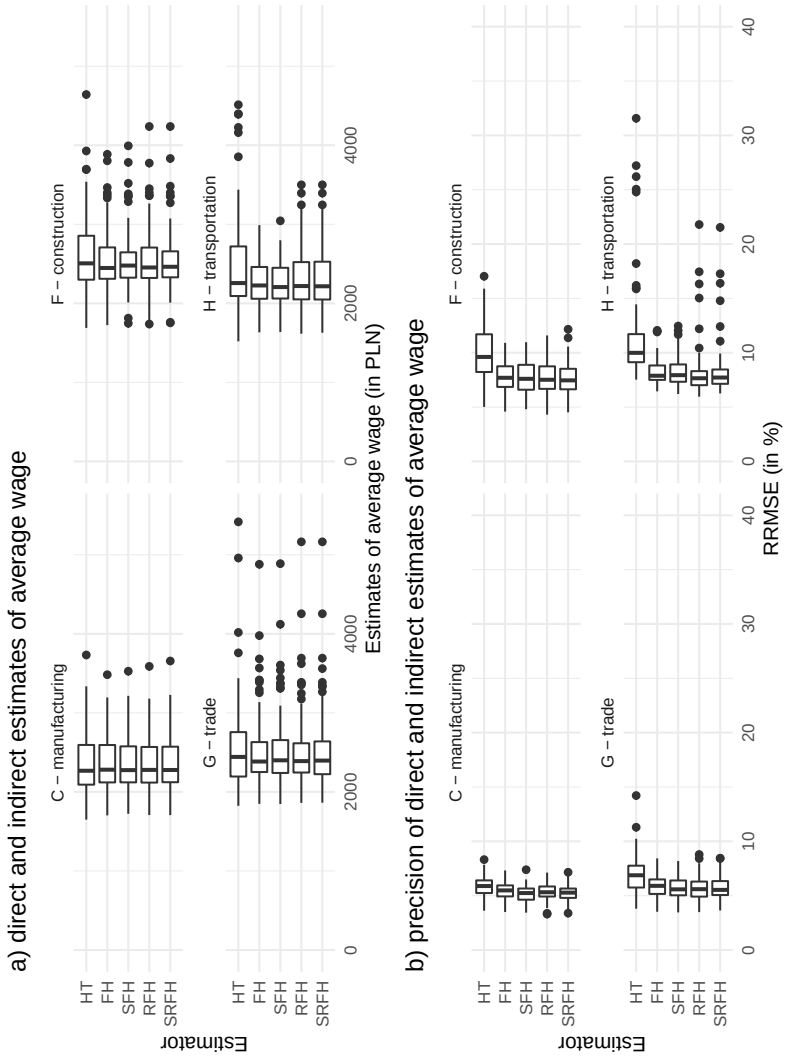


Figure 4: Estimates of the average wage produced by the direct estimator (HT), Fay-Herriot model (FH), spatial Fay-Herriot model (SFH), robust Fay-Herriot model (RFH) and spatial robust Fay-Herriot model (SRFH) along with corresponding measures of precision

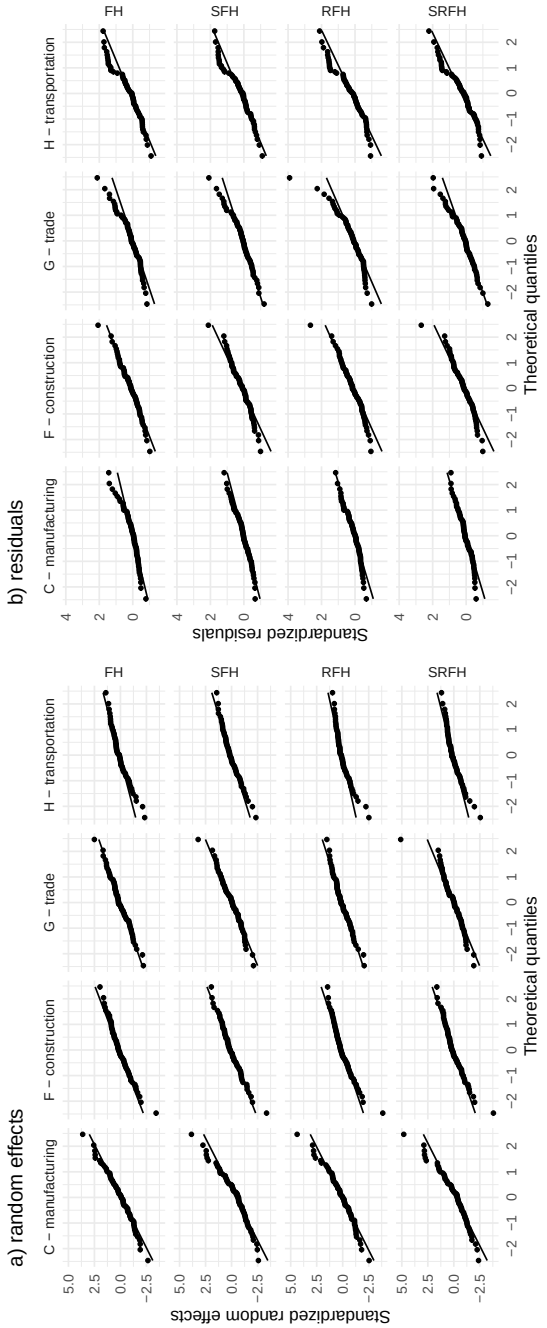
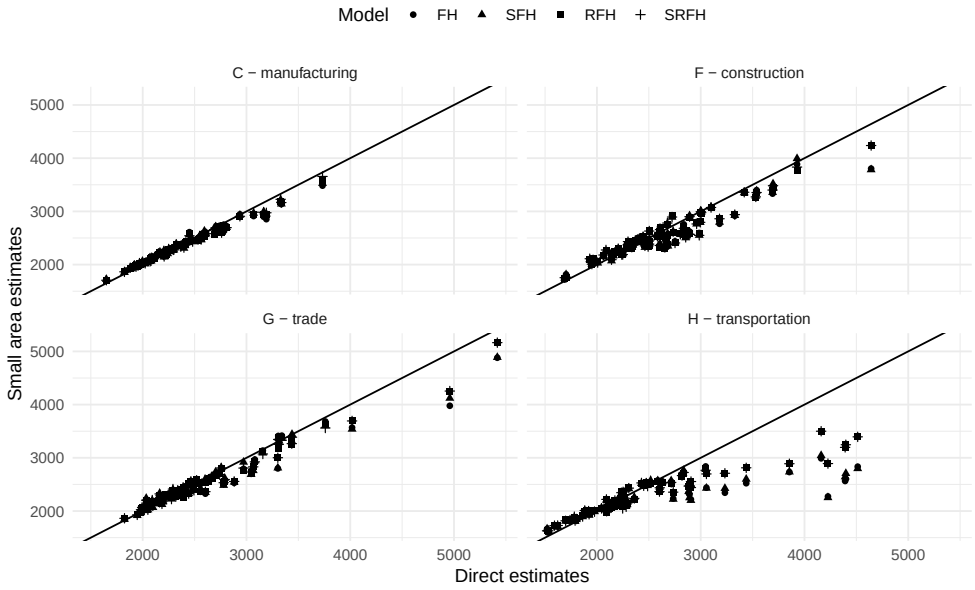


Figure 5: Quantile-quantile plots for random effects and residuals of all analyzed models

Figure 6: Comparison of direct estimates with model-based estimates



estimates of the average monthly wage for all estimators ranged from 0.46 (Transportation) to 0.85 (Trade).
We also calculated another measure that can serve as a criterion for selecting the best model – loglikelihood (see Table 5).

Table 5: Loglikelihood values of analyzed models

NACE	FH	SFH	RFH	SRFH
C: manufacturing	-486.83	-472.83	-479.04	-472.35
F: construction	-495.32	-494.21	-488.26	-487.70
G: trade	-489.80	-485.73	-476.68	-474.83
H: transportation	-601.90	-597.44	-534.65	-533.80

The highest values were obtained for the most complex approach – spatial robust Fay-Herriot model. For the manufacturing section, loglikelihood values are very close in the case of SFH and SRFH. In the rest NACE sections, there is a very small difference between these values for RFH and SRFH.

4.4 Benchmarking

In the final stage we benchmarked our estimates to known values of the average monthly wage at country level for each of the four NACE sections, using formula (13) with the 3rd benchmarking option, which makes it possible to modify the model-based estimates for all domains so as to get the same aggregate estimate for the NACE section at country level. The auxiliary benchmark information came from the publication entitled "Activity of non-financial enterprises in 2011" (Statistics Poland, 2013). Benchmarked estimates that correspond to official statistics of the average monthly wage for the four NACE sections together with mean values of direct estimates and model-based estimates are presented in Table 6.

Table 6: Mean estimates for HT, SFH and benchmarked SFH estimates

NACE	HT	SFH	Benchmark SFH
C: manufacturing	2460.43	2366.54	2460.00
F: construction	2832.83	2553.41	2383.00
G: trade	2956.26	2547.56	2402.00
H: transportation	2511.81	2236.60	2149.00

The greatest degree of similarity between the results can be seen for Manufacturing. Both direct and indirect estimates are closest to official statistics. The biggest discrepancies between the estimates can be found in Transportation, which was the least numerous section. In the case of the remaining two sections (Construction and Trade), the average monthly wage, regardless of the estimator used, was overestimated in comparison to official figures. However, SFH estimates were closer to the official values than direct HT estimates.

Benchmarked SFH estimates are shown in the choropleth map in Figure 7. Figure 7 clearly illustrates the degree of variation in average wages within provinces. However, official statistics cannot be used to conduct analyses of business indicators at this level of aggregation. In the group of manufacturing companies, the highest average monthly wage was recorded in Poland's major cities and in some of their surrounding areas. One can also notice lower levels of the average wage in companies based in the eastern part of the country. A similar situation can be observed with respect to construction companies. Relatively high values of the average wage can only be found in subregions around the main cities. The majority of subregions are home to companies where the average monthly wage is in the range between PLN 2100 and 2560.

In the case of trade enterprises, there are several subregions where the average wage is in the top range, among others Poznań, Szczecin, Warsaw, and Wrocław. In most subregions, however, the level of monthly wages is low (between PLN 1590 to 2540).

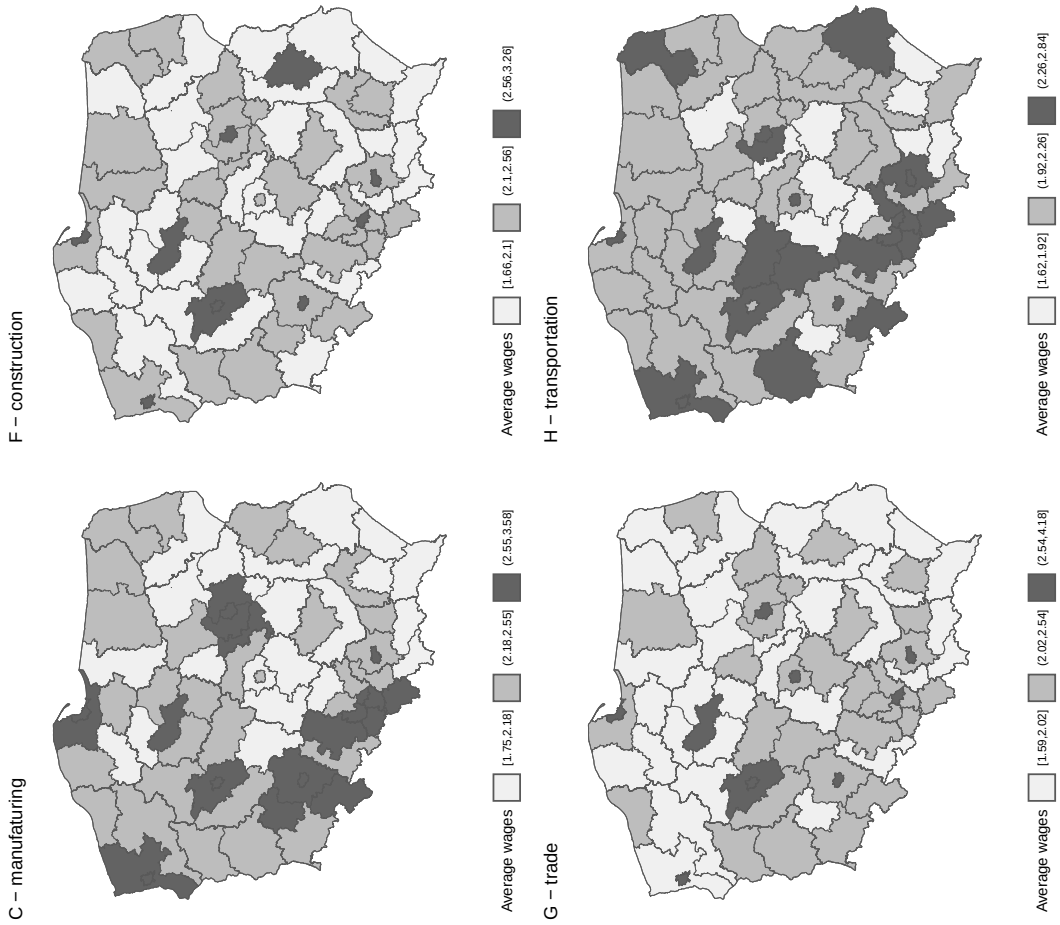


Figure 7: Spatial variation of average wage estimates by NACE sections and NUTS 3 units

Finally, as regards transportation companies, compared to the other three sections, the number of subregions with the highest level of average wages (from PLN 2260 to 2840) is the biggest.
Generally, benchmarked estimates produced by the SFH model combined with information about their geographical distribution seem to confirm the assumptions of the theory of growth poles. One can clearly see a higher level of average wages in major cities, which function as growth poles, in comparison with the neighboring subregions, where the level of average wages is clearly lower.

5 Conclusions

To the best of the authors' knowledge, the study described in this article is the first attempt to estimate one of the key characteristics of entrepreneurship – the average wage in the small business sector in Poland at the level of NUTS 3 units and by NACE section. The proposed approach helps to lower the level of aggregation for information about wages broken down by economic activity section (NACE codes), providing estimates for 73 subregions (NUTS 3 level), instead of 16 provinces (NUTS 2 level).
The study was conducted in the area of short-term statistics and was aimed at estimating the monthly wage in small companies. The application of small area estimation methods made it possible to exploit auxiliary information stored in administrative registers. While registers are the only alternative source of detailed and reliable data about companiess, the information they contain cannot be used as auxiliary variables in modeling without complex data processing.
Because of small sample sizes in the domains of interest and unreliable estimates of variance obtained using the direct approach, the authors applied the generalized variance function (GVF).
The estimation procedure involved the application of classical and extended Fay-Herriot models (robust, spatial, and robust spatial). By applying indirect estimation, it was possible to estimate the average monthly wage for domains that were not represented in the DG1 sample (transport companies in five NUTS 3 units). Estimates produced by the spatial Fay-Herriot model were found to be the best in terms of precision. Compared to HT estimates, the SFH model offers the biggest improvement in terms of RRMSE values.
In addition, the study involved the application of benchmarking in order to adjust estimates to their corresponding values at a higher level of aggregation, for which official statistics are available. The benchmarking process revealed that the use of direct estimation at a lower level of aggregation can result in considerable overestimation.
The main problem associated with the use of benchmarking is the lack of a direct estimator of the SME for benchmarked estimates. Their precision can only be estimated indirectly.

To sum up, we can state that administrative registers are a comprehensive and reliable source of auxiliary information for estimation purposes. The use of registers can considerably improve the quality of official statistics by reducing the negative impact of missing data, which is a common problem in statistical reporting. However, exclusive use of administrative data in short-term statistics is a special challenge since it requires quick and frequent access to registers, and also regular data editing and data integration.

The indirect estimates obtained in the study are an obvious improvement on direct estimates, which are produced without the use of administrative data. With constant access to administrative registers, it would be possible to produce estimates of business characteristics at lower levels of aggregation, which have not been published until now owing to insufficient precision. The proposed estimation procedure can be particularly useful for official statistics, where it is increasingly important to evaluate the quality of estimation. The Generalized Variance Function makes it possible to decrease the variability of variance of direct estimates given small sample sizes. By using benchmarking, it is possible to ensure that final estimates are consistent with those calculated for a higher level of aggregation.

The results of the study contribute to research on short-term entrepreneurship at the regional and local levels. They allow, among other things, to create choropleth maps that help to assess the wage variation across NUTS 3 units. As demonstrated in the study, the proposed methodology can be used not only to estimate the monthly wages but also to provide a relevant contribution to research regarding other short-term business characteristics like revenue, cost, income, etc. Therefore, the proposed approach involving modified extended Fay-Herriot models can be used as an evaluation tool supporting efforts to monitor progress on business unit activity and in projects such as European Funds for Modern Economy. These estimates can also be treated as a starting point of analysis for local government authorities, policymakers, and other stakeholders involved in planning an effective local/regional economic policy. The proposed estimation method is easy to adapt to existing data and its implementation should not involve many additional costs.

Our analysis confirms the need for further in-depth research on the proposed approach. Additionally, it would be advisable to consider the level of districts, since the demand for this type of data is high. However, descending to a lower level of aggregation will undoubtedly increase the methodological obstacles we faced at the NUTS3 level. A potential problem with performing analysis at such a low level of aggregation is that sample sizes in districts are relatively small (including districts with no entities eligible for sampling). To overcome this, it is necessary to look for ways in which the methods presented in this article can be further modified, e.g. by introducing stratification of companies de-pending on the number of employees or combining domains with similar characteristics in order to "increase" sample size. Our research shows that to obtain district level estimates, statistical offices must take action. It would also be useful to consider taxonomic methods, which can be used to identify similar units. Our results

can be also the first step in analyses of wage inequalities (Pereira and Galego, 2015). In future studies, we also intend to explore the possibility of using transformed values of the dependent variable (Rojas-Perilla et al., 2020) and look for other sources of auxiliary information, such as Big Data (Marchetti et al., 2015).

References

[1] Brown G., Chambers R., Heady P., Heasman D., (2001), Evaluation of small area estimation methods an–application to unemployment estimates from the UK LFS, [in:] *Proceedings of Statistics Canada Symposium*.

[2] Burgard J. P., Münnich R., Zimmermann T., (2014), The impact of sampling designs on small area estimates for business data, *Journal of Official Statistics* 30(4), 749–771.

[3] 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), 47–69.

[4] Chandra H., Chambers R., (2011), Small area estimation for skewed data in presence of zeros, *Calcutta Statistical Association Bulletin* 63(1-4), 241–258.

[5] Datta G., Ghosh M., Steorts R., Maples J., (2011), Bayesian benchmarking with applications to small area estimation, *Test* 20(3), 574–588.

[6] Dehnel G., Wawrowski Ł., (2019a), Estimation of the average wage in polish small companies using the robust approach, *Przegląd Statystyczny* 66(3), 200–213.

[7] Dehnel G., Wawrowski Ł., (2019b), Unit level models in the assessment of monthly wages of small enterprises employees, [in:] *The 13th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*, Conference Proceedings, The Socio-Economic Modelling and Forecasting, pages 35–43.

[8] Dehnel G., Wawrowski Ł., (2020), Robust estimation of wages in small enterprises: the application to poland's districts, *Statistics in Transition new series*, 21(1), 137–157.

[9] Eurostat (2008), NACE Rev. 2 – Statistical classification of economic activities in the European Community, Eurostat Methodologies and Working papers, Luxembourg, European Commission.

[10] Fabrizi E., Ferrante M. R., Trivisano C., (2018), Bayesian small area estimation for skewed business survey variables, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67(4), 861–879.

[11] Falorsi P. D., Righi P., (2015), Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys, *Survey methodology* 41(1), 215–236.

[12] Fay R., Herriot R., (1979), Estimates of income for small places: an application of james-stein procedures to census data, *Journal of the American Statistical Association* 74(366a), 269–277.

[13] Ferrante M. R., Pacei S., (2017), Small domain estimation of business statistics by using multivariate skew normal models, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180(4), 1057–1088.

[14] Fuquene J., Cristancho C., Ospina M., Morales D., (2019), Prevalence of international migration: an alternative for small area estimation, arXiv preprint, arXiv:1905.00353.

[15] González-Manteiga W., Lombardía M. J., Molina I., Morales D., Santamaría L., (2008), Analytic and bootstrap approximations of prediction errors under a multivariate fay-herriot model, *Computational Statistics & Data Analysis* 52(12), 5242–5252.

[16] Harmening S., Kreutzmann A.-K., Pannier S., Salvati N., Schmid T., (2020), A framework for producing small area estimates based on area-level models in R, *The R Journal* 15(1), 316–341.

[17] Hidiroglou M., Smith P., (2005), Developing small area estimates for business surveys at the ONS, *Statistics in Transition* 7(3), 527–539.

[18] 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), 663–685.

[19] Karlberg F., (2000), Survey estimation for highly skewed populations in the presence of zeroes, *Journal of Official Statistics* 16(3), 229–242.

[20] Karlberg F., (2015), Small area estimation for skewed data in the presence of zeroes, *Statistics in Transition, New Series* 16(4),541–562.

[21] Kreutzmann A.-K., Pannier S., Rojas-Perilla N., Schmid T., Templ M., Tzavidis N., (2019), The R package emdi for estimating and mapping regionally disaggregated indicators, *Journal of Statistical Software* 91(7), 1–33.

[22] Luzi O., Solari F., Rocci F., (2018), A study of small area estimation for italian structural business statistics, *Journal of Official statistics* 34(2), 543–555.

[23] Marchetti S., Giusti C., Pratesi M., Salvati N., Giannotti F., Pedreschi D., Rinzivillo S., Pappalardo L., Gabrielli L., (2015), Small area model-based estimators using big data sources, *Journal of Official Statistics* 31(2), 263–281.

[24] Martini B., Giannini M., (2020), Regional wage and productivity in italy: a spatio-temporal analysis, *Spatial Economic Analysis* 15(4), 392–412.

[25] Militino A., Ugarte M., Goicoa T., (2015), Deriving small area estimates from information technology business surveys, *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 178(4), 1051–1067.

[26] Molina I., Corral P., Nguyen M., (2022), Estimation of poverty and inequality in small areas: review and discussion, *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* 31(4), 1143–1166.

[27] Moura F. A., Neves A. F., Silva D. B. d. N., (2017), Small area models for skewed brazilian business survey data, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180(4), 1039–1055.

[28] Namazi-Rad M.-R., Steel D. G., (2011), Contextual effects in modeling for small domain estimation, [in:] *Proceedings of the 4th Applied Statistics Education and Research Collaboration (ASEARC) Conference*, [eds.:] E. Beh, L. Park, K. Russell (Eds.), Wollongong, University of Wollongong, 12–14.

[29] Pereira J., Galego A., (2015), Intra-regional wage inequality in portugal, *Spatial Economic Analysis* 10(1), 79–101.

[30] Pfeffermann D., (2013), New important developments in small area estimation, *Statistical Science* 28(1), 40–68.

[31] Pratesi M., Salvati N., (2008), Small area estimation: the eblup estimator based on spatially correlated random area effects, *Statistical methods and applications* 17(1), 113–141.

[32] R Core Team, (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

[33] Rao J. N. K., Molina I., (2015), *Small area estimation*, John Wiley & Sons.

[34] Riviére P., (2002), What makes business statistics special? *International Statistical Review* 70(1), 145–159.

[35] Rojas-Perilla N., Pannier S., Schmid T., Tzavidis N., (2020), Data-driven transformations in small area estimation, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183(1), 121–148.

[36] Schall R., (1991), Estimation in generalized linear models with random effects, *Biometrika* 78(4), 719–727.

[37] Schmid T., Tzavidis N., Münnich R., Chambers R., (2016), Outlier robust small-area estimation under spatial correlation, *Scandinavian Journal of Statistics* 43(3), 806–826.

[38] Sinha S. K., Rao J. N. K., (2009), Robust small area estimation, *Canadian Journal of Statistics* 37(3), 381–399.

[39] Slud E. V., Maiti T., (2006), Mean-squared error estimation in transformed fay-herriot models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(2), 239–257.

[40] Statistics Poland, (2013), *Activity of non-financial enterprises in 2011*, Statistical Publishing Establishment.

[41] Sugasawa S., Kubokawa T., (2017), Transforming response values in small area prediction, *Computational Statistics & Data Analysis* 114, 47–60.