Robust estimation of revenues of Polish small companies by NACE section and province

Grażyna Dehnel[1], Łukasz Wawrowski[2]

Abstract

Sample surveys conducted by the Central Statistical Office are currently the main source of information about revenues earned by small companies. Given the sample size, sampling scheme and the estimation method used in the survey, reliable estimates can only be produced for domains at the level of country, province or section of business classification. The market economy, however, creates a demand for local level information about businesses and economic conditions, which is provided on a regular basis at short intervals.

The article describes an empirical study designed to test a small area estimation method. The goal of the study is to apply a robust version of the Fay-Herriot model, which, unlike the classical Fay-Herriot model, makes it possible to meet the assumption of normality of random effects under the presence of outliers. These alternative models will be supplied with auxiliary variables in order to estimate revenues of small businesses (with between 10 and 49 employees). Other sources of data used in the analysis include the DG1 report, Poland's largest enterprise survey, and administrative registers. The study is expected to provide information about patterns and characteristics of the small business sector in Poland for territorial units on low level of aggregation.

Keywords: *small area estimation, indirect estimation, robust Fay-Herriot model, administrative registers, business statistics*

JEL Classification: C13, C51, M20

DOI: 10.14659/SEMF.2018.01.11

1 Introduction

The shape of Poland's modern-day economy is the result of dynamic changes during the period of economic transformation. One of the sectors that plays a significant role in the development of the economy is the sector of small companies (employing between 10 and 49 persons), which currently includes about 57,000 small businesses. Small companies are characterised by a high degree of flexibility, profitability and efficiency of economic activity. There is also a strong correlation between the development of small companies and regional development. This impact can be observed in both directions: a higher level of regional development encourages entrepreneurs to start business activity, at the same time, however,

[1] Corresponding author: Poznań University of Economics and Business, Department of Statistics, al. Niepodległości 10, 61-875 Poznań, g.dehnel@ue.poznan.pl.

[2] Poznań University of Economics and Business, Department of Statistics, al. Niepodległości 10, 61-875 Poznań, lukasz.wawrowski@ue.poznan.pl.

a growth in the number of small companies contributes to the improvement of the region's economic situation (Główny Urząd Statystyczny, 2017).

Taking into consideration the classification of economic activity, manufacturing and trade are the two most important sections. Companies conducting activities classified into these two categories account for 38% of all small businesses, and their revenues make up about 70% of all revenues in this sector. They also provide 50% of jobs that exist in the small business sector (Główny Urząd Statystyczny, 2017). An analysis of financial results of small trading and manufacturing companies is conducted and published by the Central Statistical Office only at country level. However, given the demand for more detailed information expressed by data users, the present article describes a study whose goal was to estimate certain variables at the level of province (NUTS 2). So, the target domain of estimation is province cross-classified by NACE section. Information about net revenues in the domains is not available in official publications of the Central Statistical Office (Dehnel, 2017).

The aim of the study was to estimate two variables: net revenues from the sale of goods and materials (*SH*) and net revenues from the sale of products (*SW*) for companies which employ from 10 to 49 employees. These characteristics were estimated using direct estimates from DG1 survey and auxiliary variables from administrative registers. This study is a continuation of the study described in Dehnel et al. (2017).

The article is divided into three parts. The first one provides a description of the DG1 survey. The second, theoretical part is devoted to the presentation of estimators used in the study. Estimation results are described in the third part. The article ends with conclusions and suggestions for further research.

2 DG1 survey

The study is based on data from the DG1 survey, which is the main source of information about Polish enterprises. The survey includes a 10% sample of small companies (employing more than 10 people), which are asked to complete a questionnaire about basic characteristics of the company.

By applying the Horvitz-Thompson (1952) estimator to DG1 data it is only possible to produce reliable direct estimates at province level or for NACE sections. There is, however, a growing demand for more detailed information about companies' characteristics.

3 Fay-Herriot model and its robust version

The Fay-Herriot model belongs to the class of area-level models, which means that it utilizes aggregated data instead of unit-level information. This approach was developed in 1979 as a tool for estimating income for small areas in the USA (Fay and Herriot, 1979). The construction of a Fay-Herriot model is divided into two stages. Firstly, it is assumed that the direct estimate is unbiased and can be written as the sum of the true value of the estimated parameter and random error:

$$\hat{\theta}_d = \theta_d + e_d . \tag{1}$$

Where $e_d \overset{ind}{\sim} N(0, \sigma_{ed}^2)$. In practice, variance σ_{ed}^2 is unknown and is estimated based on survey data.

In the second stage, the true value of the parameter is treated as a dependent variable in the linear model with area random effect:

$$\theta_d = x_d^T \beta + u_d \tag{2}$$

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 combing equations (1) and (2) we obtain the Fay-Herriot model given by:

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

EBLUP (Empirical Best Linear Unbiased Predictor) is the estimator of the Fay-Herriot model and is given by the following formula:

$$\hat{\theta}_d^{FH} = x_d^T \hat{\beta} + \hat{u}_d = \hat{\gamma}_d \hat{\theta}_d + (1 - \hat{\gamma}_d) x_d^T \hat{\beta} \qquad d = 1,.....D \tag{4}$$

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

EBLUP is a weighted average of the direct estimate and the regression model. Weight $\hat{\gamma}_d$ measures the uncertainty of the regression model. If sample variance $\hat{\sigma}_{ed}^2$ is small, then the larger part of the final estimate will come from the direct estimate (Boonstra and Buelens, 2011). Between-area variance $\hat{\sigma}_u^2$, like sample variance, is also unknown and must be estimated. It can be done with many techniques e.g. the Fay-Herriot method, Prasad-Rao method, ML or REML (Rao, 2015).

The robust version of the Fay-Herriot model uses Huber (1981) influence function to restrict the influence of u_d and e_d. The detailed process of robustifying all equations is

described in Sinha and Rao (2009) and Warnholz (2016). Robust EBLUP is given by the formula:

$$\hat{\theta}_d^{RFH} = x_d^T \hat{\beta}^{w} + \hat{u}_d^{w} \qquad d = 1,.....D \tag{5}$$

For unsampled domains and if between-area variance is equal to zero, indirect estimation is only based on the regression model.

The mean square error (MSE) of the parameters can be estimated by the parametric bootstrap method proposed by Gonzalez-Manteiga et. al. (2008). MSE can be used to calculate the relative root mean square error (RRMSE), which is treated as a common measure of precision for all approaches.

4 Estimation of net revenue with indirect estimation

In the study we constructed models for two dependent variables. The first one was *net revenue from the sale of products – SW*. The second variable was *net revenue from the sale of goods and materials – SH*. Data for companies representing two NACE sections were used: manufacturing and trade. In the group of manufacturing companies, the average net revenue from the sale of goods and materials (SH) was much lower than the average net revenue from the sale of products (SW), while in the group of trading companies, the relation was the opposite. The number of sampled companies in both groups was similar - 3927 (manufacturing) and 4094 (trade).

The first step of the analysis involved the direct estimation of the target variables in all target domains i.e. province by section. Fig. 1 presents the distribution of obtained estimates.

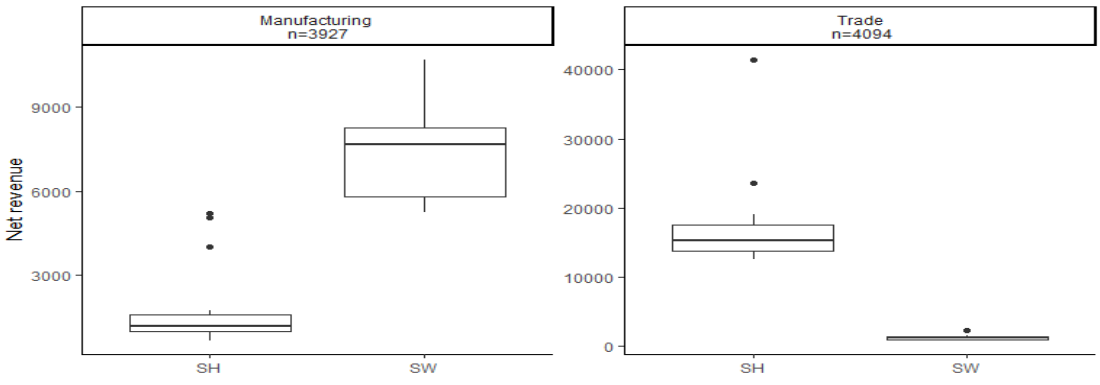

Fig. 1. Distribution of target variables by NACE sections.

The minimum value of *net revenue from the sale of goods and materials* (SH) in the manufacturing section is equal to 647,000 PLN and is observed in Zachodniopomorskie

province. Based on Fig. 1, three province outliers can be identified - Śląskie (5,200,000 PLN), Warmińsko-Mazurskie (5,039,000 PLN) and Mazowieckie (4,010,000 PLN). The maximum value of SH is close to the minimum for the net revenue from the sale of products (SW), which is equal to 5,253,000 PLN for Lubuskie province. Three provinces with the highest estimates are Warmińsko-Mazurskie (10,705,000 PLN), Wielkopolskie (9,205,000 PLN) and Zachodniopomorskie (9,133,000 PLN). The provinces vary considerably in terms of which types of business activity are identified as the main source of revenue. For example, in Warmińsko-Mazurskie province these include the food, tyre and wood industry, Śląskie province is dominated by companies mainly engaged in coal mining, steelmaking and electricity production. In Wielkopolskie province the dominant industries include the mining of salt, gypsum and lignite. Mazowieckie province is where PKN Orlen, the biggest fuel company in Poland is headquartered.

In the trade section, the sale of goods and materials (SH) is the main source of revenues. The highest estimated values are found in two provinces: Mazowieckie (41,489,000 PLN) and Małopolskie (23,585,000 PLN). The lowest estimated value is observed in the least urbanized province of Poland - Podkarpackie. The revenue from the sale of products (SW) in the trade section has a relatively marginal role - the highest value can be found in Mazowieckie province (2,377,000 PLN).

In addition to analysing the distribution of direct estimates, it is very important to consider the precision of these estimates. Table 1 contains descriptive statistics of relative root mean square errors (RRMSE) of direct estimates of net revenue.

Table 1. Descriptive statistics of RRMSE (in %) of direct estimates of net revenue.

NACE section	Variable	Minimum	Median	Mean	Maximum
Manufacturing	SH	15.8	31.4	33.9	70.9
Manufacturing	SW	9.4	11.5	14.2	37.5
Trade	SH	8.6	13.7	15.0	31.6
Trade	SW	11.7	17.1	17.7	35.0

Direct estimates of net revenue from the sale of goods and materials (SH) in the manufacturing section are characterized by very high values of RRMSE. In extreme cases, RRMSE is equal to 70.9% of the estimate (Śląskie province). For the remaining cases, the mean of RRMSE is about 15%, while the maximum exceeds 30%. The literature gives very different thresholds for precision. According to guidelines published by Eurostat for

household surveys, the precision level should depend on the survey, its purpose and the target domain. The National Institute of Statistics in Italy accepts RRMSE which does not exceed 15% for domains and 18% for small domains (Eurostat, 2013). According to the standards of the Central Statistical Office in Poland, survey results can be published if RRMSE is below 10% for target domains (Główny Urząd Statystyczny, 2013).

To obtain more precise estimates, indirect methods of estimation were applied: the Fay-Herriot model (FH) and the robust Fay-Herriot model (RFH). The models were based on information about net revenues in 2011 from the register maintained by the Ministry of Finance and the number or employees from the register of the Polish Social Insurance Institution (ZUS). Beta parameters in the models were significant and have a positive sign, which means that higher values of auxiliary variables result in higher net revenue estimates.

Fig. 2 presents the distribution of estimates obtained based on the direct Horvitz-Thompson (HT) estimator, the Fay-Herriot model (FH) and the robust Fay-Herriot model (RFH).

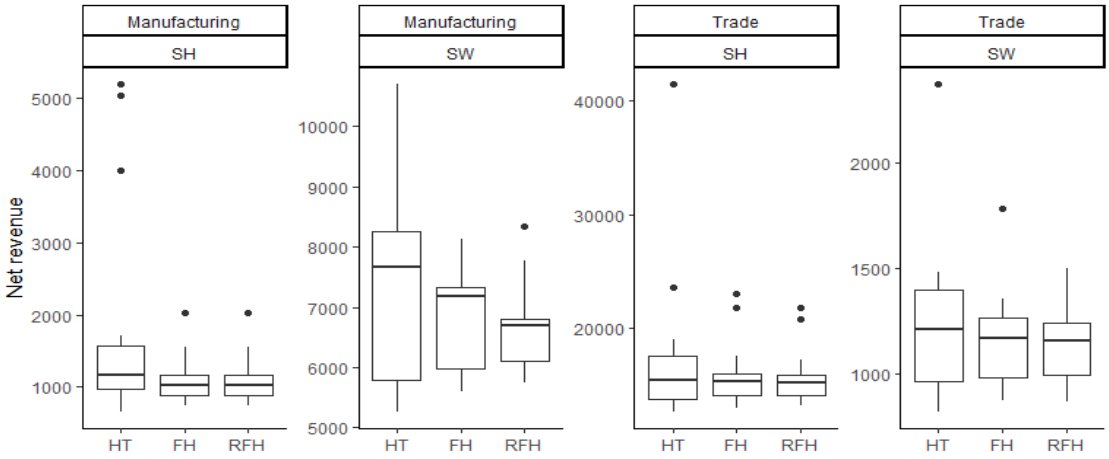

Fig. 2. Distribution of estimates by NACE section and type of net revenue.

The Fay-Herriot model belongs to the class of so-called "shrinkage" estimators, so obtained estimates have a smaller range than direct estimates. Moreover, the robust version of the Fay-Herriot model has a smaller range than the "classic" Fay-Herriot model. As regards values of net revenue from the sale of goods and materials (SH) in the manufacturing section, the maximum value is observed for Mazowieckie province and it is equal to 2,028,000 PLN for the FH model and 2,031,000 PLN for the RFH model.

The most visible change in the distribution is observed for net revenue from the sale of products (SW) in the manufacturing section. The median of direct estimates is equal to 7,664,000 PLN, 7,177,000 PLN for the Fay-Herriot model, and 6,700,000 PLN for the robust version of this model.

In the trade section estimates of both target variables obtained by applying the robust Fay-Herriot model have a smaller range in comparison to the other two approaches. With respect to net revenue from the sale of goods and materials (SH), the range of the Horvitz-Thompson estimates is equal to 28,943,000 PLN, for the Fay-Herriot model – 10,088,000 PLN, and for the robust Fay-Herriot model – 8,667,000 PLN. A very similar relation can be observed for net revenue from the sale of products (SW) in the trade section.

The next step of the study was the analysis of RRMSE. Table 2 presents values of this measure depending on section, dependent variable and estimator.

Table 2. Descriptive statistics of RRMSE (in %) of estimates by NACE section, type of net revenue and estimator.

NACE section	Variable	Estimator	Minimum	Median	Mean	Maximum
Manufacturing	SH	HT	15.8	31.4	33.9	70.9
Manufacturing	SH	FH	14.0	19.6	22.6	39.9
Manufacturing	SH	RFH	14.1	21.7	32.4	98.3
Manufacturing	SW	HT	9.4	11.5	14.2	37.5
Manufacturing	SW	FH	7.4	8.9	9.0	12.2
Manufacturing	SW	RFH	5.1	6.4	8.3	23.7
Trade	SH	HT	8.6	13.7	15.0	31.6
Trade	SH	FH	4.1	5.3	6.2	10.1
Trade	SH	RFH	4.8	7.4	8.6	15.8
Trade	SW	HT	11.7	17.1	17.7	35.0
Trade	SW	FH	10.3	13.9	14.1	19.7
Trade	SW	RFH	10.2	13.4	13.7	21.8

By applying indirect methods of estimation it was possible to reduce RRMSE of net revenue in unplanned domains, i.e. provinces. RRMSE of estimates obtained using the Fay-Herriot model is consistently lower than the precision of direct Horvitz-Thompson estimates. Estimates of net revenue from the sale of products (SW) calculated from the robust Fay-

Herriot have a better average precision than those given by the Fay-Herriot model in both sections. However, the maximum values of RRMSE are higher than in the case of the Fay-Herriot. The precision of estimating net revenue from the sale of goods and materials (SH) is better for the FH model than for the RFH model. This is associated with the result of estimating between-area variance. For this target variable and for this section the estimation algorithm of between-area variance in the case of Fay-Herriot model did not find a positive solution. As a result, the Fay-Herriot model generates synthetic estimates, which are characterized by low RRMSE. The same algorithm applied in the Robust Fay-Herriot model produces positive values of between-area variance, so the precision indicator also takes into account the uncertainty of direct estimation. RRMSE of estimates of net revenue from the sale of goods and materials (SH) in the manufacturing section obtained from the RFH model are even larger than direct estimates. In fact, there are two provinces (Warmińsko-Mazurskie and Śląskie) which are characterized by the largest RRMSE of direct estimates and large residuals in the robust Fay-Herriot model.

In addition to assessing estimation precision, estimates should also be analysed in terms of bias. Fig. 3 shows the sum of net revenue from the sale of products (SW) and net revenue from the sale of goods and materials (SH) compared to the true value of total net revenue in 2012.

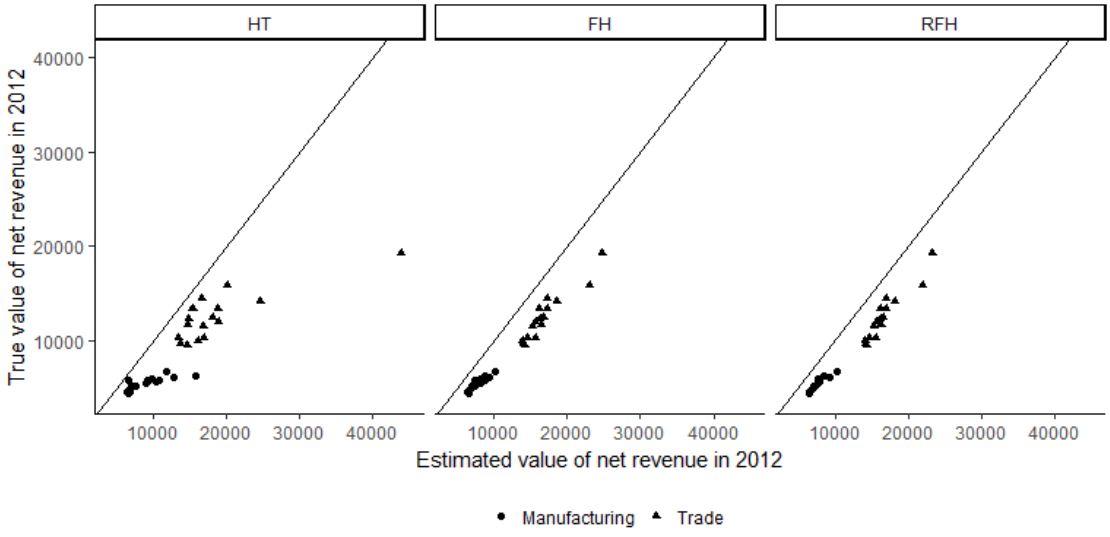

Fig. 3. Estimated and true values of net revenue in 2012.

In all cases, estimated values are overestimated in comparison with true values from the administrative register. Horvitz-Thompson estimates are characterized by the biggest bias.

Net revenue in the trade section for Mazowieckie province is overestimated by a factor of two. In the manufacturing section, Warmińsko-Mazurskie province is an outlier, with the direct estimate equal to 15,744,000 PLN, compared to the true value equal to 6,200,000 PLN. Nevertheless, the correlation between estimated and true values of net revenue is positive and strong. Spearman's correlation coefficient for the Horvitz-Thompson estimator is 0.9263, while for the indirect estimators – 0.9758 for the Fay-Herriot model and 0.9765 for its robust version. Average relative bias is equal to 55.9% for direct estimates, 40.1% for the Fay-Herriot model and 38.1% for the robust Fay-Herriot model.

Conclusions

Thanks to indirect methods of estimation, it is possible to obtain estimates of net revenue from the sale of goods and materials (SH) and net revenue from the sale of products (SW) for two NACE sections at a previously unpublished level of aggregation. Results obtained using the Fay-Herriot model and its robust version in most cases are more precise in terms of RRMSE than direct estimates. Moreover, robust estimation affects outlier values of net revenue and decreases the range of estimates. It is also worth noting that average relative bias is the smallest for estimates obtained by means of the robust Fay-Herriot model.

Further work will focus on estimating net revenue for small companies in Poland in other NACE sections. Because the level of precision of estimates generated by the Robust Fay-Herriot model is still unsatisfactory, we are considering changing the tuning factor or testing other influence functions (e.g. Tukey's, Cauchy's, Fair's, Talworth's or Welsch's) in the robust F-H model. Also, given the strategic role of the district (NUTS 4 unit), it would be interesting to apply the proposed approach could to estimate characteristics of small companies at district level

Acknowledgements

The project is financed by the National Science Centre in Poland on the basis of the decision number DEC-2015/17/B/HS4/00905.

References

Boonstra, H. J. & Buelens, B. (2011). Model-based estimation. *Statistics Netherlands.* Hague.

Dehnel, G. (2017). GREG estimation with reciprocal transformation for a Polish business survey In: Papież M. and Śmiech S. (eds) *The 11th Professor Aleksander Zelias Internetional Conference on Modelling and Forecasting of Socio-Economic Phenomena.*

Conference Proceedings (pp. 67-75). Cracow: Foundation of the Cracow University of Economics.

Dehnel, G., Pietrzak, M. & Wawrowski, Ł. (2017). Estymacja przychodu przedsiębiorstw na podstawie modelu Faya-Herriota. *Przegląd Statystyczny*, *64*(1), 79-94.

González-Manteiga, W., Lombardia, M. J., Molina, I., Morales, D. & Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, *78*(5), 443-462.

Główny Urząd Statystyczny (2013). Ludność. Stan i struktura demograficzno-społeczna. Narodowy Spis Powszechny Ludności i Mieszkań 2011. *Zakład Wydawnictw Statystycznych*. Warszawa: GUS.

Główny Urząd Statystyczny (2017). *Działalność przedsiębiorstw niefinansowych w 2015 roku.* Warszawa: GUS.

Eurostat (2013). Handbook on precision requirements and variance estimation for ESS households surveys. *European Union*.

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), 269-277.

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.

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley.

Rao, J. N. K. (2015). *Small-Area Estimation*. John Wiley & Sons, Ltd.

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

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