ARGUMENTA
OECONOMICA
CRACOVIENSIA

No 16 • 2017
ISSN 1642-168X
AOC, 2017; 16: 23–36
DOI: 10.15678/AOC.2017.1602

Grażyna Dehnel
Michał Pietrzak
Łukasz Wawrowski

AN EVALUATION OF COMPANY PERFORMANCE USING THE FAY-HERRIOT MODEL*

Abstract

Information about monthly characteristics of the small business sector is currently provided mainly by sample surveys conducted, among others, by the Central Statistical Office. Sample size enables parameters of interest to be estimated with acceptable precision only at the country or voivodeship level or by NACE section. The growing demand for reliable estimates at a low level of aggregation is the motivating force behind research into the application of indirect methods of estimation based on auxiliary sources of information. Hence, the article seeks to evaluate the possibility of applying the Fay-Herriot model to estimate one of the basic economic variables that characterise small business, i.e. revenue, based on information collected in administrative registers maintained by the Ministry of Finance.

Keywords: small area estimation, indirect estimation, administrative registers, Fay-Herriot model, economic statistics.
JEL Classification: C40, C51.

Grażyna Dehnel, Poznań University of Economics and Business, Faculty of Informatics and Electronic Economy, 61-875 Poznań, al. Niepodległości 10, Poland, e-mail: grazyna.dehnel@ue.poznan.pl

Michał Pietrzak, Poznań University of Economics and Business, Faculty of Informatics and Electronic Economy, 61-875 Poznań, al. Niepodległości 10, Poland, e-mail: michal.pietrzak@ue.poznan.pl

Łukasz Wawrowski, Poznań University of Economics and Business, Faculty of Informatics and Electronic Economy, 61-875 Poznań, al. Niepodległości 10, Poland, e-mail:lukasz.wawrowski@ue.poznan.pl

* This paper presents the results of a research project financed by the Polish National Science Centre, decision DEC- 2015/17/B/HS4/00905.

1. Introduction

Sample surveys are undoubtedly among the most important sources of monthly information and statistics about the small business sector. Such data are provided thanks to, foremost, sample surveys conducted by the Central Statistical Office (GUS). For purposes of estimation, it is necessary to consider such issues as sample size, survey design and estimation methodology, because these determine what kind of cross-classifications the estimates can be broken down into. Sample surveys conducted by GUS are based on relatively small sample sizes, which is why their results, calculated from direct estimators, are only released as aggregate estimates. The sample size in the biggest survey carried out for purposes of short-term business statistics, the DG1 survey, is sufficient to precisely estimate parameters only at the country and voivodeship level, or by section of business classification. Estimates at a lower level of aggregation, for smaller units, or for finer cross-classifications are much less reliable as a result of the high variance of estimators used to obtain them, or their pronounced bias (Dehnel 2015). Beyond that, classic estimation methods cannot be used to calculate estimates for domains with no units in the sample.

Recent years have seen a growing demand for detailed information at the level of local and regional units. This demand is expressed not only by government agencies, both local and national, but also by businesses. However, survey-based data currently provided by GUS do not meet the needs of all interested users. This is because they cannot be used to produce reliable direct estimates at a low level of aggregation, for local or regional units of territorial division or for detailed cross-classifications. While this situation could be solved by increasing the sample used in the survey, doing so would increase the cost and the duration of the survey administration. Such a modification is contrary to the principles of the modernisation of official statistics. It is therefore necessary to find tools that will enable official statistics to meet the demand expressed by users without the need to modify how the survey is administered. One possible solution is to modernise the currently used estimation methodology by relying on new techniques offered by small area estimation (SAE). SAE comprises a growing set of methods known as indirect estimation. Unlike classic survey sampling estimation methods, indirect estimators use all kinds of auxiliary information from outside a given domain or period of interest in order to improve estimation precision. This process is referred to as borrowing strength. This approach makes it possible to exploit information from a small sample, or, in special

cases, an empty sample, by supporting estimation with data from non--statistical sources without additional costs to provide reliable estimates at a low level of aggregation. Before indirect estimation methods can be widely implemented in official statistics, they must be evaluated in empirical studies to identify what is possible and beneficial in Polish conditions. The present article is intended as a contribution to this evaluation. It investigates the possibility of applying the Fay-Herriot model to estimate the mean monthly revenue of small businesses by voivodeship and NACE section, based on information collected in administrative registers.

The article has three main parts. The first describes data sources used for estimation and provides details of the empirical study. The second is devoted to methodological considerations of the analysis while the third part summarises and interprets the results of the study.

2. Data Sources

The empirical study conducted for this article was based on data from a business survey called DG1, the largest short-term survey in Poland. The target population for DG1 includes all legal persons and units without a legal personality employing at least 10 persons. It is a sample survey, with the sample including 98,000 units. Of those, medium-sized and large units employing over 49 people accounted for 18,000 companies, while the remaining 80,000 companies were small businesses employing between 10 and 49 people. All medium-sized and large units participate in the survey, while at least 10% of the sample comprised small units. Hence, the sample contains about 30,000 small businesses.

The DG1 survey is carried out to collect information about basic measures of economic activity in companies on a monthly basis. DG1 is in fact a monthly report providing essential data about each business unit, its activity and products and inventories. The data can be divided into two categories. The first contains information used to identify each business unit: its name, address, statistical ID number (REGON) and main activity by business classification. The second category contains characteristics of economic activities, such as sales revenue (from goods and services), number of employees, gross wages, wholesale and retail sales, excise tax and product--specific subsidies.

3. Description of the Study

The empirical study was limited to small businesses employing between 10 and 49 people that were active in May 2012. The dependent variable in the model was monthly revenue received by small businesses in June 2012. The auxiliary variable was revenue, whose value was determined based on information from the administrative register for December 2011. This study design was dictated (and restricted) by the availability of administrative information, which is released to the Central Statistical Office with several months' delay. Estimates for June can therefore be based on data from, at best, December of the previous year. In addition to data from administrative registers, estimation was also based mainly on a 10% sample of small businesses taking part in the DG1 survey.

Mean monthly revenue of companies employing between 10 and 49 persons was estimated for 16 voivodeships and 4 major NACE sections:

– "Manufacturing",
– "Construction",
– "Wholesale and retail trade" (Trade),
– "Transport and storage" (Transport).

To estimate the parameter of interest, an attempt was made to apply the classic Fay-Herriot model developed in small area estimation. Direct estimates obtained with the classic Horvitz-Thompson estimator were used as a point of reference for the study results. By comparing direct and indirect estimates, it was possible to evaluate the gain in precision resulting from the use of the non-standard method of estimation.

The precision of estimates obtained with the direct Horvitz-Thompson estimator and the classic Fay-Herriot model was assessed by the coefficient of variation expressed as a ratio of the standard error to the estimate. It was calculated using linearisation under simple random sampling. The theoretical basis for the classic Fay-Herriot model, the direct Horvitz-Thompson estimator and the coefficient of variation to be used as a measure of precision will be discussed in the next section of the article, which details the methodological aspects of the study.

4. Estimation Methods

4.1. Direct Estimation

The most common estimator in survey methodology is the Horvitz-Thompson estimator (hereinafter HT) (Horvitz & Thompson 1952).

Consider sample s drawn from population U, where s_d is a subsample from domain d. Counties in domains meet the condition $n_d < N_d$, where n_d denotes sample size in domain d and N_d is the population size of domain d. Under simple random sampling, each unit i in the sample has been assigned a sample weight – w_i – which is an inversion of the first order inclusion probability. The estimator of the mean in domain d is given by the formula:

$$\hat{y}_d^{HT} = \frac{1}{\hat{N}_d} \sum_{i=1}^{n_d} y_{di} w_{di},$$

where $\hat{y}_d^{HT}$ is the estimated mean of the variable of interest y in domain d and $\hat{N}_d = \sum_{i=1}^{n_d} w_{di}$.

The direct estimator is design-unbiased and design-consistent assuming that $n_d \to \infty$. Nevertheless, it is highly ineffective for domains in which n_d is very small; and it is impossible to calculate direct estimates for non-sampled domains where $n_d = 0$ (Guadarrama, Molina & Rao 2016).

To deal with the inefficiency of the direct estimator in the case of small sample size, indirect estimation methods are used. These techniques utilise data from different sources, such as censuses or administrative registers, connected with the target variable and thus improve estimation. Such a procedure is called "borrowing strength" and is the basis of small area estimation. This paper presents and examines just one of the many approaches – the Fay-Herriot model.

4.2. Fay-Herriot Model

The Fay-Herriot model was originally developed for income estimation in small domains in the USA (Fay & Herriot 1979). However, because of its simplicity and good empirical properties, it is used to estimate many other indicators, especially the poverty rate (Pratesi & Salvati 2008, Wawrowski 2014). The Fay-Herriot model is an area level model and relies on covariates measured at the domain level. Here is the general form of this model:

$$\hat{y}_d^{HT} = x'_d \beta + u_d + e_d,$$

where $\hat{y}_d^{HT}$ denotes direct estimates of the variable of interest, x'_d is a vector of auxiliary variables, β is a vector of regression parameters, u_d is a domain random effect, independent and identically distributed $(0, \sigma_u^2)$, and e_d is an independent random error $(0, \psi)$.

It is assumed that sampling variance ψ is known, though in practice it is estimated. Likewise, random effect variance σ_u^2 also needs to be estimated. This can be done using various methods, such as the moment estimator, maximum likelihood estimator (ML) or restricted maximum likelihood (REML). These methods are based on the iterative approach. If no positive solution is found, then $\hat{\sigma}_u^2 = 0$ and, thus, there are no random effects in the model. After these components are estimated, the empirical best linear unbiased predictor (EBLUP) is determined with this formula:

$$\hat{y}_d^{FH} = \hat{\gamma}_d \hat{y}_d^{HT} + (1 - \hat{\gamma}_d) x'_d \hat{\beta}.$$

The equation shows that Fay-Herriot estimates can be expressed as weighted averages of direct estimates and regression estimates. Weight:

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\psi}_d}$$

measures the uncertainty of describing the estimated variable using the regression model. If sampling variance $\hat{\psi}_d$ is small, then weight $\hat{\gamma}_d$ is large. This means that EBLUP relies on direct estimates when they are accurate and takes model estimates otherwise (Boonstra & Buelens 2011). Regression parameters are calculated in the following way:

$$\hat{\beta} = \left(\sum_{d=1}^{D} \hat{\gamma}_d x_d x'_d \right)^{-1} \sum_{d=1}^{D} \hat{\gamma}_d x_d \hat{y}_d^{HT}.$$

For non-sampled domains and when $\sigma_u^2 = 0$, a Fay-Herriot model estimate is equal only to the regression part of EBLUP.

4.3. Precision Assessment Method

The precision of estimates can be determined by calculating the mean square error. For direct estimation, it can be calculated using the Taylor series method, while for the Fay-Herriot model, the MSE estimator Rao described in Chapter 7 can be used (Rao 2003). As a comparative measure of precision, the *CV* indicator is given by:

$$CV_d = \frac{\sqrt{MSE_d}}{\hat{y}_d}.$$

This ratio indicates the share of the estimation error in the value of the target variable estimate. Smaller *CV* values are desirable.

5. Estimation Results and an Assessment of Their Precision – Income Estimation

This study sought to estimate the mean revenue of businesses by voivodeship and selected NACE sections. The analysis focused on companies employing fewer than 49 people and conducting business activity classified into 4 NACE sections: "Manufacturing", "Construction", "Wholesale and retail trade" (Trade) and "Transport and storage" (Transport).

Before performing the estimation and analysing the results, it was first necessary to calculate the descriptive statistics of the sample by voivodeship and NACE section. Table 1 contains sample sizes for each domain. All 16 of Poland's voivodeships were sampled for each section.

Table 1. Sample Size by Voivodeship and NACE Section

NACE	N	Min	Q1	Mean	Median	Q3	Max
Manufacturing	16	129	154	245	220	331	442
Construction	16	41	53	93	78	128	193
Trade	16	131	158	256	216	294	563
Transport	16	20	24	40	30	47	101

Source: based on data from the DG1 survey.

The sample distribution can be divided in two parts. For "Construction" and "Transport", the mean sample size was below 100 units, while for "Manufacturing" and "Trade" it was above 200 units. The greatest variation in sample size across voivodeships can be observed for companies classified in NACE under "Manufacturing", and the least variation for "Transport". The largest median sample size is recorded for "Manufacturing", followed by "Trade". In the case of the latter, two voivodeships were outliers with considerably larger sample sizes than the rest: Mazowieckie and Śląskie. The other two NACE sections, "Construction" and "Transport", are characterised by much smaller sample sizes. The median sample size for "Construction" is 78 units, and for "Transport" 30. For "Transport", one voivodeship – Mazowieckie, with 101 entities – had a considerably larger sample size.

Nevertheless, such sample size diversity does not necessarily mean the precision indicator will be likewise varied, as Table 2 shows.

Table 2. *CV* of Direct Estimates of Mean Revenue (in %)

NACE	N	Min	Q1	Mean	Median	Q3	Max
Manufacturing	16	9.4	11.4	17.2	13.1	16.1	48.5
Construction	16	7.1	14.3	16.3	15.9	20.3	25.4
Trade	16	8.1	10.5	15.2	13.9	17.5	30.7
Transport	16	11.9	15.5	22.7	19.9	27.8	55.8

Source: based on data from the DG1 survey.

While the trade sector was represented by the largest number of enterprises, the smallest variance of estimates was observed in the construction sector. For all NACE sections, the maximum *CV* value exceeds 20%, which the Central Statistical Office regards as the threshold above which results should be published only as aggregate data (*The Methodology…* 2011). In this case, there are 18 cross-classifications (16 voivodeships × 4 sections) in which *CV* exceeds this level, accounting for 28% of all cells. For this reason, the Fay-Herriot model will be applied as a tool to improve the reliability of mean revenue estimates.

The first step in performing small area estimation is to identify the auxiliary variables that can be utilised in the model. Based on the available data and a literature review, mean revenue from an administrative register was chosen as the predictor. In this approach, parameters of the Fay-Herriot model were estimated for each NACE section. Estimated β parameters with *p*-values (in brackets) and random effects variance are presented in Table 3.

Table 3. Estimated Parameters of the Fay-Herriot model

Parameter	Sector			
	Manufacturing	Construction	Trade	Transport
β_0	1,002.83 (0.3050)	908.14 (0.1768)	2,050.48 (0.1790)	–485.39 (0.7350)
β_1	0.52 (0.0047)	0.38 (0.0035)	0.49 (0.0002)	0.70 (0.0007)
σ_u^2	146,551.50	150,315.70	0	0

Source: based on data from the DG1 survey and the administrative register.

All β_1 parameters have a positive sign and they are statistically significant. This means that mean revenue in the administrative register is positively correlated with mean revenue in the DG1 survey. It should be noted that

random effects variance σ_u^2 could be calculated only for two sections – "Manufacturing" and "Construction". For "Trade" and "Transport", no estimation method returned a positive solution. Fay-Herriot model estimates of mean revenue in these cases will be based only on the regression component – that is, no random effect for voivodeships is observed. Such estimates will be characterised by better precision – smaller *CV* values, but at the expense of higher bias. For manufacturing and construction, it is assumed that estimated σ_u^2 can be used to obtain mean revenue estimates that are optimal in terms of precision and bias.

Figure 1 shows mean revenue estimates obtained directly, with the Horvitz-Thompson estimator, and indirectly, using the Fay-Herriot model.

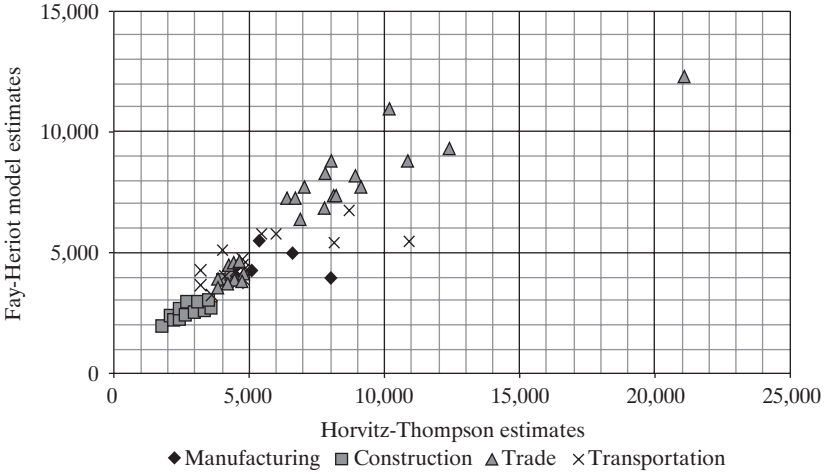

Fig. 1. Comparison of H-T Estimates and F-H Model Estimates

Source: based on data from the DG1 survey and the administrative register.

Estimated values of mean revenue are grouped by NACE section. Note that construction is characterised by the smallest revenue values, followed by transport and manufacturing, with the highest level of mean revenue estimated for trade. In general, direct and indirect estimates have similar values. Mazowieckie voivodeship is the only outlier. In this case, the direct estimate of mean revenue was 21,000 PLN, which was reduced by the Fay--Herriot model to 12,000 PLN. This property of indirect estimation is known as shrinkage. Figure 2 presents the distribution of the precision indicator *CV*.

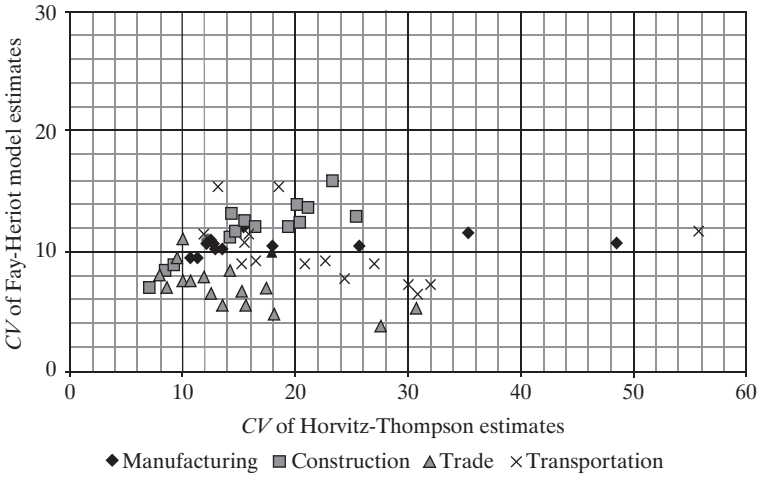

Fig. 2. Comparison of the Precision of H-T Estimates and F-H model Estimates

Source: based on data from the DG1 survey and the administrative register.

The Fay-Herriot model yields similar results but with better precision. In almost all cases, indirect estimation has contributed to reducing mean square error. The biggest difference between direct and indirect estimates – a decrease in *CV* from 28% to 4% – can be observed for Lubelskie voivodeship in the trade sector. There are three cases where the precision of the direct estimator is better than that of the indirect one: estimates for Pomorskie and Warmińsko-Mazurskie voivodeships in the trade sector and Kujawsko-Pomorskie voivodeship in the transport sector. Because these differences are very small, e.g. a rise from 13% to 15%, indirect estimates are still below the threshold of acceptability. That, in turn, is because for these NACE sections the algorithm was unable to obtain positive random effects variance. Nevertheless, *CV* values for all mean revenue estimates obtained with the Fay-Herriot model are below 20%.

The distribution of mean revenue estimates across voivodeships is shown, in Figure 3, based on the direct estimator and, in Figure 4, based on the Fay--Herriot model.

In general, the spatial distribution of estimates is very similar in both cases; however, direct estimates of average revenue are characterised by greater variability than model-based estimates. Final results obtained with the Fay-Herriot model are described below. For manufacturing, the lowest estimates can be seen in Łódzkie and Podkarpackie voivodeships – appr. 3,200 PLN. Businesses with the highest mean revenue are based in Śląskie

Fig. 3. Spatial Distribution of Direct Estimates of Mean Revenue

Source: based on data from the DG1 survey.

voivodeship (appr. 5,000 PLN) and in Mazowieckie voivodeship (appr. 5,500 PLN). In the construction sector, the lowest mean revenue estimate was calculated for Opolskie voivodeship, at 1890 PLN while the leading voivodeships in this sector are Mazowieckie (3,810 PLN) and Wielkopolskie (4,370 PLN). The highest estimates of mean revenue of all four sectors were obtained for the trade sector. Estimates exceed the 10,000 PLN threshold in two voivodeships: Zachodniopomorskie (10,888 PLN) and Mazowieckie (12,322 PLN). At 6,422 PLN, the lowest value of mean revenue was estimated in Warmińsko-Mazurskie voivodeship, which returned almost half the value calculated for Mazowieckie, despite the fact that the two regions neighbour one another. For the transport sector, the lowest mean revenue (3,344 PLN) was estimated for Podkarpackie, much of which is situated in the mountains. At 7,049, the highest revenue was estimated for Pomorskie voivodeship, with its well-developed maritime transport most likely to thank.

Fig. 4. Spatial Distribution of the Fay-Herriot Model Estimates of Mean Revenue
Source: based on data from the DG1 survey and the administrative register.

6. Conclusions and Further Research

The aim of the study reported in this article was to produce reliable estimates of the mean revenue of Polish companies employing fewer than 50 people and conducting activity classified in one of four selected NACE sections.

The estimates obtained using both approaches are very similar, but direct estimates display greater variability. Moreover, by applying the Fay-Herriot model and taking advantage of auxiliary variables from the administrative register it was possible to produce estimates at a previously unpublished level of aggregation, with acceptable precision expressed by *CV* values below 20%. The lowest mean revenue can be observed in the "Construction" section and the highest in the "Trade" section. In "Manufacturing", the highest values of revenue were estimated for Śląskie and Mazowieckie

voivodeships, in "Construction" they were for Wielkopolskie, in "Trade" again for Mazowieckie, and in "Transport" for Pomorskie.

Further research will focus on estimating other variables, such as company expenditures or losses. The authors will also consider the use of the multivariate Fay-Herriot model proposed by Benavent and Morales (2015), which can be applied when no positive solution of random effects variance can be found.

Bibliography

Benavent, R. and Morales, D. (2015) "Multivariate Fay-Herriot Models for Small Area Estimation". *Computational Statistics & Data Analysis* 94: 372–90, http://www.sciencedirect.com/science/article/pii/S016794731500170X, https://doi.org/10.1016/j.csda.2015.07.013.

Boonstra, H. J. and Buelens, B. (2011) "Model-based Estimation". *Statistics Netherlands*, https://www.cbs.nl/NR/rdonlyres/DA57C4D8-A631-4C04-B4EA-9165D264D0D6/0/2011x3706.pdf. Accessed: 2 May 2016.

Dehnel, G. (2015) "Robust Regression in Monthly Business Survey" in W. Okrasa (ed.) *Statistics in Transition – New Series* 16(1): 1–16, Warszawa, http://stat.gov.pl/en/sit-en/issues-and-articles-sit/previous-issues/volume-16-number-1-spring-2015/.

Fay, R. and Herriot, R. (1979) "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data". *Journal of American Statistical Association* 74: 269–77.

Guadarrama, M., Molina, I. and Rao, J. N. K. (2016) "A Comparison of Small Area Estimation Methods for Poverty Mapping". *Statistics in Transition. New Series and Survey Methodology* 17(1): 41–66, http://stat.gov.pl/en/sit-en/issues-and-articles-sit/current-issue/.

Horvitz, D. G. and Thompson, D. J. (1952) "A Generalization of Sampling without Replacement from a Finite Universe". *Journal of the American Statistical Association* 47: 663–85.

The Methodology of the 2011 National Population and Housing Census – Selected Aspects (2011). Warsaw: Central Statistical Office.

Pratesi, M. and Salvati, N. (2008) "Small Area Estimation: the EBLUP Estimator Based on Spatially Correlated Random Area Effects". *Statistical Methods and Applications* 17: 113–41, http://dx.doi.org/10.1007/s10260-007-0061-9.

Rao, J. N. K. (2003) *Small Area Estimation*. Hoboken, NJ: Wiley.

Wawrowski, Ł. (2014) "Wykorzystanie metod statystyki małych obszarów do tworzenia map ubóstwa w Polsce". *Wiadomości Statystyczne* 9: 46–56, http://stat.gov.pl/czasopisma/wiadomosci-statystyczne/archiwum/.

Abstract

Ocena działalności przedsiębiorstw z wykorzystaniem modelu Faya-Herriota

Informacje dotyczące miesięcznych charakterystyk sektora małych przedsiębiorstw obecnie pochodzą głównie z badań reprezentacyjnych prowadzonych m.in. przez GUS. Wielkość próby umożliwia precyzyjne oszacowanie parametrów jedynie dla całego kraju i województw bądź w przekroju sekcji PKD. Rosnąca potrzeba dostarczenia wiarygodnych szacunków na niskim poziomie agregacji skłania do prowadzenia badań dotyczących zastosowania pośrednich metod estymacji wykorzystujących dodatkowe źródła informacji. Stąd też celem artykułu jest ocena możliwości zastosowania modelu Fay-Herriota do oszacowania jednej z podstawowych charakterystyk ekonomicznych dotyczących małych przedsiębiorstw, jaką jest przychód, na podstawie informacji zawartych w rejestrach administracyjnych prowadzonych przez Ministerstwo Finansów.

Słowa kluczowe: statystyka małych obszarów, estymacja pośrednia, rejestry administracyjne, model Faya-Herriota, statystyka gospodarcza.