



Grażyna Dehnel

Poznań University of Economics and Business, Faculty of Informatics and Electronic Economy,
Department of Statistics, g.dehnel@ue.poznan.pl

Stratification of Domains Using Composite Estimation to Measure the Revenue Level of Small Businesses in Poland¹

Abstract: To meet the growing demand for detailed, precise, accurate and timely estimation of entrepreneurship and economic conditions, it is necessary to systematically extend the scope of information provided by business statistics. In view of the policy aimed at reducing survey costs and burdens for business units, the only way in which this objective can be achieved is by modernizing survey methodology. One area where this kind of research is being conducted are applications of indirect estimation based on auxiliary sources of information from administrative sources. Hence, the purpose of the study described in this article is to evaluate the precision of estimates of revenues of small businesses for domains defined by spatial aggregation and business classification by applying stratification in composite estimators based on information collected from administrative registers.

Keywords: robust estimation, business statistics, small area estimation, GREG

JEL: C40

¹ The project is financed by the Polish National Science Centre, decision DEC–2015/17/B/HS4/00905.

1. The Problem

One of the characteristics of the information society is the growing demand for accurate and timely multivariate statistics for cross-classified domains. Statistics derived from business survey data are among the most important outputs for data users. However, producing estimates about the population of enterprises poses a particular challenge owing to the unique character of that population, which is typically skewed and heavy tailed. Monthly characteristics of the small business² sector (companies employing 10–49 people) in Poland are mainly obtained from a sample business survey conducted by the Central Statistical Office (Pol. GUS). The sample size is sufficiently large to produce precise estimates of parameters only for large domains, e.g. at country and province level, or for domains defined by sections of business classification. In the face of changes in the economy, there is a growing demand for detailed information for small territorial units. This demand is expressed by government agencies, both local and national and also by businesses. Survey-based data currently provided by GUS do not meet the needs of all interested users. This is because the estimates at a lower level of aggregation, for smaller domains, or for finer cross-classifications, are much less reliable. Estimators used to obtain them have a high variance or are heavily biased (Dehnel, 2015). Additionally, the presence of outliers generally decreases the precision of population estimates based on a large survey sample and can even prevent estimation for small areas from which outliers are sourced since such estimates will be based on a much smaller sample. The problem can be overcome by increasing the sample size, but this solution is usually not pursued because it is associated with higher costs. This is what motivates the search for estimation methods that could increase the scope of statistical outputs, improve the efficiency of estimates and remain the cost of the survey on the same level. One possible solution is to modernize the currently applied methodology of estimation by using small area estimation techniques (SAE) that ‘borrow strength’ from auxiliary data such as administrative registers (Dehnel, Pietrzak, Wawrowski, 2017). In this article we propose using direct estimators in the process of composite estimation based on stratified domains to estimate characteristics of small enterprises (employing 10–49 people) operating in 2012. The objective of the study was to assess the estimation of annual revenues of small business for units defined by spatial aggregation and business classification. The domain of interest was created by cross-classifying the administrative division into provinces with the NACE category of business activity. The estimation process was supported with delayed variables from administrative registers used as auxiliary

² The term “small business” is used interchangeably with the terms “small enterprise” and “small firm”.

variables. The article refers to small area composite estimation methods which take into account the process of stratification. The variables of interest are estimated for small domains obtained by cross-classifying provinces (NUTS 2) and NACE 2 sections.

1.1. A description of the small enterprise sector

Although small enterprises account for merely 3 per cent of the entire enterprise sector, they play a significant, and in some respects crucial role in the economy. There are currently about 57,000 small companies in Poland, cf. Figure 1 (PARP, 2017). Despite making the smallest contribution to the GDP (9%), and providing only 13% of all available jobs, they are more profitable and more liquid than medium-sized (employing 50–249 people) or large (employing over 250 people) enterprises, cf. Figure 2. They have the highest rate of growth in terms of generated value added (the growth in the period 2004–2014 was 218%) and are characterised by higher export dynamics. In the period 2007–2014, exports of small enterprises increased by a factor of two, while for the remaining companies by a factor of 1.7 (PARP, 2017). In 2015 small firms spent over PLN 20 billion in investments (9% of the total value of investments in the enterprise sector). Small enterprises mainly use their own funds to finance their investments (60%). They also have the highest survival rates – on average two thirds of small companies survive their first year of operation (GUS, 2015). The survival rate tends to increase in successive years. In the group of small firms established in 2011 and still operating in 2015, the survival rate for the following year (2016) was 99.9%. On average, a small company employs 21 people (GUS, 2017).

It can be noted that the share of exports in revenues of small firms is on the rise as well. In the period 2008–2015, it grew from 6.4% (in 2008) to 9.7% (in 2015). Compared to other enterprises, however, small companies had the slowest rate of growth in terms of sales revenue (i.e. from 2008 to 2015). From 2003 to 2015, the revenue growth in this group amounted to 180%, compared to 246% achieved by large enterprises (PARP, 2017).

Because small enterprises tend to operate locally, there is a close relationship between their development and the regional development. Small business owners, who largely invest their own capitals, tend to base their companies near their places of residence, rely on local resources and focus their activities on local target markets. The scope and intensity of this process depends on the level of regional development. The dependence is mutual: the development of the small enterprise sector helps to even out regional disparities, contributes to the improvement of the living conditions of local communities, creates new workplaces, and generally fosters the region's economic growth (PARP, 2017).

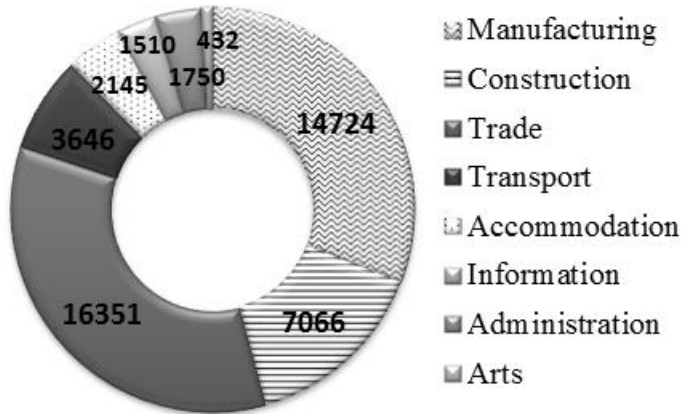


Figure 1. Number of small enterprises by activity in 2015

Source: based on the CSO study (GUS, 2017)

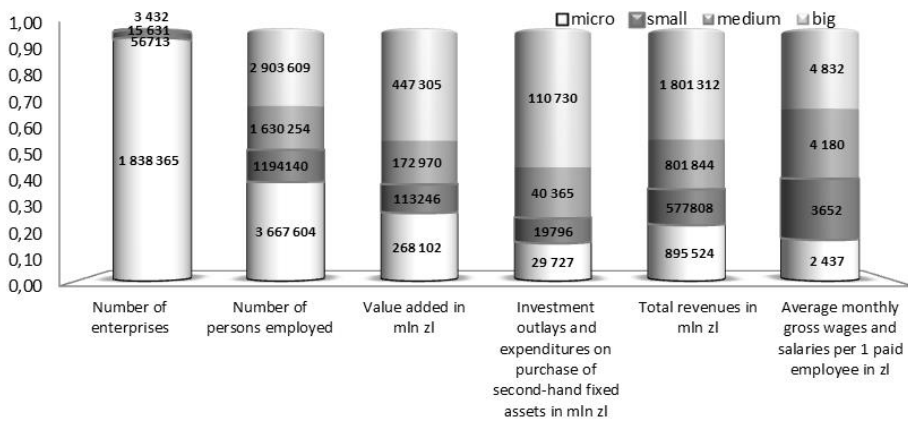


Figure 2. Enterprises' characteristics by size class in 2015 (as of 31st December)

Source: based on the CSO study (GUS, 2017)

2. The Method

Owing to the presence of outliers, it is not easy to estimate population characteristics of enterprises. The use of classical estimators at low levels of spatial aggregation or for more specific domains can lead to low precision and high bias.

A variety of methods have been devised to decrease the impact of outliers on survey estimates. Some of them involve subjective identification and removal of outlying values during survey processing (Chambers et al., 2014). We can also distinguish many objective methods for survey estimation. Some methods rely

on robust estimators, which are resistant to outliers. Others use detection techniques to decide whether an observation is a potential outlier and should be moved to a special post-stratum, where its contribution to the survey estimates is down-weighted (Chambers et al., 2014). If the number of outliers is small, estimation in the post-stratum can be difficult (Clark, Kokic, Smith, 2017). The approach proposed in this article is an attempt to cope with this problem. It consists in applying the k -means method in order to distinguish two groups of companies in each domain: the first one consists of typical observations, while the second one includes non-typical ones (outliers – observations that are numerically distant from most of the other data points in a set of data). The division is conducted based on the values of the auxiliary variables. In the first step, the division is applied to sampled units. Next, other population units (out-of-sample units) in each domain are divided into two groups depending on their distance to the cluster centres. For each group, we estimate the value of the parameter (total monthly revenue). The final estimate for a domain is given by the sum of estimates obtained for the first and the second group. Since the non-typical groups tended to contain very few observations (in some cases only 1), the estimation was made on the basis of units included in all non-typical groups identified within one NACE section. Hence, the approach used was a kind of synthetic estimation. In order to apply this approach, it was assumed that companies within one NACE section were homogeneous. The ratio estimator was selected in order to avoid the model-based approach, which tends to be unstable for non-homogenous groups (Chambers et al., 2001). The ratio estimator is given by the formula (Singh, Gambino, Mantel, 1993):

$$\hat{Y}_{SYN(R),d} = \gamma_d X_{1,d} \frac{\hat{Y}_{HT}}{\hat{X}_{1,HT}} + (1 - \gamma_d) X_{2,d} \frac{\hat{Y}_{HT}}{\hat{X}_{2,HT}} \quad 0 \leq \gamma_d \leq 1, \quad (1)$$

where:

$\hat{X}_{1,HT} = \sum_{i \in S} w_i x_{1,i}$ – direct HT estimator of the total of auxiliary variable x_1 ,

$\hat{X}_{2,HT} = \sum_{i \in S} w_i x_{2,i}$ – direct HT estimator of the total of auxiliary variable x_2 ,

γ_d – weight arbitrarily selected for a given survey, depending on the correlation between the dependent variable and auxiliary variables X_1 and X_2 . The ratio estimate is consistent and biased, although the bias is negligible in large samples (Cochran, 1977).

For groups containing typical observations, two types of the GREG estimator were used: classical, ratio GREG and modified GREG, which minimizes the impact of heteroscedasticity on the precision of estimates (Dehnel, 2017). The classical form of the GREG estimator in domain d is given by the formula (Chambers et al., 2001):

$$\hat{Y}_{GREG,d}^0 = \hat{Y}_{GREG,d} = \sum_{i \in U_d} \hat{y}_i + \sum_{i \in S_d} w_i e_i, \quad (2)$$

where $\hat{y}_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}}_d$; the vector of model parameters $\hat{\boldsymbol{\beta}}_d$ is estimated using the following formula (Rao, Molina, 2015):

$$\hat{\boldsymbol{\beta}}_d = \left(\sum_{i \in S_d} w_i \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \left(\sum_{i \in S_d} w_i \mathbf{x}_i y_i \right). \quad (3)$$

GREG estimators are asymptotically design-unbiased. However, careful model specification search plays an important role, especially for highly skewed variables and for populations with outliers, such as those analysed in business surveys (Chambers et al., 2001). In the ratio GREG estimator $\hat{Y}_{GREG,d}^{rat}$ (3a) the constant is omitted (Myrskylä, 2007).

In the modified version of the GREG estimator the vector of model parameters $\hat{\boldsymbol{\beta}}_d$ is estimated from the formula which includes an additional variable z (Chambers et al., 2001):

$$\hat{\boldsymbol{\beta}}_d = \left(\sum_{i \in S} w_i \mathbf{x}_i \mathbf{x}'_i / z'_i \right)^{-1} \left(\sum_{i \in S} w_i \mathbf{x}_i y_i / z'_i \right). \quad (4)$$

Modified GREG estimator can be given by the formula (2) (Chambers et al., 2001):

$$\hat{Y}_{GREG,d} = \sum_{i \in S_d} w_i g_i y_i, \quad (5)$$

where weight g_i depends on the value of auxiliary variable x for sampled units and is defined as:

$$g_i = 1 + \left(X_d - \hat{X}_{HT,d} \right) \left(\sum_{i \in S_d} w_i \mathbf{x}_i \mathbf{x}'_i / z'_i \right)^{-1} \left(\mathbf{x}_i / z'_i \right), \quad (6)$$

where:

U – general population (all small enterprises in Poland),

d – domain (unit obtained by cross-classifying provinces (NUTS 2) and NACE 2 sections),

g_i – weight of the i -th unit,

$\hat{Y}_{GREG,d}$ – estimate of the total in domain d given by the GREG estimator,

$\hat{X}_{HT,d}$ – direct HT estimator of the total of auxiliary variable x in domain d ,

X – total of auxiliary variable x ,

w_i – design weights,

z, x – auxiliary variables (x – revenue and the number of employees, z – the number of employees),

γ – parameter selected depending on the degree of heteroscedasticity, for $\gamma = 0$, estimator (2) has the classical form of the GREG estimator,

U_d – part of the general population containing elements belonging to domain d (part of small enterprises representing one of eight NACE sections and one of the sixteen provinces),

s_d – part of the sample containing elements belonging to domain d .

It is known from previous surveys that the value of parameter γ should be included in the interval $\langle 1, 1.5 \rangle$ (Särndal, Swensson, Wretman, 1992; Dehnel, 2017). Hence, two different settings (γ equal to 1 and 1.5) were tested in the study, which

were used in estimators $\hat{Y}_{GREG,d}^1, \hat{Y}_{GREG,d}^{1.5}$. When these estimators are used in the linear regression model, the condition of homoscedasticity does not have to be satisfied, and the level of heteroscedasticity is indicated by parameter γ .

The final estimate for domain was given by the sum of estimates obtained on the basis of the GREG estimator (or the ratio and modified GREG estimator) for the *Stratum 1* – (1) and the ratio estimator for the *Stratum 2* – (2) given by the formula (1):

$$\hat{Y}_{COM,d} = \hat{Y}_{GREG,d}^1 + \hat{Y}_{SYN(R),d}^2 \quad (7)$$

Estimation quality was assessed with reference to estimates obtained using classical direct estimators: Horvitz-Thompson (HT) The estimator of the total in domain d is given by the formula:

$$\hat{y}_d^{HT} = \sum_{i=1}^{n_d} \frac{y_{di}}{\pi_{di}}, \quad (8)$$

where:

\hat{y}_d^{HT} is the estimated mean of the variable of interest y in domain d ,

π_{di} – probability that the i -th unit belonging to domain d is in the sample,

y_{di} – value of the variable of interest for the domain d and i -th unit.

The direct estimator is design-unbiased and design-consistent assuming that $n_d \rightarrow \infty$. Nevertheless, it is very 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).

Finally, on the basis of total monthly revenue, knowing the number of sub-populations, the mean monthly revenue were computed.

2.1. Method of assessing precision and accuracy

Precision of estimates was evaluated using the extension of standard bootstrap method, taking into account sampling method (Shao, Tu, 1995: 246). Detailed discussions of this subject can be found in the monograph by J. Shao and D. Tu (1995: 238–250), J. Rao and C. Wu (1988: 231–241), E. Antal and Y. Tillé (2011: 534–543). Estimation efficiency was assessed using the coefficient of variation of the estimator based on results obtained in $B = 500$ samples ($b = 1, 2, \dots, B$), following the approach described in (Bracha, 2004: 33):

$$\hat{CV}(\hat{Y}_d) = \frac{\sqrt{\text{Var}^*(\hat{Y}_d)}}{E^*(\hat{Y}_d)} = \frac{\sqrt{\frac{1}{499} \sum_{b=1}^{500} (\hat{Y}_{b,d} - \hat{Y}_d)^2}}{E(\hat{Y}_d)}, \quad (9)$$

where:

$\hat{CV}(\hat{Y}_d)$ – the estimator of the coefficient of variation of the estimator,

$\text{Var}^*(\hat{Y}_d)$ – the variance computed based on bootstrap iterations,

$E^*(\hat{Y}_d)$ – the bootstrap estimator of the expected value.

This ratio indicates the share of the estimation error in the value of the target variable estimate. Smaller \hat{CV} values are desirable.

Accuracy of estimates was evaluated based on relative RMSE,

$$RR\hat{MSE}(\hat{Y}_d) = \frac{\sqrt{\hat{MSE}(\hat{Y}_d)}}{\hat{Y}_d} \quad (10)$$

using the formula (Rao, Molina, 2015: 44):

$$\hat{MSE}(\hat{Y}_d) = (\hat{Y}_{S,d} - \hat{Y}_{DIR,d})^2 - \text{Var}(\hat{Y}_{DIR,d}) \quad (11)$$

where:

$\hat{Y}_{DIR,d}$ – the direct estimator,

$\hat{Y}_{S,d}$ – the synthetic estimator,

$\text{Var}(\hat{Y}_{DIR,d})$ – the variance of the direct estimator.

Additionally, in order to obtain a more thorough evaluation of the estimates, the analysis focused on differences between values of the estimators obtained from a specific sample and estimates obtained from tax return data filed in December 2012. It was assumed that the following relationship holds: the ratio of *revenue* reported in tax returns by companies in the study at the province level to the value of *revenue* from the monthly enterprise survey (DG1) is constant (see Figure 4).

$$\frac{\text{revenue_AR}}{\text{revenue_DG1}} = \frac{\text{revenue_est}}{\text{revenue_DG1}}. \quad (12)$$

This approach made it possible to calculate the approximate value of *revenue* for June 2012.

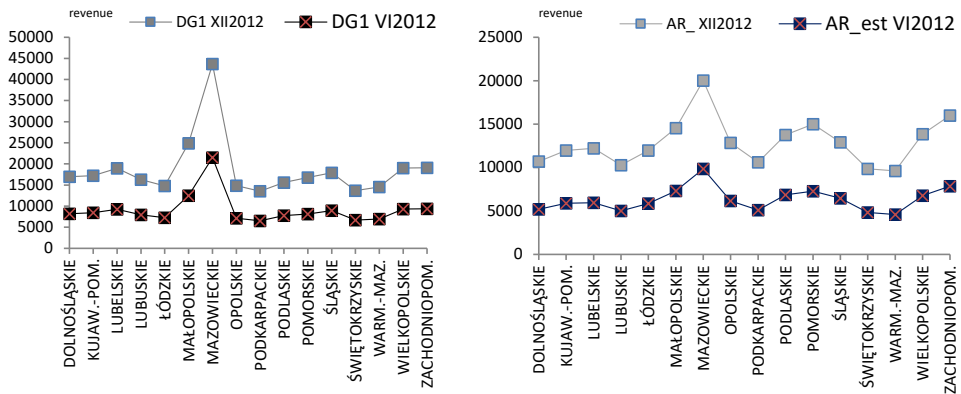


Figure 3. Value of revenue in June and December 2012 reported in the DG1 survey and in tax returns for companies classified into the Manufacturing section

Source: based on the CSO study (GUS, 2016)

3. The Data

The empirical study conducted for purposes of this article was based on data from the largest short-term business survey in Poland, called DG1. The target population includes all units employing at least 10 persons. The sample frame includes 98,000 units, of which medium-sized and large units, employing over 49 people, account for 19,000 companies, whereas the remaining 80,000 units are small businesses employing between 10 and 49 people. All medium-sized and large units participate in the survey, and at least a 10% sample of small units. The realised sample contains about 30,000 businesses (GUS, 2017).

The DG1 survey is carried out every month to collect information about basic measures of economic activity in companies, such as sales revenue (from goods and services), the number of employees, gross wages, wholesale and retail sales, excise tax and product-specific subsidies. 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 category contains information used to identify each business unit: its name, address, statistical ID number (REGON), type of main activity according to 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 (Dehnel, 2015).

3.1. Description of the study

The empirical study was limited to small businesses employing between 10 and 49 people, which were active in 2012. The dependent variable in the model was mean monthly revenue obtained by small businesses in June 2012. Two auxiliary variables were selected: *revenue* and *the number of employees* in December 2011. The first variable was taken from the register maintained by the Ministry of Finance and the other one from the ZUS register (the Social Insurance Institution). The selection of auxiliary variables was motivated by data availability. The use of administrative data in statistical practice is associated with certain limitations. One of them is the time delay which often occurs when registered data are made available for purposes of official statistics (Dehnel, 2015). Mean monthly revenue of companies employing between 10 and 49 persons was estimated for 16 provinces and 8 NACE sections: *Manufacturing, Construction, Trade, Transport, Accommodation, Information, Administration, Arts*.

The main idea of the study was to estimate the parameter of interest using composite estimation that accounts for the stratification of enterprises. Direct estimates obtained with the classic Horvitz-Thompson estimator were used as a point of reference for the results. By comparing direct and composite estimates it was possible to notice the change in estimation of mean monthly revenue resulting from the division of enterprises into more homogeneous groups.

4. The Results

The modernisation of estimation by applying new techniques offered by small area estimation (SAE) enables official statistics to meet the demand expressed by data users. 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-sta-

tistical 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 should be evaluated in empirical studies to identify what kind of possibilities and benefits can be achieved in the Polish conditions. The present article is intended as a contribution to this evaluation as it is aimed at investigating the possibility of applying stratification in composite estimators of revenues of small businesses based on information collected from administrative registers maintained by the Ministry of Finance. Before performing estimation and analyzing the results, the first and essential step of the study involved calculating descriptive statistics of the sample size and the distributions of companies depending on the variables of interest by province and NACE section. Table 1 contains information about sample sizes for domain. It can be noticed that all 16 provinces were sampled for each section.

Table 1. Sample size by NACE section

NACE	Min	Q1	Mean	Median	Q3	Max
Manufacturing	127	155	252	227	342	482
Construction	39	51	94	74	132	205
Trade	130	159	259	217	299	605
Transport	18	22	40	31	49	94
Accommodation	16	20	33	27	42	77
Information	5	8	25	17	36	102
Administration	14	16	33	27	39	87
Arts and recreation	9	13	20	19	26	38

Source: based on data from the DG1 survey.

The biggest variation in the distribution of sample size across provinces and NACE section was observed in two sections: *manufacturing* (between 127 and 482 enterprises) and *trade* (between 130 and 605). The level of variation is also considerable for *construction* (between 39 and 205 companies). For the remaining five sections, differences in the sample size between provinces were much smaller, and the sample size did not exceed 102 units. The relationship between the sample size and the size of the general population is illustrated in Figure 3. In line with the assumptions of the DG1 survey, the sample accounts for approximately 10% of the general population.

Regarding the distributions of companies depending on the auxiliary variables it can be noticed that the coefficient of variation for *revenue* varied from 64% to 1212%, and for *the number of employees* varied from 38% to 210%. The distributions were strongly asymmetric, with skewness coefficients ranging for *revenue* from 0.0 to 19 and for *the number of employees* from -2.1 to 9.9.

The hypothesis of homoscedasticity was verified using the White test and the Breusch-Pagan test. For most domains of interest (over 90%) test results confirmed

the hypothesis about the variability of the random component. This in turn justified the application of the above described approach i.e. stratification in order to distinguish more homogenous groups of companies in each domain.

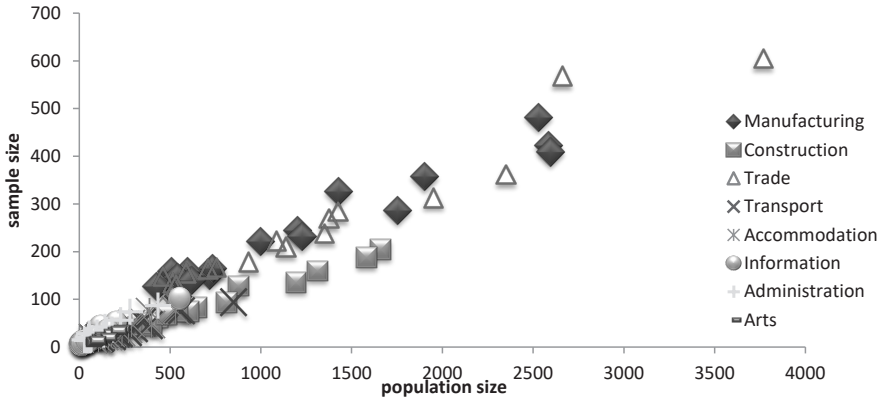


Figure 4. Scatter plot of number of firms in the region and sample size by NACE section
 Source: based on the CSO study (GUS, 2016)

Table 2 presents the characteristics of each strata. The table contains average sizes of regions’ populations and corresponding samples along with means of auxiliary variables: *revenue* and *the number of employees*. *Stratum 1* includes typical observations, while the stratum 2 includes non-typical ones. It can be noticed that the mean size of stratum 1 in each NACE is considerably larger and the means of auxiliary variables are much lower compared to stratum 2.

Table 2. Characteristics of strata for small enterprises by NACE section

NACE	Stratum 1				Stratum 2			
	\bar{N}_d	\bar{n}_d	Mean* revenue (thous. PLN)	Mean number of employees	\bar{N}_d	\bar{n}_d	Mean* revenue (thous. PLN)	Mean number of employees
Manufacturing	1358	234	6 876	32	9	4	696 569	46
Construction	692	82	6 462	29	6	3	67 088	58
Trade	1580	242	14 471	29	7	3	495 888	44
Transport	344	35	7 713	34	5	4	82 842	37
Accommodation	222	26	1 979	25	5	2	13 060	51
Information	149	22	4 705	26	3	2	87 005	46
Administration	145	28	2 736	52	7	3	35 216	58
Arts and recreation	123	17	2 351	33	3	2	31 593	55

* Annual mean per enterprise.

Source: based on data from the DG1 survey and CSO (GUS, 2016)

The study investigated the differences between the value of the estimators obtained based on a specific sample and the value based on the administrative data. The reference values of *revenue* were calculated using the ratio described above. Additionally, to obtain a more thorough evaluation, the composite estimator was compared with the HT estimator, see Figures 4 and 5. The results of this comparison indicate that the application of the composite estimation has reduced the values of estimates in comparison with HT. For nearly all domains of interests (provinces) the *mean revenue* for the HT estimator is overestimated; in contrast, the parameter of interest for GREG estimator for some domains is underestimated. It is composite estimation which produces results that are closest to the reference values.

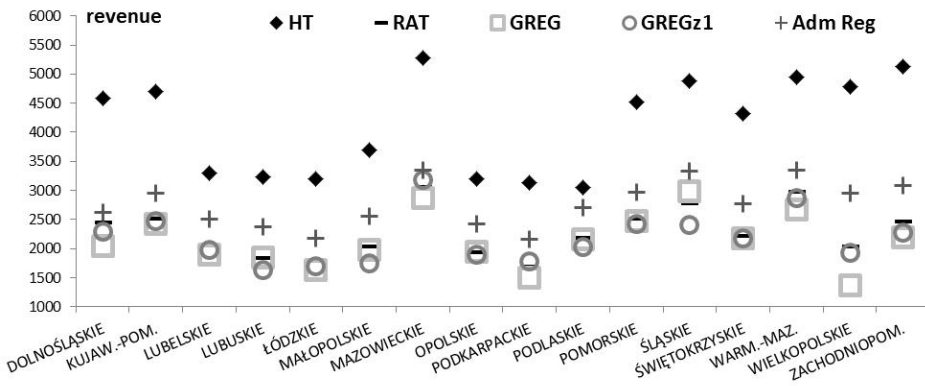


Figure 5. Comparison of estimates of revenue in June 2012 for Manufacturing

Note: HT – \hat{y}_d^{HT} , RAT – $\hat{Y}_{COM,d}^{rat}$, GREG – $\hat{Y}_{COM,d}^{GREG}$, GREGz1 – $\hat{Y}_{COM,d}^{GREG-1}$, Adm Reg – Administrative Register

Source: based on the CSO study (GUS, 2016)

Spatial variation in differences between the estimates and real values is illustrated in maps showing the level of deviation across provinces (see Figure 7). The results of the study indicate that direct HT estimates for most domains are overestimated compared with values from administrative registers. The biggest discrepancies can be observed for the HT estimator. These conclusions confirm the properties of classical Horvitz-Thomson direct estimation. If the sample size is too small and if the sample contains outliers, the direct estimator is very ineffective and may grossly underestimate or overestimate the population totals.

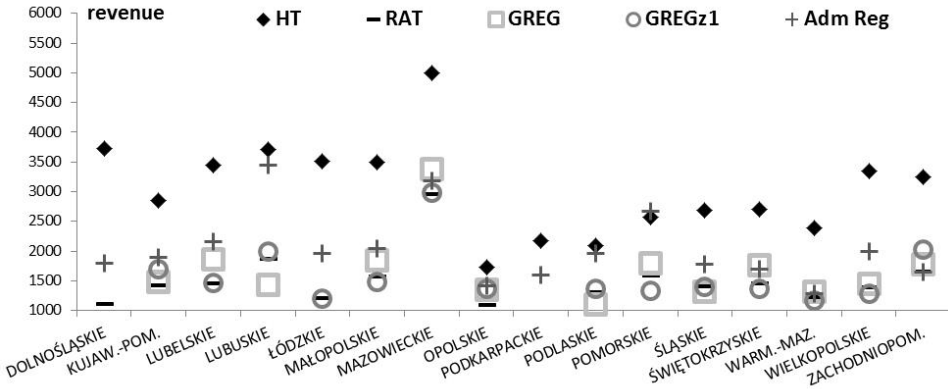


Figure 6. Comparison of estimates of revenue in June 2012 for Construction

Note: HT – \hat{y}_d^{HT} , RAT – $\hat{Y}_{COM,d}^{rat}$, GREG – $\hat{Y}_{COM,d}^{GREG}$, GREGz1 – $\hat{Y}_{COM,d}^{GREG-1}$, Adm Reg – Administrative Register

Source: based on the CSO study (GUS, 2016)

The distribution of relative differences between values of HT estimator and composite estimation, which represent the sum of $\hat{Y}_{SYN(R),d}^{(2)}$ and one of $\hat{Y}_{GREG,d}^{rat(1)}$, $\hat{Y}_{GREG,d}^{(1)}$, $\hat{Y}_{GREG,d}^{(1)}$ estimators of revenues obtained based on data from tax returns is shown in Figure 7. The estimation approach based on the k -means method has generally improved the estimates (in terms of comparison with tax registers), but the degree of gain varied. Differences in estimation quality depend on the type of estimator used for composite estimation and the domain defined by province and NACE section. The reduction of relative differences between estimates of revenues obtained based on data from tax returns and values of the composite estimators compared to HT estimates is greater for domains of a relatively smaller size and when the correlation between the estimated and the auxiliary variable is stronger.

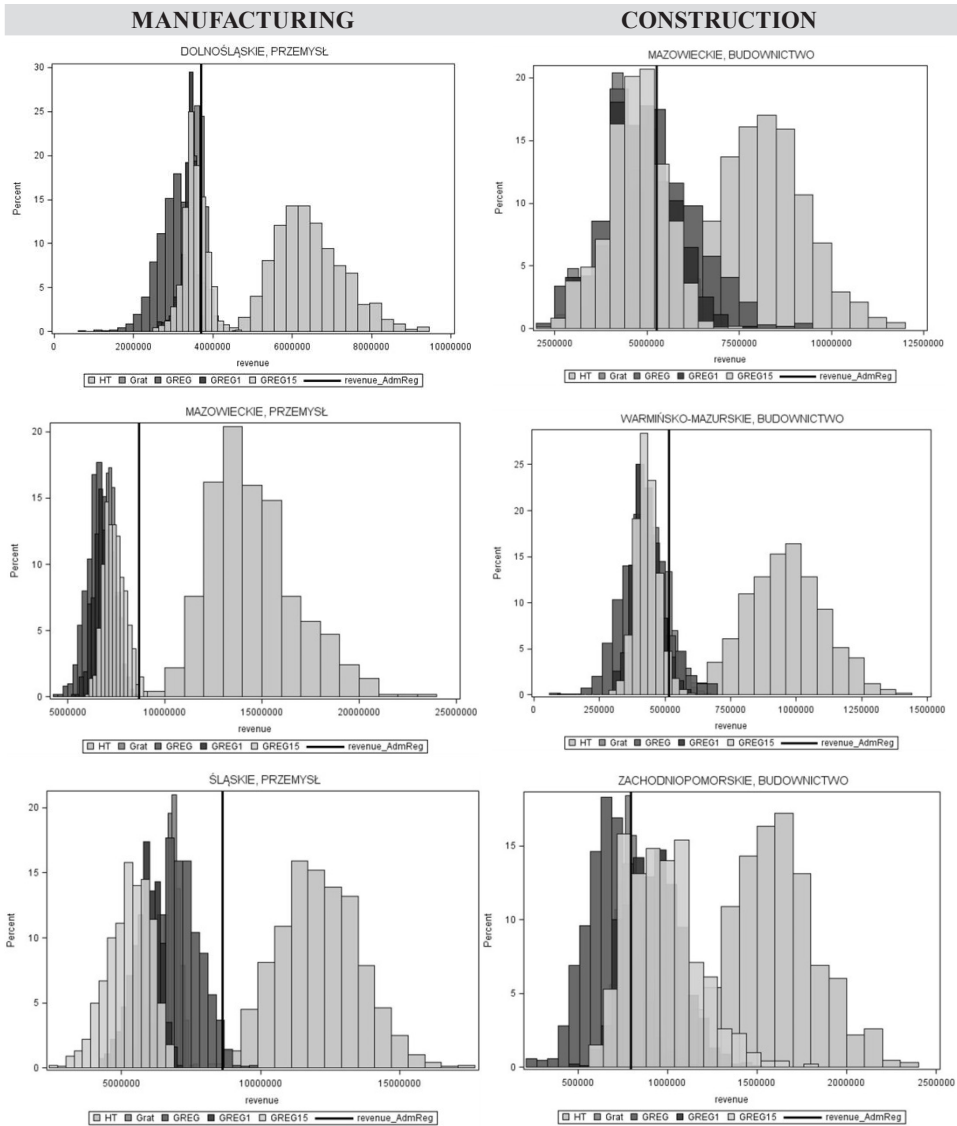


Figure 7. Distribution of estimates of mean revenue for selected provinces and Trade

Note: $HT - \hat{Y}_{HT,d}$, $Grat - \hat{Y}_{COM,d}^{rat}$, $GREG - \hat{Y}_{COM,d}^{GREG}$, $GREG1 - \hat{Y}_{COM,d}^{GREG-1}$, $GREG1,5 - \hat{Y}_{COM,d}^{GREG-1,5}$

Source: based on the CSO study (GUS, 2016)

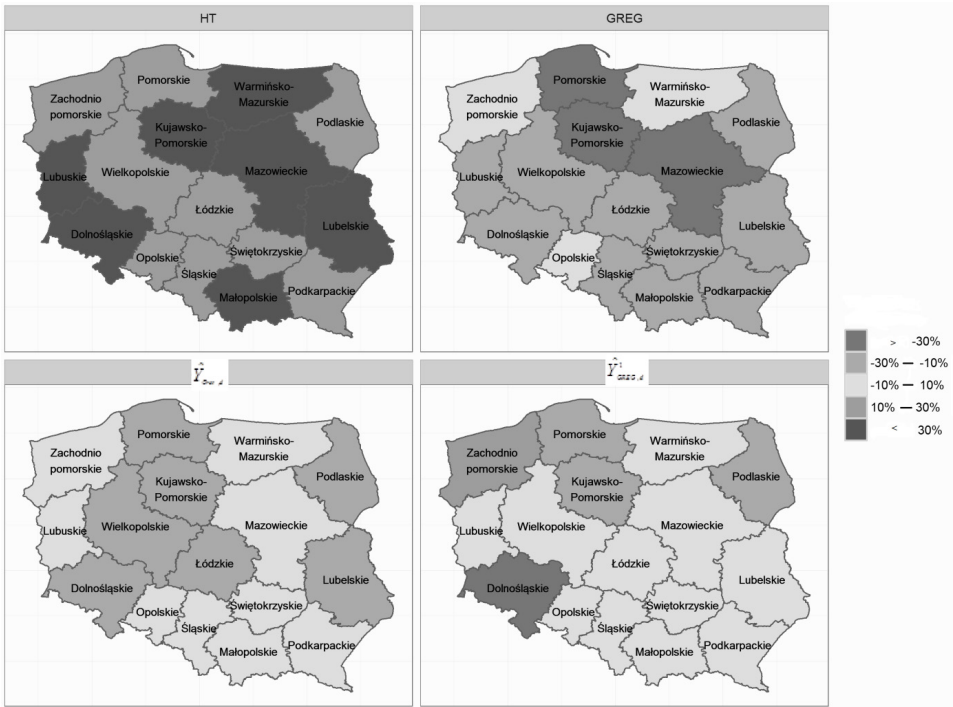


Figure 8. Spatial distribution of relative differences between values of HT estimator and composite estimation, which represent the sum of $\hat{Y}_{SYN(R),d}^{(2)}$ and one of $\hat{Y}_{GREG,d}^{rat(1)}$, $\hat{Y}_{GREG,d}^{(1)}$, $\hat{Y}_{GREG,d}^{1(1)}$ estimators of revenues obtained based on data from tax returns

Note: HT – $\hat{Y}_{HT,d}$, Grac – $\hat{Y}_{COM,d}^{rat}$, GREG – $\hat{Y}_{COM,d}^{GREG}$, GREG1 – $\hat{Y}_{COM,d}^{GREG-1}$

Source: based on data from the DG1 survey and CSO (GUS, 2016)

The quality of estimates obtained using HT estimator ($\hat{Y}_{HT,d}$) and the proposed approach (composite estimators which represent sum of $\hat{Y}_{SYN(R),d}$ and one of $\hat{Y}_{GREG,d}^{rat}$, $\hat{Y}_{GREG,d}^{(1)}$, $\hat{Y}_{GREG,d}^{1(1)}$ (3)) was also assessed by comparing the distributions of coefficients of variation (see Table 3) and relative RMSE (see Table 4). In case of $\hat{Y}_{SYN(R),d}$ estimator weights γ_d were selected arbitrarily: 0.9 for revenue and 0.1 for the number of employees. The values of weights were selected taking into account the strength of the relationship between the dependent variable y and auxiliary variables. The use of composite estimation has improved the estimated precision for the majority domains. The median CV value for seven sections (except for *Arts and recreation*) does not exceed 20%. In addition, composite estimation is characterised for a considerably lower variation in the value of the precision measure (except for *Construction* and *Arts and recreation*). The smallest estimated variation of estimates was observed in the manufacturing and trade sectors represented by the largest number of enterprises.

Table 3. CV of HT and composite estimates of revenue (in %) by province and NACE section

Estimator	Min	Mean	Median	Sx	Max	Min	Mean	Median	Sx	Max
	<i>Manufacturing</i>					<i>Construction</i>				
$\hat{Y}_{HT,d}$	7,6	11,6	11,6	3,4	19,4	8,7	14,8	15,1	5,3	24,8
$\hat{Y}_{COM,d}^{rat} = \hat{Y}_{GREG,d}^{rat(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	3,2	5,7	5,2	2,5	12,8	5,6	19,7	16,3	15,6	53,4
$\hat{Y}_{COM,d}^{GREG} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	4,6	8,3	7,6	3,0	13,1	11,0	30,4	20,0	32,6	103,5
$\hat{Y}_{COM,d}^{GREG-1} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	4,1	6,7	5,8	2,4	11,7	9,5	20,6	18,1	13,6	50,2
$\hat{Y}_{COM,d}^{GREG-1,5} = \hat{Y}_{GREG,d}^{1,5(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	4,1	7,6	7,0	3,2	14,4	9,9	20,0	18,2	10,0	39,4
	<i>Trade</i>					<i>Transport</i>				
$\hat{Y}_{HT,d}$	6,1	10,4	10,8	2,1	14,0	12,1	16,7	15,4	4,5	24,0
$\hat{Y}_{COM,d}^{rat} = \hat{Y}_{GREG,d}^{rat(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	1,8	6,2	4,5	4,5	16,9	5,3	7,5	7,3	2,1	10,9
$\hat{Y}_{COM,d}^{GREG} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	2,7	7,1	5,5	4,1	16,7	7,6	12,9	12,1	4,2	20,9
$\hat{Y}_{COM,d}^{GREG-1} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	2,2	7,2	6,3	4,0	15,8	6,5	11,1	9,9	5,4	23,3
$\hat{Y}_{COM,d}^{GREG-1,5} = \hat{Y}_{GREG,d}^{1,5(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	3,2	8,4	7,6	4,0	15,8	6,5	11,3	9,5	4,9	21,8

Estimator	Min	Mean	Median	Sx	Max	Min	Mean	Median	Sx	Max
	<i>Manufacturing</i>					<i>Construction</i>				
	<i>Accommodation</i>					<i>Information</i>				
$\hat{Y}_{HT,d}$	8,7	12,4	12,3	3,1	16,3	13,2	21,6	21,0	6,6	33,8
$\hat{Y}_{COM,d}^{rat} = \hat{Y}_{GREG,d}^{rat(1)} +$ $\hat{Y}_{SYN(R),d}^{(2)}$	8,5	16,6	13,9	10,1	30,0	4,7	10,8	9,0	6,5	23,5
$\hat{Y}_{COM,d}^{GREG} = \hat{Y}_{GREG,d}^{(1)} +$ $\hat{Y}_{SYN(R),d}^{(2)}$	9,0	17,2	14,3	10,6	31,4	6,0	14,6	13,4	7,7	27,3
$\hat{Y}_{COM,d}^{GREG-1} = \hat{Y}_{GREG,d}^{(1)} +$ $\hat{Y}_{SYN(R),d}^{(2)}$	7,5	16,6	13,7	11,4	31,7	5,7	14,3	13,0	7,8	27,3
$\hat{Y}_{COM,d}^{GREG-1,5} = \hat{Y}_{GREG,d}^{1,5(1)} +$ $\hat{Y}_{SYN(R),d}^{(2)}$	7,8	16,9	13,7	11,6	32,1	5,8	14,8	13,2	7,4	27,3
	<i>Administration</i>					<i>Arts and recreation</i>				
$\hat{Y}_{HT,d}$	13,3	19,6	19,0	3,9	26,5	35,2	42,0	43,1	7,3	53,0
$\hat{Y}_{COM,d}^{rat} = \hat{Y}_{GREG,d}^{rat(1)} +$ $\hat{Y}_{SYN(R),d}^{(2)}$	5,9	11,5	10,9	3,6	17,8	17,3	58,7	35,6	64,1	172,4
$\hat{Y}_{COM,d}^{GREG} = \hat{Y}_{GREG,d}^{(1)} +$ $\hat{Y}_{SYN(R),d}^{(2)}$	6,9	13,3	13,4	3,5	19,2	33,9	75,6	55,0	66,3	192,9
$\hat{Y}_{COM,d}^{GREG-1} = \hat{Y}_{GREG,d}^{(1)} +$ $\hat{Y}_{SYN(R),d}^{(2)}$	5,7	13,2	13,3	4,9	21,9	25,0	70,5	44,6	73,6	200,2
$\hat{Y}_{COM,d}^{GREG-1,5} = \hat{Y}_{GREG,d}^{1,5(1)} +$ $\hat{Y}_{SYN(R),d}^{(2)}$	5,9	14,3	14,8	5,9	27,2	22,1	70,9	45,7	76,4	204,8

Source: based on data from the DG1 survey and CSO (GUS, 2016)

In terms of accuracy the composite estimators are considerably worse compared to direct ones (or: to the direct one). For most values of estimated parameters, RRMSE performance of composite estimators are significantly higher than the CV of the direct estimator (see Tables 3 and 4). The results of the study indi-

cate that the bias of composite estimators is high, especially in the case of the domains represented by the smallest number of enterprises – *Arts and recreation* and *Administration* (see Table 2).

Table 4. Relative RMSE and composite estimates of revenue (in %) by province and NACE section

Estimator	Min	Mean	Median	Sx	Max	Min	Mean	Median	Sx	Max
	<i>Manufacturing</i>					<i>Construction</i>				
$\hat{Y}_{COM,d}^{rat} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	8,3	23,6	24,5	10,7	46,6	19,3	64,4	33,4	62,7	171,7
$\hat{Y}_{COM,d}^{GREG} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	12,0	25,9	26,9	11,1	48,6	14,2	110,3	35,2	142,5	356,8
$\hat{Y}_{COM,d}^{GREG-1} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	10,7	24,7	25,6	11,5	46,2	11,5	59,8	37,3	55,3	153,1
$\hat{Y}_{COM,d}^{GREG-1,5} = \hat{Y}_{GREG,d}^{1,5(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	9,7	21,6	17,4	12,6	49,6	7,3	27,2	21,7	18,9	52,6
	<i>Trade</i>					<i>Transport</i>				
$\hat{Y}_{COM,d}^{rat} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	12,5	24,6	26,2	5,5	31,5	11,0	47,5	49,7	22,8	83,0
$\hat{Y}_{COM,d}^{GREG} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	17,1	26,3	25,6	5,7	33,5	20,8	51,8	50,5	22,9	97,1
$\hat{Y}_{COM,d}^{GREG-1} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	17,3	27,4	26,8	7,4	41,5	17,4	45,6	37,7	26,0	101,0
$\hat{Y}_{COM,d}^{GREG-1,5} = \hat{Y}_{GREG,d}^{1,5(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	12,8	27,3	23,7	10,8	46,6	24,6	48,0	34,7	25,4	99,7

Estimator	Min	Mean	Median	Sx	Max	Min	Mean	Median	Sx	Max
	Manufacturing					Construction				
	Accommodation					Information				
$\hat{Y}_{COM,d}^{rat} = \hat{Y}_{GREG,d}^{rat(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	22,6	41,1	41,1	18,5	59,6	18,3	31,7	33,6	8,5	41,4
$\hat{Y}_{COM,d}^{GREG} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	19,6	40,2	40,2	20,5	60,7	19,0	34,4	35,2	11,0	48,1
$\hat{Y}_{COM,d}^{GREG-1} = \hat{Y}_{GREG,d}^{1(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	19,0	40,9	40,9	21,9	62,7	20,2	35,4	34,4	13,1	52,8
$\hat{Y}_{COM,d}^{GREG-1,5} = \hat{Y}_{GREG,d}^{1,5(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	20,1	42,4	42,4	22,2	64,6	21,4	34,4	30,0	13,9	56,1
	Administration					Arts and recreation				
$\hat{Y}_{COM,d}^{rat} = \hat{Y}_{GREG,d}^{rat(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	20,1	56,2	56,5	23,7	108,6	17,6	203,6	161,8	172,6	505,8
$\hat{Y}_{COM,d}^{GREG} = \hat{Y}_{GREG,d}^{(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	26,3	56,6	45,7	25,6	114,8	25,3	194,7	158,0	145,4	404,1
$\hat{Y}_{COM,d}^{GREG-1} = \hat{Y}_{GREG,d}^{1(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	25,6	55,8	46,6	28,7	121,5	28,1	165,0	146,0	128,2	350,9
$\hat{Y}_{COM,d}^{GREG-1,5} = \hat{Y}_{GREG,d}^{1,5(1)} + \hat{Y}_{SYN(R),d}^{(2)}$	8,1	54,1	46,6	32,6	124,8	29,7	189,0	183,6	119,0	359,0

Source: based on data from the DG1 survey and CSO (GUS, 2016)

The spatial distribution of *mean revenue* estimates across provinces for four NACE sections is shown in Figure 9 – based on the composite estimates ($\hat{Y}_{COM,d}^{GREG}$) estimator.

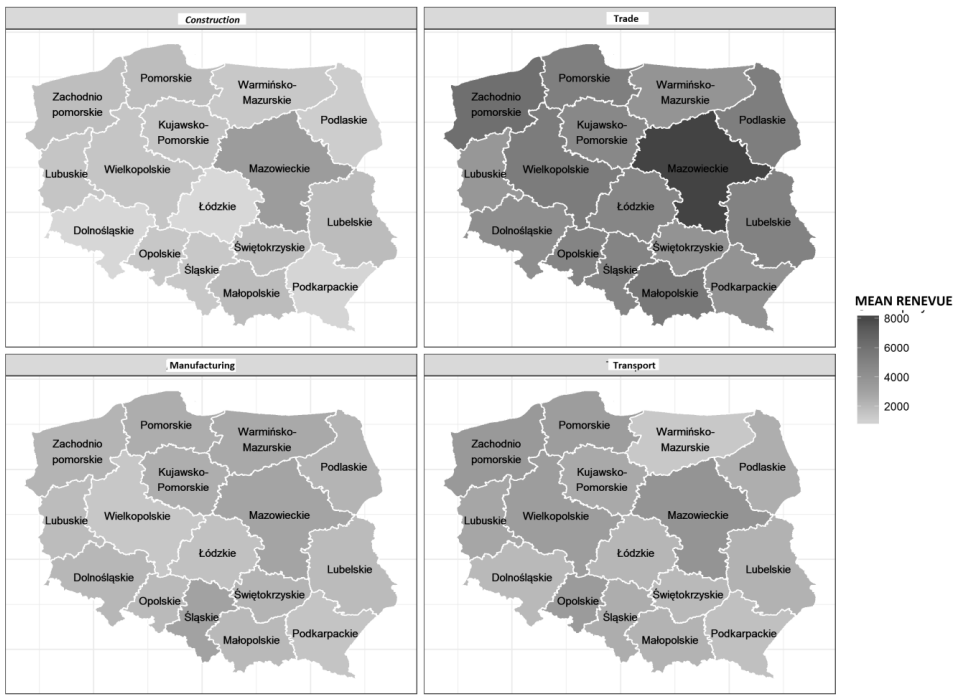


Figure 9. Spatial distribution of composite estimates ($\hat{Y}_{COM,a}^{GREG}$) of mean revenue by province and for 4 NACE sections

Source: based on data from the DG1 survey and CSO (GUS, 2016)

5. Conclusions

The aim of the study reported in this article was to produce reliable estimates of mean revenue of Polish companies employing fewer than 50 people and conducting activity classified into eight NACE sections by applying stratification in composite estimators based on information collected from administrative registers maintained by the Ministry of Finance.

The article is intended as a contribution to evaluation of applying new techniques offered by small area estimation (SAE) to business survey. This approach is a proposal of a new tool enabling official statistics to meet the demand expressed by data users by providing reliable estimates at a low level of aggregation. By relying on auxiliary variables from the administrative register and applying the k -means method to composite estimation, it was possible to produce estimates for most domains at a previously unpublished level of aggregation, with acceptable accuracy, measured in terms of RRMSE values. Regarding the precision, it can

be noticed that there is a relationship between section size and estimation precision: for sections represented by more companies, estimation precision tends to be higher. However, the gain in estimation precision within one section is greater for domains of a relatively smaller size and when the correlation between the estimated and the auxiliary variable is stronger.

A significant improvement in estimation quality could be obtained by selecting an adequate estimator with properly designated strata. In the case of a large number of small domains, the required diagnostic process decreases the practical usefulness of the proposed approach.

References

- Antal E., Tillé Y. (2011), *A Direct Bootstrap Method for Complex Sampling Designs From a Finite Population*, "Journal of the American Statistical Association", vol. 106(494), pp. 534–543.
- Bracha C. (2004), *Estymacja danych z badania aktywności ekonomicznej ludności na poziomie powiatów dla lat 1995–2002*, GUS, Warszawa.
- Chambers R., Chandra H., Salvati N., Tzavidis N. (2014), *Outlier robust small area estimation*, "Journal of the Royal Statistical Society: Series B", vol. 76(1), pp. 47–69.
- Chambers R.L., Falvey H., Hedlin D., Kocic P. (2001), *Does the Model Matter for GREG Estimation? A Business Survey Example*, "Journal of Official Statistics", vol. 17, no. 4, pp. 527–544.
- Clark R.G., Kocic P., Smith P.A. (2017), *Comparison of two Robust Estimation Methods for Business Surveys*, "International Statistical Review", vol. 85, no. 2, pp. 270–289, <http://dx.doi.org/10.1111/insr.12177>.
- Cochran W.G. (1977), *Sampling Techniques*, John Wiley and Sons, New York.
- Dehnel G. (2015), *Robust regression in monthly business survey*, [in:] W. Okrasa (ed.), *Statistics in Transition – new series*, vol. 16, no. 1, Warsaw, pp. 1–16, <http://stat.gov.pl/en/sit-en/issues-and-articles-sit/previous-issues/volume-16-number-1-spring-2015/> [accessed: 29.10.2018].
- Dehnel G. (2017), *GREG estimation with reciprocal transformation for a Polish business survey*, [in:] M. Papież, S. Śmiech (eds.), *Proceedings of the 11th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*, Foundation of the Cracow University of Economics, Crakow, pp. 67–75.
- Dehnel G., Pietrzak M., Wawrowski Ł. (2017), *An Evaluation of Company Performance Using the Fay-Herriot Model*, "Argumenta Oeconomica Cracoviensia", no. 16, pp. 23–36. <http://dx.doi.org/10.15678/AOC.2017.1602>.
- Guadarrama M., Molina I., Rao J.N.K. (2016), *A comparison of small area estimation methods for poverty mapping*, "Statistics in Transition New Series and Survey Methodology", vol. 17, no. 1, pp. 41–66, <http://stat.gov.pl/en/sit-en/issues-and-articles-sit/previous-issues/volume-17-number-1-march-2016/> [accessed: 29.10.2018].
- GUS (2015), *Małe i średnie przedsiębiorstwa niefinansowe w latach 2009–2013*, Warsaw.
- GUS (2016), *Report "Use of administrative data in the survey: Assessment of current business activity of enterprises"*, Warsaw.
- GUS (2017), *Działalność przedsiębiorstw niefinansowych w 2015 roku*, Warsaw.
- Myrskylä M. (2007), *Generalised Regression Estimation for Domain Class Frequencies*, Tilastokeskus – Statistikcentralen – Statistics Finland, Helsinki.
- PARP (2017), *Raport o stanie sektora MSP w Polsce 2017*, Warsaw.
- Rao J.N.K., Molina I. (2015), *Small area estimation. Wiley series in survey methodology*, 2nd ed., Wiley, Hoboken.


- Rao J.N.K., Wu C.F.J. (1988), *Resampling Inference With Complex Survey Data*, "Journal of the American Statistical Association", vol. 83(401), pp. 231–241.
- Särndal C.E., Swensson B., Wretman J. (1992), *Model Assisted Survey Sampling*, Springer Verlag, New York.
- Shao J., Tu D. (1995), *The jackknife and bootstrap*, Springer Verlag, New York.
- Singh M.P., Gambino J.G., Mantel H. (1993), *Issues and options in the provision of small area statistics*, [in:] G. Kalton, J. Kordos, R. Platek (eds.), *Proceedings of the International Scientific Conference on Small Area Statistics and Survey Designs*, vol. 1, Central Statistical Office, Warsaw, pp. 37–75.

Warstwowanie domen z użyciem estymacji złożonej w ocenie małych przedsiębiorstw w Polsce

Streszczenie: Aby sprostać rosnącemu zapotrzebowaniu na szczegółową, dokładną i terminową ocenę dotyczącą przedsiębiorczości i gospodarki, konieczne jest między innymi systematyczne rozszerzanie zakresu informacji dostarczanych przez statystykę gospodarczą. Prowadzona obecnie polityka, mająca na celu zmniejszenie kosztów badań i obciążeń sprawozdawczych podmiotów gospodarczych, wymusza działania polegające na modernizacji metodologii badań. Jeden z obszarów, w których prowadzone są tego rodzaju analizy, to aplikacje w zakresie estymacji pośredniej wykorzystującej dane o zmiennych pomocniczych pochodzących z systemów administracyjnych. Stąd też celem artykułu jest ocena jakości szacunku przychodów dla małych przedsiębiorstw przy wykorzystaniu zaproponowanego podejścia metodycznego polegającego na możliwości włączenia stratyfikacji do estymacji złożonej, w oparciu o informacje pochodzące z rejestrów administracyjnych.

Słowa kluczowe: estymacja odporna, statystyka przedsiębiorstw, statystyka małych obszarów, GREG

JEL: C40

	<p>© by the author, licensee Łódź University – Łódź University Press, Łódź, Poland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license CC-BY (http://creativecommons.org/licenses/by/3.0/)</p> <p>Received: 2017-12-30; verified: 2018-07-28. Accepted: 2018-12-07</p>
---	--