Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663
www.czasopisma.uni.lodz.pl/foe/
2(334) 2018

DOI: http://dx.doi.org/10.18778/0208-6018.334.03

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

The Use of the Robust GREG Estimator to Estimate Small Trade Firms[1]

Abstract: In the face of dynamic changes in the economy, there is a growing demand for multivariate statistics for cross-classified domains. In economic statistics, this demand poses a particular challenge owing to the unique character of the population of enterprises, which is what motivates the search for estimation methods that can exploit administrative sources to a greater extent. The adoption of new solutions in this area is expected to increase the scope of statistical outputs and improve the efficiency of estimates. The purpose of the presented study is to test the application of the robust GREG estimator based on the LS method and least median of squares regression to estimate characteristics of small trade firms operating in 2012. The estimation process is supported with delayed variables from administrative registers used as auxiliary variables. The paper refers to small area estimation methods. The variables of interest are estimated at the low level of aggregation represented by cross-section province and NUTS 2.

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. Introduction

The Polish economy has undergone dramatic changes in the recent years. Its present form has been shaped by a series of dynamic economic transformations. Its main driving force is generated by small and medium-sized companies. While 90% of these entities are micro-companies, one cannot overlook the role played by small companies, i.e. those employing between 10 and 49 people. These companies are characterised by a considerable degree of flexibility and enterprise. They are able to compete with the largest companies, thanks to tight cost control, quick responsiveness to changing market requirements and the ability to quickly implement innovation. Their revenues account for about half of the entire SME sector (46.8% in 2013). From the perspective of business classification, one of the most important sections in this sector, and the most numerous, is *Trade*. The section comprises companies involved in retail and wholesale trade and firms specialising in the repair of motor vehicles and motorcycles (according to Polish Business Classification). Trade firms account for 30% of all small companies (see Figure 1) and their revenue makes up 20% of revenue generated by all small and medium-sized enterprises (SMEs) in the Trade section (GUS, 2014).

Trade firms do not require large investments to start and conduct business activity, which is the case in other areas, and they can expect to see positive financial results after a relatively short time (GUS, 2015). These characteristics favour the creation of new firms, especially in the sector of retail trade. However, owing to intense competition, especially from well-developed retail chains, death rates in this group of businesses are very high. Only 28.7% of trade companies established in 2009 survived until 2013, which corresponds to a considerably lower survival rate than the average for the entire SME group (35.6%).

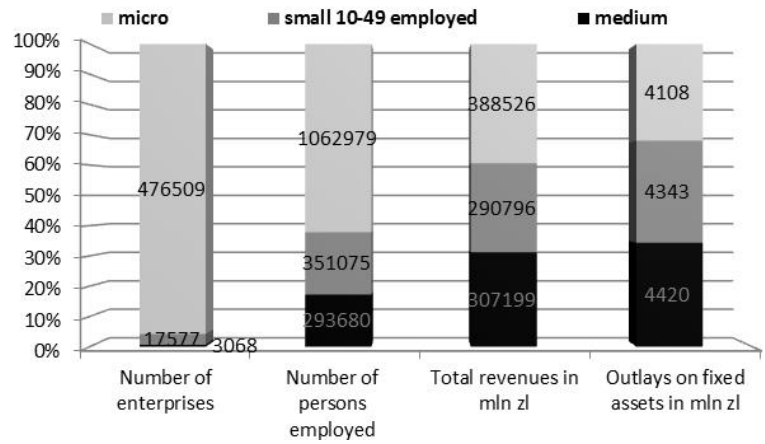

Figure 1. SMEs in the Trade section by size class in 2013

Source: based on the CSO study (GUS, 2015)

Small trade firms employ about 29% of all people employed in small companies. In the period 2009–2013, the group of small trade enterprises saw quite dynamic growth. The number of small enterprises rose faster than the number of companies in other size categories of the trade section, with the average annual growth rate of 4%, see Figure 2, Table 1. This increase contributed to a rise in the number of people employed in small trade firms, but at a slower annual rate of only 1.5%. In 2013, people employed in small enterprises accounted for 16.2% of all employees working in the trade industry, compared to 14.7% in 2009. In 2013, the average employment in this group was 20 people (GUS, 2015).

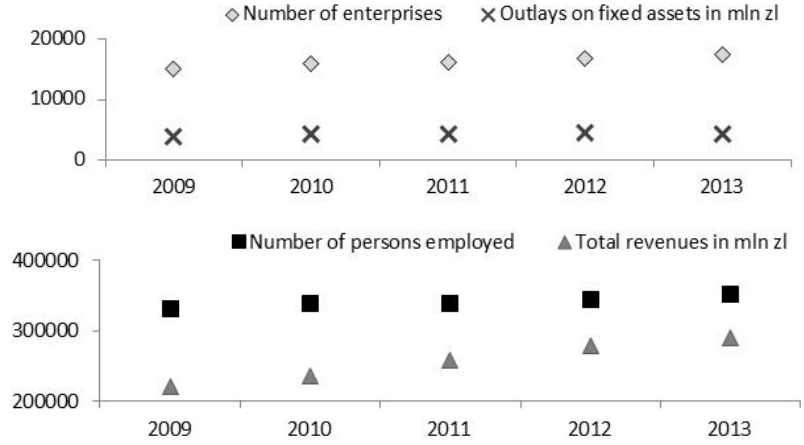

Figure 2. Basic information on enterprises by principal activity and by size class in 2009–2013
Source: based on the CSO study (GUS, 2015)

Table 1. Basic information on enterprises by principal activity and by size class in 2009–2013

Years	Number of persons employed	Total revenues in mln zl	Number of enterprises	Outlays on fixed assets in mln zl
2009	331 400	219 953	15 195	3841
2010	337 888	236 522	15 962	4190
2011	339 161	258 911	16 134	4279
2012	343 576	278 603	16 731	4398
2013	351 075	290 796	17 577	4343

Source: based on the CSO study (GUS, 2015)

In 2009–2013, small trade enterprises registered the highest revenue increase in the SME sector (by 32.2%): the average total revenue per one entity rose from 14.5 million PLN in 2009 to 16.5 million PLN in 2013, see Table 2, Figure 3.

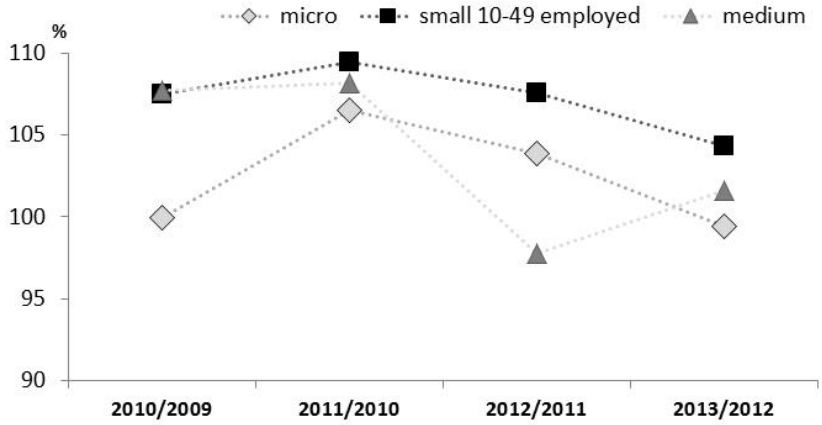

Figure 3. Changes in total revenues of SMEs in the trade section in 2009–2013
Source: based on the CSO study (GUS, 2015)

Table 2. Total revenues of SMEs in the trade section in 2009–2013

Total revenues in mln zl	2009	2010	2011	2012	2013
Micro	353 097	353 036	376 008	390 660	388 526
Small 10–49 employed	219 953	236 522	258 911	278 603	290 796
Medium	265 030	285 627	309 094	302 216	307 199
SME	838 080	875 185	944 013	971 479	986 521
Enterprises total	1 154 922	1 215 612	1 317 624	1 370 226	1 399 443

Source: based on the CSO study (GUS, 2015)

An analysis of changes in financial results of small trade enterprises on the basis of information published by the Central Statistical Office (CSO) can only be conducted at the national level. The territorial variation across the provinces can only be investigated by applying small area estimation methods. The presented article describes the results of a study aimed at estimating enterprise characteristics cross-classified by activity type (section) and territorial domain. The main objective of the study was to estimate mean vrevenue of small trade firms across the provinces using robust GREG estimation, which takes advantage of the LS method and least median of squares (LMS) regression. Auxiliary variables for the estimation were delayed variables from administrative registers.

2. Estimation methods

2.1. Direct estimation

The most common estimator in survey methodology is the Horvitz-Thompson (1952) estimator, further abbreviated as HT. Consider sample s drawn from population U, where s_d is a sub-sample from domain d. Counties in domains meet the condition $n_d < N_d$, where n_d denotes the 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. The estimator of the mean in domain d is given by the formula:

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

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

The direct HT estimator is design-unbiased and design-consistent assuming that $n_d \to \infty$. 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 $n_d = 0$. In the study, the HT estimator was used as a benchmark for evaluating other estimation methods.

2.2. The robust GREG estimator

The estimation of variables for small domains associated with business entities poses a particular challenge owing to the characteristic distribution of units with respect to the variables of interest. One general approach proposed as a way of tackling this problem involves modifying the sample, and thus creating an estimator which is resistant to large residuals. In this procedure, sampled units for which the variable of interest lies outside certain cut-off points are modified. One example of this approach is winsorization (Chambers et al., 2000). The method consists in splitting the sample into two groups using cut-off points. One group contains observations used to build a model, which are left unmodified. The other group contains outliers, which are included in the sample after modification. Model parameters are estimated on the basis of the modified sample using the winsorized estimator. The classic form of the GREG estimator is given by the formula (Rao, Molina, 2015):

$$\hat{Y}_{\text{GREG},d} = \sum_{i \in U_d} \hat{y}_i + \sum_{i \in s_d} w_i e_i = \sum_{i \in s_d} w_i g_i y_i \quad \hat{y}_i = x'_i \hat{\beta}_d \,. \tag{1}$$

In addition to the variable of interest, the GREG estimator requires auxiliary variables x_i.

The winsorized estimator is given by:

$$\hat{Y}_{\text{win}} = \sum_{i \in s_d} \tilde{w}_i y_i^* = \sum_{i \in s_d} w_i g_i y_i^*, \tag{2}$$

where modified values of study variable y_i^* are calculated by the following formula (Gross, Bode, Taylor, Lloyd-Smith, 1986; Dehnel, 2014):

$$y_i^* = \begin{cases} \left(\frac{1}{\tilde{w}_i}\right) y_i + \left(1 - \frac{1}{\tilde{w}_i}\right) K_{\text{Ui}} & \text{if} \quad y_i > K_{\text{Ui}} \\ y_i & \text{if} \quad K_{\text{Li}} \le y_i \le K_{\text{Ui}} \\ \left(\frac{1}{\tilde{w}_i}\right) y_i + \left(1 - \frac{1}{\tilde{w}_i}\right) \text{K}_{\text{Li}} & \text{if} \quad y_i < K_{\text{Li}} \end{cases} \tag{3}$$

$$g_i = \left(1 + x_i' \left(\sum_{i \in s_d} w_i x_i x_i'\right)^{-1} \left(t_x - \sum_{i \in s_d} w_i x_i\right)'\right), \tag{4}$$

where:

s_d – population parameter for domain d,

$U = \{1, \ldots, i, \ldots, N\}$ – general population of size N,

$s (s \subseteq N)$ – sample,

$\tilde{w}_i = w_i g_i$,

$w_i = 1/\pi_i$ – sampling weights,

g_i – weights dependent on the value of the vector of auxiliary variables for the sampled units,

$x_i = (x_{1i}, \ldots, x_{ki}, \ldots, x_{Ki})'$ – vector of auxiliary variables,

$t_x = \sum_{i \notin U} x_i$ – population total,

K_{Ui} – upper cut-off value,

K_{Li} – lower cut-off value.

In order to compute cut-off values $\hat{K}_{Ui}$ and $\hat{K}_{Li}$, various methods presented in the publications of Preston and Mackin (2002) and Dehnel (2014) can be used. Two of them were used in this study: the LS method and the LMS method (Pres-

ton, Mackin, 2002). They were described by Rousseeuw and Leroy (2003). The LMS technique should be more robust than LS because the OLS regression model is fitted in the absence of outlying units, without totally removing them (Preston, Mackin, 2002).

3. Description of the study

The target population for the empirical study consisted of small companies (10–49 employees) conducting activity classified into the *Trade* section. The analysed model consisted of the dependent variable – *mean revenue* obtained by companies in June 2012. Various variables derived from different administrative registers were considered as auxiliary variables. Finally, 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 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). The estimation was conducted at the level of provinces.

4. Assessment of estimates obtained in the study

The precision of the obtained estimates was tested using the bootstrap method. 1000 bootstrap samples were drawn from the original sample and used to estimate the value of *revenue* for June 2012 across domains of interest. The efficiency of estimation was evaluated by calculating the coefficient of variation for the estimator (Bracha, 2004):

$$\mathrm{CV}\left(\hat{Y}_d\right) = \frac{\sqrt{\mathrm{Var}\left(\hat{Y}_d\right)}}{\hat{Y}_d}. \tag{5}$$

To estimate bias, it is necessary to know the value of the estimated parameter for the general population. In the absence of this information, it was estimated indirectly, based on data from tax returns 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 (DG–1) is constant (see Figure 4).

$$\frac{\text{revenue\_AR}}{\text{revenue\_DG1}} = \frac{\text{revenue\_est}}{\text{revenue\_DG1}}. \tag{6}$$

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

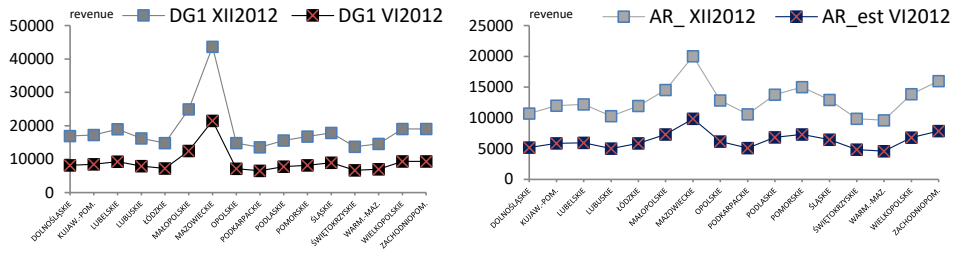

Figure 4. Value of revenue in June and December 2012 reported in the DG–1 survey and in tax returns for companies classified into the Manufacturing section

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

5. Results of the empirical study

The point of reference for the evaluation of estimation precision was the estimate obtained using the classic HT estimator and GREG. The measure of efficiency (CV) indicates that the HT estimator is characterised by the highest variation (see Figure 5). The variation is lower for the GREG estimator based on auxiliary variables from administrative registers and in the case of the GREG estimator based on least median of squares regression. A considerable reduction in variation is also achieved by the application of the GREG estimator based on the LS method. The biggest difference between the HT and GREG estimates can be observed for the Lubelskie and Małopolskie provinces – a decrease in CV from 32% to 7% and from 28% to 5% respectively.

The reference values for the estimation of *revenue* were calculated using the ratio described above. Additionally, to obtain a more thorough evaluation, the winsorized GREG estimator was compared with the HT and GREG estimators, see Figure 6. The results of this comparison indicate that the application of robust GREG estimation has considerably improved the accuracy of estimates in comparison with HT or GREG estimation. For nearly all domains of interests (provinces), the HT estimates of *mean revenue* are significantly overestimated; in contrast, the GREG estimator underestimates the parameter of interest for some domains. The largest discrepancy between estimates for different domains can be observed for the provinces for which the auxiliary variables used in the model were characterised by the highest dispersion.

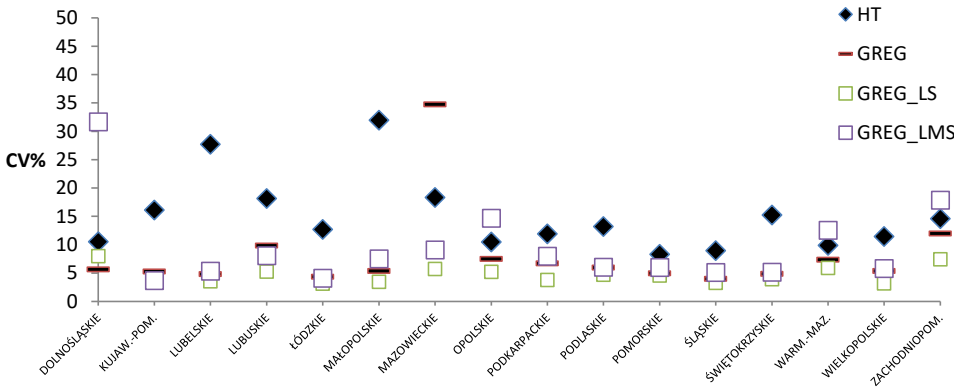

Figure 5. Comparison of estimation precision for Trade

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

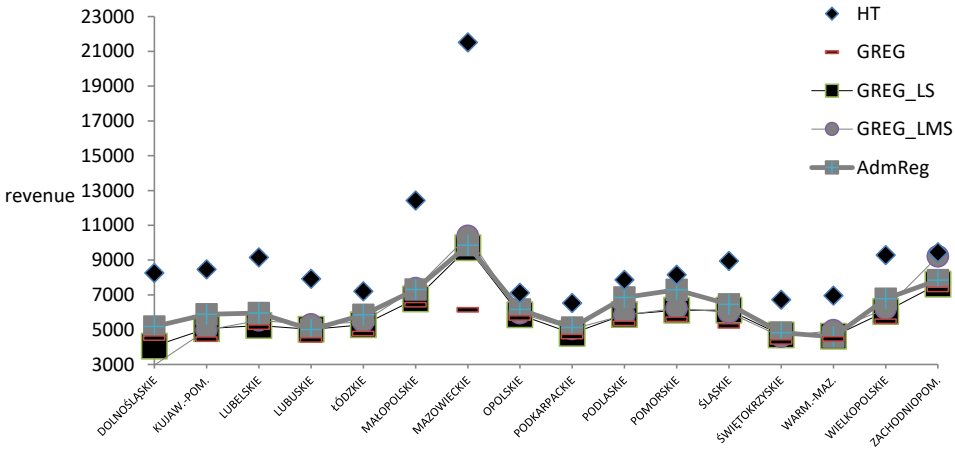

Figure 6. Comparison of estimates of revenue in June 2012 for Trade

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

Interesting conclusions can be drawn from analysing the distributions of estimates obtained by the bootstrap method (see Figure 4). According to the survey sampling theory, the HT and GREG estimators are unbiased by definition. The empirical results, however, indicate that they are, in fact, biased. This can be explained by the sample bias and influence of outliers. While the inclusion of auxiliary variables in GREG estimation does improve estimation precision, it is winsorized GREG estimation that produces results that are closest to the reference values.

The distribution of *mean revenue* estimates across the provinces is shown in Figure 7 – based on the HT, GREG and winsorized GREG estimator. In general, direct estimates of *mean revenue* are characterised by greater variability than the GREG and winsorized GREG estimates (see Figure 8). The final results obtained with the winsorized GREG_LS estimates (with the highest estimation precision)

are described below. The highest estimates of *mean revenue* were obtained for the Mazowieckie (9,705 PLN) and Zachodniopomorskie (7,616 PLN) Provinces. The lowest value of *mean revenue* was estimated for the Dolnośląskie (4,059 PLN) and Warmińsko-Mazurskie (4,630 PLN) provinces – which is almost half of the value calculated for the Mazowieckie Province, despite the fact that these are neighbouring regions.

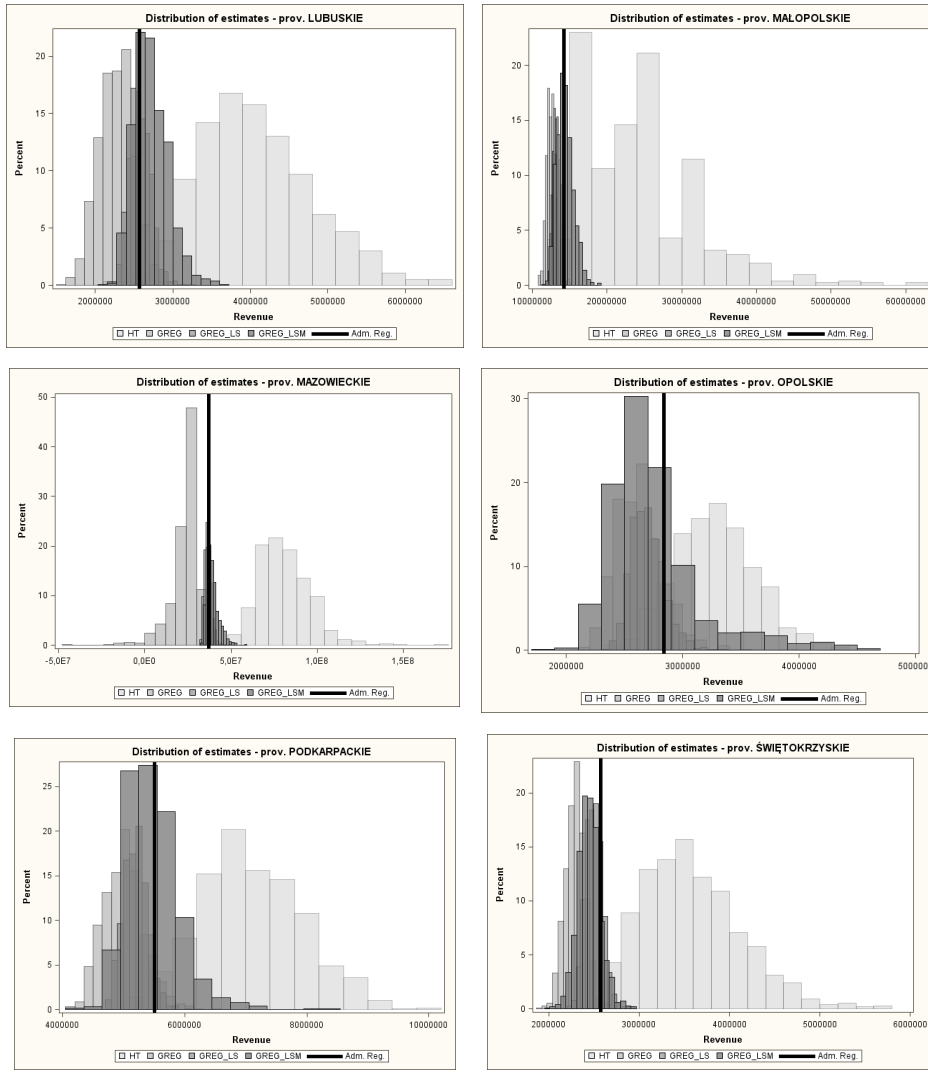

Figure 7. Distribution of estimates of mean revenue for selected provinces and Trade

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

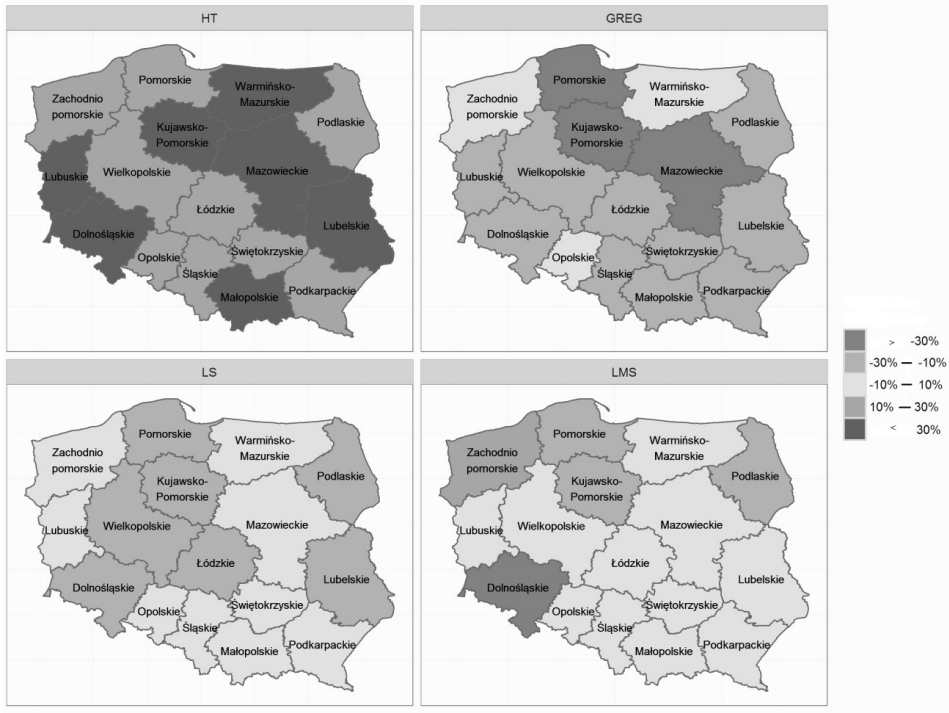

Figure 8. Spatial distribution of relative bias for the HT, GREG, LS and LMS estimates of mean revenue compared to data from tax returns

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

6. Conclusions

The analysis of the regional variation between small trade firms was conducted using methods provided by small area estimation. Additionally, the study made use of robust regression. The application of delayed auxiliary variables from administrative registers improved estimation precision, both in the case of the classic GREG estimator and its modified version based on winsorization. The use of winsorization not only resulted in a precision gain but also improved accuracy. Each of the three methods of robust regression were characterised by similar estimation quality. The observed differences resulted from how resistant each method was to the influence of outliers. More robust techniques produced a greater gain in efficiency. The evaluation of estimation quality in terms of accuracy has revealed that sample modification using winsorization helps to reduce bias resulting from the presence of outliers.

References

Bracha C. (2004), *Estymacja danych z badania aktywności ekonomicznej ludności na poziomie powiatów dla lat 1995–2002*, GUS, Warszawa.

Chambers R., Kokic P., Smigh P., Cruddas M. (2000), *Winsorization for Identifying and Treating Outliers in Business Surveys, Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria.

Dehnel G. (2014), *Winsorization Methods in Polish Business Survey*, "Statistics in Transition – New Series", vol. 15, no. 1, pp. 97–110, http://pts.stat.gov.pl/czasopisma/statistics-in-transition/ [accessed: 25.11.2017].

Dehnel G. (2015), *Rejestr podatkowy oraz rejestr ZUS jako źródło informacji dodatkowej dla statystyki gospodarczej – możliwości i ograniczenia*, [in:] K. Jajuga, M. Walesiak (eds.), *Taksonomia 24. Klasyfikacji i analiza danych – teoria i zastosowania*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Gross W.F., Bode G., Taylor J.M., Lloyd-Smith C.W. (1986), *Some finite population estimators which reduce the contribution of outliers*, Proceedings of the Pacific Statistical Conference, 20–24 May 1985, Auckland.

GUS (2014), *Działalność przedsiębiorstw niefinansowych w 2012 roku*, Warszawa.

GUS (2015), *Małe i średnie przedsiębiorstwa niefinansowe w latach 2009–2013*, Warszawa.

GUS (2016), *Wykorzystanie danych administracyjnych w badaniu: Ocena bieżącej działalności gospodarczej przedsiębiorstw*, Warszawa.

Horvitz D.G., Thompson D.J. (1952), *A Generalization of Sampling without Replacement from a Finite Universe*, "Journal of the American Statistical Association", vol. 47, pp. 663–685.

Preston J., Mackin C. (2002), *Winsorization for Generalised Regression Estimation*, Paper for the Methodological Advisory Committee, Australian Bureau Of Statistics, Canberra.

Rao J.N.K., Molina I. (2015), *Small Area Estimation*, Wiley, Hoboken, doi: 10.1002/9781118735855.

Rousseeuw P.J., Leroy P.M. (2003), *Robust Regression and Outlier Detection*, Wiley-Interscience, Hoboken.

Odporny estymator GREG w ocenie małych przedsiębiorstw handlowych

Streszczenie: Dynamiczne zmiany w gospodarce spowodowały wzrost zapotrzebowania na dane statystyczne zarówno co do liczby cech, jak i rodzajów przekrojów. W statystyce gospodarczej sprostanie temu wyzwaniu jest szczególnie trudne ze względu na specyfikę populacji przedsiębiorstw. Wymusza ono poszukiwanie metod szacunku zmierzających w kierunku zwiększenia stopnia wykorzystania źródeł administracyjnych. Adaptacja nowych rozwiązań ma przyczynić się zarówno do rozszerzenia zakresu informacji, jak i do zwiększenia efektywności prowadzonych szacunków. Celem niniejszego badania jest próba wykorzystania odpornego estymatora GREG uwzględniającego KMNK i metody najmniejszej mediany kwadratów w szacunku charakterystyk dotyczących małych przedsiębiorstw handlowych działających w 2012 roku. W estymacji jako zmienne pomocnicze uwzględnione zostały zmienne opóźnione w czasie, pochodzące z rejestrów administracyjnych. W artykule odwołano się do metod estymacji reprezentowanych przez statystykę małych obszarów. Badanie prowadzone jest na niskim poziomie agregacji. Domenę studiów stanowi sekcja PKD z uwzględnieniem przekroju województw.

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

JEL: C40

Received: 2016-12-17; verified: 2017-10-13. Accepted: 2018-01-13