STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 287—312

ON SOME ROBUST ESTIMATORS FOR POLISH BUSINESS SURVEY[1]

Grażyna Dehnel, Elżbieta Gołata

ABSRACT

The paper presents attempts to use administrative data and non-standard techniques to estimate basic economic information about small business in the joint cross-section of Polish Classification of Economic Activities and regions. Due to many outliers and significant fraction of entities for which many variables are equal to zero, the considered distributions are heterogeneous. Therefore, Horvitz-Thompson estimates are compared with the robust ones: modified GREG, Winsor and local regression. Results obtained in the study present practical possibilities of adopting robust estimation techniques to small business data in Poland. Properties of the estimators are discussed.

Key words: Domain estimation, Robust estimators, Small business statistics.

1. Aim of the study, data and assumptions of the research

The objective of the referred study was to provide economic information about small business enterprises (employing up to 9 workers) across regions depending on the type of economic activity. Like in other countries, enterprises in Poland can be divided into three categories: big, medium and small, depending on their size measured by the number of employees[2]. Statutory duty to provide information on economic effects differs between these groups as concerns scope and frequency of supplying the reports to Central Statistical Office (CSO). According to the regulations, small business entities are exempt from statistical reporting and take part in a yearly survey which enables estimation for the whole country. The majority of small business entities are self-employed individuals, especially in commerce, which operate mainly on local markets. Although small

[1] The paper was presented during the 57th Session of ISI held in Durban, South Africa in August 2009, on IPM56: Modelling Economic Data to Produce Small Area Estimates.

[2] According to Polish regulations, the category of big enterprises comprises all those which employ at least 50 persons. Enterprises employing from 10 to 50 people are referred to as medium. And all those employing up to 9 people (inclusive) are called *small business*.

economic entities constitute the basis of local economy[1], there is no information about this group available at the regional level. The development of market economy stimulates information demand for more and more detailed data about small business at the regional level (Falorsi *et al*. (2000), Klimanek and Paradysz (2006), Pawlowska (2005)). In order to meet this demand, our research was aimed at widening the range of estimates about small business by providing information for regions and type of economic activity.

Domain estimation in business statistics is rather difficult due to undesirable properties of the estimated characteristics. Variables that describe small businesses (those which are estimated as well as those which are used as covariates) are usually of non-homogeneous distributions: extremely right skewed with high dispersion and strong kurtosis. The population of small business is heterogeneous also because of the existence of many outliers, in addition to the presence of quite a significant fraction of entities for which many variables are equal to zero. Outliers have a significant effect on estimation results, especially at a low level of aggregation. Under such conditions the properties of estimators like unbiasedness or effectiveness are not preserved. Although for some observations the variables take extreme values, they need not necessarily be false. Extremely large observations are a natural component in business surveys. That is why we tried to explore some alternative estimation techniques which are less sensitive to outliers. The following methods of estimation were analysed: generalised regression estimator GREG, its modification proposed by Chambers *et al.* (2001a), Winsor estimator as discussed by Mackin and Preston (2002) and local regression presented by Kim *et al.* (2001). Another detail objective of the paper was to examine the precision of small domain estimates in respect to different non-standard estimation techniques when applied to Polish data for small business.

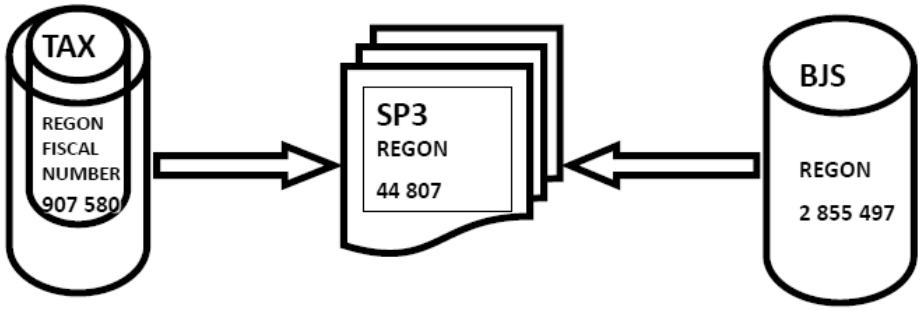
As concerns data, an attempt was made to apply auxiliary information from administrative files for more effective use of small businesses survey (SP-3). The analysis refers to data from the year 2001 for which databases were made available by Central Statistical Office (CSO). The main data source about small business units used in the study is SP-3 survey conducted by CSO on a 5% random sample chosen using a stratified sampling design. 114 thousand units were chosen, but only 44 807 responded (39.3% response rate). The information obtained in the survey comprise, among others, the following topics: costs, revenue, financial result, employment, investments, etc. It is assumed that the sample is drawn from an updated frame - the register called Database of Statistical Units (BJS). But in reality a considerable part of entities does not run their business, for example because of bankruptcy.

[1] According to Central Statistical Office there were 3.325 million enterprises in Poland in 2001. Most of them, i.e. 1.6 million, belonged to the small business category, employing up to 9 people. Most small economic entities, that is 1.545 million, are self-employed individuals. And only less than 4% of small business companies possess legal corporate personality. Over 3.2 million people, that is about 20% of the Polish labour force were employed in small business companies.

Since small business statistics is connected with a serious problem of non-response and incompleteness, our intention was to use supplementary data from registers providing complete and reliable information. Auxiliary variables used as covariates were taken from two administrative registers: Database of Statistical Units (BJS) and the Ministry of Finance's Tax Register (see figure 1). The Database of Statistical Units (BJS) was specially created by CSO to serve as a frame in all surveys conducted on the population of economic enterprises. BJS was constructed on the basis of Economic Units Register called REGON and includes additional information from other data sources. It is updated every year and in 2001 contained 2 855 497 records.

The Ministry of Finance's Tax Register called POLTAX served as the second auxiliary data source and was considered as especially reliable. Confidentiality of the data was specially secured. In the year 2001 for which data from SP-3 survey was available, the tax register contained information on tax statements from less than 1 million economic entities. The data base comprised 844 675 records referring to self-employed individuals (Personal Income Tax - PIT) and 62 905 records for business that possess legal corporate personality (Corporate Income Tax - CIT).

Figure 1. Integration of data from different sources



Source: Own compilation.

It was assumed that integration of all the three data sources: SP-3, BJS and POLTAX will enable a more detailed and complete analysis. A deterministic approach using a unique linkage key was applied (REGON number). But the percentage of matches was only 60. Differences in the number of observations in data sources used in the study were due to a high fraction of non-response in SP-3 survey, as well as incompleteness of the tax register and outdated information in BJS. Over 40% of units sampled could not be matched with the records in the tax register. The high fraction of missing linkage caused the danger of selectivity. It was very likely that the unmatched records constituted a specific group of enterprises, which differed significantly from the average in terms of the most important characteristics. This supposition was confirmed by the analysis of

difference in estimates of costs and revenue based on *(i)* integrated database, *(ii)* firms drawn to the sample but unmatched with the Tax Register and *(iii)* taking into account the whole sample (matched and unmatched records - see table 1).

Table 1. Comparison of cost and revenue estimates based on different data, small business, Poland, 2001

Variable (thousands PLN)	Estimation based on Integrated database *(i)*	Estimation based on firms not included in the Integrated database *(ii)*	Estimation based on the whole SP-3 survey *(iii)*
Cost per firm	318.8	155.4	253.8
Revenue per firm	369.1	184	282.1

Source: Own estimation based on SP3 Survey and Integrated database.

The overestimation obtained upon the integrated database made us apply a special procedure relying on assigning to unmatched firms values obtained from the survey. The correlation coefficients estimated for matched units showed a strong positive relation between variables form SP-3 survey and Tax Register $(r \in (0{,}75;0{,}99))$. This imputation procedure allowed us to take all records from SP-3 survey integrated with Tax Register data for further analysis.

Estimation was conducted for the following economic variables: average number of employees under a contract employed in a company, the amount of average monthly salary paid over one year per company, average revenue per company and average costs per company. As auxiliary variables we used information from additional data sources: BJS and Tax Register. For the two estimated variables: Revenue and Costs we applied Costs and Revenue from Tax Register as covariates. For the remaining two target variables: average number of employed and Gross Wage, as auxilary information we implemeted different combinations of: Number of Employees, Gross Wage, Revenue and Costs from Tax Register. The choice was made mostly upon the strength of relation between the variables.

The small domains were constructed on the basis of regions and types of economic activity and were defined as an 'intersection' of the joint distribution of economic entities by voivodships and sections of Polish Economic Activity classification (PKD section). Estimation was conducted for 176 domains: 17 regions (R) and 11 PKD sections (S) (**R**egion&**S**ection). Since the number of domains taken into account in the study is quite considerable, we present only selected results for one of the estimated variables - Gross Wage in the region of *Zachodniopomorskie voivodship* and in selected sections: *Industry, Construction, Trade, Hotels and Restaurants.* However, to evaluate results obtained in the study, synthetic characteristics across all domains were provided.

The results provided by the study were compared with the original GREG approach. In order to determine the estimation precision an approximated method based on samples and bootstrap was applied. The bootstrap method presented by

Efron (1979) for simple random sampling needs modifications when applied for complex samples. We applied the approach described in detail by Shao and Tu (1995). The simulation study consisted of 500 iterations. In each one, a subsample of size *n-1* was drawn and for each iteration sample modified weights were distinguished to estimate the unknown parameter $\hat{Y}_{*b}$, where $(*)$ stands for different robust estimators used in the study. The empirical variance was obtained according to the standard approach

$$Var(\hat{Y}_*) = \frac{1}{500-1}\sum_{b=1}^{500}(\hat{Y}_{Rb} - \hat{Y}_*)^2 \tag{1}$$

Estimation precision was evaluated on the basis of estimator's coefficient of variation:

$$CV(\hat{Y}_*) = \frac{\sqrt{Var(\hat{Y}_*)}}{(\hat{Y}_*)} \tag{2}$$

Coefficient $RedCV(\hat{Y}_*)$ was used to measure the degree of *CV* reduction obtained by robust estimators applied:

$$RedCV(\hat{Y}_*) = \frac{CV(\hat{Y}_*) - CV(\hat{Y}_{DIR})}{CV(\hat{Y}_{DIR})} \ . \tag{3}$$

The term CV was used instead of REE as we perceive that variation obtained in the bootstrap method does not include the bias in contrast to MSE.

$$deff\left(\hat{Y}\right) = \frac{Var\left(\hat{Y}_*\right)}{Var\left(\hat{Y}_{DIR}\right)} \tag{4}$$

Additionally, the *design effect* coefficient *deff* (proposed by Kish (1965), (1995)) was provided to measure the effectiveness of the examined estimator in comparison with the direct one. All the values smaller than unity indicate that the method applied is more effective. In the course of our research correlation analysis as well as model evaluation specially detecting residuals were naturally implemented. We made also an attempt to provide validation of the estimates obtained against information from other sources and compared the results for other levels of aggregation.

2. The GREG estimator and its modifications

In business statistics which is characterised by huge diversification, direct estimation procedures do not provide satisfactory results. Applying GREG estimation that use auxiliary information is perceived as a solution, which usually

increases precision considerably. Despite this advantage D. Hedlin (2004) draws attention to how important good modelling is. He indicates that for a set of real data, different GREG estimators might produce wildly different results. The difference between them lies entirely in the choice of the model. The modification proposed by Chambers *et al.* (2001a) refers to GREG estimators assuming heteroscedasticity. They introduce transformation that depends on including additional auxiliary variable – '*z*' to the regression model of *y* on *x*:

$$\hat{Y}_{GREG,d} = \sum_{i \in S_d} w_i g_i y_i \tag{4}$$

where:

d – domain,

g_i – weight of *i*-th individual observation defined as:

$$g_i = 1 + \left(X_d - \hat{X}_{HT,d}\right)' \left(\sum_{i \in s_d} w_i \mathbf{x}_i \mathbf{x}'_i / z_i^{\gamma} \right)^{-1} \left(\mathbf{x}_i / z_i^{\gamma} \right) \tag{5}$$

X – auxiliary variable which, depending on the approach, is defined as: *(i)* number of employees or *(ii)* revenue (see table 2)

z – auxiliary variable which, depending on the approach, is defined as: *(i)* number of employees or *(ii)* revenue (see table 2)

$\hat{Y}_{GREG,d}$ – estimate of Gross wage in domain *d* obtained by applying the GREG estimator

$\hat{X}_{HT,d}$ – direct Horvitz-Thompson estimate of the total value of the auxiliary variable *x* in domain *d*

X_d – total value of the auxiliary variable *x* in domain *d,*

γ – coefficient characterising the degree of heteroscedasticity, for $\gamma = 0$ the original GREG estimator is obtained.

As it is known from other surveys, γ coefficient should be included in the interval $1 \le \gamma \le 2$ (Särndal (1992) p.255); the following estimators were analyzed:

1) $\hat{Y}_{DIR}$ – direct Horvitz-Thompson estimator (HT),
2) $\hat{Y}^0_{GREG}$ – $\gamma = 0 \Rightarrow z_i^0$ regression estimator based on linear regression model assuming homoscedasticity (GREG estimator),
3) $\hat{Y}^1_{GREG}$ – $\gamma = 1 \Rightarrow z_i^1$,
4) $\hat{Y}^{1,5}_{GREG}$ – $\gamma = 1{,}5 \Rightarrow z_i^{1,5}$,
5) $\hat{Y}^2_{GREG}$ – $\gamma = 2 \Rightarrow z_i^2$.

The estimators denoted by numbers *(3) (4)* and *(5)* are regression estimators based on linear regression model assuming heteroscedasticity. The direct estimator was considered as the reference value when comparing the effectiveness of GREG estimator and its modifications. The second one of the estimators *(2)* - $\hat{Y}^0_{GREG}$ takes the form of GREG estimator, but its value differs slightly in comparison with the original GREG formula. The difference results from the fact that not all of the units drawn to the sample take part in the estimation. Those observations, for which auxiliary variable '*z*' is equal to zero[1] are omitted. In the case of the remaining three estimators, the linear regression model assumes heteroscedasticity of a degree indicated by γ coefficient. The modification assumes that each of the auxiliary variables might take both roles, of '*x'* and '*z'* with the warranty that '*z'* does not equal zero. Many combinations were considered and finally the following approaches were proposed (see table 2). The estimation was conducted for Gross Wage *(y)*. As an auxiliary variable *(x)* and *(z)* we used: *(i)* the number of employees from BJS and *(ii)* revenue from Tax Register.

Table 2. Auxiliary variables *'x'* and *'z'* used in GREG's modification, small business, Poland, 2001

Approach	Auxiliary variable '*x*'/ *data source*	Auxiliary variable '*z*' / *data source*
1	Revenue / *Tax Register*	Revenue / *Tax Register*
2	Revenue / *Tax Register*	Number of employees / *BJS*
3	Number of employees / *BJS*	Number of employees / *BJS*

Source: Own estimation based on SP3 Survey and Integrated database.

The synthetic characteristics summarising estimates of Gross Wage in the cross-section of all domains are presented in table 3. They show that the highest precision was obtained for $\hat{Y}^0_{GREG}$, the estimator ignoring observations for which "*z*" was equal to zero. With the increase of γ, usually the maximum, median and the average value of *CV* also increases. The biggest proportion of domains for which the relative dispersion of the estimator was smaller than in the case of the direct one, was also observed for $\hat{Y}^0_{GREG}$, and then other estimators $\hat{Y}^1_{GREG}$, $\hat{Y}^{1.5}_{GREG}$ and $\hat{Y}^2_{GREG}$ were classified. Only for the model with *Revenue* as an auxiliary variable was the order of the estimators changed, which can be explained by a weaker relation of the estimated and auxiliary variables (see table 4).

[1] The variable '*z'* is assumed not to take zero value. Nevertheless, it happens in practice that zero values occur for variables for which they are not expected (for example the revenue might be equal to zero for an enterprise running a business).

We cannot unequivocally determine which of the estimators is characterized by the highest precision. In each case an appropriate model and the definition of '*x*' and '*z*' is necessary. Model misspecification may lead to negative estimates. Such an untypical situation was observed for domain *Trade* in *Dolnoslaskie voivodship* (see figure 2. B).

Table 3. Characteristics of the CV distribution, GREG's modification estimates of Gross Wage, all domains, small business, Poland, 2001

Characteristics	Estimator				
	$\hat{Y}_{DIR}$	$\hat{Y}^{0}_{GREG}$	$\hat{Y}^{1}_{GREG}$	$\hat{Y}^{1,5}_{GREG}$	$\hat{Y}^{2}_{GREG}$
x – number of employees z - number of employees					
min	0.0721	0.0665	0.0670	0.0671	0.0673
max	1.0964	1.0402	1.0522	1.1754	1.3406
average	0.3153	0.3019	0.3020	0.3030	0.3061
median	0.2861	0.2739	0.2773	0.2799	0.2811
proportion of domains for which $CV < CV_{DIR}$ *(%)*		73.30	72.16	71.59	68.18
x - revenue z - number of employees					
min	0.073	0.070	0.073	0.073	0.073
max	0.779	0.990	0.863	0.795	0.979
average	0.304	0.297	0.306	0.310	0.312
median	0.281	0.266	0.283	0.286	0.286
proportion of domains for which $CV < CV_{DIR}$ *(%)*		68.75	56.25	46.02	45.45
x - revenue z - revenue					
min	0.070	0.070	-1.731	-1.388	-1.332
max	0.793	13.556	5.093	3.203	11.370
average	0.305	0.421	0.358	0.316	0.401
median	0.280	0.274	0.276	0.280	0.302
proportion of domains for which $CV < CV_{DIR}$ *(%)*		57.95	65.34	63.07	55.11

Source: Own estimation based on SP3 survey data, BJS, Tax Register.

Table 4. Correlation coefficients between the estimated variable – Gross Wage and auxiliary variables from SP3 survey, BJS and Tax Register, 2001, Poland Zachodniopomorskie voivodship

Gross Wage by PKD sections	Auxiliary variables	
	Number of employees (BJS)	Revenue (Tax Register)
Industry	0,538	0,479
Construction	0,587	0,301
Trade	0,579	0,517
Hotels and restaurants	0,536	0,402

Source: Own estimation based on SP3 survey data and Tax Register.

Defining auxiliary variables '*x*' and '*z*' as Revenue, the estimators $\hat{Y}^{1,5}_{GREG}$ and $\hat{Y}^{2}_{GREG}$ provided negative estimates. As can be seen from figure 2.A, outliers in both dimensions (of the estimated as well as of the auxiliary variable) influence estimation results. The regression lines corresponding to $\hat{Y}^{0}_{GREG}$, $\hat{Y}^{1}_{GREG}$, $\hat{Y}^{1,5}_{GREG}$, $\hat{Y}^{2}_{GREG}$, presented on the diagram show the following regularity (which was observed also for other domains). The γ coefficient determines the degree to which outliers in the dimension of *'y'* are included in the model. The linear regression $\hat{Y}^{0}_{GREG}$ with γ=0 includes outliers in the dimension of *'x'*. With the increase of γ, outliers in the dimension of *'y'* are considered to a greater degree by assigning to them greater weights g_i, and less importance is attached to outliers in the dimension of *'x'*. The value of γ influences also the range of weights. The smallest dispersion of weights refers to γ =0 *(* $g_i 0 \in$ *(-0,13 ; 2,4))*, and the biggest range is observed for γ=2 *(* $g_i 2 \in$ *(-35 ; 55))* (see also tab. 5).

Table 5. Characteristics of weight's distribution:

A) domain defined as Industry in Dolnoslaskie voivodship

x –revenue z - revenue	Min	Max	Weight for outlier *(x)*	Average	Standard deviation
g_i *0*	0.912	1.997	1.249	1.039	0.123
g_i *1*	-4.705	8.384	1.290	1.057	0.595
g_i *1,5*	-16.917	25.764	1.087	1.090	1.607
g_i *2*	-35.176	50.416	1.016	1.129	3.008

B) domain defined as Trade in Dolnoslaskie voivodship

x –revenue *z - revenue*	Min	Max	Weight for outlier *(x)*	Average	Standard deviation
g_i *0*	0.964	2.448	0.984	1.020	0.067
g_i *1*	0.117	54.323	0.117	1.042	1.631
g_i *1,5*	-1.456	56.063	-1.456	1.042	1.687
g_i *2*	-3.170	55.812	-3.170	1.040	1.684

Source: Own estimation based on SP3 survey data and Tax Register.

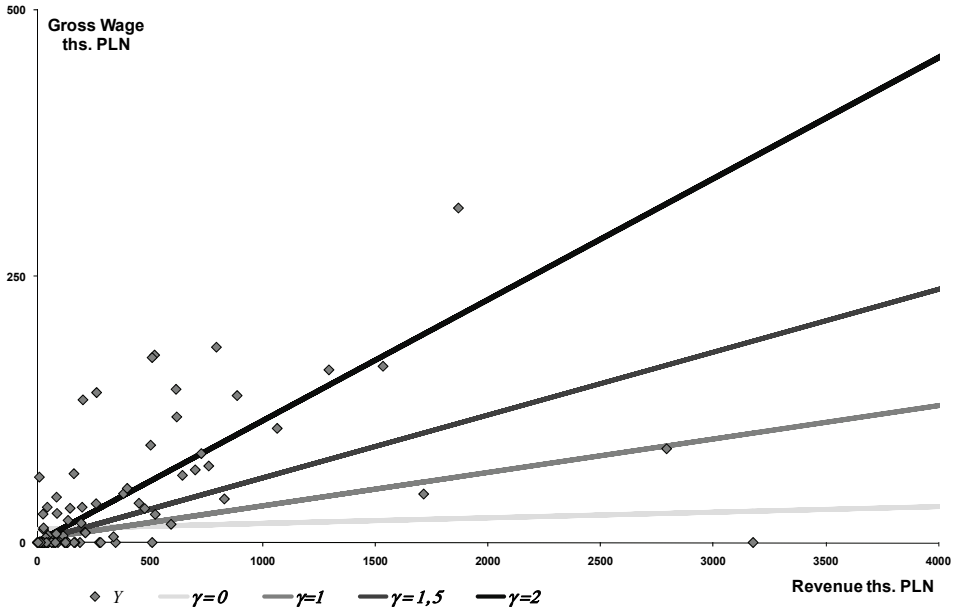
The g_i weights are sample dependent and it is not desirable that the weights are far from unity. Negative weights are particularly undesirable as they tend to increase the estimator variance. That is why some methods for restricting g_i weights were developed. There is also a close relationship between g_i weights of a sample unit and its influence on Beta. The most common measure of this influence is DFBETA defined by a change in the estimate of Beta when a unit is excluded from the sample:

$$DFBETA_i = \left(\sum_{i \in S} x_i x_i^{'} / z_i^{\gamma} \right)^{-1} \left(\frac{x_i}{z_i^{\gamma}} \right) \left(\frac{e_i}{1 - h_i} \right) \tag{6}$$

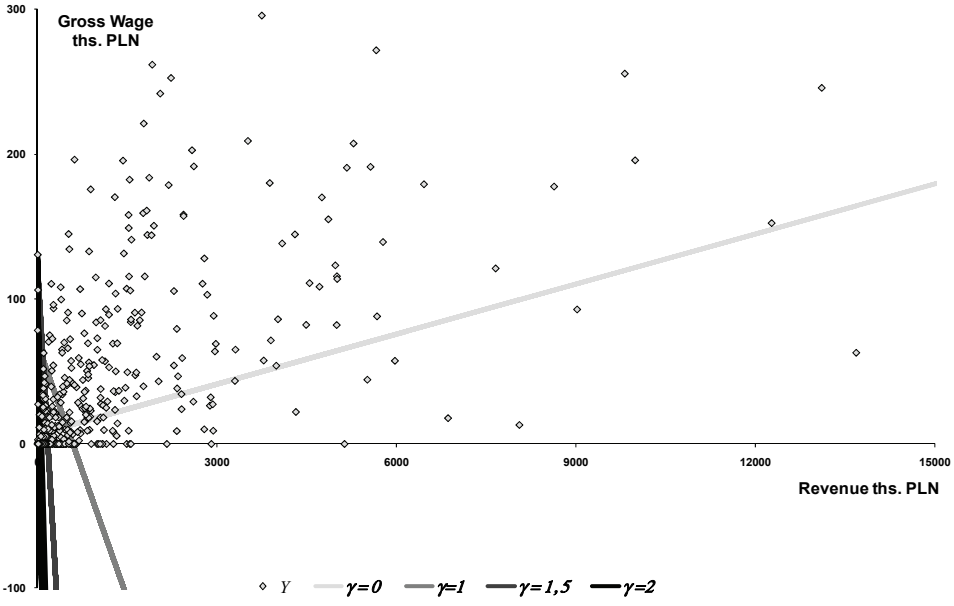
The main reason for negative estimates are outliers, i.e. those companies in which the Gross Wage is very big and the Revenue very small. They influence the model parameters and result in incomparably large weights assigned to those observations. DFBETA was used to identify the influential units. The outlying observations were modified before conducting further estimation.

Figure 2. Regression lines for $\hat{Y}_{GREG}^{\gamma}$ estimators in Dolnoslaskie voivodship:

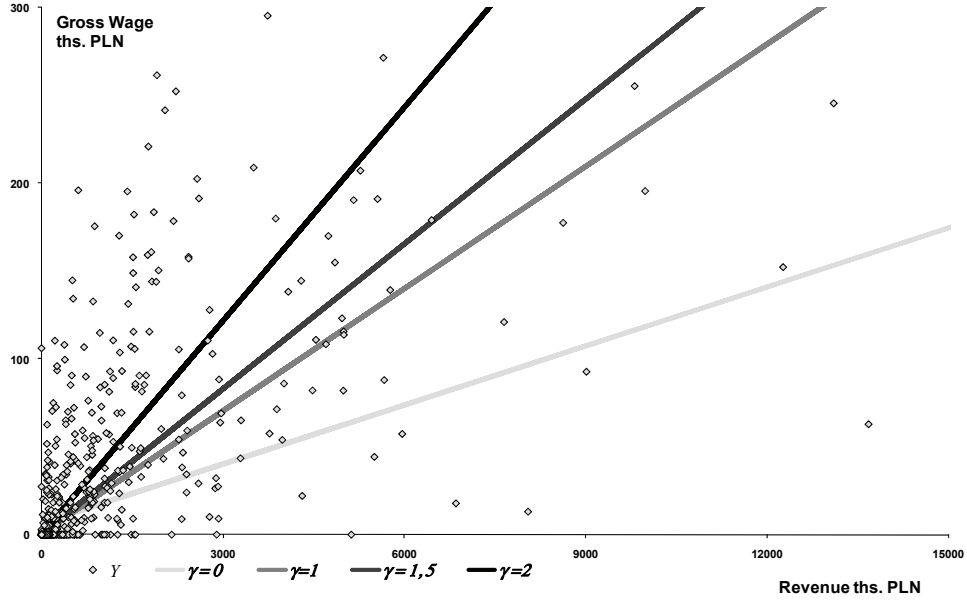
A) domain defined as Industry



B) domain defined as Trade



C) domain defined as Trade after substituting the outliers with median



Source: Own estimation based on SP3 survey data and Tax Register.

To avoid negative estimates, one proposed solution is to substitute outliers with median or implement post-stratification (Chambers *et al.* (2001a)). The observations were divided into three groups: outliers in the dimension of *'x'*, in the dimension of *'y'* and the remaining ones (this group desirably should be the largest). The results obtained by substitution are presented on the figure 2.C and by post-stratification in table 6.The modification considerably changed the model parameters for $\hat{Y}^{1,5}_{GREG}$ and $\hat{Y}^{2}_{GREG}$. Adapting both methods improved the estimation precision as well as reduced the variation of the estimator.

Table 6. Estimation precision for Gross Wage in Trade section of PKD in Dolnoslaskie voivodship

Estimator Coefficient of Variation – CV					*deff coefficient*			
$CV(\hat{Y}_{DIR})$	$CV(\hat{Y}^{0}_{GREG})$	$CV(\hat{Y}^{1}_{GREG})$	$CV(\hat{Y}^{1,5}_{GREG})$	$CV(\hat{Y}^{2}_{GREG})$	$deff(\hat{Y}^{0}_{GREG})$	$deff(\hat{Y}^{1}_{GREG})$	$deff(\hat{Y}^{1,5}_{GREG})$	$deff(\hat{Y}^{2}_{GREG})$
before modification								
0,104	0.086	1.729	-1.377	-1.045	0.8724	6.9087	33.7324	162.167
after substitution of outliers by median								
0,097	0.089	0.088	0.093	0.126	0.9635	1.0119	1.0895	1.5613
after post-stratification								
0,1063	0.1104	0.1111	0.1186	0.1753	1.1393	1.3669	1.5692	2.5939

Source: Own estimation based on SP3 survey data, BJS and Tax Register.

The analysis conducted so far allows us to draw the following conclusions:

- Implementation of Chambers *et al.* (2001a) modification provides the results which are very much differentiated depending on the auxiliary variables and the type of the model,
- Introducing the additional auxiliary variable '*z*' makes it possible to move the estimation problem from disjunctive values of the auxiliary variable to disjunctive values of the estimated variable,
- Application of the additional variable '*z*' creates also the possibility to estimate the basic economic information for small business at a lower aggregation level,
- Weights from the modified model may be characterized by great variation, which is contradictory to the general assumption that the product of weights g_i and w_i should be close to w_i,
- Due to outliers in the sample data, weights designated from the modified model, may obtain values less than zero, which might result in negative estimates,
- Gain in estimation precision is greater for domains of a smaller size (number of observations) and in the case of a stronger correlation between the estimated and the auxiliary variable,
- A significant improvement in estimation precision may be obtained by selecting an adequate model including variables '*x*' and '*z*', with properly designated auxiliary variables, which in the case of a large number of small domains makes application of the modified model rather difficult,
- Weights assigned to the GREG estimator are strongly related to the distance measure DFBETA, which enables identification of outliers,
- Post-stratification is proposed for domains, in case of which selecting a suitable model is difficult.

3. The Winsor estimation

The precision of GREG estimator is influenced by characteristics of enterprises in terms of the distribution of estimated variables and those which serve as auxiliary ones. The existence of outliers in the sample has also an undesirable effect. According to the modification proposed by Chambers *et al.* (2001a), the proportion of disjunctive observations may be reduced by their substitution or post-stratification. Both methods are rather time-consuming and require, in each case, identification of the outlying observations. That is why we also considered other methods, which apply a less individual procedure to solve the problem of outliers. As an example one may consider a group of methods aimed at making the estimator less sensitive to large residuals, which was described by Kokic and Bell (1994) or Chambers (1996). The individuals drawn into the sample, for which the variable takes a value beyond specified border points are changed.

The sample modification may be provided in different ways. One of the methods depends on modification of weights in such a way, that the proportion of outliers in estimation is very small (Hidiroglou, Srinath (1981)). Another method proposes the modification of the value of the estimated variable, for example the Winsor estimation. This estimation technique was introduces by Searls (1966).

The algorithm applied may be divided into several steps. The initial stage comes down to designating some border points K_{Ui} and K_{Li}, which serve to delimitate the individuals into two groups: of typical and untypical observations. In the second step the untypical observations are modified to obtain values which are close to the border points:

$$y_i^* = \begin{cases} \left(\frac{1}{\widetilde{w}_i}\right) y_i + \left(1 - \frac{1}{\widetilde{w}_i}\right) K_{Ui} & \text{if} \quad y_i > K_{Ui} \\ y_i & \text{if} \quad K_{Ui} \geq y_i \geq K_{Li} \\ \left(\frac{1}{\widetilde{w}_i}\right) y_i + \left(1 - \frac{1}{\widetilde{w}_i}\right) K_{Li} & \text{if} \quad y_i < K_{Li} \end{cases} \tag{7}$$

The last stage is estimation of the response variable using any estimation technique upon data from the modified sample:

$$\hat{Y}_{win} = \sum_{i \in s} w_i g_i y_i^* = \sum_{i \in s} \widetilde{w}_i y_i^* \tag{8}$$

where: $\widetilde{w}_i = w_i g_i$.

The biggest problem in Winsor estimation is the designation of adequate border points, which are crucial in indicating the outliers. In practice many different methods are proposed, for example robust regression. It is important to

underline that if the procedure applied in this step leads to proper delimitation, the estimation precision might be significantly improved. We decided to determine the limits by using four different techniques of robust regression: *(i) Trimmed Least Squares (TLS), (ii) Trimmed Least Absolute Value (LAV), (iii) Sample Splitting (TSS), (iv) Least Median of Squares (LMS)*. It was also assumed that the proportion of individuals removed from the sample was constant and in the case of each technique equal to 5%. The sample modification was conducted according to Winsor type II estimation and estimates from modified sample were provided by GREG estimator.

The *Trimmed least squares (TLS)* removes from the sample units with the biggest squared residuals. The *Trimmed Least Absolute Value (LAV)* method designates units characterised by the maximal absolute residuals. According to *Sample Splitting (TSS) Technique* the choice of individuals removed from the sample involves some random mechanism. In this method two regression models are estimated for observations drawn to the sample after a random division into two groups. The residuals for one group are calculated upon the model obtained for the other group. The final regression model is constructed after rejecting the observations with the biggest residuals. This approach makes the *TSS* technique more robust in comparison with others like *TLS* or *LAV*. The *Least Median of Squares (LMS)* is based on the bootstrap idea. It depends on selecting subsamples of size $n-1$ according to a simple random scheme with replacement. For each subsample a model is provided, estimates and squared residuals which serve to designate the median. Finally, the model with the smallest median of squared residuals was chosen.

Evaluating the influence of robust regression technique on the estimation precision we considered two approaches: determining both limits: upper and lower, and designation of one limit only - the upper one. The first proposition means that the modification refers to both types of outliers: in *'y'* as well as in *'x'* dimension. While in the second approach the individuals were modified with respect to the outliers in *'y'* dimension only. The estimation was conducted for Gross Wage *(y)*. As an auxiliary variable *(x)* we used Revenue from Tax Register.

Table 7. Comparison of Winsor and direct estimators precision using techniques: *TLS, LAV, SST, LMS*

Type of Winsor estimator $\hat{Y}_{Win}$ by robust regression technique	Indicator $W_{CV} = \frac{\overline{CV\left(\hat{Y}_{Win}\right)}}{\overline{CV\left(\hat{Y}_{DIR}\right)}}$	
	two border points	one border point
Trimmed least squares TLS	0.804	0.893
Trimmed least absolute value LAV	0.798	0.887
Sample Splitting TSS	**0.386**	**0.475**
Least median of squares LMS	0.843	0.960

Source: Own estimation based on SP3 survey data and Tax Register.

Assigning two border points provides a bigger gain in precision (see tab. 7). The greatest average reduction in *CV* was obtained for *TSS* technique. The W_{CV} ratios for *TLS* and *LAV* techniques are very similar. They indicate about 20% smaller variation than in the case of direct estimator (for two border points). It was found that *Sample Splitting Technique* $\hat{Y}_{TSS}$ provided the most efficient estimates.

To evaluate results of the research, similarly as in the case of GREG modification, the reference values were provided by direct Horvitz-Thompson (HT) estimates. Comparison analysis of Winsor estimators with GREG's modification is presented in Table 8. This time the estimated variable was defined as Revenue. To preserve comparability, we applied an identical set of variables in both GREG modification and Winsor estimator. The results show that the most efficient GREG estimates are obtained for the weight γ=1.5. Both mean and median values of the coefficient of variation are the lowest. But for bigger values of the parameter γ=2 efficiency is decreasing. If median is taken into account, a very high precision (close to the one for γ=1.5) is observed also for γ=0 and γ=1. Regardless of γ parameter, GREG estimator provides more precise estimates (in view of the median) than the direct one. The highest fraction of domains for which GREG estimates were more efficient is for γ=1 but with increase of γ this percentage decreases.

Basic distribution characteristics referring to the variation of Winsor estimator across all domains (**R**egion&**S**ection) show similar *CV* for those using techniques *TLS, LAV* and *LMS* and direct HT estimator. The mean belongs to an interval (0.263 – 0.303) and the median (0.210 – 0.232). The smallest variation is observed for *TSS* technique (mean equals to 0.225 and median 0.172). Also for *TSS* the highest fraction of domains for which it provided more efficient estimates was observed (almost 74%). This record is quite close to GREG with γ=1, and much better than for GREG with γ=1.5. It provides also absolutely better

estimates in terms of other characteristics of *CV distribution* like *min, average and median.*

Table 8. Characteristics of CV distribution: Winsor and modified GREG estimators, revenue, all domains, 2001 r.

Characteristics	*GREG Estimator*				
	$\hat{Y}_{DIR}$	$\hat{Y}^{0}_{GREG}$	$\hat{Y}^{1}_{GREG}$	$\hat{Y}^{1,5}_{GREG}$	$\hat{Y}^{2}_{GREG}$
y - Revenue					
x – Cost from TAX register			***z - Revenue from TAX register***		
min	0,088	0,074	0,075	-7,340	-3,834
max	0,902	1,407	1,417	8,148	4,474
average	0,276	0,273	0,278	0,230	0,298
median	0,232	0,212	0,213	0,210	0, 226
proportion of domains for which $CV < CV_{DIR}$ *(%)*		72,16	74,43	70,45	59,09
Characteristics	*WINSOR Estimator*				
	$\hat{Y}_{DIR}$	$\hat{Y}_{TLS}$	$\hat{Y}_{LAV}$	$\hat{Y}_{TSS}$	$\hat{Y}_{LMS}$
y - Revenue					
x – Cost from TAX register					
min	0,088	0,071	0,070	0,033	0,072
max	0,902	1,371	1,352	1,485	1,578
average	0,276	0,263	0,270	0,225	0,303
median	0,232	0,211	0,210	0,172	0,228
proportion of domains for which $CV < CV_{DIR}$ *(%)*		64,77	64,20	73,86	50,00

Source: Own estimation based on SP3 survey data and Tax Register.

Evaluating the precision of Winsor estimates we analysed a ratio relating mean and median of its *CV* to variation of direct HT estimator. Also *deff* coefficients relating variance of both estimators were considered (see tab. 9). Values that are less than one indicate improvement in estimation precision obtained as a result of implementing the robust approach.

Table 9. Synthetic measures of estimation effectiveness: Winsor and modified GREG estimators of revenue, all domains, 2001 r.

Estimation technique	$W_{\overline{CV}} \frac{\overline{CV\left(\hat{Y}_*\right)}}{\overline{CV\left(\hat{Y}_{DIR}\right)}}$	$W_{Me(CV)} \frac{Me\left(CV(\hat{Y}_*)\right)}{Me\left(CV(\hat{Y}_{DIR})\right)}$	$deff = \frac{Var(\hat{Y}_*)}{Var(\hat{Y}_{DIR})}$
Modified GREG			
$g_i(0)$	0,9891	0,9138	0,9913
$g_i(1)$	1,0072	0,9181	1,0080
$g_i(1,5)$	0,8333	0,9052	0,8334
$g_i(2)$	1,0797	0,9741	1,0795
Winsor Estimator			
TLS	0,9529	0,9095	0,9526
LAV	0,9783	0,9052	0,9786
TSS	0,8152	0,7414	0,8315
LMS	1,0978	0,9828	0,9913

Source: Own estimation based on SP3 survey data and Tax Register.

Comparing $W_{\overline{CV}}$, $W_{Me(CV)}$ and *deff* of different techniques, one should notice the highest reduction obtained for *Sample Splitting Technique*. It is also worth noticing that in terms of median the improvement was observed for all robust techniques, which is not always the case for mean and variance. Estimates obtained for $\hat{Y}^1_{GREG}$, $\hat{Y}^2_{GREG}$ and Winsor using *LMS* are less efficient than direct estimation. Interestingly enough, earlier estimates obtained by Dehnel (2008) for other variables provided other results. This shows ambiguity in the evaluation of robust estimation techniques with relation to estimated variables, availability of auxiliary information etc.

The appraisal of Winsor estimation showed that:

- Depending on the type of estimated characteristic, different methods might be applied to determine the border points used in the delimitation process: two (upper and lower) or one (upper). Both cases were examined showing that the estimation precision is bigger for two limits,
- Simulation research demonstrated the relation between efficiency and type of robust regression technique used. The more robust regression technique was applied, the more efficient estimates were produced. This was shown by comparison of TSS estimates with those obtained by TLS and LAV,
- Limits used in the delimitation process and therefore the type of robust regression technique, influence the precision of Winsor estimator. The highest precision in our study was observed for Sample Splitting Technique (TSS).

4, Local regression estimation

Apart from GREG modification or Winsor estimation, among other methods that are less sensitive to outliers, the local regression technique is often discussed. Its important advantage is its applicability in the case of a nonlinear relation between the estimated and auxiliary variables. Therefore, an attempt was made to present the possibility to implement local regression to estimate economic characteristics of small business for the joint distribution of regions and type of sections of economic activity (Region&Section). The research was conducted using the following estimator (Chambers, Dorfman, Wehrly (1993), Dorfman (2000)):

$$\hat{y}_{loc,i} = \mathbf{c}'_j \left(\mathbf{D}'_i \mathbf{W}_i \mathbf{D}_i\right)^{-1} \mathbf{D}'_i \mathbf{W}_i \mathbf{y}_s \quad i=1, 2, \ldots, N \tag{9}$$

where: U denotes population and s is the sample

$\mathbf{c}'_j$ is the vector of ones at the j-th position while the remaining are equal to zero,

$\mathbf{D}_i$, $i = 1, 2, \ldots, N$, are matrices of the dimension $n \times 2$ each, with $\begin{bmatrix} 1 & (x_j - x_i) \end{bmatrix}$ in the j-th row, $j = 1, 2, \ldots, n$,

$\mathbf{W}_i$, for $i = 1, 2, \ldots, N$, are diagonal matrices of the dimension $n \times n$ and $w_i b_i^{-1} K\left[(x_j - x_i) b_i^{-1}\right]$ at (j, j) position, where $K(\cdot)$ is the kernel function and b_i is the bandwidth for the i-th observation.

The main problems considered in local regression estimation are: to choose the kernel function and to determine the appropriate bandwidth. Once again many methods and suggestions can be found in this field (Chambers, Dorfman, Wehrly (1993), Chambers (1996), Kim, Breidt, Opsomer (2001)). In our study we decided to use the Epanechikov kernel (see also Hedlin, (2004)):

$$K(u_{ji}) = \max\left[0, \frac{3}{4}\left(1 - u_{ji}^2\right)\right] = \max\left[0, \frac{3}{4}\left(1 - \left(\frac{(x_j - x_i)}{b_i}\right)^2\right)\right]. \tag{10}$$

Different definitions were used for the bandwidth, for which the kernel function was determined. One of them was to assign one constant bandwidth for the whole sample according to: $b_i = \frac{1}{4}(x_{\max} - x_{\min})$ => $\hat{Y}_{loc}(\max, \min)$. In other cases, the bandwidth was determined by the value of an auxiliary variable x. Two types of *nearest neighbour bandwidth* suggested by Chambers (1996) were applied: substituting b_i with the difference between the value of the auxiliary variable[1] for the i-th observation (x_i) and for the observation identified by

[1] Units in the sample file were sorted by x_k in ascending order.

i + *20* (x_{i+20}) or *i* + *40* (x_{i+40}): $b_i = x_{i+20} - x_{i-20}$ => $\hat{Y}_{loc}(20)$ and $b_i = x_{i+40} - x_{i-40}$ => $\hat{Y}_{loc}(40)$. Additionally, we assumed the bandwidth to be more narrow than proposed by Chambers, that is : $b_i = x_{i+10} - x_{i-10}$ => $\hat{Y}_{loc}(10)$.

The comparison of robust estimation in business survey was supplemented by the analysis of how the bandwidth definition influences the estimation precision. To preserve comparability, we applied an identical set of variables as in the case of GREG modification and Winsor estimation. The estimation was conducted for Gross Wage *(y)*. As an auxiliary variable *(x)* we used Revenue from Tax Register.

Table 10. Characteristics of the CV distribution, local regression estimator, all domains, 2001

Characteristics	Estimator				
	$\hat{Y}_{DIR}$	$\hat{Y}_{loc}(10)$	$\hat{Y}_{loc}(20)$	$\hat{Y}_{loc}(40)$	$\hat{Y}_{loc}(\max, \min)$
min	0.10	0.09	0.08	0.09	0.08
max	0.63	0.60	1.77	1.78	0.73
average	0.27	0.25	0.36	0.36	0.27
median	0.25	0.23	0.21	0.21	0.24
proportion of domains for which $CV < CV_{DIR}$ *(%)*		73.30	71.59	72.16	59.66

Source: Own estimation based on SP3 survey data and Tax Register.

The characteristics of relative dispersion distribution for direct and local estimators are comparable, except the maximum. The relatively highest precision is observed for $\hat{Y}_{loc}(10)$ estimator, in terms of the characteristics of the distribution and as concerns the proportion of domains for which *CV* is smaller than obtained for the direct estimator.

The analysis of local regression estimation allowed us to make the following insights:

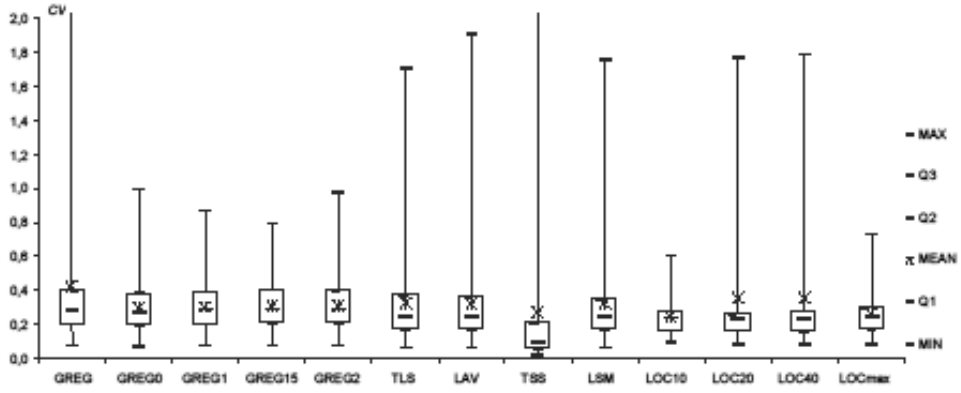
- As the bandwidth increases, the local regression estimates resemble GREG estimates more closely.
- The highest precision, in terms of CV, was obtained for local regression with bandwidth $\hat{Y}_{loc}(10)$ and $\hat{Y}_{loc}(20)$
- For a narrow bandwidth, the estimation is based on many local models, which lengthens the computing time
- As the bandwidth gets increasingly narrow, the local change in the estimated variable is taken into account to a greater degree. As the bandwidth increases, the smoothing effect is more significant

- The weights designated by the kernel function do not depend on the values of the estimated variable but on the auxiliary variables. This means that they can be applied for many estimated variables, in case the set of auxiliary variables is unchanged.

5. Comparison of the results obtained and conclusions

The GREG estimation is very popular and one of the most frequently used. But its application to business statistics is connected with a danger arising from outliers, whose presence results in a significant bias of the estimates (see Hedlin (2004)). Several modifications belonging to 'closer' and 'further' family of GREG estimators were applied and discussed. Among those closely related: estimation based on the model using inverse transformation and Winsor estimation introducing 'border' points to distinguish the outlier observation, were analysed. As concerns the 'further' family, we examined local regression, which has the ability to accommodate local departures from the linear model implemented. The estimation precision was evaluated on the basis of estimator's coefficient of variation *CV*, *RedCV* – coefficient measuring the degree of *CV* reduction and *deff* coefficient (see figures 3–5).

Figure 3. Characteristics of CV distribution across all domains, different robust estimators, 2001



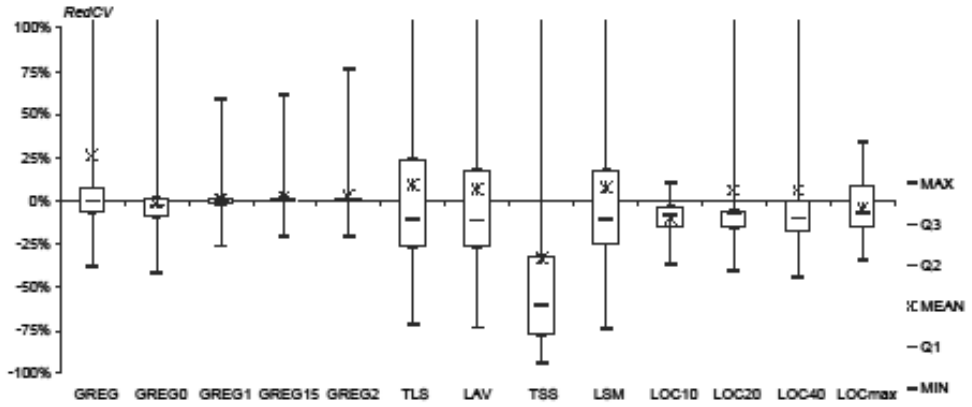
Source: Own estimation based on SP3 survey data, BJS and Tax Register.

The highest estimation precision expressed by the average value and the quartiles of the estimator's CV across all domains, is observed for the Winsor $\hat{Y}_{TSS}$ and local $\hat{Y}_{loc}(10)$ estimators. This remark does not take into account the extreme value observed for Winsor *TSS*, which was due to high dispersion in the

case of one of the bootstrap sub-samples. For the remaining Winsor estimators, local regression and GREG's modification, the relative dispersion, as measured by the quartiles and the average, is just a little bit smaller than for the original GREG estimator, but very similar.

The narrowest range of the CV is observed for two local estimators: $\hat{Y}_{loc}(10)$ and $\hat{Y}_{loc}(\max,\min)$ as well as the GREG modification by Chambers *at al*. For the remaining set of estimators examined, the variation is much more spread, but does not exceed the maximum range observed for the original GREG estimator (*CV=12%*).

Figure 4. Characteristics of *RedCV* distribution across all domains, different robust estimators, 2001



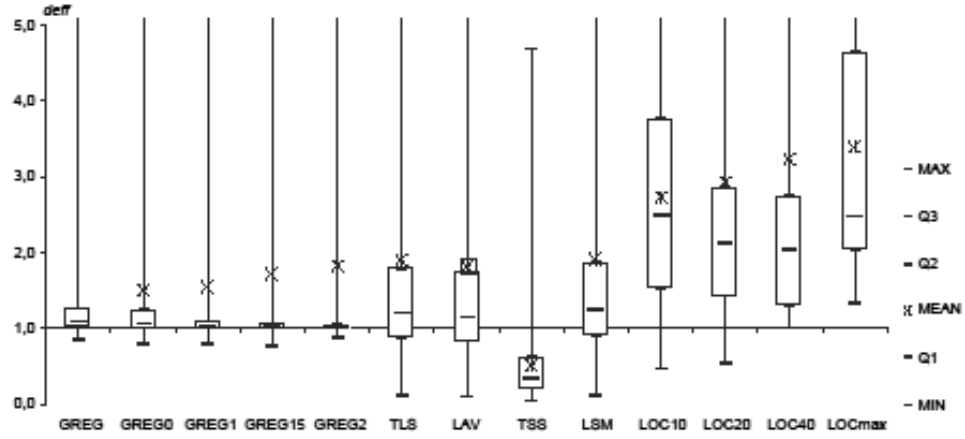
Source: Own estimation based on SP3 survey data, BJS and Tax Register.

Coefficient *RedCV (%)* represents the degree to which the variation of the estimators is reduced in comparison with the direct HT estimator (see figure 4). If the average value across all domains of study, or the quartiles of the distribution are considered, the reduction was obtained for the Winsor *TSS* estimator (on average by 33%, but for median by 61%) and the local regression estimators: $\hat{Y}_{loc}(10)$ (average - by 10%, median - by 9%) and $\hat{Y}_{loc}(\max,\min)$ (average - by 4%, median – by 8%). For the remaining estimation techniques applied, the average value of *CV* is bigger than in the case of direct HT estimator. This means that the techniques used did not result in improving the estimation precision. Generally, it is due to extreme values of *CV* observed for odd domains, which influence the mean. If the median, as a measure independent of outliers, was considered, reduction should be noticed for all Winsor and local regression estimators.

Another coefficient used to evaluate the gain in efficiency of an estimator is *deff*, which compares the variation of the estimator considered with respect to the

direct estimator (see figure 5). If its value is smaller than one, it means that the estimation technique applied is more efficient than the direct one. Evaluating the precision in terms of *deff* coefficient, it can be observed that the greatest gain was obtained by Winsor *sample splitting* estimator *TSS* (on average by 50%). The average variation across all domains, for the whole set of remaining estimators, is greater than that obtained using the direct one. Especially high value of *deff* coefficient is observed for all the local regression estimators. To explain this, we should bear in mind that although local models are constructed, in the estimation process all the observations are taken into account, including outliers. This has a direct influence on high variation of local estimators.

Figure 5. Characteristics of *deff* coefficient distribution across all domains, different estimators, 2001



Source: Own estimation based on SP3 survey data, BJS and Tax Register.

To complete the evaluation of the estimators, the bias analysis was conducted (see Chambers, Brown, Heady, Heasman (2001)). The procedure applied enabled the verification of the random character of the bias of the estimators in comparison with the unbiased direct estimates. Once again, the best results were obtained for the Winsor estimator.

Table 11 presents relation between direct and robust estimates obtained for revenue for a higher aggregation level. The estimates refer to domains defined as PKD Section of economic activity. It shows that local regression (Loc10) and *TSS* provided estimates with the smallest relative difference in case of all domains but one. This refers to GREG (γ =0) in section: *Human health and social work* .

Table 11. Comparison of the results for other levels of aggregation, Revenue, Type of Economic Activity (PKD Sections), 2001

Sections of Economic Activity	Relation between robust and direct estimates								
	GREG	$GREG^1$	$GREG^{1.5}$	$GREG^2$	TLS	LAV	TSS	LSM	LOC_{10}
Manufacturing	109,64	112,42	280,26	84,86	110,47	110,56	108,25	109,71	113,60
Construction	103,15	100,89	277,23	53,58	98,80	97,76	102,24	104,98	99,55
Wholesale and retail trade;	96,28	94,78	-15,30	118,78	96,49	96,45	96,92	95,89	98,75
Accommodation and food service activities	111,01	112,64	194,24	114,63	111,95	118,35	113,11	113,66	108,71
Transportation and storage	101,53	106,72	240,87	92,50	101,41	103,21	101,08	101,28	99,53
Financial and insurance activities	102,92	103,17	206,74	14,03	105,09	104,45	98,30	104,20	93,81
Real estate activities	108,12	113,14	389,46	82,06	105,27	104,22	108,78	107,62	96,98
Education	103,16	105,67	275,31	80,26	98,15	97,49	99,24	99,79	90,75
Human health and social work activities	101,51	102,32	185,47	92,34	105,55	107,75	106,20	103,33	107,17
Other service activities	103,50	105,55	26,16	150,02	113,44	113,99	101,13	106,16	101,39
Number of sections with smallest difference	1	0	0	0	0	0	**3**	1	**5**

Source: Own estimation based on SP3 survey data and Tax Register.

The idea of Winsor estimators differs in comparison with others used in this study in the sense that it modifies the outliers in the direction of the border points, whereas in the GREG's modification, as well as in the case of local regression, values of the auxiliary variables for the outlying observations are not modified. What is changed is their influence on the estimated variable thorough the value of weights. The results of the research show a negative influence of the outlying observations on estimation precision. These conclusions were supported by the correlation analysis between direct estimates and the estimates obtained by applying other robust techniques. In the case of Winsor estimators, a strong correlation is observed ($r \in (0{,}991;0{,}998)$), similarly for GREG modification but for local regression estimators the correlation is much weaker ($r \in (0{,}88;0{,}89)$).

Summing up the evaluation analysis, exploring the distribution of the coefficients *CV*, *RedCV* and *deff* across all domains, together with the bias analysis, the best record was obtained by the Winsor estimator using *Sample*

Splitting Technique (TSS). It is rather difficult to provide the ranking of the remaining estimators, as it changes depending on the criterion. It would be much easier to indicate the 'best' estimator in each group. As concerns the GREG modification it would be $\hat{Y}_{GREG}^{1,5}$, the one with *'z'* variable taking into account heteroscedasticity of the regression of *y* on *x* proportionally to $z_i^{1,5}$. But a well-known drawback of GREG – that it can often provide negative weights – should be borne in mind. Among local regression estimators we would point out $\hat{Y}_{loc}(10)$ with the narrowest bandwidth for which kernel functions were assigned. It was independently determined for each *i*-th observation by the *nearest neighbour bandwidth* technique within $(i \pm 10)$ observations. Characterizing the local regression it should be stressed that the bandwidth significantly influences the computing time. For Winsor estimators, the one implementing *Sample Splitting Technique* $\hat{Y}_{TSS}$ provided the most efficient estimates. In this method two regression models were estimated for sample observations randomly divided into two groups. The evaluation of such models in terms of residuals made us reject the most outlying observations. This approach makes the *TSS* technique more robust in comparison with others like *TLS* or *LAV*.

REFERENCES

BREIDT, F.J., OPSOMER, J.D. (2000) *Local Polynomial Regression Estimation in Survey Sampling.* The Annals of Statistics, **28**, 1026–1053.

CHAMBERS, R.L. (1996) *Robust case-weighting for multipurpose establishment Surveys*, Journal of Official Statistics, Vol.12, No.1, 3–32.

CHAMBERS, R., DORFMAN, A.H., WEHRLY, T.E. (1993) *Bias Robust Estimation in Finite Populations Using Nonparametric Calibration.* Journal of the American Statistical Association, **88**, 268–277.

CHAMBERS, R., KOKIC, P., SMITH, P. and CRUDDAS, M. (2000) *Winsorization for Identifying and Treating Outliers in Business Surveys*, Proceedings of the Second International Conference on Establishment Surveys (ICES II), 687–696.

CHAMBERS R., BROWN G., HEADY P., HEASMAN D. (2001) *Evaluation of Small Area Estimation Methods – an Application to Unemployment Estimates from the UK LFS*, Proceedings of Statistics Canada Symposium 2001, Achieving Data Quality in a Statistical Agency: a Methodological Perspective.

CHAMBERS, R.L, FALVEY, H., HEDLIN, D., KOKIC P. (2001a) *Does the Model Matter for GREG Estimation? A Business Survey Example*, Journal of Official Statistics, Vol.17, No.4, 527–544.

DEHNEL G. (2008) *Estymator GREG a estymacja typu Winsora w badaniach mikroprzedsiębiorstw (GREG and Winsor estimation in small business survey), [in:] Statystyka wczoraj, dziś i jutro (Statistics yesterday, today and tomorrow)*, Warszawa, Główny Urząd Statystyczny i Polskie Towarzystwo Statystyczne (Central Statistical Office and Polish Statistical Association), 58–71, Summ. - Bibliogr. ISBN 978-83-7027-431-3 (in Polish).

DORFMAN, A.H. (2000) *Non-Parametric Regression for Estimating Totals in Finite Populations.* Proceedings of the Survey Research Methods. American Statistical Association, 47–54.

EFRON, B. (1979) *Bootstrap methods: Another look at the jackknife*, [in:] Annals of Statistics 7, 1979, 1–26.

FALORSI P. D., FALORSI S., RUSSO A., PALLARA S. (2000) Small Domain Estimation Methods For Business Surveys, Statistics in Transition, June 2000, Vol. 4, No.5, 745–751.

HEDLIN D. (2004) *Business Survey Estimation*, R&D, Sweden.

HIDIROGLOU, M.H., SRINATH, K.P. (1981) *Some estimators of population total from simple random samples containing large units*, JASA, **76**, 690–695.

KIM, J.Y., BREIDT, F.J. and OPSOMER, J.D. (2001) *Local polynomial regression estimation in two-stage sampling.* Proceedings of the Section on Survey Research Methods, American Statistical Association, 55–61.

KISH L. (1965) *Survey Sampling,* Wiley.

KISH L. (1995) *Methods for design effects*, Journal Official Statistics*,* **11,** 55–77.

KLIMANEK T., PARADYSZ J. (2006) *Adaptation of EURAREA experience in business statistics,* "Statistics in Transition", Vol.7, No. 4.

KOKIC, P.N., BELL, P.A. (1994) *Optimal winsorizing cutoffs for a stratified finite population estimator*, Journal of Official Statistics, **10**, 419–435.

MACKIN, C., PRESTON J. (2002) *Winsorization for Generalised Regression Estimation*, Australian Bureau of Statistics.

PAWLOWSKA Z. (2005) *Role of small and medium enterprises in creating a demand on work*, [in:] "Wiadomosci Statystyczne"**,** No.2, 34–46 (in Polish).

SÄRNDAL C.E., SWENSSON B., WRETMAN J. (1992) *Model Assisted Survey Sampling*, Springer Verlag, New York.