A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 286, 2013

Grażyna Dehnel[*], *Tomasz Klimanek*[**], *Jacek Kowalewski*[***]

INDIRECT ESTIMATION ACCOUNTING FOR SPATIAL AUTOCORRELATION IN ECONOMIC STATISTICS

Abstract. The authors present results of a study which attempted to use indirect estimation methods (including a method accounting for spatial correlation) to estimate certain characteristics enterprises. The study relied on data from the DG-1 survey conducted by the Statistical Office in Poznan, which provides the basis for most of the short-term indicators used to describe enterprise activity in Poland. The DG-1 survey is a monthly report about economic activity, which collects crucial information about economic entities, their activity, production and stocks. The survey is addressed to enterprises employing over 9 people.

Key words: indirect estimation, spatial autocorrelation, economic statistics.

I. INTRODUCTION

Modern surveys in the field of economic statistics make wide use of classical methods of estimation. They are intended to estimate values of basic economic indicators for businesses for large domains, such as provinces or classes of economic activities. The growing demand for information for small domains, however, has called for new estimation methods that could meet the requirements specified by consumers of information.

Nowadays one can observe the development of such methods as statistical data integration, calibration, imputation or indirect estimation. In the case of economic statistics, the estimation of key variables proves particularly challenging owing to problems such as strong asymmetry, high variation and concentration, since it is difficult to retain the properties of classical estimators used in sample surveys, such as unbiasedness, or high effectiveness.

To overcome this problem, attempts are being made to apply indirect estimation techniques, which, under the above mentioned circumstances, could provide more reliable estimates than those obtained by direct estimation, thus "strengthening" estimates by, among others, exploiting the so-called auxiliary

[*] Ph.D., Poznan University of Economics, Department of Statistics.
[**] Ph.D., Poznan University of Economics, Department of Statistics.
[***] Ph.D., Statistical Office in Poznan.

variables from additional data sources. One suggested approach in indirect estimation for surveys involves using information about spatial correlation as an auxiliary variable. So far, attempts at incorporating spatial information in non-classical estimation have been made in the field of agriculture [Klimanek, Szymkowiak, 2010], labour market [Klimanek, 2012] or the residential property market [Beręsewicz, Klimanek, 2013]. This paper presents the results of a study where spatial correlation was exploited to support indirect estimation of economic statistics.

The study involved enterprises employing from 10 to 49 persons[1]. The main aim of the study was to test the usefulness of spatial correlation for indirect estimation of enterprises from the DG-1 survey. The second aim was to evaluate the effect of incorporating spatial information in the model on estimation precision. That aim could be reached by proving that application of the estimators taking into account the spatial autocorrelation could be justified by performing the formal test for the existence of spatial dependency (e.g. global Moran's I test).

II. DATA SOURCE

The study relied on information from a DG-1 survey conducted by the Statistical Office in Poznan. The survey is administered on a monthly basis. Its objective is to collect up-to-date information about basic indicators of economic activity of large and medium-sized enterprises, such as *revenue from sales (of products and services), number of employees, gross wages, volume of wholesale trade and retail sales, excise tax, specific subsidies.*

Data from the DG-1 survey are also used to estimate the majority of short-term indicators that characterise the socio-economic situation of the country and provinces. They are disseminated in reports released by the President of the Main Statistical Office (GUS), periodicals published by GUS and they are delivered to national and international institutions, such as the National Bank of Poland, IMF, UN, OECD and Eurostat.

Reporting in the DG-1 survey is obligatory for all large enterprises (employing more than 49 persons) and a 10% sample of medium-sized enterprises. The percentage of enterprises in different categories of the Polish Classification of Business Activity (PKD) selected for the sample is set to reflect the structure of enterprises in the province. Economic units are divided into strata by ownership status and PKD category.

The sample frame includes 98,000 units, of which 19,000 are large enterprises (with over 49 employees), 80,000 are medium-sized enterprises (from 10 to 49 employees).

[1] In Polish public statistics, for purposes of statistical reporting, this subpopulation is defined as medium-sized enterprises (from 10 to 49 employees).

In effect, about 30,000 units (both large and small) participate in the survey every month.

III. DESCRIPTION OF THE STUDY

The study was limited to medium-sized enterprises that were active in August of 2012. *Gross wages* were the target variable, while *revenue from sales of products (goods and services)* was the auxiliary variable.

Table 1. Descriptive statistics on the distribution of the target variables

Descriptive statistics	Net revenue from sales of products (goods and services produced by the company)	Net revenue from sales of goods and materials	Total net revenue from sales	Gross wages listed in field 07
	in thousand PLN			
min	0	0	0	0
max	3918065	5468647	94368.3	9386712
Q_1	72.2	0	64	357
Q_2	416.95	56.7	149.7	1080.1
Q_3	1612.975	800	351.2	3342.8
mean	3578	3116	442	6694
s(x)	37284	46298	1606	73740
$V_{s(x)}$ (%)	1042	1486	364	1102

Source: own calculations based on DG-1 survey data from August 2012.

The general population included all medium-sized enterprises that participated in the DG-1 survey. This choice enabled access to detailed information about the target and auxiliary variables. With the general population defined in this way, it was possible to conduct a simulation study, which was then used to evaluate estimation precision.

Gross wages were estimated by region and PKD category. The domain adopted for purposes of estimation was a unit created by combining territorial information at the subregional level (NUTS 3) with the economic classification defined in terms of PKD categories. The population was thus broken down into 990 domains (66 regions x 15 PKD categories).

It seemed quite natural for real life problems to relax the strong assumption of independence connected with effects for every two regions (assumption underlying the simplest version of EBLUP estimator). One could expect that neighbouring regions or regions with relatively short distance between each

other are more similar than regions separated by a larger distance. This was the motivation for the authors to include also the analysis of spatial correlation and to examine its influence on the behavior of the estimation in the study.

To detect the spatial dependency the Moran's I test was used under the randomization approach where the spatial weights matrix was row standardized. The computations were conducted in **R** environment with the use of ***spdep*** package. The results are presented in the table below, suggesting that only for section F (*construction*), there exists spatial autocorrelation as far as the *gross wages* are considered. The phenomena could be also seen on the map presenting the spatial distribution of the variable under study. For section F (construction) we can see clusters of NUTS3 areas with similar level of gross wages (Fig. 2) while for section C (*manufacturing*) and G (*trade*) something resembling a mosaic could be seen (Fig. 1, Fig. 3).

Table 1. Global Moran's I autocorrelation coefficient for gross wages for sections: C (*manufacturing*), F (*construction*) and G (*trade*) under randomization approach

Section	Moran's I statistics	Moran's I statistic deviate	*p*-value
Manufacturing	–0.0472	–0.3732	0.6455
Construction	0.1486	1.9234	0.02722
Trade	–0.0220	–0.0773	0.5308

Source: Own calculations.

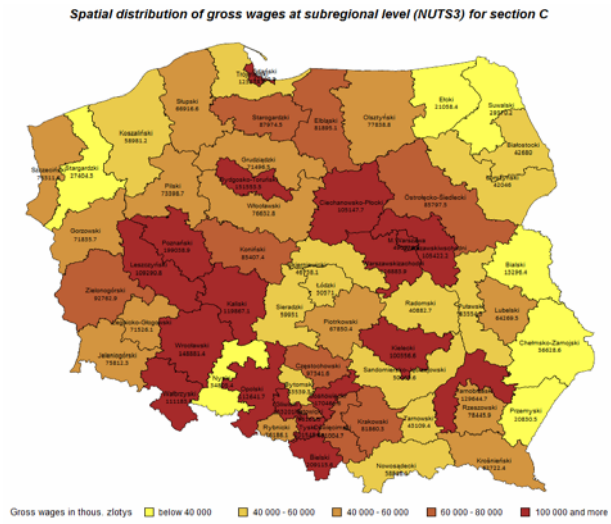

Fig. 1. Spatial distribution of gross wages at subregional level (NUTS3) *manufacturing*

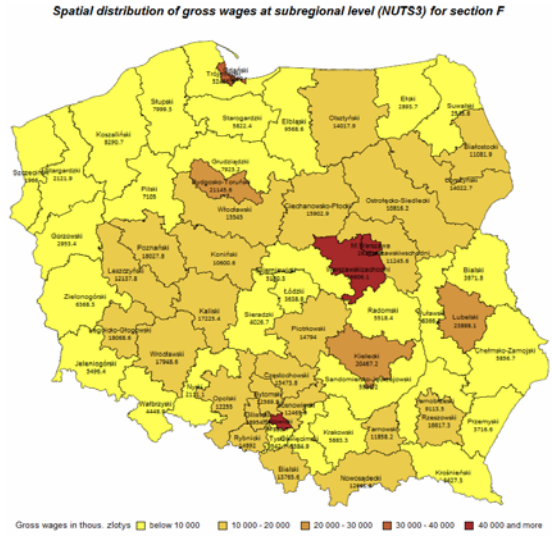

Fig. 2. Spatial distribution of gross wages at subregional level (NUTS3) for *construction*

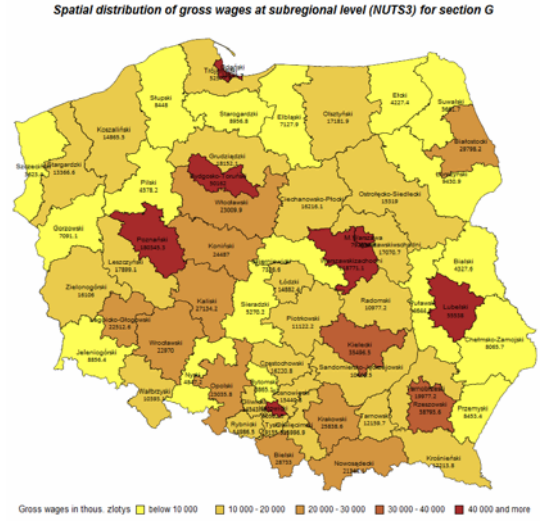

Fig. 3. Spatial distribution of gross wages at subregional level (NUTS3) for *trade*

The precision of estimators analysed in the study was evaluated using the bootstrap method. 1000 iterations of drawing 5% samples were made, which were then used to calculate:

- Mean absolute relative bias (for model based estimators: synthetic, eblup and spatial eblup)

$$ARB\left(\hat{Y}_d\right)=\frac{1}{1000}\left|\sum_{b=1}^{1000}\frac{\hat{Y}_{b,d}-Y_d}{Y_d}\right| \tag{1}$$

where Y_d - population parameter for domain *d*,

- Relative root mean square error (for model based estimators: SYNTH, EBLUP and SEBLUP)

$$RMSE\left(\hat{Y}_d\right)=\frac{\sqrt{\frac{1}{1000}\sum_{b=1}^{1000}\left(\hat{Y}_{b,d}-Y_d\right)^2}}{Y_d} \tag{2}$$

Owing to the large volume of estimation results, the following presentation is limited to estimates for the variable of *gross wages* for 3 PKD categories: *manufacturing, construction, trade*.

IV. ESTIMATION METHODS

The following five estimators were used[2]:
– direct (DIRECT)
– generalized regression estimator (GREG)
– synthetic regression estimator (SYNTH)
– empirical best linear unbiased predictor (EBLUP)
- empirical best linear unbiased predictor with spatial autocorrelation (SEBLUP)The direct estimator (Horvitz–Thompson)

$$\hat{\bar{Y}}_d^{DIRECT}=\frac{1}{\hat{N}_d}\sum\nolimits_{i\in u_d} w_{id}\,y_{id} \tag{3}$$

[2] Formulas used to estimate MSE are omitted here for the sake of brevity. However, they can be found in the documentation of the EURAREA Project on the website of UK's ONS http://www.statistics.gov.uk/eurarea.

where $\hat{N}_d = \sum_{i \in u_d} w_{id}$ and $w_{id} = \frac{1}{\pi_{id}}$ assuming that inclusion probability $\pi_{id,jd} = 0$ for all $d \neq d'$ or $i \neq j$.

- GREG estimator

$$y_{id} = \mathbf{x}_{id}^T \boldsymbol{\beta} + \varepsilon_{id} \tag{4}$$

where $E(\varepsilon_{id}) = 0,\ Var(\varepsilon_{id}) = \sigma_\varepsilon^2$

$$\hat{Y}_d^{GREG} = \frac{1}{\hat{N}_d} \sum_{i \in s_d} \frac{y_i}{\pi_i} + \left(\overline{\mathbf{X}}_d^T - \frac{1}{\hat{N}_d} \sum_{i \in s_d} \frac{\mathbf{x}_i}{\pi_i} \right)^T \hat{\boldsymbol{\beta}} \tag{5}$$

where $\hat{N}_d = \sum_{i \in s_d} \frac{1}{\pi_i}$ and $\hat{\boldsymbol{\beta}}$ are estimated using the method of weighted least squares and applying design–based weights:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in u_d} w_{id} x_{id} x_{id}^T \right)^{-1} \sum_{i \in u_d} w_{id} x_{id} y_{id} \tag{6}$$

- SYNTH estimator

Standard two level linear model:

$$\boldsymbol{y_{id}} = \boldsymbol{x_{id}^T \beta} + \boldsymbol{u_d} + \boldsymbol{e_{id}} \tag{7}$$

$$u_d \sim iid\ N(0, \sigma_u^2), \quad e_{id} \sim iid\ N(0, \sigma_e^2)$$

$$\hat{\overline{Y}}_d^{SYNTH} = \overline{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}} \tag{8}$$

with $\overline{\boldsymbol{X}}_{.\boldsymbol{d}} = (\overline{\boldsymbol{X}}_{.\boldsymbol{d},1}, \ldots, \overline{\boldsymbol{X}}_{.\boldsymbol{d},\boldsymbol{p}})^{\boldsymbol{T}}$. Estimator does not respect sampling weights.

- EBLUP estimator [Chambers, Saei 2004],

$$\hat{\overline{Y}}_d^{EBLUP} = \sum_{i=1}^{n_d} y_{id} + \mathbf{X}_r^* \hat{\boldsymbol{\beta}} + \mathbf{Z}_r^* \mathbf{T}_s^* (\mathbf{y}_s^* - \mathbf{X}_s^* \hat{\boldsymbol{\beta}}) \tag{9}$$

where: $\mathbf{X}_r^*$ is the matrix of covariates not observed in the sample, which values are known from the population, $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s^T \sum_s^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^T \sum_s^{-1} \mathbf{y}_s$ is Aitken's generalized least squares (GLS) estimator of $\boldsymbol{\beta}$, $\mathbf{Z}_r^*$ is incidence matrix for the random effects not observed in the sample $\mathbf{T}_s^* = \Omega^{-1} \mathbf{Z}_s^T \mathbf{W}_s^{-1} \sum_s^{-1} \mathbf{Z}_s$, $\mathbf{y}_s^*$ is vector of $\mathbf{y}$ values observed in the sample, $\mathbf{X}_S^*$ is the matrix of covariates observed in the sample, which values are known from the population.

All parameters are estimated using restricted maximum likelihood (REML) method. The predicted values contain weighted fixed and random effects.

- SEBLUP estimator accounting for autocorrelation of random effects connected with the location of domains in space [Chambers, Saei 2004, D'Alò, Falorsi, Solari 2004].

In matrix notation the model can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{10}$$

where: $\mathbf{y}$ is a vector of the dependent variable, $\mathbf{X}$ and $\mathbf{Z}$ are known matrices of the orders, respectively: $N \times P$ (the number of observations times the number of auxiliary variables) and $N \times P$ (the number of observations times the number of small areas). Matrix $\boldsymbol{Z}$ is an incidence matrix defined in the following way:

$$\boldsymbol{Z} = \begin{bmatrix} 1_{N_1} & . & . & 0 \\ 0 & . & . & 0 \\ 0 & . & . & 0 \\ 0 & . & . & 1_{N_D} \end{bmatrix},$$

where 1_{N_d} is a vector $\boldsymbol{N_d}$, with all elements equal to 1.

$\mathbf{u}$ and $\mathbf{e}$ are vectors of random variables with expected values equal to 0 and a covariance–variance matrix respectively $\boldsymbol{N \sim [0, \sigma_U^2 A]}$ and $\boldsymbol{N \sim [0, \sigma^2 I_N]}$, elements $\boldsymbol{a_{(dd')}}$ of matrix $\boldsymbol{A}$ are given by the formula:

$$a_{(dd')} = \left[1 + \delta_{(dd')} \exp\left(\frac{dist(dd')}{\alpha} \right) \right]^{-1}, \tag{11}$$

where $dist(dd')$ denotes the distance between small areas $\boldsymbol{d}$ and $\boldsymbol{d'}$,

$$\delta_{(dd')} = \begin{cases} 0 \text{ for } & d = d' \\ 1 \text{ for } & d \neq d' \end{cases} \text{ and } \alpha \text{ is a parameter of scale.} \tag{12}$$

V. RESULTS

Because of the limited space of the article, only a small part of the distribution of the estimators for domains is presented[3] (see Fig. 4-6).

The results presented above show that design based estimators although unbiased are in many cases unsatisfactory because of the large variance of the estimates. On the other hand, the distribution of model based estimates is more leptokurtic and in many cases it follows the normal distribution while the distribution of DIRECT or GREG estimators sometimes is multimodal or highly skewed. It is very difficult to point out which of model based estimators has better properties based on the presented figures.

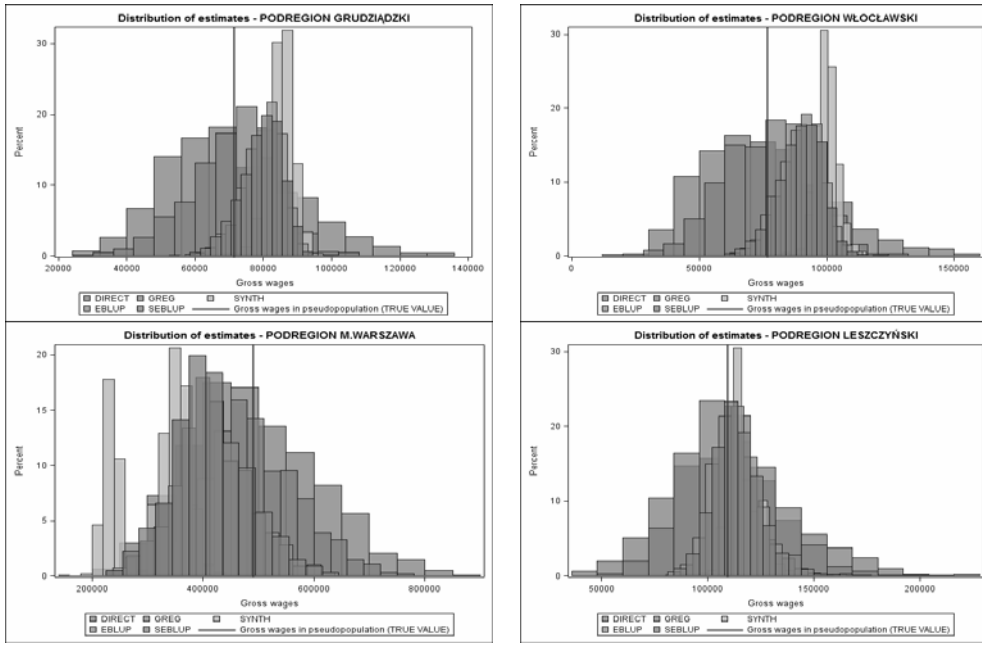

Fig. 4. Distribution of estimates for selected NUTS3 areas in section C (manufacturing)
Source: own calculations based on DG–1 survey data from August 2012.

[3] 66 figures were produced for every section. Total number of figures is then 3 x 66 = 198.

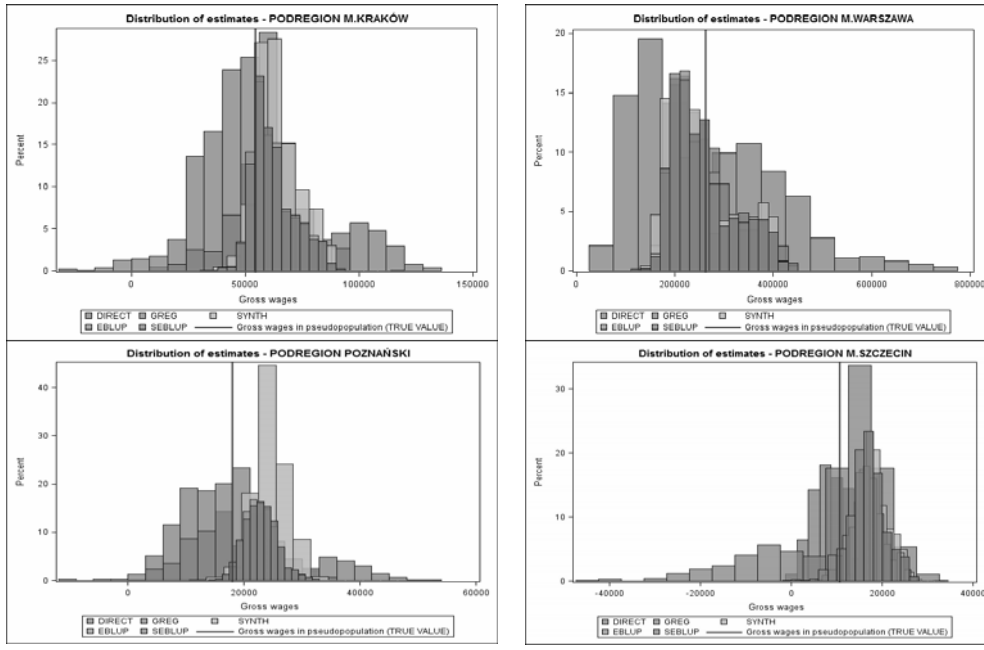

Fig. 5. Distribution of estimates for selected NUTS3 areas in section F (construction)
Source: own calculations based on DG–1 survey data from August 2012

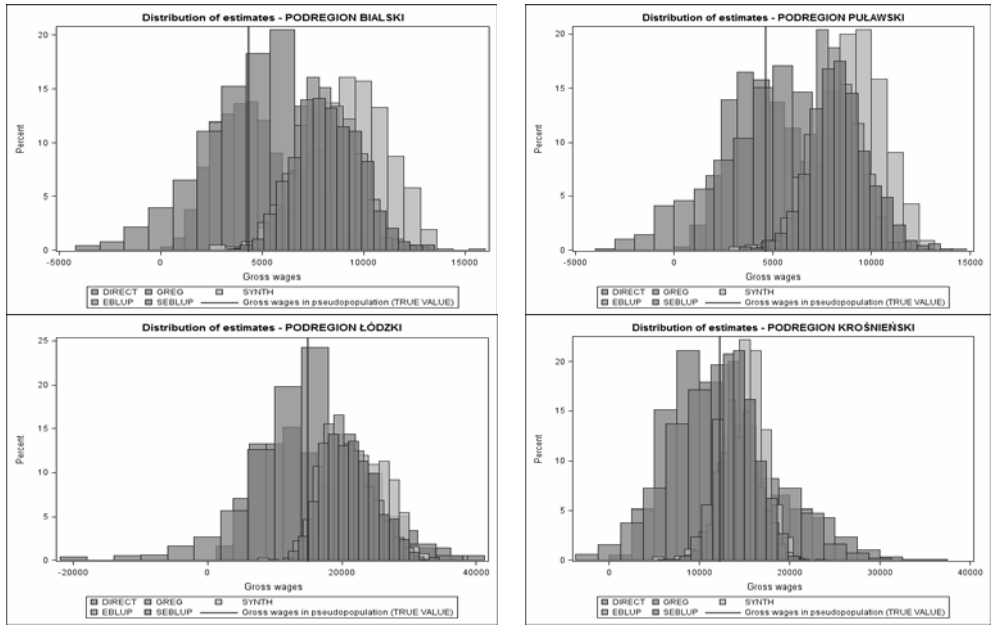

Fig. 6. Distribution of estimates for selected NUTS3 areas in section G (trade)
Source: own calculations based on DG–1 survey data from August 2012.

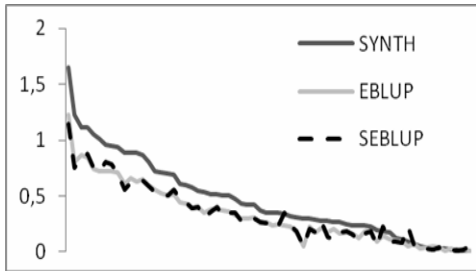

Fig. 7. Distribution of mean absolute relative bias for manufacturing

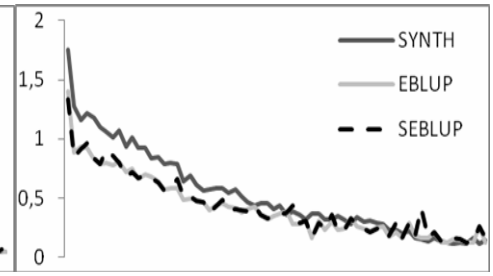

Fig. 8. Distribution of relative root mean square error for manufacturing

Source: Own calculations. NUTS3 subregions are ordered according to ARB of SYNTH estimator.

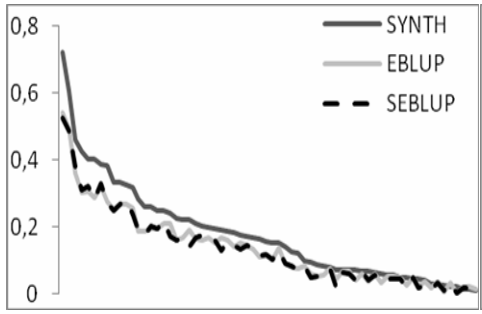

Fig. 9. Distribution of mean absolute relative bias for construction

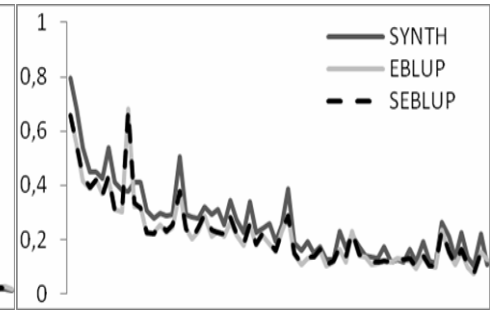

Fig. 10. Distribution of relative root mean square error for construction

Source: Own calculations. NUTS3 subregions are ordered according to ARB of SYNTH estimator.

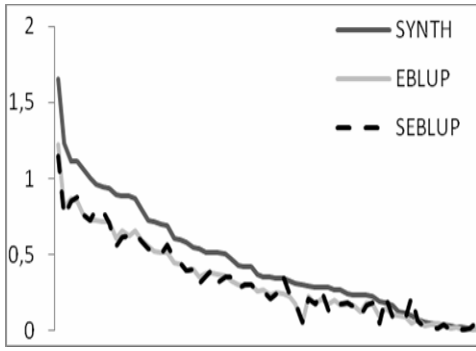

Fig. 11. Distribution of mean absolute relative bias for trade

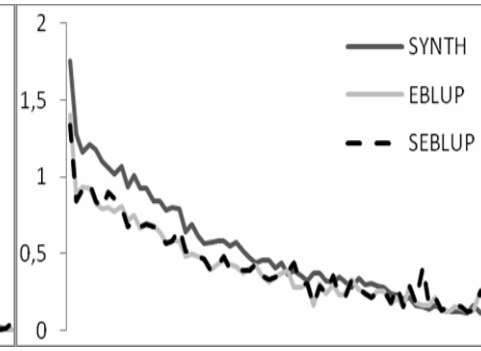

Fig. 12. Distribution of relative root mean square error for trade

Source: Own calculations. NUTS3 subregions are ordered according to ARB of SYNTH estimator.

Figures 7-12 present the distribution of two performance criteria: mean absolute relative bias and relative root mean square error for three analyzed sections. In every case the order of the subregions (NUTS3 areas) was presented according to the decreasing value of mean absolute relative bias of synthetic estimation. One could see that two versions of EBLUP estimation have in general better properties in relation to the above mentioned criteria than SYNTH estimator. There are very few cases when the synthetic estimation is better than EBLUP. Still there is the problem to claim that spatial version of EBLUP is better than just simple version of EBLUP. The authors decided to present also a table showing the share of cases when SEBLUP was better than EBLUP (see table 2). So two conditions were created: one for ARB criterion and the second for RMSE. In more than half of cases SEBLUP outperformed EBLUP for section C (*manufacturing*) where Moran's I test showed no significant spatial autocorrelation. Quite similar situation was for section G (trade) – SEBLUP was better in relation to ARB criterion but worse when analyzing RMSE. Quite inspiring results were obtained for section F (construction). Although Moran's I test detected the presence of significant spatial autocorrelation, SEBLUP estimation was better in the context of absolute relative bias, but in 85% of cases RMSE was lower for EBLUP estimator. Such results suggest that a more detailed approach here is needed – e.g. local Moran's I statistics. The value of this measure could be a kind of recommendation for determining types of small domains and then for building separate models for these types.

Table 2. Better performance of SEBLUP estimator in relation to EBLUP (in %)

Section	ARB	RMSE
Manufacturing	53.0	54.5
Construction	60.6	15.2
Trade	60.6	42.4

Source: Own calculation.

VI. CONCLUSIONS

Several remarks could be formed based on the results of the study:

- direct estimator, although design unbiased has two disadvantages in the case of small domain estimation:

1. the variance of estimates and of course estimation error are in most cases unacceptable,
2. when sample size in a domain is zero, no estimates could be generated

- generalized regression estimator has also unacceptable variance of the estimates,
- model based estimators (synthetic, empirical best linear unbiased predictor and empirical best linear unbiased predictor with autocorrelation structure) have smaller variance but they are biased,
- although in many cases the estimation which takes into account the spatial autocorrelation has better properties compared with other estimators, there is no clear evidence that better properties are related to the measure of the global Moran's I statistics,
- there is the need to conduct a similar study which will take into account ESDA techniques (e.g. local Moran's I statistics).

REFERENCES

D'Alò M., Falorsi S., Solari F. (2004) *EURAREA Documentation on SAS/IML program on Linear Mixed Model with Spatial Correlated Area Effects in Small Area Estimation*, EURAREA Deliverable 3.3.2, *EURAREA EBLUPGREG Software Documentation*, Statistics Finland EURAREA Consortium, Deliverables D2.3.2, D3.3.2.

Klimanek T., *Wykorzystanie estymacji pośredniej, uwzględniającej korelację przestrzenną w analizie rynku pracy* (2012) [w:] Analiza wielowymiarowa w badaniach społeczno–ekonomicznych, Wydawnictwo UE w Poznaniu (UEP), Poznań, s. 126-139.

Klimanek T., Szymkowiak M., *Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy* (2012) [w:] „Taksonomia 19 : klasyfikacja i analiza danych – teoria i zastosowania", Wydawnictwo UE we Wrocławiu, s. 601-609.

Klimanek T., Beręsewicz M., *Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w badaniach rynku nieruchomości* (2013) [w:] Taksonomia 20 : klasyfikacja i analiza danych – teoria i zastosowania, Wydawnictwo UE we Wrocławiu, *w druku.*

Saei A., Chambers R., (2004) *Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects*, University of Southampton.

Grażyna Dehnel, Tomasz Klimanek, Jacek Kowalewski

ESTYMACJA POŚREDNIA Z UWZGLĘDNIENIEM KORELACJI PRZESTRZENNEJ W STATYSTYCE GOSPODARCZEJ

W referacie zostaną przedstawione wyniki badania, w którym podjęto próbę wykorzystania metod estymacji pośredniej (w tym także metody, która uwzględnia korelację przestrzenną) do oszacowania pewnych charakterystyk przedsiębiorstw. W badaniu wykorzystano informacje pochodzące z badania DG–1 prowadzonego przez Urząd Statystyczny w Poznaniu, stanowiącego podstawę do opracowywania większości wskaźników krótkookresowych dotyczących działalności przedsiębiorstw w Polsce. Badanie DG–1 to miesięczny meldunek o działalności gospodarczej, który zawiera najważniejsze informacje dotyczące podmiotów gospodarczych, ich działalności oraz produkcji wyrobów i zapasów. Obejmuje ono swym zasięgiem przedsiębiorstwa, w których liczba pracujących przekracza 9 osób.