PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

TOMASZ KLIMANEK

USING INDIRECT ESTIMATION WITH SPATIAL AUTOCORRELATION IN SOCIAL SURVEYS IN POLAND[1]

1. BACKGROUND

First attempts at applying various approaches to parameter estimation for small areas in Poland were undertaken after the International Conference on Small Area Statistics held in Warsaw in 1992 (eds. Kalton, Kordos, Platek, 1993). There were only a few applications of small area estimation (SAE) methods to measure the scope of unemployment, poverty, household structure and in agriculture related surveys (Kordos, Paradysz, 2000). Further applications and examination of "standard" indirect estimators properties were undertaken within the EURAREA project[2] after the IASS Satellite Conference held in Riga in 1999. The first important study to apply SAE methodology in LFS was conducted in 2003. The authors (Bracha, Lednicki and Wieczorkowski, 2003) estimated totals and rates of several labour market characteristics by region, subregion and poviat (NUTS2, NUTS3 and NUTS4). They used direct, synthetic and composite estimators.

The second important study was conducted in 2004 by E. Gołata. The study was intended to rely on EURAREA project experiences. The Polish database - the so-called super-population labelled POLDATA - was created on the basis of 3 data sources: the 1995 Micro-census, the 1995 Household Budget Survey and the Local Data Bank. POLDATA represented as closely as possible the characteristic of Poland in 1995 with respect to the new administration division of the country which was introduced in January 1999. For the purposes of applying the standard estimators, the proportion of ILO unemployed (in the whole population over 15) in an area was estimated (Gołata, 2004).

One should also mention the study conducted in 2004 by Kubacki. The parameter of interest in the study was unemployment size for NUTS 2 and NUTS 4 level. Registered unemployment constituted the additional data source used in the study (with covariates: the number of unemployed persons, the number of employed persons, the

[1] The results presented in the paper are the outputs of Work Package 5 (Case studies) of EUROSTAT project ESSnet on SAE 61001.2009.003-2009.859 (2009-2012).

[2] The EURAREA project no. IST-2000-26290 entitled *Enhancing Small Area Estimation Techniques to meet European needs* is part of 5th framework programme for research, technological development and demonstration of EU. Its main co-ordinator is ONS – Office for National Statistics, UK.

number of economically inactive persons, the number of dwellings and the number of persons aged 15 and above, Kubacki, 2004).

Both design and model based types of estimators were applied:

– design based estimators including post stratification methods (both ratio and regression estimator), synthetic estimator (both ratio and regression estimator),
– model based estimators include empirical Bayes (EB) estimator and hierarchical Bayes (HB) estimator.

Recent years have seen a growing interest in new possibilities and tools developed to meet the growing demand for estimates at local level. Special projects were carried in the Central Statistical Office (CSO) in cooperation with university researchers. The projects referred to different subjects, for example: labour market, household structure, business statistics including small business. From the perspective of the Population Census 2011, which was in progress at the time, special research was undertaken within the Central Statistical Office and newly established Centre for Small Area Statistics in Poznan. It was aimed at examining administration registers, their quality and usefulness as sources of auxiliary variables in small domain estimation. But practical application of SAE methods of official statistics in Poland is still not part of normal practice.

2. GENERAL SETUP

The aim of the Author was to continue explorations of estimating labour market characteristics for small domains. First, data infrastructure referring to economic activity and unemployment is presented and discussed. The intention was to include all variable categories, which experience has shown to be effective. In the case of ILO[3] unemployment the 'standard variables' are: age, sex, education, employment status and housing. In the study data from the following sources were used:

– Register of Unemployed,
– Vital Statistics Register,
– Tax Register data will be used in an indirect form via Commuting Survey which was based on data from Tax Register – the first edition of the survey is available for 2006.

Taking into account special features of the labour market in Poland, especially its high territorial differentiation, various estimation techniques will be analysed. Theoretical approaches to estimation with spatial effects proposed by R. Chambers & A. Saei (2003) together with the SAS software provided by EURAREA project are considered

[3] ILO unemployment – unemployment defined according to International Labour Organization, which is comparable among European and non-European countries. According to this definition the unemployed in general comprise all persons above a specified age who during the reference period were: without work, that is, were not in paid employment or self employment during the reference period; currently available for work, that is, were available for paid employment or self-employment during the reference period; and seeking work, that is, had taken specific steps in a specified recent period to seek paid employment or self-employment.

and adjusted to fit specific arrangements. The standard EURAREA estimators used in the study are: direct (for comparative reasons), generalised regression estimator – GREG, synthetic and EBLUP. Also EBLUPGREG, which takes into account spatial correlation structure, was applied.

The structure of economically active population and, especially, unemployment and its structure are of exceptional social interest in Poland. Unemployment, since the very beginning of 90's has assumed alarming dimensions and is characterised by great territorial differentiations at the national as well as regional level. This characteristic is due to structural differences in economy and regional inequalities shaped by distinct historical experiences as well as the transformation process. Regularities observed at the national level, in most cases, cannot be generalised and differ from region to region. For example, in June 2011 the highest registered unemployment rate in Poland was observed in the Warmia-Masuria Province – 19,5%, and the lowest in the Wielkopolska Province – 9,1% (province (voivodship) refers to the NUTS2 level according to Eurostat territorial division).This situation requires advanced studies reflecting regional specificities.

Data available from the Labour Force Survey enable estimation of employed and unemployed for the whole country by age, sex and place of residence: urban and rural areas. But at the regional level (NUTS2) only aggregated data can be obtained from LFS differentiated by sex and place of residence (into urban and rural areas), but not by age.

The goal is then to estimate the percentage of unemployed people in the population of 15 and older[4] at the NUTS3 level based on data from the Labour Force Survey 1^{st} quarter of 2008. Although there had been previous attempts to estimate unemployment at the NUTS4 but they were not very satisfactory.

3. DESCRIPTION OF THE SURVEY AND EMPLOYMENT-RELATED POPULATION FLOW STUDY

Labour Force Survey

The LFS methodology is based on the definitions of the economically active population, the employed and the unemployed adopted by the Thirteenth International Conference of Labour Statisticians (October 1982) and recommended by the International Labour Office. The survey concentrates on the situation from the point of view of economic activity of population, i.e. the fact of being employed, unemployed or economically inactive in the reference week. The labour force survey is a probability sample survey.

Sampling for the LFS follows the two-stage household sampling. The primary sampling units subject to the first stage selection, are census units called census clasters

[4] It is different from the unemployment rate but such an assumption will significantly simplify the computations, especially as far as the MSE of the estimators is concerned.

– CCs in towns, while in rural areas they are enumeration districts – EDs[5]. Second stage sampling units are dwellings[6]. The primary sampling units (PSU) are sampled with the application of the so-called stratification. Strata correspond to provinces (voivodships). Strata within provinces were created depending on the size of a place; rural areas were included into the smallest ones.

The estimation process consists of defining the appropriate generalizing factors, referred to as weights. This is achieved in three steps. The first step provides primary weights, which basically are the reciprocals of selection probabilities for ultimate sampling units (i.e. dwellings), which compensates for the disproportionate construction of the sample. The secondary weights are calculated in the next step by dividing primary weights by R, where R rate depends on the category of a place of residence of a given dwelling (the rural area or one of the five town classes mentioned above). The secondary weights are also final for the results concerning households and families. Final weights for the results concerning the population are calculated in the third step. The purpose of this step is to adjust the LFS results to the current demographic estimates. It is given by calculating the so-called modifiers for each of 48 categories defined by the place-of-residence (urban/rural), sex and 12 age groups. Final weights result from multiplying secondary weights by adequate modifiers.

The variances of complex estimators obtained in LFS cannot be estimated with ordinary methods and special, approximate procedures must be employed. Since 2003 one of the most popular approximate methods has been chosen for this purpose, based on the resampling and bootstrap rule. The detailed description of the bootstrap procedure applied in the variance estimation for LFS estimates in Poland is presented in details in Bracha et al. (2003).

Tax Register – general characteristics of an employment-related population flow study in 2006

The use of administrative registers in Polish public statistics is at the initial stage. The only larger study in which they played a supporting role was the 2011 National Census, which relies on information from various sources in order to collect certain data (reducing the burden on respondents), update the sampling frame for sample surveys and to update the database of buildings and dwellings. Wider access to administrative databases provides an opportunity for Polish public statistics to develop methods of

[5] In rural areas application of smaller first stage sampling units is useful for organizational reasons, but negatively influences precision of results for these areas. In order to improve this, the principle of the so-called overrepresentation of rural areas was applied, i.e. the number of dwellings sampled from rural areas is about 10% higher than the number resulting from the so-called proportional allocation (related to the number of dwellings in the whole population).

[6] Sampling of primary sampling units and dwellings is conducted on the basis of the Domestic Territorial Division Register, including among others a list of territorial statistical units and a list of dwelling addresses within CC's and EDs.

how they can be used in statistical reporting, and consequently, to upgrade the statistical infrastructure.

A study of employment-related population flow was conducted on the basis of data from the tax system collected in the POLTAX database in the Statistical Office in Poznan in 2009. The study was intended to provide estimates about the volume and directions of commuter traffic involving people in paid employment, using data from 31 December 2006. The results and the methodological details of the study have been made partially available in the Regional Data Bank and in the book entitled "Commuting in Poland", edited by K. Kruszka, Poznan, in October 2010.

Registers, after verification and cleaning, turned out to be a good source of information about the structure of economic activity from the territorial perspective. Additionally, the set included characteristics of sex and age (variable derived from the variable ”birth date"), which enabled another breakdown. One disadvantage is their incompleteness. In addition, they do not cover all the characteristics reported in other studies (such as education, class of places of residence, etc., which are included in LFS); this means that these datasets cannot fully replace the previously used measurement tools. Nevertheless, registers can play a supporting role and provide a good source for auxiliary variables in indirect estimation of labour market characteristics.

4. APPROACH TO THE PROBLEM

The problem was to estimate the percentage of unemployed people at the lower level of aggregation than presented in the CSO's publications. Having in mind that data available from the Labour Force Survey enable direct estimates of employed and unemployed for the whole country by age, sex and place of residence and for aggregated data at the regional level (NUTS2), it was decided to try to get small area estimates at the NUTS3 level.

Another problem was to evaluate the results. The applied software of course enables us to compute MSEs for the studied estimators, but there was a serious difficulty to validate the estimates against population values. It was decided that despite of small differences in the definition of LFS and registered unemployment the latter will be used as a kind of benchmark.

The natural choice was to use the EURAREA code.

The variable *status on the labour market* was recoded into binary variable getting 1 if the person was unemployed and 0 otherwise. This way target variable was created.

As potential covariates in the model we chose:

– commuting to work,
– place of residence,
– sex,
– 6 age groups.

All covariates were recoded into binary variables producing nine variables:

X1 – commuting to work (1 if a person commutes to work, 0 otherwise),

X2 – place of residence (1 if a person lives in rural area, 0 otherwise),
X3 – sex (1 if a person is a male, 0 otherwise),
X4 – group of age (1 if a person is up to 20 years of age, 0 otherwise),
X5 – group of age (1 if a person is 20-24 years of age, 0 otherwise),
X6 – group of age (1 if a person is 25-34 years of age, 0 otherwise),
X7 – group of age (1 if a person is 35-44 years of age, 0 otherwise),
X8 – group of age (1 if a person is 45-54 years of age, 0 otherwise),
X9 – group of age (1 if a person is over 55 years of age, 0 otherwise).

Then we used stepwise selection to get the following model (two variables were excluded from the model: X6 was not significant and X9 because of collinearity):

Table 1.

Model parameter estimates (domain level) after excluding X6 and X9

Variable	Parameter estimate	Standard error	t Value	Pr > |t|
Intercept	0.22376	0.18512	1.21	0.2317
X1	-0.26409	0.09227	-2.86	0.0058
X2	0.00630	0.02145	0.29	0.7700
X3	-0.50223	0.45085	-1.11	0.2699
X4	2.40624	0.63456	3.79	0.0004
X5	-1.11075	0.53435	-2.08	0.0421
X7	-1.25566	0.30082	-4.17	0.0001
X8	1.00178	0.24339	4.12	0.0001

And its goodness of fit is as follows

Table 2.

Model goodness of fit (domain level) after excluding X6 and X9

Root MSE	0.01236	R-Square	0.7191
Dependent Mean	0.05522	Adj R-Sq	0.6852
Coeff Var	22.37504		

Although another two variables (X2 and X3) are not significant we decided to include them in the model as in our opinion they are of special importance (place of residence and sex).

5. METHODS

The seven “standard” estimators were applied:
– synthetic population level estimator NSMEAN,
– direct estimator,

– GREG with a standard linear regression model,

– synthetic estimator considered under two different models:

a) a linear two-level model with individual data,

b) a linear model with area-level covariates and a pooled sample estimate of within-area variance,

– EBLUP estimator using models:

a) a linear two-level model with individual data,

b) a linear model with area-level covariates and a pooled sample estimate of within-area variance.

The eighth method was also EBLUP estimator however based on the assumption of the existence of spatial autocorrelation – EBLUPGREG SPATIAL – a linear two-level model with individual data taking into account the spatial correlation structure.

Direct estimator

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

where: $\hat{N}_d = \sum_{i\in u_d} w_{id}, \quad w_{id} = 1/\pi_{id}$.

MSE Estimator:

$$M\hat{S}E(\hat{\hat{Y}}_d^{DIRECT}) = \left(\frac{1}{\hat{N}_d}\right)^2 \sum_{i\in u_d} w_{id}(w_{id}-1)(y_{id}-\hat{\hat{Y}}_d^{DIRECT})^2 \tag{2}$$

(assuming, that $\pi_{id,jd'} = 0$, for all $d \neq d'$ or $i \neq j$.

GREG Estimator

$$y_{id} = \mathbf{x}_{id}^T\boldsymbol{\beta} + \varepsilon_{id}, \tag{3}$$

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

$$\hat{Y}_d^{GREG} = \frac{1}{\hat{N}_d}\sum_{i\in s_d}\frac{y_i}{\pi_i} + \left(\bar{\mathbf{X}}_d^T - \frac{1}{\hat{N}_d}\sum_{i\in s_d}\frac{\mathbf{x}_i}{\pi_i}\right)^T \hat{\boldsymbol{\beta}}, \tag{4}$$

where $\hat{N}_d = \sum_{i\in s_d}\frac{1}{\pi_i}$ and $\hat{\boldsymbol{\beta}}$ are estimated by using LSM weighted by weights resulting from the sampling process:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i\in u_d} w_{id}x_{id}x_{id}^T\right)^{-1} \sum_{i\in u_d} w_{id}x_{id}y_{id}, \tag{5}$$

$$\mathrm{E}(\varepsilon_{id}) \;=\; 0, \qquad \mathrm{Var}\ (\varepsilon_{id}) = \sigma_\varepsilon^2.$$

Assuming: $r_{id} = y_{id} - \mathbf{x}_{id}^T\hat{\boldsymbol{\beta}}$ and $\hat{Y}_d^{GREG} = \sum_{i\in u_d} w_{id}g_{id}y_{id}$, with weights g : $g_{id} = 1 + (\bar{X}_{.d} - \bar{x}_{.d})^T (\sum_{i \in u_d} w_{id} x_{id} x_{id}^T)^{-1} x_{id}$,

$$M\hat{S}E(\hat{Y}_d^{GREG}) = \sum_{i \in u_d} \sum_{j \in u_d} \frac{\pi_{ijd} - \pi_{id}\pi_{jd}}{\pi_{ijd}\pi_{id}\pi_{jd}} g_{id} r_{id} g_{jd} r_{jd}. \qquad (6)$$

SYNTHETIC ESTIMATORS

MODEL A

Standard two level linear model:

$$y_{id} = x_{id}^T \boldsymbol{\beta} + u_d + e_{id}, \qquad (7)$$

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

$$\hat{Y}_d^{SYNTH} = \bar{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}}, \qquad (8)$$

z $\bar{X}_{.d} = (\bar{X}_{.d,1}, ..., \bar{X}_{.d,p})^T$

Estimator does not respect sampling weights

$$M\hat{S}E(\hat{Y}_d^{SYNTH}) = \hat{\sigma}_u^2 + \bar{\mathbf{X}}_{.d} \hat{\mathbf{V}} \bar{\mathbf{X}}_{.d}^T, \qquad (9)$$

where $\hat{\mathbf{V}}$ is the covariance matrix of covariates.

MODEL B

Model for domain is as follows:

$$\bar{y}_{.d} = \bar{x}_{.d}^T \boldsymbol{\beta} + \xi_d, \qquad (10)$$

where $\xi_d \sim iid\ \ N(0, \sigma_u^2 \ + \frac{\sigma_e^2}{n_d})$ and n_d denotes sample size for area d.

Variance σ_e^2 is estimated according to the formula:

$$\hat{\sigma}_e^2 = \frac{1}{n - na} \sum_i \sum_d (y_{id} - \bar{y}_{.d})^2, \qquad (11)$$

where: $n-$ sample size;

na – number of domains in the sample.

One level regression model with β i σ_u^2 estimated itreratively from:

$$\hat{\boldsymbol{\beta}} = (x^{\mathbf{T}} \mathbf{D}^{-1} x)^{-1} x^{\mathbf{T}} \mathbf{D}^{-1} y, \qquad (12)$$

where $y-$ vector of the sample elements y,

x – matrix with rows consisting of x_d^T,

D – diagonal matrix with iteratively updated values $(\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_d})$ on the diagonal.

$$\hat{Y}_d^{SYNTH} = \bar{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}}, \qquad (13)$$

$$M\hat{S}E(\hat{\hat{Y}}_d^{SYNTH}) = \hat{\sigma}_u^2 + \bar{\mathbf{X}}_{.d}^T\hat{\mathbf{V}}\bar{\mathbf{X}}_{.d}, \tag{14}$$

where $\hat{\mathbf{V}}$ is the estimate of the covariance matrix $(x^{\mathbf{T}}\mathbf{D}^{-\mathbf{1}}x)^{-\mathbf{1}}$.
EBLUP ESTIMATORS

MODEL A

$$\hat{\hat{Y}}_d^{EBLUP} = w_d^{EBLUP}\hat{\hat{Y}}_d^{GREG} + (1 - w_d^{EBLUP})\hat{\hat{Y}}_d^{SYNTH}, \tag{15}$$

$$w_d^{EBLUP} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_d}. \tag{16}$$

In more details the models may be presented as follows:

EBLUP using MODEL A

$$\hat{\hat{Y}}_d = \gamma_d(\bar{y}_{.d} - \bar{\mathbf{x}}_{.\mathbf{d}}^{\mathbf{T}}\hat{\beta}) + \bar{\mathbf{X}}_{.d}^T\hat{\boldsymbol{\beta}}, \tag{17}$$

where:

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_d}, \tag{18}$$

$\bar{y}_{.d}$, $\bar{\mathbf{x}}_{.\mathbf{d}}^{\mathbf{T}}$, are corresponding sample means of y and the covariates for domain d.
$\hat{\boldsymbol{\beta}}, \hat{\sigma}_e^2, \hat{\sigma}_u^2$are parameters estimated using standard two-level linear model.

MSE Estimators.
ESTIMATOR 1

$$M\hat{S}E(\hat{\hat{Y}}_d) = \frac{\gamma_d\hat{\sigma}_e^2}{n_d} + (1 - \gamma_d)^2\bar{\mathbf{X}}_{.d}^T\hat{\mathbf{V}}\bar{\mathbf{X}}_{.d} \tag{19}$$

ESTIMATOR 2

$$\begin{aligned} M\hat{S}E(\hat{\hat{Y}}_d) = {} & \frac{\gamma_d\hat{\sigma}_e^2}{n_d} + (1 - \gamma_d)^2\left(\bar{\mathbf{X}}_{.d}^T\hat{\mathbf{V}}\bar{\mathbf{X}}_{.d}\right) + \\ & 2 \times \left(\frac{\hat{\sigma}_e^2}{n_d}\right)^2\left(\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_d\right)^{-3}\left(V\hat{a}r(\hat{\sigma}_u^2) + \frac{\hat{\sigma}_u^4}{\hat{\sigma}_e^4}V\hat{a}r(\hat{\sigma}_e^2) - 2\frac{\hat{\sigma}_u^2}{\hat{\sigma}_e^2}C\hat{o}v(\hat{\sigma}_u^2, \sigma_e^2)\right), \end{aligned} \tag{20}$$

where:

$$\begin{aligned} V\hat{a}r(\hat{\sigma}_e^2) &= \frac{2\hat{\sigma}_e^4}{m_1 - m_2 - p}, \\ C\hat{o}v(\hat{\sigma}_e^2, \hat{\sigma}_u^2) &= \frac{2\hat{\sigma}_e^2\hat{\sigma}_u^2}{m_1 - m_2 - p}, \\ V\hat{a}r(\hat{\sigma}_u^2) &= V\hat{a}r(\hat{\rho})V\hat{a}r(\hat{\sigma}_e^2) + \hat{\sigma}_e^4V\hat{a}r(\hat{\rho}) + \hat{\rho}^2V\hat{a}r(\hat{\sigma}_e^2), \end{aligned}$$

m_1– number of selected units,
m_2– number of domains,
$\rho = \sigma_u^2/\sigma_e^2$ is the variance ratio,

$$V\hat{a}r(\hat{\rho}) = \frac{2}{\sum_d n_d^2 / (1 + n_d\hat{\rho})^2},$$

where $\hat{\mathbf{V}}$ is the covariance matrix of covariates.

$V\hat{a}r(\hat{\sigma}_e^2)$ is estimated variance $\hat{\sigma}_e^2$ and $V\hat{a}r(\hat{\sigma}_u^2)$ is estimated variance $\hat{\sigma}_u^2$. Estimator 1 is a corresponding approximation if the number of domains is large. Estimator 2 may be applied in any case.

MODEL B

$$\hat{\hat{Y}}_d^{EBLUP} = w_d^{EBLUP}\hat{\hat{Y}}_d^{DIRECT} + (1 - w_d^{EBLUP})\hat{\hat{Y}}_d^{SYNTH}, \tag{21}$$

$$w_d^{EBLUP} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\psi}_d}. \tag{22}$$

EBLUP using MODEL B.

$$\hat{\hat{Y}}_d^{EBLUP} = \gamma_d\hat{\hat{Y}}_d^{direct} + (1 - \gamma_d)\bar{\mathbf{X}}_{.d}^T\hat{\boldsymbol{\beta}}, \tag{23}$$

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}, \tag{24}$$

where

$$\hat{\beta} = (x^{\mathbf{T}}\mathbf{D}^{-\mathbf{1}}x)^{-\mathbf{1}}x^{\mathbf{T}}\mathbf{D}^{-\mathbf{1}}y, \tag{25}$$

where:
y– vector of the sample elements y,
x – matrix with rows consisting of $\bar{x}_{.d}^T$,
D – diagonal matrix with iterative updated values $(\hat{\sigma}_u^2 + \hat{\sigma}_e^2)$ on the diagonal.

MSE Estimators

ESTIMATOR 1

$$M\hat{S}E(\hat{\hat{Y}}_d) = \gamma_d\hat{\psi}_d + (1 - \gamma_d)^2\left(\bar{\mathbf{X}}_{.d}^T\hat{\mathbf{V}}\bar{\mathbf{X}}_{.d}\right), \tag{26}$$

$\hat{\psi}_d$ is an estimator of residual variance inside domains $\hat{\psi}_d = \dfrac{\sigma_e^2}{n_d}$.

ESTIMATOR 2

$$M\hat{S}E(\hat{\hat{Y}}_d) = \gamma_d\hat{\psi}_d + (1 - \gamma_d)^2\left(\bar{\mathbf{X}}_{.d}^T\hat{\mathbf{V}}\bar{\mathbf{X}}_{.d}\right) + \hat{\psi}_d^2\left(\hat{\sigma}_u^2 + \hat{\psi}_d\right)^{-3}V\hat{a}r(\hat{\sigma}_u^2) \tag{27}$$

ESTIMATOR 3

$$M\hat{S}E(\hat{\hat{Y}}_d) = \gamma_d\hat{\psi}_d + (1-\gamma_d)^2\left(\bar{\mathbf{X}}^T_{.d}\hat{\mathbf{V}}\bar{\mathbf{X}}_{.d}\right) + 2\times\hat{\psi}^2_d\left(\hat{\sigma}^2_u + \hat{\psi}_d\right)^{-3} V\hat{a}r(\hat{\sigma}^2_u), \tag{28}$$

where $\hat{\mathbf{V}}$ is the covariance matrix of covariates and $V\hat{a}r(\hat{\sigma}^2_u)$ is an estimate of $\hat{\sigma}^2_u$.

EBLUP using spatial correlation (EBLUPGREG_SPATIAL)

Software prepared by ISTAT was based on re-formulation of the expressions contained in Saei and Chambers (2003) in order to obtain a more efficient SAS code.

The considered model is the general linear mixed model

$$y = X\boldsymbol{\beta} + Zu + e, \tag{29}$$

where:

X and **Z** are known matrices of order N×P and N×DOM respectively;

X is the matrix of the population values of the covariates and **Z** is the incidence matrix for the spatial random area effect;

e and **u** are vectors of random variables with mean and variance and covariance matrices expressed respectively by the couples:

$$N \sim [0,\ \sigma^2 I_N],$$

$$N \sim [0,\ \sigma^2_U A],$$

I_N being the identity matrix of order N and A square matrix of order DOM allowing a spatial correlation structure to be included in the model. The generic element of A $\mathrm{a}_{dd'}$ is given by

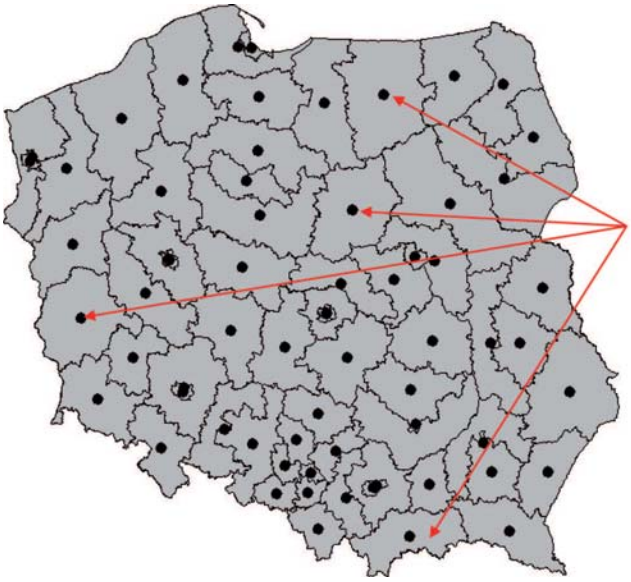
$$\mathbf{a}_{(dd')} = \left[1 + \delta_{(dd')}\exp\left(\frac{\boldsymbol{dist(d,d')}}{\alpha}\right)\right]^{-1}, \tag{30}$$

$$\delta_{(dd')} = \begin{cases} 0\ \boldsymbol{for} & \boldsymbol{d = d'}, \\ 1\ \boldsymbol{for} & \boldsymbol{d \neq d'}, \end{cases}$$

$\boldsymbol{dist(d,d')}$ is the Euclidean distance between centroids of area d and d'.

6. SOFTWARE

The software used in the study was SAS. We started from using PROC SURVEYMEANS to compute the *direct* estimator. Then the EURAREA code was implemented. Not only were Seven Standard Estimators computed but the EBLUPGREG software with its options was tested to find out whether to allow for spatial autocorrelation or not. The SAS software was also used for computing coordinates of the centroids which are of special importance in estimation conducted via EBLUPGREG taking into account the spatial dependence.



The coordinates of centroids (marked by dots) will be inputted in EBLUPGREG code, allowing for spatial correlation structure to be included in the model. All centroids are marked by black dots however, only four of them (just as examples) are pointed by red arrows.

Figure 1. Centroids of small domains – NUTS4

7. RESULTS

Figures 2 and 3 show that the spatial pattern of the direct estimates for NUTS3 level fits quite well to the pattern of the values coming from the administrative registers.

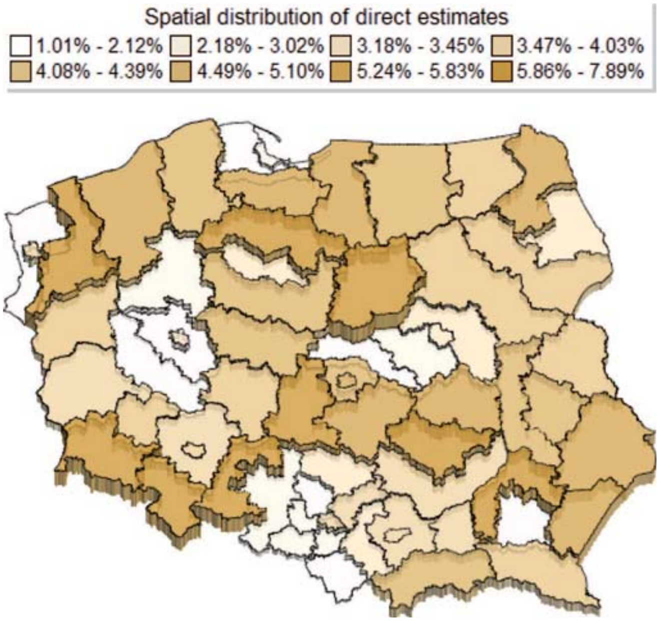


Figure 2. Direct estimates

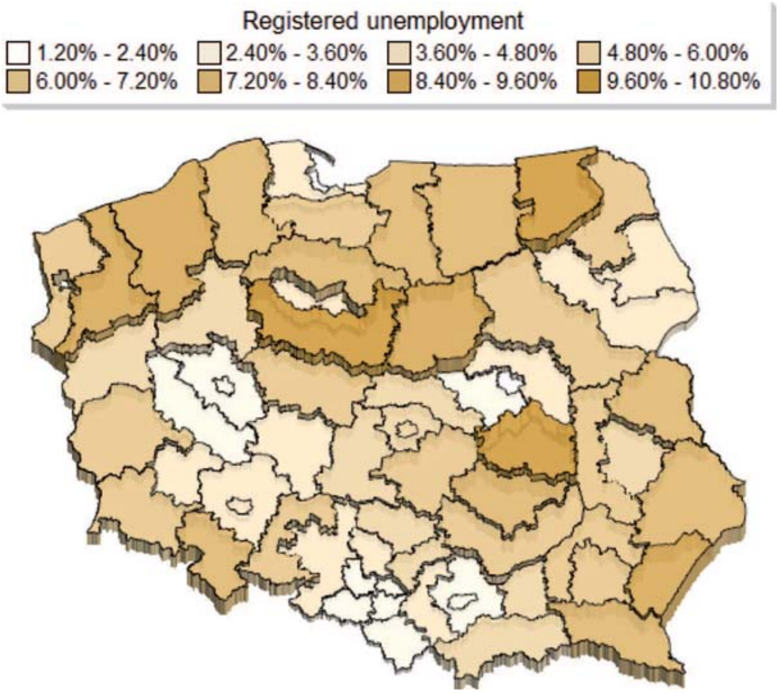


Figure 3. Registered unemployment

The following figures (Fig. 4 – Fig.10) give some overview on how the applied estimators reproduce the registered unemployment. The Figure 4 showed that direct estimates of ILO (International Labor Bureau) unemployment are in most of the cases lower for the NUTS3 areas than registered unemployment. Quite different situation is presented on the Figure 5. In case of model assisted estimator (GREG) the estimated unemployment from LFS is higher than registered. However if the ranges of unemployment are compared one could see that direct estimates have slightly more narrow range when compared with GREG. The registered unemployment ranges from 1.41% to 10.77%, direct estimates range from 1.01% to 7.89% and finally GREG from 2.92% to 10.89%.

Synth_A estimator has the smallest variance but the bias in this case is unacceptable. Synth_B estimates are quite similar to direct ones. However the range for the estimates obtained while applying Synth_B is very short – from 2.35% to 5.94%. So one could conclude that smoothing of the estimates is too strong.

The comparison of the EBLUP estimators based on different assumptions revealed that they should be of a special interest in applications of small area methodology in Labor Force Survey. The results obtained suggest their bias is relatively small with relatively small variation. For instance the range of estimates produced by EBLUP-GREG_SPATIAL (which takes into consideration the spatial structure of data) is from 4.50% to 10.34%. One should notice that even some basic assumptions relating to the

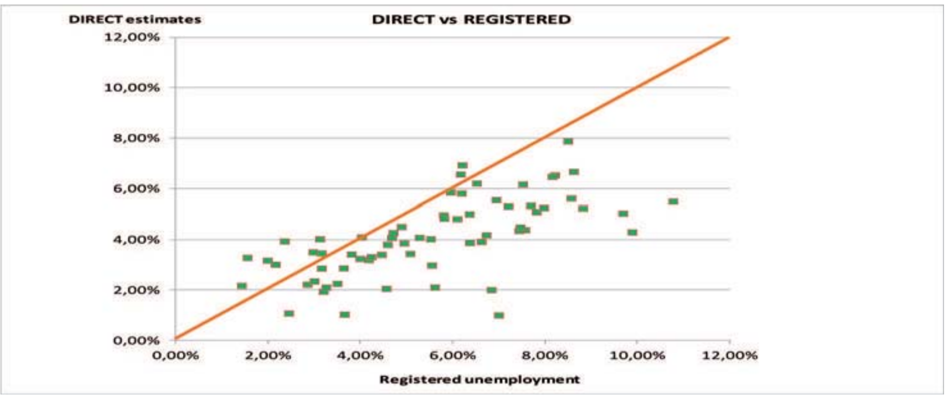


Figure 4. Comparison of registered unemployment rate and direct estimates

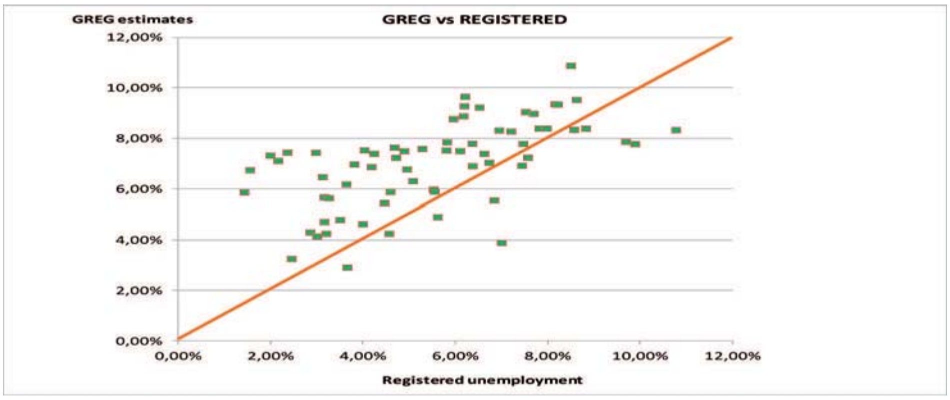


Figure 5. Comparison of registered unemployment rate and greg estimates

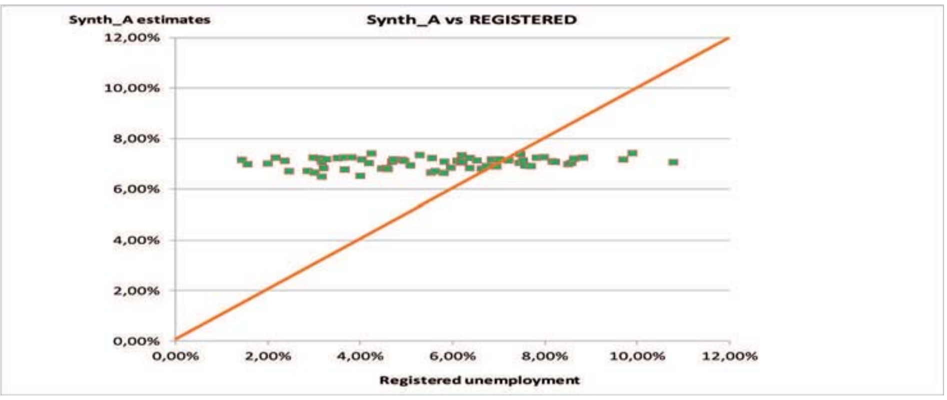


Figure 6. Comparison of registered unemployment rate and Synth_A estimates

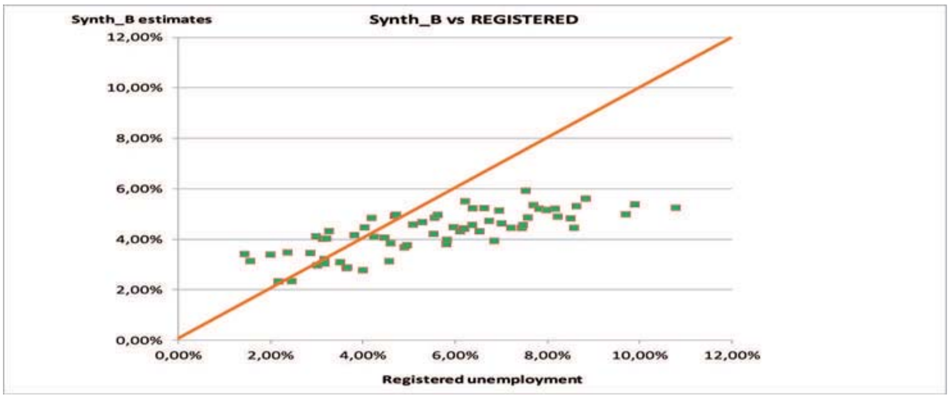


Figure 7. Comparison of registered unemployment rate and Synth_B estimates

nature of variable are violated (number of unemployed people aging 15 and more is not a continuous variable and the distribution could not be normal), the behavior of the EBLUPs is quite well.

The estimates obtained by application of EBLUP_A and EBLUP_B models show the same patters as their direct components. EBLUP_A is a linear combination of GREG and SYNTH_A while EBLUP_B is a linear combination of DIRECT and SYNTH_B.

We also compared the MSE of the studied estimators. The software produced in EURAREA project computes the MSE of seven standard estimators and also for spatial version of EBLUPGREG estimator. However the approach presented by us has two simplifications. In fact for DIRECT and GREG estimators we have mainly variance as these two estimators are design unbiased. The second problem while applying EURAREA code the reader should realize is the fact, that the quality assessment measures are computed assuming simple random sampling. In fact the sampling design applied in LFS surveys is rather stratified and two-stage in most cases.

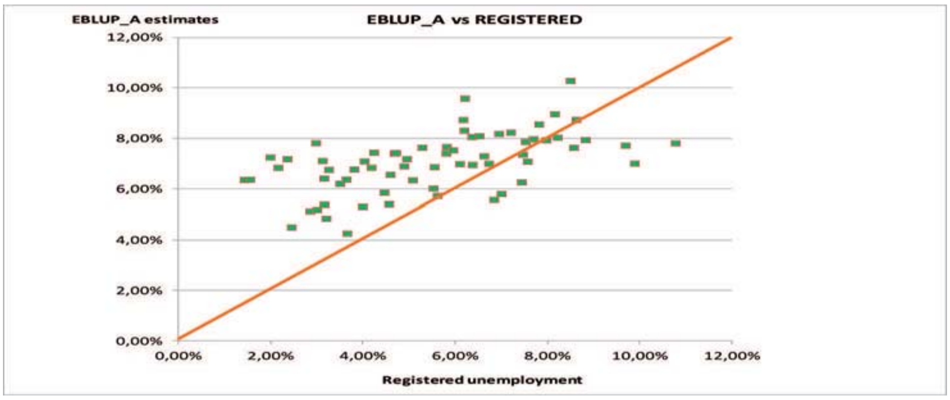


Figure 8. Comparison of registered unemployment rate and EBLUP_A estimates

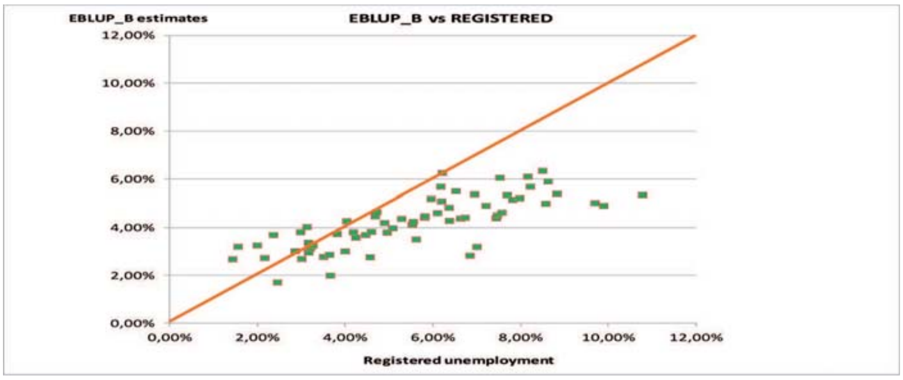


Figure 9. Comparison of registered unemployment rate and EBLUP_B estimates

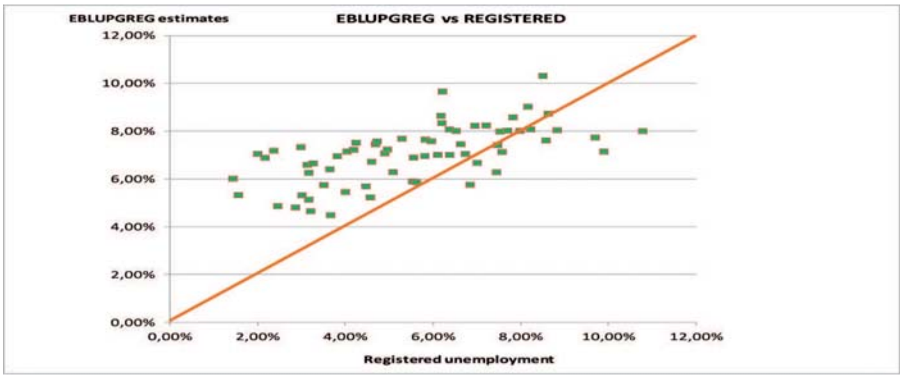


Figure 10. Comparison of registered unemployment rate and EBLUPGREG_SPATIAL estimates

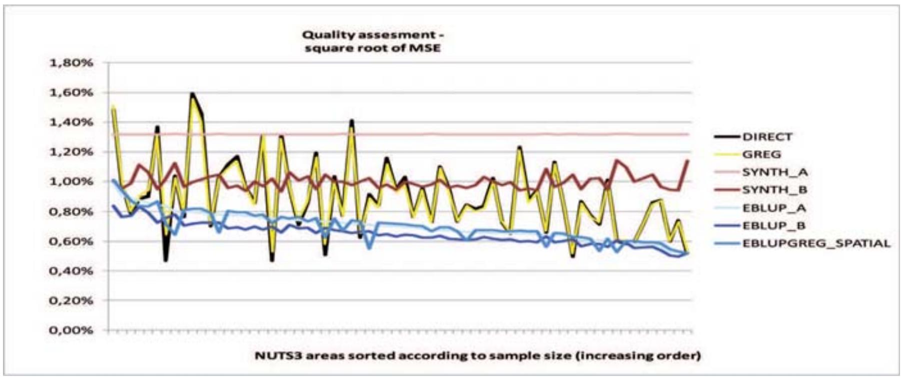


Figure 11. Distribution of MSEs (NUTS3 ordered according to increasing sample size)

The last Figure (11) presents the distribution of MSEs where the NUTS3 were ordered according to the increasing sample size. The highest values of MSEs are connected with SYTH_A estimator. In this case the important input in its value is the input of bias. The design unbiased estimators – DIRECT and GREG show a quite high amount of variance which is slightly smaller in the case of GREG. However when sample size increases the variance of the DIRECT estimator is decreasing. In the case of SYNTH the variance is rather constant. The best performance as far as the behavior of the estimation is concerned is connected with EBLUPs. They are quite similar and have the smallest MSE.

Poznan University of Economy

LITERATURA

[1] Bracha, C., Lednicki, B. and Wieczorkowski, R., (2003), *Estimation of Data from the Polish Labour Force Surveys by poviats (counties) in 1995—2002* (in Polish), Central Statistical Office of Poland, Warsaw.
http://www.stat.gov.pl/cps/rde/xbcr/gus/PUBL_estymacja_danych_z_bad_na_poziomie_pow_dla_lat_1995_2002.pdf

[2] Chandra H., Salvati N., Chambers R., (2009), *Small Area estimation for Spatially Correlated Populations – A Comparison of Direct and Indirect Model-Based Methods,* Southampton Statistical Sciences Research Institute, Methodology Working Paper M07/09, University of Southampton.

[3] 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.

[4] *EURAREA_Project_Reference_Volume,* http://www.statistics.gov.uk/eurarea.

[5] *EURAREA EBLUPGREG Software Documentation*, Statistics Finland EURAREA Consortium, Deliverables D2.3.2, D3.3.2, 2004.

[6] Gołata E., (2004), *Estymacja pośrednia bezrobocia na lokalnym rynku pracy*, Wydawnictwo AE w Poznaniu, Poznań.

[7] Kruszka K., (ed.), (2010), *Commuting in Poland*, Statistical Office in Poznań, Poznań.

[8] Kubacki J., (2004), *Application of the Hierarchical Bayes Estimation to the Polish Labour Force Survey*, Statistics in Transition, 6 (5), 785-796, Warsaw.

[9] Saei A., Chambers R., (2003), *Small Area Estimation: A Review of Methods Based on the Application of Mixed Models*, University of Southampton.

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

WYKORZYSTANIE ESTYMACJI POŚREDNIEJ UWZGLĘDNIAJĄCEJ KORELACJĘ PRZESTRZENNĄ W BADANIACH SPOŁECZNYCH W POLSCE

Streszczenie

Artykuł przedstawia propozycję wykorzystania metod estymacji pośredniej (w tym także tej metody, która uwzględnia korelację przestrzenną) do oszacowania pewnych charakterystyk rynku pracy w populacji osób w wieku 15 lat i więcej w przekroju podregionów w Polsce w 2008 roku. Jest to bardziej szczegółowy poziom agregacji przestrzennej niż ten prezentowany w publikacjach Głównego Urzędu Statystycznego opartych na wynikach Badania Aktywności Ekonomicznej Ludności. Drugim celem jest porównanie miar precyzji estymatora bezpośredniego z precyzją estymatora typu EBLUP (empirical best linear unbiased predictor) oraz estymatora typu EBLUPGREG_SPATIAL (uwzględniającego korelację przestrzenną).

Słowa kluczowe: statystyka małych obszarów, autokorelacja przestrzenna, bezrobocie, Badanie Aktywności Ekonomicznej Ludności (BAEL)

USING INDIRECT ESTIMATION WITH SPATIAL AUTOCORRELATION IN SOCIAL SURVEYS IN POLAND

Abstract

The article presents possible application of indirect estimation methods (including the method accounting for spatial correlation) to estimate some characteristics of labor market in the population of people aged 15 and over at the level of NUTS3 in Poland in 2008. This is a more detailed spatial aggregation of data compared with that found in publications of the Central Statistical Office based on Labour Force Survey results. The second aim of the article is to compare the precision measures of the direct estimator with those of the EBLUP estimator (*empirical best linear unbiased predictor*) and the EBLUPGREG_SPATIAL estimator (which takes into account spatial correlation).

Key words: small area statistics, spatial autocorrelation, unemployment, Labour Force Survey (LFS)