METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
X, 2009, str. 121 - 130

APPLICATION OF EBLUP ESTIMATION TO THE ANALYSIS OF SMALL AREAS ON THE BASIS OF POLISH HOUSEHOLD BUDGET SURVEY

Alina Jędrzejczak[1,2], Jan Kubacki[2]
[1]Katedra Metod Statystycznych UŁ, [2]USLORB
e-mail: j.kubacki@stat.gov.pl

Abstract: In the paper the results of small area estimation using empirical best linear unbiased predictor (EBLUP) for the data coming from Polish Household Budget Survey are presented. The results were obtained using small area models of household expenditures for regions. Estimation of sampling errors was conducted by means of the balanced repeated replication (BRR) technique. The estimation of EBLUPs and their corresponding mean square errors (MSE) was carried out using variance components technique. To calculate MSE of EBLUP the maximum likelihood method (ML) and restricted maximum likelihood method (REML) were used. The computation was made using SAE package designed for R-project.

Key words: small area estimation, empirical best linear unbiased predictor (EBLUP), household budget survey, variance estimation

INTRODUCTION

The main objective of many modern sample surveys is to provide estimates of totals, means and other parameters not only for the population but also for subpopulations (or domains) such as geographic areas and socio-economic groups. Direct estimates of domain parameters are based only on domain specific data. It is seldom possible to have overall sample size large enough to support reliable direct estimates for all the domains of interest. Therefore, it is often necessary to use indirect estimates that „borrow strength" by using values of the variables from related areas and thus increase the „effective" sample size.

The paper is focused on the following two areas:

- To present shortly the theoretical background of the EBLUP estimation for the basic area level models.

- To apply the EBLUP theory in order to improve the results of the estimation in the case of the household expenditures in Poland.

EBLUP ESTIMATION

Small area means or totals can be expressed as linear combination of fixed and random effects. Best linear unbiased prediction (BLUP) estimators of such parameters can be obtained in a classical way using BLUP estimation procedure. BLUP estimators minimize Mean Square Error (MSE) within the class of linear unbiased estimators and do not depend on the normality of random effects. Maximum likelihood (ML) or restricted maximum likelihood (REML) methods can be used to estimate the variance and covariance components, assuming normality.

The EBLUP procedure was applied in many important statistical surveys conducted all over the world. The pioneer work in this area was that of Fay and Herriot [Fay and Herriot 1979], where EBLUP technique was used for evaluating per capita income and some other statistics obtained for counties. Application of EBLUP estimators for the Survey on Life Conditions in Tuscany by Pratesi and Salvati [Pratesi and Salvati 2008] introduced the spatial EBLUP. EBLUP estimators were also discussed in detail in the EURAREA project [EURAREA consortium, 2004], where one of the discussed variables was household equivalent income. A report entitled "Social Exclusion and Integration in Poland: An Indicators-based Approach" [UNDP, 2006] prepared for UNDP (United Nations Development Programme) presented the benefits obtained from the EBLUP technique applied for average equivalent income.

The examples described below are based on a special kind of the general linear mixed model which is widely known as basic area level model [Rao, 2003]

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v} + \mathbf{e} \tag{1}$$

where: $\mathbf{y}$ - is n×1 vector of sample observations, $\mathbf{X}$ - known matrix of explanatory variables, $\boldsymbol{\beta}$ - is vector of linear regression coefficients, $\mathbf{v}$ - area-specific random effect vector, $\mathbf{e}$ - sampling error vector.

It is usually assumed that $\mathbf{v}$ and $\mathbf{e}$ are independently distributed with mean 0 and covariance matrices $\mathbf{G}$ and $\mathbf{R}$ respectively.

EBLUP estimator for the small area model given by (1) has the following form :

$$\boldsymbol{\theta}_{EBLUP} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}\mathbf{G}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{2}$$

where:

$$\boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}\,. \tag{3}$$

M is the identity matrix, and **G** is the matrix with non-zero diagonal and its values are equal to σ_v^2 which is the model variance. It is usually computed using special iterative procedure that applies the Fisher algorithm.

Fisher "scoring" algorithm is a form of Newton- Raphson method used to solve maximum likelihood equations numerically. In the case of basic area level models, the variance component σ_v^2 can be obtained by means of the following iterative formula:

$$\sigma_v^{2(a+1)} = \sigma_v^{2(a)} + [I(\sigma_v^{2(a)})]^{-1} s(\tilde{\boldsymbol{\beta}}^{(a)}, \sigma_v^{2(a)}) \tag{4}$$

where: $I(\sigma_v^2) = \frac{1}{2}\sum_{i=1}^{m} \frac{1}{(\sigma_v^2 + \psi_i)^2}$ denotes the Fisher information,

$s(\tilde{\boldsymbol{\beta}}, \sigma_v^2) = \frac{1}{2}\sum_{i=1}^{m} \frac{1}{\sigma_v^2 + \psi_i} + \frac{1}{2}\sum_{i=1}^{m} \frac{(\hat{\theta}_i - \mathbf{z}_i^T \tilde{\boldsymbol{\beta}})^2}{(\sigma_v^2 + \psi_i)^2}$ is a score function, ψ_i - sampling variance for the region *"i"*.

PRECISION OF THE EBLUP ESTIMATION

Mean square error estimate (MSE) of EBLUP can be obtained from the following formula:

$$MSE(\theta_{EBLUP}) = g_1(\hat{\delta}) - b_\delta^T(\hat{\delta})\nabla g_1(\hat{\delta}) + g_2(\hat{\delta}) + 2g_3(\hat{\delta}) \tag{5}$$

where δ is a variance dependent parameter. Using this formula we usually assume that the mean square error of EBLUP is the sum of three main elements g_1, g_2 and g_3 which are described by the following equations:

$$g_1(\hat{\delta}) = diag(\mathbf{G} - \mathbf{G}\mathbf{V}^{-1}\mathbf{G}) \tag{6}$$

$$g_{2i}(\hat{\delta}) = (\mathbf{X_i} - \mathbf{m}_i^T\mathbf{G}\mathbf{V}^{-1}\mathbf{X})(\mathbf{X}\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X_i} - \mathbf{m}_i^T\mathbf{G}\mathbf{V}^{-1}\mathbf{X}) \tag{7}$$

$$g_{3i}(\hat{\delta}) = (\mathbf{m}_i^T(\mathbf{V}^{-1} - \mathbf{G}(\mathbf{V}^{-1}\mathbf{V}^{-1}))\mathbf{V}(\mathbf{m}_i^T(\mathbf{V}^{-1} - \mathbf{G}(\mathbf{V}^{-1}\mathbf{V}^{-1}))^{\mathbf{T}})\mathbf{I} \tag{8}$$

where $\mathbf{m_i}$ is a vector with zeroes for all elements with exception for the element having an index *i* while **I** is the inversed Fisher information matrix.

The additional bias factor $b_\delta^T(\hat{\delta})\nabla g_1(\hat{\delta})$ emerges only in case of the maximum likelihood (ML) estimation and can be calculated using some special formulas (see: Rao, 2003, page 129). This drawback of the ML method, connected with not taking into account the loss in the degrees of freedom coming from the estimation of regression parameters $\boldsymbol{\beta}$ can be omitted using REML.

Restricted (or residual) maximum likelihood (REML) is the method for fitting linear mixed models that produces unbiased estimates for variance and covariance parameters. In the case of REML the computation algorithm is almost identical except that in the procedure that evaluates σ_v^2 a special **P** matrix instead of $\mathbf{V^{-1}}$ matrix is used as follows:

$$\mathbf{P} = \mathbf{V}^{-1} - (\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}) \tag{9}$$

The MSE for REML variant is computed using a simpler formula. Here the bias factor is neglected:

$$MSE(\theta_{EBLUP-REML}) = g_1(\hat{\delta}) + g_2(\hat{\delta}) + 2g_3(\hat{\delta}) \tag{10}$$

APPLICATION ON THE BASIS OF THE POLISH HOUSEHOLD BUDGET SURVEY

Sampling frame for Polish HBS is a list of census enumeration areas (CEAs) prepared by means of the census data. It is updated annually according to the increase of dwellings due to the completion of new buildings, and the decrease of dwellings due to the demolition of buildings and changes in administration division of the country. In some cases the Census Enumeration Areas are linked together, to achieve the minimum primary sampling units' size, that is set as 250 dwellings for urban areas and 150 for rural areas. The two-stage sampling plan is used and primary sampling units (PSU) are selected with probabilities proportional to their size using Hartley-Rao method. In each PSU 24 dwellings are selected (with 2 dwellings for each survey month and this dwellings are interviewed also in the same month next year). Additionally, in each PSU 150 dwellings are selected as a reserve sample, which is used in the case of non-response. The stratification is done by 16 voivodships, and in each voivodship, according to the class of size of localities and urban-rural criteria. Large cities constitute separate data. The number of strata in each voivodship ranges from 3 to 12. Altogether there are 96 strata.

In order to reduce the effect of unequal selection probabilities for Primary Sampling Units (PSU) the survey results are specially weighted. The starting point of each dwelling weight is the inverse proportion to its inclusion probability. The non-response coefficients in Polish HBS are relatively high and considerably affect the socioeconomic structure of households in the sample. To reduce the non-response bias the special weights from internal, as well as from external sources (e.g. Labour Force Survey), are applied [Kordos et al., 2002] . Balanced half-samples technique (BRR) is used to determine the standard error.

In the paper the results for some models using data from Polish HBS are presented. These models were constructed using different sources of data that come from Polish Public Statistics and also from administrative data. Firstly we present the region (voivodship) model for per capita expenditures. The exploratory variable

is per capita value of GDP for regions. The ordinary linear regression model $y = a_o + a_1 x$ can be summarized as follows:

$$\hat{y} = 411{,}70 + 2{,}35x$$

$$t_\alpha \quad (12{,}01) \quad (6{,}46)$$

The value of determination coefficient (R^2) is 0,748, the value of corrected R^2 is equal to 0,730, the F statistics is 41,749 and standard estimation error is equal to 28,994.

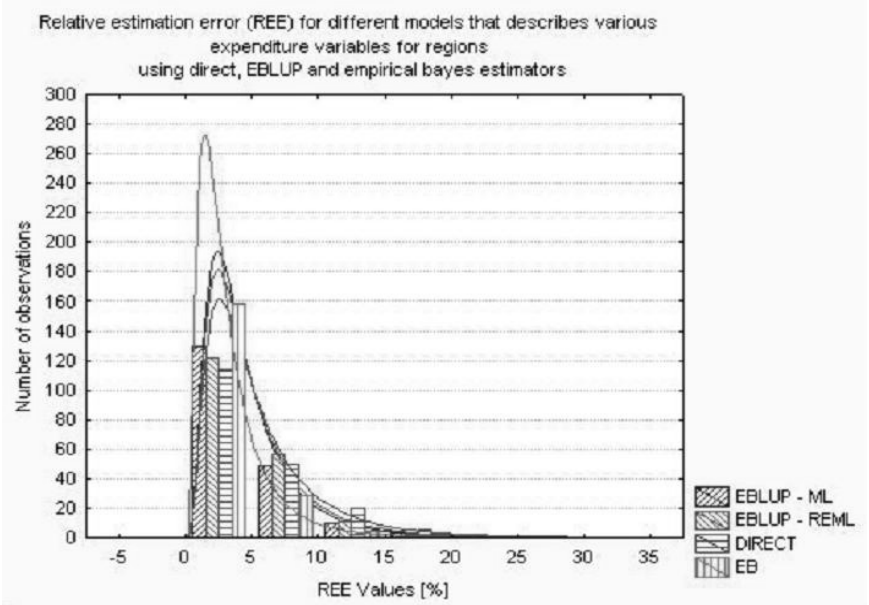
As it is showed in table 1, EBLUP estimator gives systematically lower MSE values and they are slightly greater for the REML variant. The reduction of relative estimation error (REE) for the EBLUP estimates is presented in figure 2. The figure presents the results for various models with different characteristics (including models for expenditures of consumer goods and services- food and non-alcoholic beverages, clothing and footwear, housing water, electricity, household equipment and routine maintenance of the house, health, transport, communication, recreation and culture etc).

Table 1. Per capita expenditures for regions- direct and EBLUP estimates

Region	Direct estimates	Variance for direct estimates	EBLUP - ML estimates	MSE EBLUP - ML	EBLUP-REML - estimates	MSE EBLUP-REML
Dolnośląskie	647,79	518,93	650,41	299,38	650,14	328,68
Kujawsko-Pomor.	590,69	401,36	605,40	258,07	603,75	279,49
Lubelskie	593,77	107,58	590,99	98,19	591,41	100,17
Lubuskie	630,53	724,69	621,54	345,22	622,25	386,97
Łódzkie	654,95	164,02	648,37	137,85	649,32	142,92
Małopolskie	615,67	79,74	615,46	73,92	615,48	75,14
Mazowieckie	792,14	203,72	787,08	200,98	787,81	203,33
Opolskie	655,04	314,92	632,88	224,96	635,53	239,90
Podkarpackie	566,18	170,67	569,54	145,62	569,03	150,60
Podlaskie	592,42	44,06	592,29	42,56	592,31	42,88
Pomorskie	661,88	86,81	659,07	79,86	659,52	81,31
Śląskie	655,84	180,23	659,15	150,91	658,68	156,59
Świętokrzyskie	569,74	598,05	584,30	323,31	582,91	357,60
Warmińsko-Mazur.	570,39	295,84	580,77	216,06	579,45	229,61
Wielkopolskie	597,73	164,07	613,17	139,03	610,92	143,93
Zachodniopomorskie	654,69	415,63	645,63	263,35	646,59	285,77

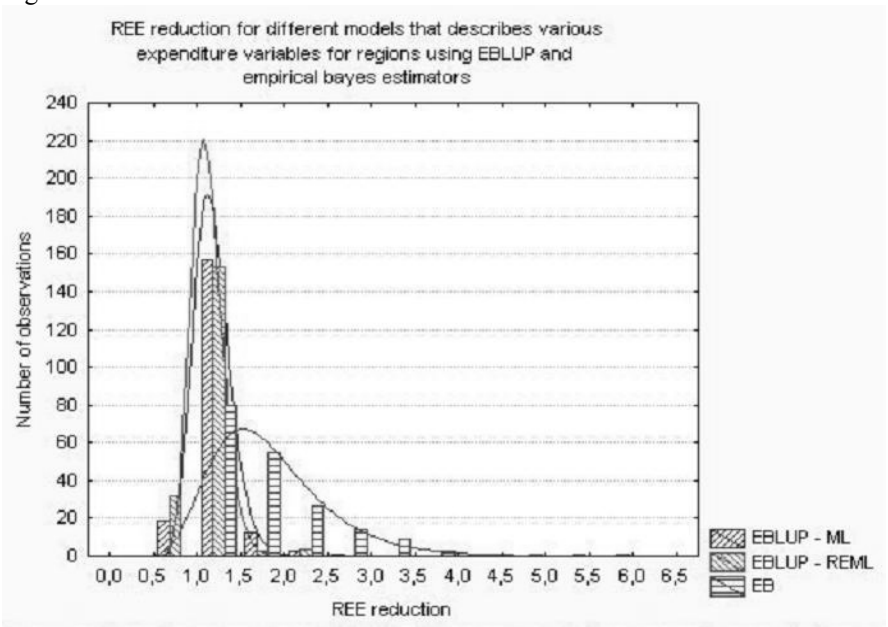
Source:own calculations using HBS data, WesVar software and SAE package for R-project

Figure 1. Relative estimation error values for direct, EBLUP and EB estimators



Source: own calculations using HBS data, WesVar and Statistica software and SAE package for R

Figure 2. Relative estimation error reduction for EBLUP and EB estimators



Source: own calculations using HBS data, WesVar and Statistica software and SAE package for R

In order to assess the variability of variance elements g_1, g_2 and g_3 new measures of such variability are introduced, that allow to evaluate the share of each element in the overall MSE value. For g_1 element it can be expressed as follows:

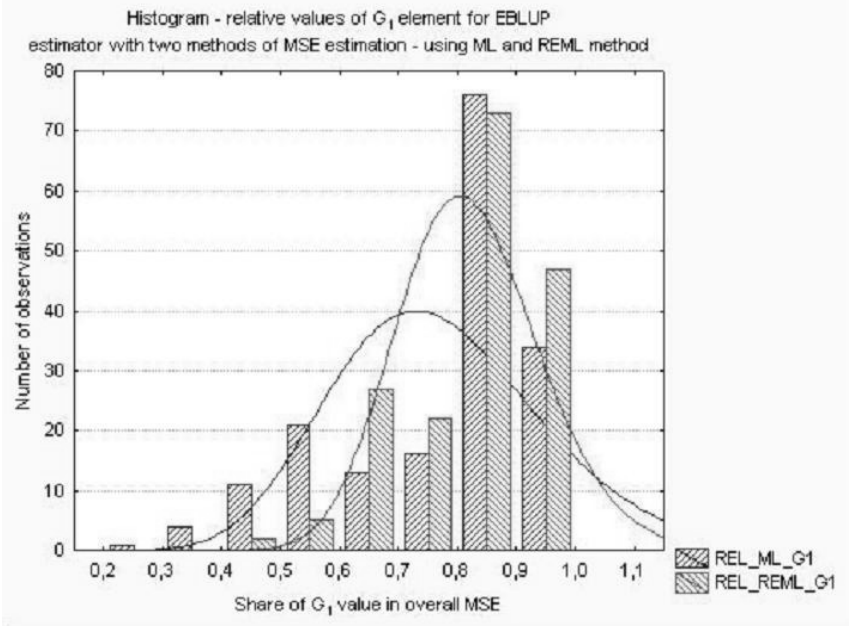
$$\mathrm{Re}\,\mathrm{lg}_1 = g_1 / MSE(\theta_{EBLUP}) \tag{11}$$

Analogical measures can be obtained for g_2 and g_3

According to the results presented in this paper, using EBLUP technique improves the quality of the estimation. Observed reduction of MSE proves this. However, because of the method of MSE estimation, this results can be rather treated as good approximation.

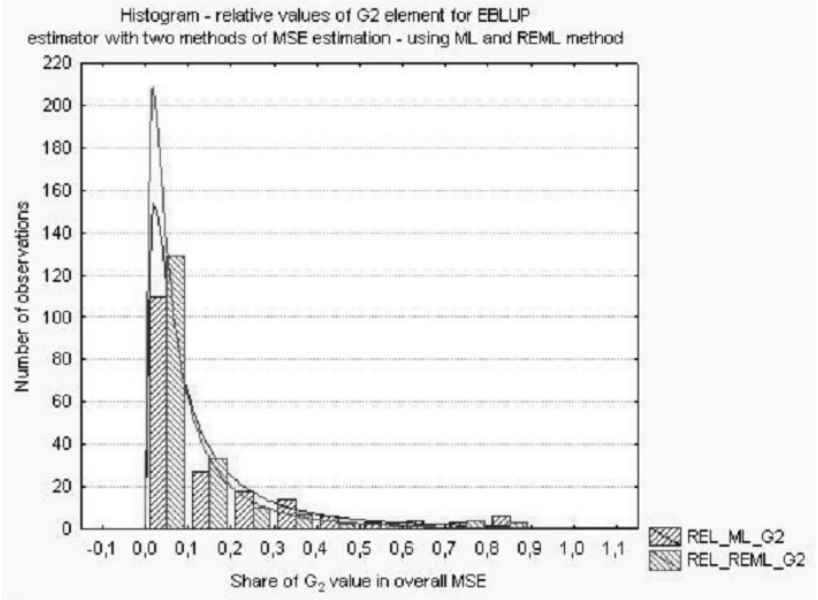
Values of MSE obtained using EBLUP technique are more adequate than the values obtained using the naive EB procedure. This result is consistent with the theory. More details concerning the procedure for empirical Bayes (EB) estimation can be found in [Bracha et al, 2004]. It can be supposed that in the future some kind of compromise will be reached to become certain about the reliability of the methods discussed here. The comparison of different implementation methods of estimation may also be useful in this field.

Figure 3. Histogram of relative G1 element values for two variants of EBLUP estimation



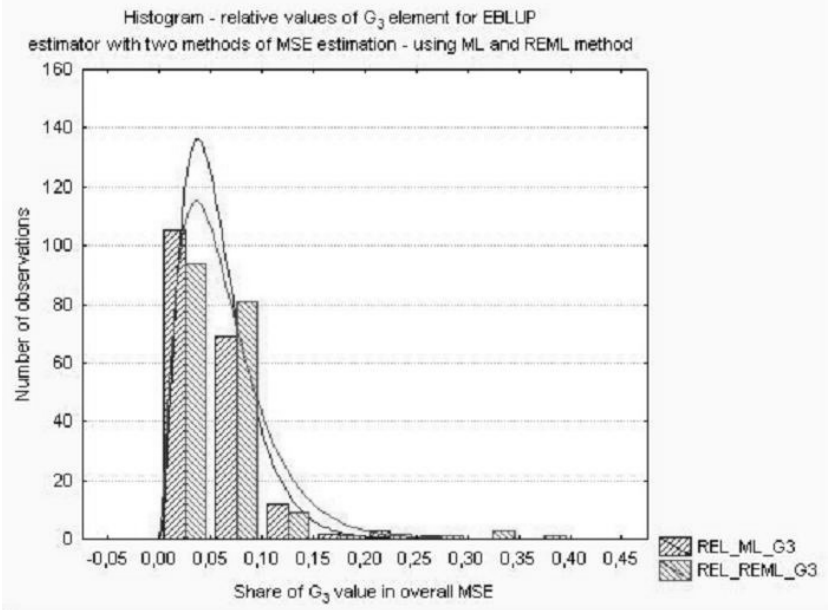
Source: own calculations using HBS data, WesVar and Statistica software and SAE package for R

Figure 4. Histogram of relative G2 element values for two variants of EBLUP estimation



Source: own calculations using HBS data, WesVar and Statistica software and SAE package for R

Figure 5. Histogram of relative G3 element values for two variants of EBLUP estimation



Source: own calculations using HBS data, WesVar and Statistica software and SAE package for R

5. CONCLUSIONS

EBLUP and other indirect estimators are likely to be more efficient than the corresponding direct ones as a result of "borrowing strength" from other domains in time and in space.

The EBLUP estimation procedure, based on a general linear mixed model, has an additional advantage of taking into account the between-area variation beyond that explained by the auxiliary variables included in a regression model.

Analyzing the empirical results presented in the paper one can easily notice that the EBLUP technique improves significantly the estimation quality. The mean square error reduction, as compared to the direct estimates, proves this. However, because of the method of MSE estimation, these results can rather be treated as good approximations.

The problem of effective MSE estimation is of great importance and should be analyzed in detail. In the paper the empirical distributions of mean square error components for EBLUP have been presented and discussed.

The obtained dependencies reveal that g_1 component has the main share in the overall MSE value. In some cases this share is greater than 90%, what may indicate that the implementation of MIXED procedure in SAS system, where only the first two elements are taken into account - see [Rao 2003, Part 6.2.7, p 105] - may give also proper results.

In most cases the share of g_2 component is not greater than a few percent but it is worth mentioning that for some models its contribution to the overall MSE value is much higher- even 90%. This result suggests that g_2 can also play an important role in MSE estimation.

The g_3 element's share is usually smaller than 5% but the most frequent values are smaller than 1% what easily explains its supplementary role. However, also because of this – it shouldn't be neglected in MSE estimation.

REFERENCES

Bracha, Cz., Lednicki, B., Wieczorkowski, R. (2004) Wykorzystanie złożonych metod estymacji do dezagregacji danych z Badania Aktywności Ekonomicznej Ludności w 2003 roku. „Z Prac Zakładu Badań Statystyczno-Ekonomicznych" z. 299

The EURAREA Consortium (2004), Project Reference, Volume I and III

Fay, R.E., Herriot, R.A. (1979), Estimation of Income from Small Places: Application of James Stein Procedure to Census Data. Journal of the American Statistical Association, 74, pp. 269-277

Kordos, J. (2005), Household surveys in transition countries. Chap. XXV in Household Sample Surveys in Developing and Transition Countries, United Nations, New York

Kordos, J., Lednicki, B., Żyra, M. (2002), The household sample surveys in Poland. Statistics in Transition, vol. 5, No. 4, pp. 555-589
Pratesi, M., Salvati, N. (2008), Small area estimation: the EBLUP estimator based on spatially correlated random area effects, Statistical Methods and Applications, 17, No 1, pp. 113-141
Rao, J.N.K. (2003), Small Area Estimation, Wiley, London
Software for ecological inference (2007) http://www.bias-project.org.uk/software/
UNDP Polska (2006), Report "Social Exclusion and Integration in Poland: An Indicators-based Approach", Warsaw

Zastosowanie estymatorów eblup do analizy małych obszarów na podstawie badania budżetów gospodarstw domowych

Streszczenie: W artykule przedstawiono wyniki estymacji dla małych obszarów w Polsce otrzymane z wykorzystaniem estymacji bezpośredniej oraz metody EBLUP (*empirical best linear unbiased prediction*). Analizy prowadzone były na podstawie próby pochodzącej z Badania Budżetów Gospodarstw Domowych oraz informacji dodatkowych pochodzących ze źródeł administracyjnych. W celu przeprowadzenia estymacji EBLUP oszacowane zostały modele wydatków gospodarstw domowych, uwzględniające również zmienność pomiędzy obszarami. Oszacowania wariancji dla estymatorów bezpośrednich otrzymano za pomocą metody replikacyjnej BRR (*balanced repeated replication*). Estymatory EBLUP wraz z ich błędami średniokwadratowymi oszacowane zostały za pomocą techniki komponentów wariancyjnych. W celu uzyskania ocen błędów średniokwadratowych dla EBLUP zastosowano metodę ML (największej wiarygodności) oraz metodę REML (największej wiarygodności z restrykcjami). Obliczenia prowadzone były w pakiecie SAE działającym w środowisku R.

Keywords: statystyka małych obszarów, estymacja EBLUP, estymacja wariancji, badanie budżetów gospodarstw domowych