ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 3(322)2016

http://dx.doi.org/10.18778/0208-6018.322.05

Alina Jędrzejczak [*], *Jan Kubacki* [**]

ESTIMATION OF MEAN INCOME FOR SMALL AREAS IN POLAND USING RAO-YU MODEL

Abstract. Modelling and estimating relationships that combine time series and cross-sectional data is often discussed in the statistical literature but in these considerations sampling errors are seldom taken into account. In the paper the application of the Rao-Yu model involving both- autocorrelated random effects between areas and sampling errors-has been presented. On the basis of this model the empirical best linear unbiased predictor (EBLUP) with time correlation has been obtained. As an example the application of several income-related variables for the Polish voivodships (regions) and the years 2003–2011 was used on the basis of the Polish Household Budget Survey and selected explanatory variables obtained from Polish Local Data Bank. The computations were performed using *sae2* and *sae* packages for *R-project* environment and WesVAR software. The precision of the direct estimates was obtained using Balanced Repeated Replication (BRR) technique.

For most investigated cases, the proposed methods based on the Rao-Yu model yielded the significant improvement of small area estimates due to substantial reduction of their relative estimation errors as compared to the ordinary EBLUP technique. For some income variables examined within the study very high values of time-related autocorrelation coefficient were observed. These values were in some cases higher than 0.9, what can be – in our opinion – a good illustration of income growth tendency observed in Poland in the period under consideration.

Keywords: small area estimation, EBLUP estimator, Rao-Yu model, nonlinear analysis
JEL: C22, C23, C83, D31, R11

1. INTRODUCTION

Most large-scale sample surveys are designed to provide reliable estimates for large geographical regions and large subgroups of a population. For example, the well-known Household Budget Survey conducted by the Polish Central Statistical Office provides reliable estimates of incomes and expenditures of households for the overall Polish population and for macro regions, while for smaller areas and/or more detailed income variables the sampling errors can be very large indeed (see: Jędrzejczak, Kubacki 2014)

Recently, the demand for estimates at a small level of aggregation has significantly increased in response to growing demands of policy-makers, in contrast to national estimates that were commonly used in the past. To fulfil

[*] Ph.D., Associate Professor, Department of Statistical Methods, University of Lodz.
[**] Ph.D., Department of Statistical Methods, University of Lodz.

growing requirements for detailed high-quality statistics it is necessary to employ indirect estimation methods that "borrow strength" from related areas in time and/or in space. Small area models can – for the purpose of quality improvement – apply the dependencies related to the spatial (see: Kubacki Jędrzejczak (2014), Dehnel et al. (2013)) and time relationships (see for example: Żądło (2012)). These models, incorporating sample and auxiliary information from other domains as well as other time periods, can yield to substantial quality improvements as compared to ordinary small area models, where only explanatory variables from administrative sources and other statistical surveys are used. It is also related to introduction of some constraints that can positively affect the quality of obtained estimates. The models using time-related dependencies can additionally be helpful in the analysis of the dynamics of the observed phenomena, what can be supplementary related to the econometric models, including the panel models.

The traditional Fay-Herriot model, which is the small area model most frequently used, is based on cross-sectional data only, so it neglects the information coming from other time-points. On the contrary, the Rao-Yu model for small areas (Rao, Yu (1992, 1994)) simultaneously involves random effects as well as time-correlated errors. A good illustration of its applications related to the victimization surveys in the USA can be found in Fay and Diallo (2012), in Fay and Li (2012) or in Li, Diallo and Fay (2012). The abovementioned works are related to the recently published package *sae2* for *R-project* environment (see Fay and Diallo (2015)) which has been implemented for the purpose of this work.

The main objective of the paper is to assess the possible efficiency gains coming from the application of the models combining time-series and cross-sectional data as compared to the traditional Fay-Herriot approach. Finally, the indirect (model-based) estimates will be compared to the direct estimates which are the ones most frequently applied in practice. The illustration of the methods is based on various income distributions coming from the Polish Household Budget Survey for the years 2003–2011. The administrative registers and selected explanatory variables obtained from Polish Local Data Bank have been applied as auxiliary data.

2. RAO-YU MODEL FOR TIME-CORRELATED RANDOM EFFECTS

Rao-Yu model, which is an extension of the Fay-Herriot (1979) model, was first described by its authors in two consecutive papers (Rao and You (1992, 1994)). Both these approaches assume standard relationships between direct survey-based estimates y_{it} for the i^{th} area at time *t*, and their expected population values, θ_{it}, that can be presented below

$$y_{it} = \theta_{it} + e_{it} \tag{1}$$

where: $i = 1, ..., m,\ t = 1, ..., T;\ e_i = (e_{i1}, \dots, e_{iT})^T$ is the error component related to the survey design (for whom the normal distribution is assumed) with zero mean and known covariance matrix Σ_i. The random effects are assumed to be independent between areas.

Rao-Yu model for the population values takes the following form:

$$\theta_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta} + v_i + u_{it} \tag{2}$$

wherein

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it} \tag{3}$$

where:

$\mathbf{x}_{it} = (x_{it1}, ..., x_{itp})^T$ is the vector of explanatory variables for area *i* and time *t*.

$\boldsymbol{\beta}$ is the vector of regression coefficients.

$v_i \sim N(0, \sigma_v^2)$ for $i = 1, ..., m$ are independent and identically distributed (iid) random effects that describe time-independent differences between areas,

ρ is a temporal correlation parameter and

$\varepsilon_{it} \sim N(0, \sigma^2)$ are independent and identically distributed random effects, which describes the variability of the series u_{it}, wherein $t = 1,..., T$.

By combining the Rao-Yu population model (2) with the sampling model (1), one can obtain the Rao-Yu model for small areas. It is worth mentioning that Rao and Yu considered only the case for which $|\rho| < 1$ and hence assumed the stationarity of the series explained by the equation (3), what determines the following relationships for all *i* and *t*:

$$Var(u_{it}) = \sigma^2 / (1 - \rho^2) \tag{4}$$

In the case where $\rho = 1$ and the assumption about stationarity (4) is ignored, the dependency (3) describes the well- known random walk process. Thus some discontinuity in the above model exists for $\rho = 1$.

A two-stage predictor of small-area means at a given time point can be obtained under the proposed model by first deriving the Best Linear Unbiased Predictor (BLUP) and then replacing the unknown variance components by their consistent estimators. Assuming that the parameters σ^2, σ_v^2 and ρ that determine the variance-covariance structure of the model (2) are all known, the BLUP estimator for the area i and time t has the following form

$$\widetilde{\theta}_{it} = \mathbf{x}_{it}^T\widetilde{\boldsymbol{\beta}} + (\sigma_v^2\gamma_{t,v} + \sigma^2\gamma_{t,u})^T(\boldsymbol{\Sigma}_i + \sigma^2\boldsymbol{\Gamma}_u + \sigma_v^2\boldsymbol{\Gamma}_v)^{-1}(\mathbf{y}_i - \mathbf{X}_i\widetilde{\boldsymbol{\beta}}) \tag{5}$$

where:

$\boldsymbol{\Gamma}_u$ is $T \times T$ matrix with the elements equal to $\rho^{|i-j|}/(1-\rho^2)$,

$\boldsymbol{\Gamma}_v$ is $T \times T$ matrix the elements of which are equal to 1,

$\mathbf{V}_i = \boldsymbol{\Sigma}_i + \sigma^2\boldsymbol{\Gamma}_u + \sigma_v^2\boldsymbol{\Gamma}_v = Cov(\mathbf{y}_i)$,

$\mathbf{V} = diag(\mathbf{V}_i) = Cov(\mathbf{y})$,

$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ $T \times T$ is the generalized least squares estimator of the regression coefficients β , and

$\boldsymbol{\Gamma}_{\text{t}}$ is the t^{th} column of matrix $\boldsymbol{\Gamma}_u$.

The Empirical Best Linear Unbiased Predictor (EBLUP) can be easily obtained using the formula (5), when instead of the unknown parameters σ^2, σ_v^2 and ρ, their consistent estimators $\hat{\sigma}^2$, $\hat{\sigma}_v^2$ and ρ are included. Mean squared error (MSE) for the EBLUP estimator (5) is derived on the basis of the variance components vector $\boldsymbol{\delta} = (\sigma_v^2, \sigma^2, \rho)$ using the second order approximation, what can be expressed as follows

$$MSE(\widetilde{\theta}_{iT}(\hat{\delta})) = g_{1iT}(\delta) + g_{2iT}(\delta) + 2g_{3iT}(\delta) \tag{6}$$

where

$$g_{1iT}(\delta) = \mathbf{m}_i^T(\mathbf{G}_i - \mathbf{G}_i\mathbf{V}_i^{-1}\mathbf{G}_i)\mathbf{m}_i \tag{7}$$

$$g_{2iT}(\delta) = d_i^T\left(\sum\nolimits_{i=1}^m X_i^T V_i^{-1} X_i\right)^{-1} d_i \tag{8}$$

$$g_{3iT}(\delta) = tr\left[\left(\frac{\partial b_{iT}^T}{\partial\delta}\right)V_i\left(\frac{\partial b_{iT}^T}{\partial\delta}\right)^T \overline{V}(\hat{\delta})\right] \tag{9}$$

wherein $\boldsymbol{\gamma_v}$ and $\boldsymbol{\gamma}_{\text{u}}$ are t-th columns of matrices Γ_v and $\Gamma_{v=u}$ respectively that satisfy the relationships

$$G_i = G_i(\delta) = \sigma_v^2 \Gamma_v(\rho) + \sigma_v^2 \Gamma_u(\rho),$$

$$d_i^T = x_{it}^T - b_i^T X_i^T,$$

$$b_i^T = m_i^T G_i V_i^{-1} = (\sigma_v^2 \gamma_{t,v} + \sigma^2 \gamma_{t,u}) V_i^{-1},$$

and further $\bar{V}(\hat{\delta}) = \Im^{-1}(\delta)$ is the information matrix $\Im$ inverse of dimension 3×3:

$$\Im_{jk}(\delta) = \tfrac{1}{2} tr(V^{-1} V_{(j)} V^{-1} V_{(k)})$$

where $V_{(r)} = \frac{\partial V}{\partial \delta_j}$.

In particular, where the REML method is used in order to obtain the variance components, the information matrix $\Im$ can be described in the form

$$\Im_{jk}^{REML}(\delta) = \tfrac{1}{2} tr(P V_{(j)} P V_{(k)})$$

where $P = V^{-1} - V^{-1} X (X^T V^{-1} X) X^T V^{-1}$.

At the computing stage (as it was mentioned in the introduction), the package *sae2* using *R-project* environment has been applied. Its implementation to the data sets considered in the paper required some modifications which, in a simplified form, can be presented as follows:

```
    library(RODBC)
    library(sae2)
    channel1 < – odbcConnectExcel(“Input.xls”)
    command < – paste(“select * from [Sheet1$] order by region,year”, sep=”)
    base < – sqlQuery(channel1, command)
    T < – length(unique(base$YEAR))
    D < – 16
    resultT.RY < – eblupRY(DOCHG_AVG ~ PKB_PC, D, T, vardir =
diag((base$DOCHG_SD)^2),data=base, ids=base$REGION)
```

After performing the calculations, the process of saving the results to the file is done. Here standard *cut* and *format* functions have been used. For simplification, the presented macro does not include the other variables except for *available income* and includes the example of calling only the REML version for eblupRY function. The computation for ML variants are similar and can be obtained by setting the option method to = “ML”.

3. RESULTS AND DISCUSSION

The primary goal of the paper was the estimation of mean available income and its main components obtained on the basis of the Polish Household Budget Survey for the years 2003–2011. The available income is defined as a sum of household's current incomes from various sources reduced by prepayments on personal income tax made on behalf of a tax payer by tax-remitter (this is the case of *income from hired work* and *social security benefits* and *other social benefits*) by tax on income from property, taxes paid by self-employed persons, including those in free professions and individual farmers and by social security and health insurance premiums. Its main components are the following:

- income from hired work,
- income from a private farm in agriculture,
- income from self-employment other than a private farm in agriculture, from free profession,
- income from property,
- income from rental of a property or land,
- social insurance benefits,
- other social benefits,
- other income (including gifts and alimonies).

Please note that the estimation of some of these components, selected on the basis of their incidence in the whole sample, is also considered in the paper. More details can be found for example in the publication of CSO entitled "Household budget survey in 2014" available at http://stat.gov.pl/en/topics/living-conditions/ living-conditions/household-budget-survey-in-2014,2,9.html.

The estimation results have been presented in tables and on figures, separately for two main variables of interest: *available income* and *income from self-employment*. The basic estimation outcomes have been outlined in Tables 2 and 3, comprising the direct and indirect estimates for all the regions in the selected years, as well as their relative estimation errors REE. The REE values have been evaluated by dividing MSEs by their corresponding income estimates and then expressing the results in %. The last two columns of Tables 2 and 3 present the REE reduction for EBLUP and for Rao-Yu, respectively. To obtain these results, the precision measures of indirect estimators were related to the corresponding indirect ones (i.e. for the first column: variances of direct estimators were divided by MSEs of EBLUPs; for the second column: variances of direct estimators were divided by MSEs of Rao-Yu-EBLUPs).

Analyzing the results presented in Tables 2 and 3 one can easily come to the general conclusion that the presented Rao-Yu model improves the precision of small-area estimates not only in relation to direct estimates, what is easy to obtain, but also in comparison with other indirect techniques based on small-area models. In particular, for the first year of the analysis the average efficiency gain connected with indirect estimation of *available income* instead of the direct one for some cases exceeded 10% (REE reduction was 1.064 on average), while for some regions with high variances REE reduction was huge (e.g. 1.275 for *pomorskie*). The application of the Rao-Yu method resulted in further improvement as the additional REE reduction for *availablе incom*e was 1.195 on average (Table 2).

This regularity was even more evident for the variables presenting slightly higher relative estimation error levels (i.e. the variables that are more "rare" than the overall *available income*). For example, in the case of *income from self-employment* (tab. 3), the average REE reduction due to traditional EBLUP estimation was 1.133 while for the EBLUPs based on the Rao-Yu model it was much higher and equaled 1.416. Thus the direct estimator variance was by 41.6% higher than the mean squared error of Rao-Yu-EBLUP.

In the case of some major or universal variables (e.g.: *available income* or *income from hired work*) the mean squared errors of Rao-Yu-EBLUP estimates turned out to be more similar to the ones obtained by means of the classical small-area model proposed by Fay and Herriot. This tendency seems particularly clear in Fig. 5 presenting the distribution of REE reduction for Rao-Yu EBLUP estimators due to the time-related effects referenced to the ordinary EBLUP estimators for different categories of income. It should also be noted that for some very rare variables (e.g.: *unemployment benefits*), the improvement of quality is not always achieved due to the limitations connected with the goodness-of-fit of the underlying models.

Table 1. Characteristics of Rao-Yu model for regions for *available income* and *income from self- employment*

Year / variable	Covariance structure parameters			LogLikelihood
	σ^2	σ_v^2	ρ	
Available income	1066.48	0.0013	0.9699	–737.37
Income from self-employment	41.56	71.219	0.8809	–554.97

Source: authors' calculations.

For most of the estimated models relatively high autocorrelation ρ values have been observed, what suggests that the time-related dependencies between areas are strong (see: Table 1). Moreover, for some well-fitted models (including *available income*) the values of this parameter visibly exceed 0.9, whereas the model diagnostics indicate that the parameter σ_v^2 makes only a small contribution to the model variability (see.: eq. (2), (3)). It leads to the conclusion that the between-area variability is mostly determined by time-related component, what can be a confirmation of reliability and adequacy of the methods applied within the study. For the other models the values of parameters σ_v^2 and σ^2 have a more even character, so the model variability comes from both between-area and between-period variation.

It is worth noting that for some small-areas the estimates based on the Rao-Yu model show some additional advantage of reflecting the real dynamics of a particular income component (as it can be observed for *income from self-employment*, *pomorskie* region for 2011), where the direct estimate was established at the level of 173 zł, while the Rao-Yu estimate was equal to 135 zł, which seemed more realistic. It can be easily confirmed by comparing these results to the direct values for the next year (154 zł as it can be obtained from official publications). Hence, it is possible that for such cases Rao-Yu model explains the dynamics of observed variables better. It obviously improves the time-relationships for such situations.

Table 2. Estimation results for *available income* by region (direct estimates, ordinary EBLUP and Rao-Yu EBLUP)

Region/Year	Available income							
	Direct estimator		Ordinary EBLUP estimator – REML		EBLUP estimator Rao-Yu model – REML		REE reduction for EBLUP	REE reduction for Rao-Yu
	Estimate zł	REE %	Estimate zł	REE %	Estimate zł	REE %		
1	2	3	4	5	6	7	8	9
2004								
Dolnośląskie	748.86	1.47	748.19	1.44	746.07	1.37	1.023	1.071
Kujawsko-Pomorskie	652.84	2.08	657.67	1.99	652.72	1.83	1.045	1.134
Lubelskie	676.70	1.29	674.48	1.27	674.35	1.22	1.010	1.051
Lubuskie	654.37	3.03	663.04	2.76	669.83	2.41	1.098	1.256
Łódzkie	730.53	2.62	726.20	2.45	731.42	2.11	1.071	1.241
Małopolskie	717.61	3.50	708.35	3.12	702.99	2.58	1.122	1.357
Mazowieckie	940.18	1.43	938.29	1.42	930.55	1.29	1.002	1.104
Opolskie	744.64	1.30	741.17	1.28	747.36	1.22	1.014	1.063
Podkarpackie	585.43	1.58	587.91	1.55	586.21	1.48	1.019	1.064
Podlaskie	644.88	3.47	645.46	3.14	651.12	2.53	1.102	1.368
Pomorskie	753.32	4.87	741.94	3.82	765.91	2.83	1.275	1.723
Śląskie	748.29	0.93	749.22	0.92	750.60	0.90	1.010	1.038
Świętokrzyskie	615.46	3.02	623.15	2.79	615.33	2.51	1.083	1.200
Warmińsko-Mazur.	657.68	1.47	657.64	1.45	658.59	1.38	1.018	1.067
Wielkopolskie	700.85	2.88	713.40	2.61	707.86	2.24	1.102	1.284
Zachodniopomorskie	762.04	1.66	756.74	1.62	752.88	1.51	1.025	1.102

Table 2 (cont.)

1	2	3	4	5	6	7	8	9
2011								
Dolnośląskie	1282.93	2.68	1287.97	2.50	1321.88	1.99	1.075	1.352
Kujawsko-Pomorskie	1108.94	2.17	1112.18	2.09	1111.89	1.95	1.040	1.116
Lubelskie	1025.80	2.07	1028.84	2.02	1027.82	1.81	1.029	1.149
Lubuskie	1189.89	1.55	1186.85	1.52	1192.57	1.38	1.019	1.122
Łódzkie	1203.19	2.62	1202.11	2.46	1224.93	2.00	1.064	1.309
Małopolskie	1156.79	2.53	1157.34	2.40	1167.22	2.02	1.056	1.254
Mazowieckie	1622.96	2.02	1615.41	2.01	1669.56	1.59	1.005	1.266
Opolskie	1181.90	1.88	1177.09	1.83	1178.66	1.64	1.026	1.146
Podkarpackie	937.85	2.52	950.29	2.41	945.67	2.11	1.046	1.197
Podlaskie	1224.92	1.45	1216.42	1.44	1208.41	1.34	1.011	1.083
Pomorskie	1286.94	3.09	1268.70	2.83	1298.67	2.20	1.090	1.406
Śląskie	1215.44	0.95	1217.14	0.95	1220.96	0.91	1.009	1.052
Świętokrzyskie	1062.78	2.37	1066.79	2.27	1057.54	2.05	1.043	1.158
Warmińsko-Mazur.	1096.87	2.63	1095.26	2.50	1111.93	2.17	1.049	1.210
Wielkopolskie	1135.02	2.73	1155.10	2.53	1170.17	2.09	1.079	1.306
Zachodniopomorskie	1231.10	3.16	1212.66	2.91	1226.36	2.27	1.084	1.388

Source: authors' calculations

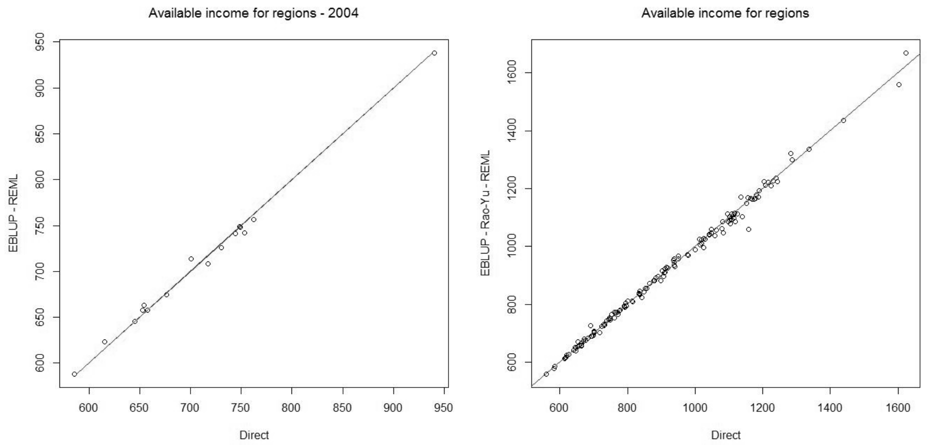

Figure 1. Direct vs. Indirect available income estimates [indirect= EBLUPs based on Fay-Herriot (top) or Rao-Yu (bottom)]

Source: own elaboration.

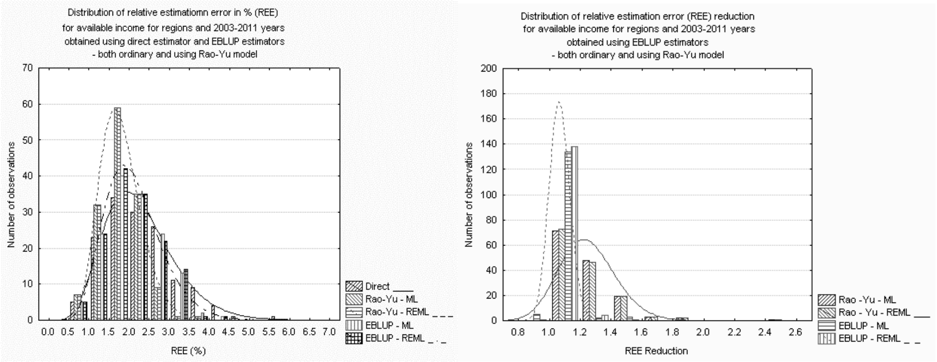

Figure 2. Distribution of relative estimation error (REE) in % and REE reduction (direct estimator and EBLUP estimators: ordinary and using Rao-Yu model)

Source: own elaboration.

Table 3. Estimation results for *income from self-employment* by region (direct estimates, ordinary EBLUP and Rao-Yu EBLUP)

Region/Year	Income from self-employment							
	Direct estimator		Ordinary EBLUP estimator – REML		EBLUP estimator Rao-Yu model – REML		REE reduction for EBLUP	REE reduction for Rao-Yu
	Estimate zł	REE %	Estimate zł	REE %	Estimate zł	REE %		
1	2	3	4	5	6	7	8	9
2004								
Dolnośląskie	58.69	8.80	57.42	8.32	57.36	7.30	1.057	1.205
Kujawsko-Pomorskie	38.43	7.38	39.38	7.06	39.86	6.52	1.046	1.132
Lubelskie	53.29	16.10	52.66	13.10	45.30	10.55	1.230	1.526
Lubuskie	53.83	12.07	52.07	11.03	58.50	7.50	1.094	1.609
Łódzkie	34.89	9.80	36.65	9.05	38.22	7.90	1.084	1.241
Małopolskie	64.02	10.88	60.96	9.88	61.63	7.24	1.101	1.501
Mazowieckie	93.34	9.25	96.01	9.11	95.97	5.51	1.015	1.679
Opolskie	48.23	16.39	48.81	13.46	46.17	11.43	1.218	1.434
Podkarpackie	31.54	15.32	34.73	13.04	34.03	11.08	1.174	1.382
Podlaskie	60.65	6.39	59.19	6.29	58.42	5.83	1.017	1.097
Pomorskie	83.77	14.16	69.83	11.68	85.11	6.98	1.213	2.027
Śląskie	46.51	7.13	47.12	6.83	47.45	6.15	1.043	1.159
Świętokrzyskie	44.26	11.84	45.29	10.70	41.33	9.89	1.106	1.197
Warmińsko-Mazur.	60.42	15.05	53.93	13.31	62.39	8.45	1.130	1.780
Wielkopolskie	56.29	5.06	56.35	4.95	57.16	4.51	1.022	1.121
Zachodniopomorskie	75.39	12.77	64.75	11.35	71.35	8.18	1.126	1.561

Table 3 (cont.)

1	2	3	4	5	6	7	8	9
2011								
Dolnośląskie	104.30	4.91	104.07	4.81	111.32	3.88	1.021	1.267
Kujawsko-Pomorskie	84.86	11.05	84.98	10.17	92.61	7.23	1.087	1.528
Lubelskie	70.18	5.56	71.25	5.40	71.04	4.97	1.028	1.118
Lubuskie	115.25	6.40	109.40	6.44	113.86	5.27	0.994	1.213
Łódzkie	89.97	13.09	89.85	11.47	101.03	7.54	1.142	1.736
Małopolskie	115.47	13.03	107.46	11.25	114.24	6.88	1.158	1.894
Mazowieckie	178.75	3.85	178.65	3.91	177.50	3.26	0.984	1.182
Opolskie	84.16	18.49	89.47	13.79	86.71	9.68	1.341	1.911
Podkarpackie	57.45	13.96	62.24	12.16	61.66	9.22	1.148	1.515
Podlaskie	88.87	10.99	90.76	9.81	85.79	7.98	1.120	1.377
Pomorskie	173.42	15.28	122.63	12.59	135.46	7.05	1.213	2.168
Śląskie	85.72	6.60	86.80	6.34	86.90	5.44	1.042	1.215
Świętokrzyskie	71.87	10.71	75.08	9.71	81.09	7.63	1.104	1.405
Warmińsko-Mazur.	90.84	8.02	89.81	7.74	95.59	6.11	1.036	1.312
Wielkopolskie	105.94	13.01	105.52	10.87	112.37	7.19	1.197	1.810
Zachodniopomorskie	112.87	11.68	108.18	10.28	114.32	6.53	1.136	1.789

Source: authors' calculations.

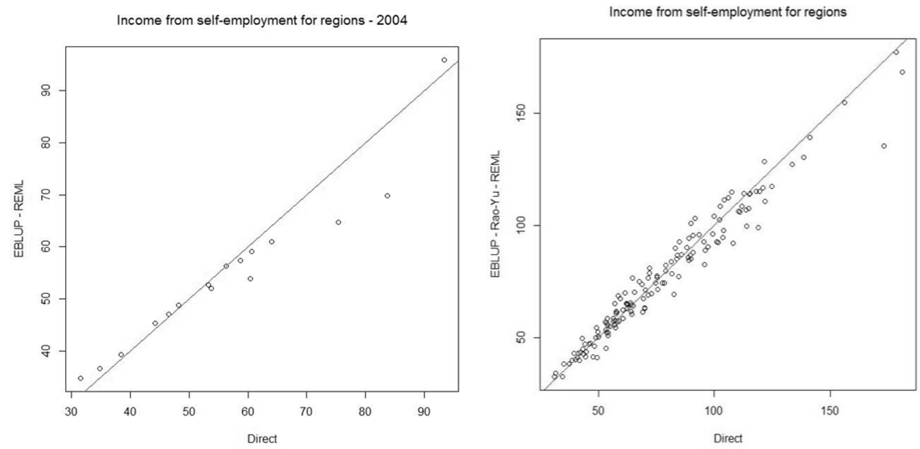

Figure 3. Direct vs. Indirect estimates of self-employment income [indirect= EBLUPs based on Fay-Herriot (top) or Rao-Yu (bottom)]

Source: own elaboration.

.

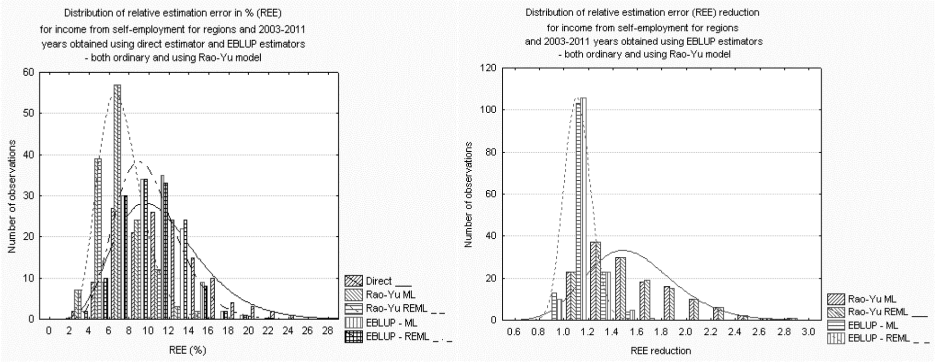

Figure 4. Distribution of relative estimation error (REE) in % and REE reduction for self-employment income in 2003–2011 (direct estimator and EBLUP estimators – both ordinary and using Rao-Yu model)

Source: own elaboration.

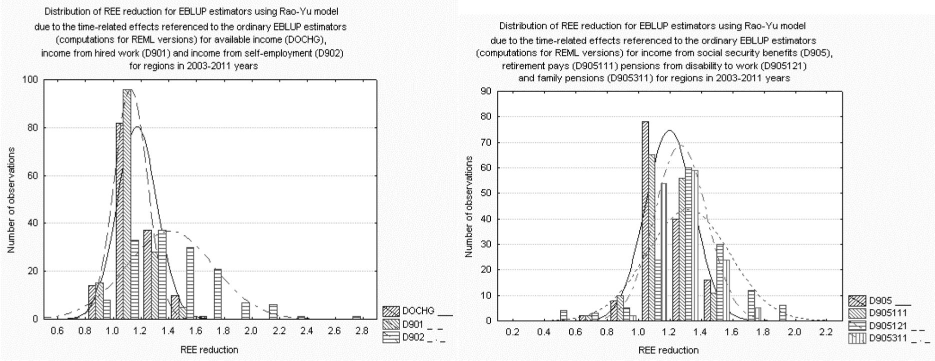

Figure 5. Distribution of REE reduction for Rao-Yu EBLUP estimators due to the time-related effects (referenced to the ordinary EBLUP estimators for different categories of income)

Source: own elaboration

Table 4. REE reduction for Rao-Yu EBLUP estimators due to time-related effects (referenced to the ordinary EBLUP estimators for different categories of income)

Region	Income						
	available	from hired work	from self-employment	social security benefits	retirement pays	pensions from inability to work	family pensions
1	2	3	4	5	6	7	8
2004							
Dolnośląskie	1.0473	1.0141	1.1401	1.1626	1.1100	1.3906	1.1014
Kujawsko-Pomorskie	1.0848	1.0135	1.0822	1.1938	1.0558	1.3406	1.1275

Table 4 (cont.)

1	2	3	4	5	6	7	8
Lubelskie	1.0405	1.0173	1.2412	1.1870	1.1666	1.8319	1.0952
Lubuskie	1.1445	1.0098	1.4713	1.1177	1.0531	1.2849	1.0954
Łódzkie	1.1581	1.0317	1.1455	1.1657	1.1007	1.0731	1.0326
Małopolskie	1.2094	1.0567	1.3634	1.3371	1.3246	1.6089	1.0415
Mazowieckie	1.1024	1.1137	1.6538	1.1961	1.1530	1.2329	1.2149
Opolskie	1.0483	1.0366	1.1776	1.1255	1.2248	1.5757	1.1765
Podkarpackie	1.0439	1.0664	1.1774	1.2700	1.2014	1.2720	1.1399
Podlaskie	1.2417	1.0189	1.0795	1.4519	1.3132	1.8526	1.1694
Pomorskie	1.3507	1.0945	1.6715	1.1442	1.1277	1.6763	0.9667
Śląskie	1.0276	1.0251	1.1111	1.1627	1.2169	1.4802	1.2972
Świętokrzyskie	1.1081	1.0103	1.0820	1.2102	1.1542	1.1778	1.0899
Warmińsko-Mazur.	1.0483	1.0417	1.5747	1.2013	1.1053	1.3607	1.0965
Wielkopolskie	1.1647	1.0182	1.0963	1.1123	1.1487	1.6113	1.1190
Zachodniopomorskie	1.0744	1.0575	1.3869	1.1395	1.0759	1.3530	1.0803
2011							
Dolnośląskie	1.2573	1.1775	1.2410	1.1394	1.2624	1.1278	1.4181
Kujawsko-Pomorskie	1.0731	1.1767	1.4058	1.1190	1.1325	1.1340	1.4646
Lubelskie	1.1163	1.0773	1.0866	1.0773	1.0882	1.3511	1.0413
Lubuskie	1.1013	1.2587	1.2201	1.1184	1.2789	0.9967	1.3526
Łódzkie	1.2308	1.0866	1.5204	1.1645	1.1975	1.2047	1.2787
Małopolskie	1.1875	1.1079	1.6361	1.0764	1.1189	1.0795	1.2818
Mazowieckie	1.2595	1.2654	1.2020	1.2149	1.1683	1.2402	1.2049
Opolskie	1.1164	1.2402	1.4254	1.1481	1.4563	1.4058	1.6729
Podkarpackie	1.1448	1.0602	1.3191	1.1155	1.2070	1.0824	1.5990
Podlaskie	1.0708	1.0308	1.2297	1.0033	1.0287	1.1357	1.3168
Pomorskie	1.2900	1.2851	1.7873	1.1203	1.3347	1.0441	1.5156
Śląskie	1.0430	1.0468	1.1657	1.1376	1.1090	1.1813	1.4049
Świętokrzyskie	1.1104	1.1972	1.2727	1.1132	1.2295	1.0149	1.4322
Warmińsko-Mazur.	1.1535	1.0607	1.2661	1.1388	1.4067	1.0095	1.2847
Wielkopolskie	1.2101	1.1271	1.5119	1.1000	1.1856	1.1832	1.3327
Zachodniopomorskie	1.2805	1.1478	1.5750	1.0830	1.2619	1.0023	1.3734

Source: authors' calculations.

The existence of such a strong time-dependency for the variables under consideration can also be confirmed using other statistical tools. One of these approaches was the application of a non-linear estimation in order to prepare a model incorporating higher-order components. The preliminary model of this type is presented in Fig. 6 and describes available income dependencies over the period 2000–2012, including price changes. The model has the following form:

$$y_{i,t} = a + bx_{i,t} + ct + dt^2 + et^3$$

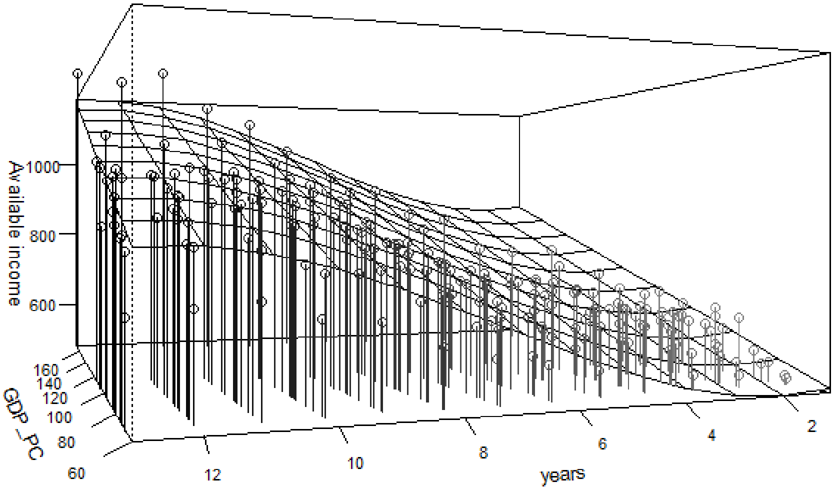

Figure 6. Non-linear model for *available income* including price changes for HBS for the years 2000–2012 and GDP per capita (Poland=100%)

Source: own elaboration.

The variable t changes from 1 to 13 for the years 2000–2012, $x_{i,t}$ determines GDP per capita values from the region i and time t. The summary of the model can be presented as follows.

Table5. Diagnostics of non-linear model for *available income* including price changes for Polish Household Budget Survey and the years 2000 –2012 against GDP per capita (Poland=100%)

Parameter	Parameter value	Standard error	t-statistics	p-value	Significance
R^2= 0.8901, Corrected R^2= 0.8880					
a	311.425	23.9893	12.9819	<0.00001	***
b	3.54047	0.160436	22.0679	<0.00001	***
c	–42.5037	11.2172	–3.7892	0.0002	***
d	12.1258	1.82691	6.6373	<0.00001	***
e	–0.55875	0.085999	–6.4972	<0.00001	***

Source: authors' calculations.

As it has been shown in table 5, the influence of the model parameters on the estimated small-area means was proven significant; also the value of R^2 coefficient remained satisfactory. Moreover, the positive influence of *d* variable and the negative influence of the variable *e* (because of some minor reductions of real incomes observed for the year 2012) are visible. Such a dependency, detected also for the periods going beyond the considered time-series (comprising only the years 2003–2011), may additionally indicate that the application of the models based on cross-sectional and time-series data is reasonable.

One can also use the dynamic small-area model, as it has been described in Fay and Diallo (2012),which incorporates time-related dependencies in a slightly different form. In the quoted work (Fay and Diallo, 2012, p. 3747), the following opinion can be found: „the dynamic model does not assume stationarity, does not constrain ρ to be less than 1, and avoids a discontinuity at ρ = 1. In fact values greater than 1 can reflect systematically increasing divergence, a phenomenon called "Matthew effect" in some works. Such a situation can be observed for some more detailed variables related to household income, i.e. *income from hired work* and *income from self-employment* (but is not observed for the other income-related variables considered in the study), and can also be determined when the time period (sliding-span) is used and is equal to 6 years and the starting point of the observation is moved for every year from 2003 to 2006. In the case of sliding span, for some income components, relatively large values of ρ parameter have been observed, being substantially higher than for the whole 2003–2011 period. It would also be interesting to verify whether such a situation can be confirmed using more traditional econometric models or panel models. The preliminary analysis, which has been conducted for such models, does not clearly indicate that for the Polish regions the divergence phenomena can be observed. However, the further considerations on this issue go beyond the scope of this article and require a separate, more complete analysis.

4. SUMMARY

The results obtained within this study confirm that the efficient estimation of income distribution parameters can be a serious problem, especially for small areas and rare variables- the estimators may be seriously biased and their standard errors far beyond the values that can be accepted by social policy-makers for making reliable policy decisions Indirect estimation methods that increase the effective sample size by using information from other domains or from other periods of time can be used to solve this problem.

The presented Rao-Yu model improves the precision of small-area estimates not only in relation to direct estimates, what is easy to obtain, but also in comparison with other indirect techniques based on small-area models. The small area model approach, including EBLUP and Spatial EBLUP procedures based on a general linear mixed model, presents a well-known advantage of taking into account the between-area variation beyond that explained by the auxiliary variables included in classical regression models. The application of the Rao-Yu model, where time-dependent effects are also taken into account, may significantly improve the quality of estimates for small areas, given that there is evident dependency of the observed values over time. The estimation quality improvement is more evident for variables which are related to minor domains, as is „*income from self-employment*". For "rare" variables, however, which are observed for only a few sampling units within the domains of interest, the application of Rao-Yu model can be difficult.

Further benefits can be expected when time-dependent nonlinear relationships are taken into account. The primary analysis of nonlinear models presented above may be a starting point for detailed comparisons between Rao-Yu method, nonlinear models and econometric panel models.

REFERENCES

Caselli F., Esquivel G., Lefort F. (1996), *Reopening the convergence debate: A new look at cross-country growth empirics*, "Journal of Economic Growth" vol. 1, no. 3, pp. 363–389.

Dehnel G., Klimanek T., Kowalewski J. (2013), *Indirect Estimation Accounting for Spatial Autocorrelation in Economic Statistics*, "Acta Universitatis Lodziensis Folia Oeconomica", vol. 286, pp. 293–305.

Fay R.E., Diallo M. (2012), *Small Area Estimation Alternatives for the National Crime Victimization Survey*, [in:] "Proc. Survey Research Methods Section of the American Statistical Association", pp. 3742–3756, https://www.amstat.org/sections/SRMS/ Proceedings/ y2012/Files/304438_73111.pdf

Fay R.E., Diallo M. (2015), *sae2: Small Area Estimation: Time-series Models*, package version 0.1-1, https://cran.r-project.org/web/packages/sae2/index.html

FayR.E., Herriot R.A. (1979), *Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data*, "Journal of the American Statistical Association", 74, pp. 269–277, http://www.jstor.org/stable/2286322

Fay R.E., Li J. (2012), *Rethinking the NCVS: Subnational Goals through Direct Estimation*, presented at the 2012 Federal Committee on Statistical Methodology Conference, Washington, DC, Jan. 10–12, 2012 https://fcsm.sites.usa.gov/files/2014/05/Fay_2012FCSM_I-B.pdf

Jędrzejczak A., Kubacki J. (2014), *Problemy jakości danych statystycznych w przypadku badania cech rzadkich*, "Wiadomości Statystyczne", no. 6, pp. 11–26.

Kubacki J., Jędrzejczak A. (2014), *Small area estimation under spatial SAR model*, Small Area Estimation 3–5 September 2014, Poznań.

http://sae2014.ue.poznan.pl/presentations/process.php?id=26

Molina I., Marhuenda Y. (2015), *sae: An R Package for Small Area Estimation*, "The R Journal", vol. 7, no. 1, pp. 81–98, http://journal.r-project.org/archive/2015-1/molina-marhuenda.pdf

Li J., Diallo M.S., Fay R.E. (2012), *Rethinking the NCVS: Small Area Approaches to Estimating Crime*, presented at the Federal Committee on Statistical Methodology Conference, Washington, DC, Jan. 10 –12, 2012, https://fcsm.sites.usa.gov/files/2014/05/Li_2012FCSM_I-B.pdf

Rao J.N.K. (2003), *Small Area Estimation*, Wiley Interscience, Hoboken, New Jersey.

Rao J.N.K., Yu M. (1992), *Small area estimation combining time series and cross-sectional data.* "Proc. Survey Research Methods Section. Amer. Statist. Assoc.", pp. 1–9 https://www.amstat.org/sections/SRMS/Proceedings/papers/1992_001.pdf

Rao J.N.K., Yu M. (1994), *Small-Area Estimation by Combining Time-Series and Cross-Sectional Data*, "The Canadian Journal of Statistics", vol. 22, no. 4, pp. 511–528, http://www.jstor.org/stable/3315407

R Core Team, (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org

Westat (2007), WesVar® 4.3 User's Guide.

Yu M. (1993), *Nested error regression model and small area estimation combining cress-sectional and time series data*, A thesis submitted to the Faculty of the Graduate and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Carleton University, Ottawa, Canada.

Żądło T. (2012), *On Accuracy of Two Predictors for Spatially and Temporally Correlated Longitudinal Data,* "Studia Ekonomiczne", no. 120, pp. 97–10.

Alina Jędrzejczak, Jan Kubacki

SZACOWANIE ŚREDNIEGO DOCHODU DLA MAŁYCH OBSZARÓW W POLSCE Z WYKORZYSTANIEM MODELU RAO-YU

Streszczenie. Modelowanie i szacowanie zależności, które uwzględniają szeregi czasowe oraz dane przekrojowe jest często dyskutowane w literaturze statystycznej, ale na ogół w takich pracach nie są brane pod uwagę błędy losowe. W pracy przedstawiono zastosowanie modelu Rao-Yu uwzględniającego zarówno autokorelację między obszarami efektów losowych zjawisk w czasie, jak i błędy losowe oszacowane na podstawie próby. Na podstawie modelu otrzymano empiryczny najlepszy nieobciążony predyktor liniowy (EBLUP), uwzględniający korelację zjawisk w czasie. Jako przykład wybrano aplikację dla kilku zmiennych dochodowych wyznaczonych dla województw dla lat 2003–2011 na podstawie Badania Budżetów Gospodarstw Domowych wraz z wybranymi zmiennymi objaśniającymi pochodzącymi z Banku Danych Lokalnych. GUS. Obliczenia wykonano w systemie R-*project* z użyciem pakietów *sae2* i *sae* oraz programu WesVar. Precyzję dla szacunków bezpośrednich wyznaczono z użyciem metody półprób zrównoważonych (BRR).

Dla większości rozważanych przypadków zaproponowana metoda, stosująca model dla małych obszarów typu Rao-Yu, skutkuje znaczącą poprawą szacunków średniego dochodu gospodarstw domowych w Polsce, o czym świadczą oceny błędów szacunku porównane do zwykłej estymacji EBLUP. Dla części otrzymanych modeli stwierdzono istnienie wysokiej autokorelacji związanej ze składnikiem losowym dla czasu ρ (o wartościach niekiedy wyższych od 0.9), co dobrze ilustruje tendencje wzrostowe dla dochodów gospodarstw domowych w Polsce w rozważanym okresie.

Słowa kluczowe: Estymacja dla małych obszarów, estymator EBLUP, model Rao-Yu, analiza nieliniowa