STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 725–742, DOI 10.21307/stattrans-2017-009

ESTIMATION OF SMALL AREA CHARACTERISTICS USING MULTIVARIATE RAO-YU MODEL

Alina Jędrzejczak [1,2], Jan Kubacki [2]

ABSTRACT

The growing demand for high-quality statistical data for small areas coming from both the public and private sector makes it necessary to develop appropriate estimation methods. The techniques based on small area models that combine time series and cross-sectional data allow for efficient "borrowing strength" from the entire population and they can also take into account changes over time. In this context, the EBLUP estimation based on multivariate Rao-Yu model, involving both autocorrelated random effects between areas and sampling errors, can be useful. The efficiency of this approach involves the degree of correlation between dependent variables considered in the model. In the paper we take up the subject of the estimation of incomes and expenditure in Poland by means of the multivariate Rao-Yu model based on the sample data coming from the Polish Household Budget Survey and administrative registers. In particular, the advantages and limitations of bivariate models have been discussed. The calculations were performed using the *sae* and *sae2* packages for R-project environment. Direct estimates were performed using the WesVAR software, and the precision of the direct estimates was determined using a balanced repeated replication (BRR) method.

Key words: small area estimation, EBLUP estimator, Rao-Yu model, multivariate analysis.

1. Introduction

The motivation for the paper is twofold. First, the growing demand for high-quality statistical data at low levels of aggregation, observed over the last few decades, has attracted much attention and concern amongst survey statisticians, but only a few works have been devoted to the small area estimation involving the combination of cross-sectional and time-series data. Second, the evidence on income distribution and poverty gathered for OECD countries in the latter part of

[1] Institute of Statistics and Demography, Faculty of Economics and Sociology, University of Łódź. E-mail: jedrzej@uni.lodz.pl.

[2] Centre of Mathematical Statistics, Statistical Office in Łódź. E-mail: j.kubacki@stat.gov.pl.

the first decade of the 2000s confirms that there has been an significant increase in income inequality, which has grown since at least the mid-1980 and there are still substantial differences in regional income levels (see: *Growing Unequal?*, OECD 2008; *Divided We Stand. Why Inequality Keeps Rising*. OECD 2011). Due the problem of high disparities between regions it is becoming crucial to provide reliable estimates of income distribution characteristics for small areas. The task is rather difficult as heavy-tailed and extremely asymmetrical income distributions can yield many estimation problems even for large domains. For some population divisions (by age, occupation, family type or geographical area) the problem becomes more severe and estimators of income distribution characteristics can be seriously biased and their standard errors far beyond the values that can be accepted by social policy-makers for making reliable policy decisions. That latter case is the area of applications for small area estimation.

Within the framework of survey methodology and small area estimation one can apply several methods to improve the estimation quality. Making use of auxiliary data coming from administrative registers or censuses within the traditional framework of survey methodology (ratio and regression estimators) can obviously improve the quality of estimates. However, the most important issue is the synthetic estimation that moves away from the design-based estimation of conventional direct estimates to indirect (and usually model-dependent) estimates that „borrow strength" from other small areas or other sources in time and/or in space. The term „borrowing strength" means increasing the effective sample size and is related to using additional information from larger areas, which can be applied for both interest (Y) and auxiliary variables (X). A large variety of small-area techniques, including small area models, have been described in Rao (2003), Rao, Molina (2015). In the paper we are especially interested in the multivariate case of the Rao-Yu model, the extension of the Fay-Herriot model, which "borrows strength" from other domains and over time.

Multivariate models can account for the correlation between several dependent variables and can specifically be applied to the situations when correlated income characteristics are involved. Multivariate models, being extensions of basic small area models, have been studied in some papers within the framework of small area estimation literature. In particular, interesting studies concerning multivariate linear mixed models can be found in the papers by Fay (1987) and Datta et al. (1991). In Datta et al. (1996) one can find the application of multivariate Fay-Herriot model in the context of hierarchical model with the application to estimating the median income of four-person families in the USA. Recently, some papers have been published where the multivariate linear mixed models were employed, including the works by Benavent and Morales (2016), Porter et al. (2015). The interesting applications related to the victimization surveys in the USA can be found in Fay and Diallo (2012), in Fay and in Li, Diallo and Fay (2012). Also, some applications of Rao-Yu model have been published. Here, we can mention the works by Janicki (2016) and Gershunskaya (2015). One of the applications for the univariate case of the Rao-Yu model can be found in the previous paper of the authors (Jędrzejczak, Kubacki (2016)). The increase in the number of applications in this area can also be related to the recently published package sae2 for R-project environment (Fay, Diallo (2015)).

The aim of this paper is to present the method for estimating small area means on the basis of sample and auxiliary data coming from other areas and different periods of time. The authors' proposition is to use two-dimensional models which can be applied to simultaneously estimate correlated income variables. The example of the application is based on the micro data coming from the 2003–2011 Polish Household Budget Survey on income and expenditure assumed as dependent variables, and administrative registers. In the application two-dimensional Rao-Yu model is compared with simpler estimation techniques.

2. Univariate and multivariate Rao-Yu model

Various small area models can be utilized in order to improve the quality of estimation in the presence of insufficient sample sizes. They can account for between-area variability beyond that explained by traditional regression models and thus make it possible to adjust for specific domains. Most of these models are special cases of the general linear mixed model.

General linear mixed model is a statistical linear model containing both fixed and random effects, which can be described as follows (see e.g.: Rao (2003), Chapter 6.2):

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

In the equation given above $\mathbf{y}$ is a $n \times 1$ vector of the observations that can come from a sample survey, $\mathbf{X}$ and $\mathbf{Z}$ are known $n \times p$ and $n \times h$ matrices that can represent auxiliary data, $\mathbf{v}$ and $\mathbf{e}$ are independently distributed random variables with covariance matrices $\mathbf{G}$ and $\mathbf{R}$ respectively, related to the model variance components. Depending on the variance-covariance structure many variants of the model (1) can be specified, among them the model with block-diagonal covariance structure, which has been the basis for many small area models, including the popular Fay-Herriot model or the Rao-Yu model. They are the examples of area-level model in contrast to the unit-level models that are not considered in the paper.

Univariate model

Rao-Yu small area model, which incorporates time series and cross-sectional data, is a special case of the general linear mixed model with block diagonal covariance structure as described in Rao and Yu (1994) and in Rao (2003). A linear mixed model for the population values, θ_{it}, for the domain i (*i=1,…m*) in time t (t=1,…,T) is the following

$$\theta_{\mathrm{it}} = \mathbf{x}_i^T\boldsymbol{\beta} + \mathrm{v}_\mathrm{i} + \mathrm{u}_{it} \tag{2}$$

where:

x_i^T is a row vector of known auxiliary variables,

β is a vector of fixed effects,

v_i is a random effect for the area i, $v_i \overset{iid}{\sim} N(0,\sigma_v^2)$,

u_it is a random effect for the area i and time t, representing the stationary time-series described by AR(1) process

$$u_{it} = \rho u_{i,t-1} + \epsilon_{it}$$

with constraint |ρ|<1 and $\varepsilon_{\mathrm{it}} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$.

Based on the model (2) we can obtain the corresponding model for the observed sample values, y_{it}, which takes the form:

$$y_{it} = \theta_{it} + e_{it} = \mathbf{x}_i^T\boldsymbol{\beta} + v_i + u_{it} + e_{it} \tag{3}$$

where:

e_{it} is a random sampling error for the area i and time t, with

$$\mathbf{e}_i = (e_{i1}, \dots, e_{iT})^T$$

following T-variate normal distribution with the mean 0 and known covariance matrix **Σ.**

It is worth noting that the random variables v_i , ε_i and $\mathbf{e}_\mathrm{i}$ are mutually independent and the matrix **Σ** with diagonal elements equal to sampling variances for the domain *i* corresponds to the matrix **R** from the model (1).

The crucial role in the model is played by the random terms v and u. They are two components constituting the total random effect of the Rao-Yu model. The first one (*v*) accounts for the between-area variability while the second one (*u*) accounts for the variability across time. In particular: v_i's are independent and identically distributed random effects that describe time-independent differences between areas; the u_i's follow the autoregressive process with ρ being temporal correlation parameter for all the areas of interest.

Multivariate model

Assume $\boldsymbol{\theta}_{it} = \left(\theta_{it,1}, \dots, \theta_{it,r}\right)^T$ as a vector of unknown population parameters. Let $\mathbf{y}_{it}$ be a vector of direct estimators of *r* parameters of interest related to sample observations which can be expressed as $\mathbf{y}_{it} = \left(y_{it,1}, \dots, y_{it,r}\right)^T$. The multivariate population model for the *j*-th variable of interest (*j*=1,...,*r*) takes the following form (similar model can be found in Fay et al. (2012)):

$$\theta_{it,j} = \mathbf{x}_{it,j}^T\boldsymbol{\beta}_j + v_{i,j} + u_{\mathrm{it,j}} \tag{4}$$

where:

$\mathbf{v}_i = \left(v_{i,1}, \dots, v_{i,r}\right)^T \overset{iid}{\sim} N_r(0, \boldsymbol{\sigma}_v^2)$ is a vector of random effects for the area i ,

u_{it} is a random effect for the area i and time t, representing the stationary time-series described by AR(1) process

$$u_{it,k} = \rho u_{i,t-1,k} + \epsilon_{it,k}$$

with constraint |ρ|<1 and $\boldsymbol{\varepsilon}_{it} = (\varepsilon_{it,1}, \dots, \varepsilon_{it,r})^T \overset{iid}{\sim} \mathrm{N}_r(0, \boldsymbol{\sigma}^2)$.

It is worth noting that the model (4) also posits a single autoregression parameter ρ and the random variables $\mathbf{v}_i$, $\boldsymbol{\varepsilon}_i$ and

$$\mathbf{e}_i = (e_{i1,1}, e_{i2,1}, \dots, e_{iT,1}, \dots, e_{i1,r}, e_{i2,r} \dots, e_{iT,r})^T$$

are mutually independent.

The sampling model corresponding to the formula (4) can be written as

$$y_{it,j} = \theta_{it,j} + e_{it,j} = \mathbf{x}_{it,j}^T \boldsymbol{\beta}_j + v_{i,j} + u_{\mathrm{it,j}} + e_{it,j} \tag{5}$$

with the covariance matrix of random effects, linking the matrices $\boldsymbol{\sigma}^2$ and $\boldsymbol{\sigma}$, equal

$$\mathbf{G} = \mathbf{M} \otimes \left[\left((\boldsymbol{\sigma}\boldsymbol{\sigma}^{\boldsymbol{T}})\mathbf{u}_c\right) \otimes \boldsymbol{\Gamma}_u + \left((\boldsymbol{\sigma}_v \boldsymbol{\sigma}_v^{\boldsymbol{T}})\mathbf{u}_c\right) \otimes \boldsymbol{\Gamma}_v\right],$$

where: $\boldsymbol{\Gamma}_u$ is covariance matrix of $\mathbf{u}_i = (u_{i1}, \dots, u_{iT})^T$ with the elements equal to $\rho^{|t-s|}/(1-\rho^2)$ for an entry (t,s) that represent the AR(1) model for $u_{it} = \rho u_{i,t-1} + \epsilon_{it}$, with constraint |ρ|<1. Vectors $\mathbf{v}_\mathrm{i}$ represent the random effects, reflecting time-independent differences between areas. The vectors $\boldsymbol{\sigma}_v$ and $\boldsymbol{\sigma}$ represent the model errors connected with the random effects **u** and **v**, respectively, and have r elements each. The matrix $\mathbf{u}_\mathrm{c}$ is $r \times r$ matrix of $\rho_{u,jk}$ values with the diagonal elements equal to 1 and for the remaining elements $(j \neq k)$, related to the correlation of the random effects **u** with respect to the multidimensional structure specified within the model. **M** is $m \times m$ diagonal matrix with elements equal to 1.

Using the multivariate Rao-Yu model given by (5) we can formulate the **best linear unbiased predictor (BLUP) estimator** of a small area parameter θ_{it} as a linear combination of fixed and random effects:

$$\tilde{\theta}_{iT} = \mathbf{x}_{iT}^T \tilde{\boldsymbol{\beta}} + \mathbf{m}_i^T \mathbf{G}_\mathbf{i} \mathbf{V}_i^{-1} (\mathbf{y}_\mathbf{i} - \mathbf{X}_\mathbf{i} \tilde{\boldsymbol{\beta}}) \tag{6}$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^\mathbf{T}\mathbf{V}^{-\mathbf{1}}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{V}^{-\mathbf{1}}\mathbf{y}$ is the generalized least squares estimator of $\boldsymbol{\beta}$ and $\mathbf{m}_\mathrm{i}$ is a vector with values equal to 1 for the area i for j-th variable and T-th period of time and zeroes for the other elements and $\mathbf{V}_\mathbf{i} = \mathbf{R}_\mathbf{i} + \mathbf{Z}_\mathbf{i}\mathbf{G}_\mathbf{i}\mathbf{Z}_\mathbf{i}^\mathbf{T}$. Note that in the multidimensional case, the i subscript is connected with r-dimensional vectors, where r is the number of dependent variables in the multidimensional model.

The procedure of obtaining EBLUP (Empirical BLUP) estimates is involved in the replacement of several variance components by their consistent estimators using Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) procedures (see e.g.: Rao and Molina (2015), pp.102–105).

Assuming that the vector of the estimators of the model variance parameters is $\tilde{\boldsymbol{\delta}} = (\tilde{\sigma}^{\mathbf{2}}, \tilde{\sigma}_{v}^{\mathbf{2}}, \tilde{\rho})$, the second-order approximation of **mean square error (MSE) of the EBLUP estimator** can be obtained using the following general formula (see e.g.: Rao, 2003, eq.(6.3.15)):

$$\widehat{MSE}\left(\tilde{\theta}_{it}(\tilde{\boldsymbol{\delta}})\right) = g_{1iT}(\tilde{\boldsymbol{\delta}}) + g_{2iT}(\tilde{\boldsymbol{\delta}}) + 2g_{3iT}(\tilde{\boldsymbol{\delta}})$$

where

$$g_{1it}(\tilde{\boldsymbol{\delta}}) = \mathbf{m}_{\mathbf{i}}^{\mathbf{T}}(\mathbf{G}_{\mathbf{i}} - \mathbf{G}_{\mathbf{i}}\mathbf{V}_{\mathbf{i}}^{-\mathbf{1}}\mathbf{G}_{\mathbf{i}})\mathbf{m}_{\mathbf{i}}$$

$$g_{2iT}(\tilde{\delta}) = \mathbf{d}_{\mathbf{i}}^{\mathbf{T}}\left(\sum_{i=1}^{m} \mathbf{X}_{\mathbf{i}}^{\mathbf{T}}\,\mathbf{V}_{\mathbf{i}}^{-\mathbf{1}}\mathbf{X}_{\mathbf{i}}\right)^{-1} \mathbf{d}_{\mathbf{i}}$$

$$g_{3iT}(\tilde{\delta}) = tr\left[\left(\frac{\partial \mathbf{b}_{iT}^{T}}{\partial \boldsymbol{\delta}}\right)\mathbf{V}_{i}\left(\frac{\partial \mathbf{b}_{iT}^{T}}{\partial \boldsymbol{\delta}}\right)^{T} \bar{V}(\widehat{\boldsymbol{\delta}})\right]$$

where

$$\mathbf{d}_{\mathbf{i}}^{\mathbf{T}} = \mathbf{x}_{\mathbf{iT}}^{\mathbf{T}} - \mathbf{b}_{\mathbf{i}}^{\mathbf{T}}\mathbf{X}_{\mathbf{i}}^{\mathbf{T}}$$

$$\mathbf{b}_{\mathbf{i}}^{\mathbf{T}} = \mathbf{m}_{\mathbf{i}}^{\mathbf{T}}\mathbf{G}_{\mathbf{i}}\mathbf{V}_{\mathbf{i}}^{-\mathbf{1}}$$

The detailed expressions of the derivatives $\mathbf{b}_{\mathrm{i}}$ can be found in Diallo (2014) and in Fay and Diallo (2012). For the multidimensional case one can also check the sae2 source code (Fay and Diallo (2015)) available at http://cran.r-project.org .

3. Results and discussion

In the application we were interested in the simultaneous estimation of *per capita income* (Y_1) and *expenditure* (Y_2) in Poland by region NUTS2, based on the sample data coming from the Polish Household Budget Survey. Multivariate models can fit to this kind of situations as they account for the correlation between several dependent variables. To improve the estimation quality we decided to formulate a bivariate small area model where the explanatory variables (X_1, X_2) were GDP per capita for regions coming from administrative registers. To obtain better estimates for the year 2011, we decided to utilize historical data coming from the years 2003-2011, which enabled "borrowing strength" not only across areas but also over time. This was possible by using the multivariate Rao-Yu model (5) based on cross-sectional and time-series data and obviously making use of the correlation between the predicted variables. The results obtained on the basis of these model were compared to the ones obtained from the respective univariate models for each response variable and to the classical Fay-Herriot model. The basis for the calculations was the micro data coming from the Polish Household Budget Survey and regional data from the GUS Local Data Bank.

At the first stage, direct estimates of both parameters of interest for 16 regions were calculated from the HBS sample together with their standard errors obtained by means of the Balanced Repeated Replication (BRR) technique. At the second stage the models were formulated and estimated from the data and finally EBLUP estimates were obtained as well as their MSE estimates. In order to evaluate the possible advantages of the estimators obtained by means of the bivariate Rao-Yu model (5) for j=1,2, we also estimated the parameters of simpler small area models and their corresponding EBLUPs. In particular, we additionally estimated the parameters of:

- the traditional Fay-Herriot model, “borrowing strength” only from other areas,
- univariate Rao-Yu model (eq. 3), “borrowing strength” from areas and over time.

In the computations conducted in R-project environment the packages *sae* and *sae2* have been applied. The sae2 package includes the implementation of the estimation procedure for the Rao-Yu model, which provides an extension of the basic type A model to handle time series and cross-sectional data (Rao (2003)). A special R macro has been developed that simplifies the reading of the input data from Excel spreadsheets, performing calculations for ordinary EBLUP models and Rao-Yu models for both uni- and two-dimensional cases. This macro has been helpful in obtaining the following: the diagnostics for EBLUP models, diagnostic charts for relative estimation errors (REE), relative estimation error reduction (REE reduction) and REE reduction due to time relationships. The macro presented in the appendix describes simple calculations for 3-dimensional Rao-Yu model using sae2 package and eblupRY function.

In Table 1 we show estimation results obtained for the two-dimensional model (5). For each dependent variable the estimates of fixed effects and the parameters of variance-covariance structure of the model, σ^2, σ_v^2 and ρ, are presented.

Table 1. Diagnostics of Rao-Yu two-dimensional model of *available income* and *expenditure* based on sample and administrative data

Variable	Coefficient estimates	Standard error	t-Statistics	P value
Submodel 1: **Y_1- Avail. Income 2003-2011**	$\boldsymbol{\sigma_1^2}$**= 1309.49** $\boldsymbol{\sigma_{1v}^2}$**=0.002 ρ=0.959 LogL=-1415.140**			
Intercept	76.455	49.170	1.555	0.120
X_1 GDP per capita	0.030	0.001	21.293	0.000
Submodel 2: **Y_2- Expenditure 2003-2011**	$\boldsymbol{\sigma_2^2}$**= 620.050** $\boldsymbol{\sigma_{2v}^2}$**=0.001 ρ=0.959 LogL=-1415.140**			
Intercept	226.620	34.046	6.656	0.000
X_2 GDP per capita	0.021	0.001	21.131	0.000

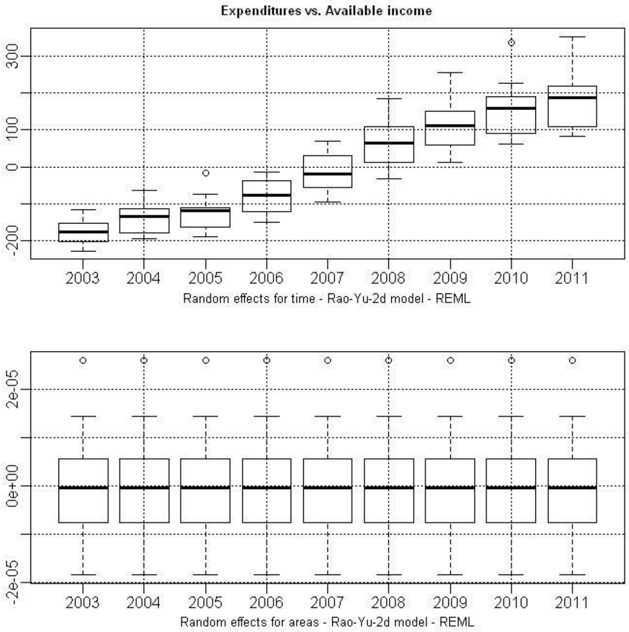

Figure 1. Distributions of random effects obtained for *available income/expenditure* 2-dimensional Rao-Yu model (top- **time effects**, bottom-**area effects**)

Source: Own calculations.

The model diagnostics indicate that the parameter σ_v^2 has only a small contribution to the variability of the model, which is mostly determined by time-related component. Figure 1 additionally shows the decomposition of random effects of the model (5) into two components: area effects (v_i) and time-area effects (u_{it}). In the figure it is possible to observe the impact and distribution of these effects over time. The random effects are consumed by time-related component while the influence of time-independent ones remains negligible.

Tables 2 and 3 show estimation results obtained for 16 NUTS2 regions in Poland. To assess the average relative efficiency and efficiency gains for each pair of estimators we utilized the following formulas (see: Rao (2003)):

$$\overline{EFF}_{est1/est2} = \frac{\overline{REE}(EST_1)}{\overline{REE}(EST_2)}, \quad \text{where: } \overline{REE}(EST) = \sum_{i=1}^{m} REE_i$$

Table 2 comprises the estimates of both variables of interest: per-capita *available income* and *expenditure* for regions, obtained using direct estimator, Rao-Yu EBLUP and Rao-Yu two-dimensional EBLUP. Each estimate is accompanied by its estimated precision: relative estimation error (REE) defined as the relative root MSE. The results obtained for income are in general better than the corresponding ones obtained for expenditure, which can be explained by higher dispersion of income. The improvement is also more evident for regions with poor direct estimates.

Table 2. Estimation results for *available income* and *expenditure* by region in the year 2011 (direct estimates and Rao-Yu EBLUPs – uni- and two-dimensional in PLN)

Region	Direct		Rao-Yu model		2d Rao-Yu model		Efficiency gains due to time effects [%]	
	Para-meter estimate	REE [%]	Para-meter estimate	REE [%]	Para-meter estimate	REE [%]	1D model	2D model
	Available income							
Dolnośląskie	1282.93	2.68	1321.88	1.99	1305.90	1.87	125.7	133.7
Kujawsko-Pomor.	1108.94	2.17	1111.89	1.95	1114.76	1.63	107.3	128.1
Lubelskie	1025.80	2.07	1027.82	1.81	1017.38	1.65	111.6	121.9
Lubuskie	1189.89	1.55	1192.57	1.38	1182.99	1.31	110.1	116.0
Łódzkie	1203.19	2.62	1224.93	2.00	1219.33	1.77	123.1	138.8
Małopolskie	1156.79	2.53	1167.22	2.02	1165.11	1.85	118.7	129.3
Mazowieckie	1622.96	2.02	1669.56	1.59	1649.08	1.42	126.0	141.3
Opolskie	1181.90	1.88	1178.66	1.64	1182.39	1.55	111.6	117.7
Podkarpackie	937.85	2.52	945.67	2.11	946.37	1.77	114.5	136.2
Podlaskie	1224.92	1.45	1208.41	1.34	1202.31	1.33	107.1	108.1
Pomorskie	1286.94	3.09	1298.67	2.20	1298.66	1.84	129.0	154.0
Śląskie	1215.44	0.95	1220.96	0.91	1222.23	0.84	104.3	112.1
Świętokrzyskie	1062.78	2.37	1057.54	2.05	1045.38	1.79	111.0	126.8
Warmińsko-Maz.	1096.87	2.63	1111.93	2.17	1099.61	2.01	115.4	124.8
Wielkopolskie	1135.02	2.73	1170.17	2.09	1148.01	1.84	121.0	137.3
Zachodniopomor.	1231.10	3.16	1226.36	2.27	1210.95	2.08	128.1	140.2
Average	**1185,21**	**2,28**	**1195,89**	**1,85**	**1188,15**	**1,66**	**117.4**	**130.6**
	Expenditure							
Dolnośląskie	1057.49	2.91	1086.02	2.06	1077.05	1.71	127.3	153.2
Kujawsko-Pomor.	922.75	1.16	924.81	1.10	924.16	1.03	104.1	111.7
Lubelskie	856.17	2.03	860.19	1.79	868.50	1.49	110.1	132.1
Lubuskie	975.64	2.13	983.78	1.77	996.11	1.39	115.7	146.9
Łódzkie	1042.70	1.96	1055.32	1.62	1049.65	1.41	116.4	133.3
Małopolskie	982.59	2.62	989.86	1.99	986.73	1.63	123.0	150.3
Mazowieckie	1308.35	1.62	1339.86	1.35	1331.49	1.19	119.4	135.6
Opolskie	1048.57	2.63	1048.66	1.99	1043.37	1.55	124.2	159.7
Podkarpackie	843.00	1.44	845.14	1.33	842.73	1.20	107.0	118.1
Podlaskie	903.42	4.58	889.59	2.70	947.43	1.71	142.4	124.6
Pomorskie	1061.25	1.85	1058.78	1.58	1058.49	1.42	113.2	125.9
Śląskie	1039.73	0.95	1043.57	0.89	1037.58	0.79	104.9	118.2
Świętokrzyskie	848.58	1.84	851.34	1.63	859.73	1.41	109.8	126.3
Warmińsko-Mazur.	870.30	3.06	880.69	2.42	888.67	1.92	116.8	146.7
Wielkopolskie	913.66	2.03	930.58	1.70	928.85	1.49	113.4	129.1
Zachodniopomor.	972.04	2.78	979.81	2.08	992.95	1.77	123.6	145.4
Average	**977,89**	**2,22**	**985,50**	**1,75**	**989,59**	**1,44**	**118.9**	**144.7**

Source: Own calculations.

The last two columns of Table 2 demonstrate "*efficiency gains due to time effects*" obtained as REE reduction for the Rao-Yu models with respect the ordinary EBLUP estimators based on Fay-Herriot model. It can be noticed that the proposed method overwhelms the classical approach by 30.6% for *available income* and by 44.7% for *expenditure*. This improvement was possible due to time relationships incorporated into Rao-Yu models which are not included into the classical Fay-Herriot ones.

As it can be noticed in Table 2, the average efficiency gains coming from time-correlation between random effects are on average doubled when the bivariate Rao-Yu model is taken into account - for *available income* they exceed 30 %, for *expenditure* are almost 45% (the corresponding values for the univariate Rao-Yu model were 14.4% and 18.9%). This improvement comes from the bivariate approach making use of the correlation between several dependent variables.

Table 3 presents in detail the efficiency gains coming from the application of 2d Rao-Yu model for both variables of interest. The EBLUPs based on this model were compared to the direct approach and to the EBLUPs obtained on the basis of simpler model-based approaches. Even with respect to the univariate Rao-Yu model one can observe substantial increase in precision (for income by 11.2% and for expenditure by 21.2%). Figures 2 and 4 present the empirical distributions of REEs for different small area estimators applied in the study while the distributions of REE reduction by means of the proposed model are presented in Figures 3 and 5. As it can be seen in the illustrations the bivariate approach can significantly improve the precision of the estimates.

Table 3. Relative efficiency [in%] for *available income* and *expenditure* in 2011

Region	$\text{EFF}_{\text{direct/Rao-Yu2d}}$		$\text{EFF}_{\text{EBLUP/Rao-Yu2d}}$		$\text{EFF}_{\text{RaoYu/Rao-Yu2d}}$	
	Available income	Expen-diture	Available income	Expen-diture	Available income	Expen-diture
Dolnośląskie	143.7	169.9	133.7	153.2	106.3	120.3
Kujawsko-Pomorskie	133.2	113.1	128.1	111.7	119.4	107.3
Lubelskie	125.4	136.5	121.9	132.1	109.2	119.9
Lubuskie	118.2	153.3	116.0	146.9	105.4	126.9
Łódzkie	147.7	138.7	138.8	133.3	112.8	114.5
Małopolskie	136.6	161.1	129.3	150.3	108.9	122.2
Mazowieckie	142.0	136.0	141.3	135.6	112.2	113.5
Opolskie	120.9	169.5	117.7	159.7	105.5	128.5
Podkarpackie	142.4	120.1	136.2	118.1	118.9	110.4
Podlaskie	109.4	268.2	108.1	124.6	101.0	157.7
Pomorskie	167.8	130.5	154.0	125.9	119.4	111.2
Śląskie	113.1	119.4	112.1	118.2	107.5	112.6
Świętokrzyskie	132.2	130.3	126.8	126.3	114.2	115.1
Warmińsko-Mazurskie	130.9	159.2	124.8	146.7	108.2	125.6
Wielkopolskie	148.2	136.0	137.3	129.1	113.5	113.9
Zachodniopomorskie	152.0	157.0	140.2	45.4	109.5	117.6
Average efficiency gain	**135,2**	**149,9**	**130.6**	**144.7**	**111.2**	**121.2**

Source: Own calculations.

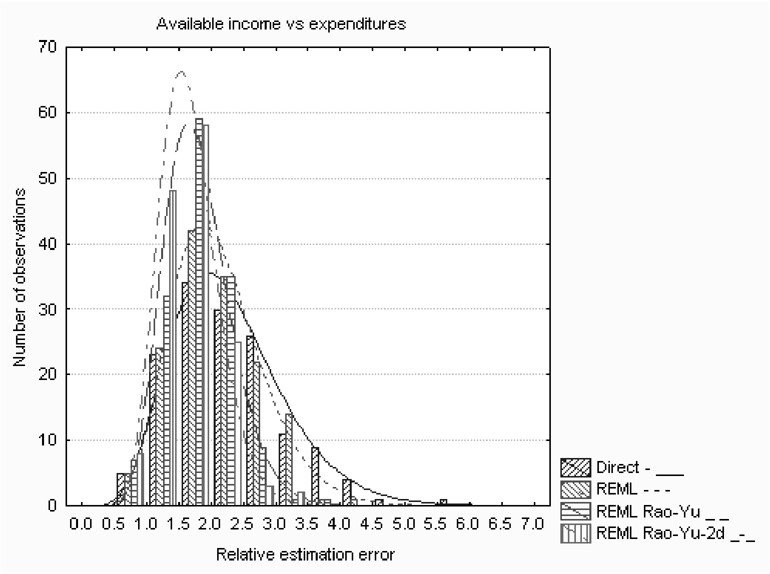

Figure 2. Distribution of REE for *available income* estimates in % in the years 2003-2011 (direct estimator and EBLUPs: ordinary and using Rao-Yu model – both 1 and 2-dimensional)

Source: Own calculations.

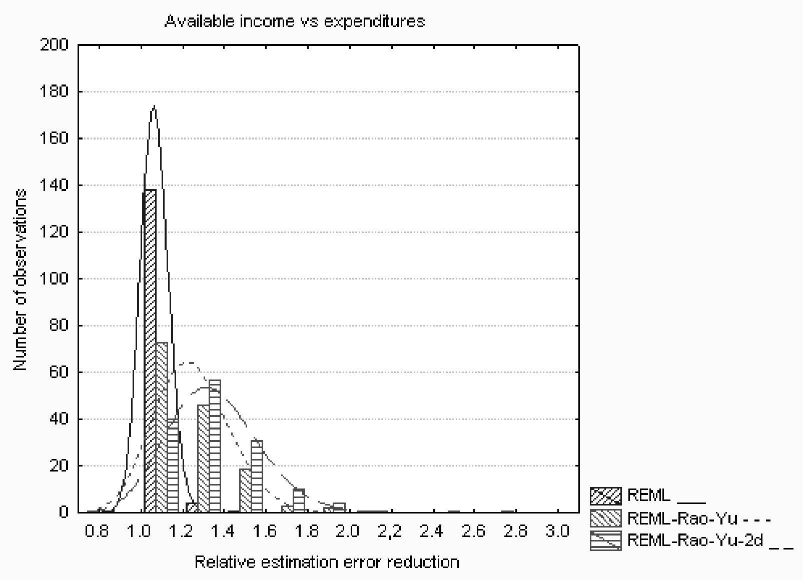

Figure 3. Distribution of REE reduction for *available income* estimates in the years 2003-2011 (direct estimator and EBLUPs: ordinary and using Rao-Yu model – both 1 and 2-dimensional)

Source: Own calculations.

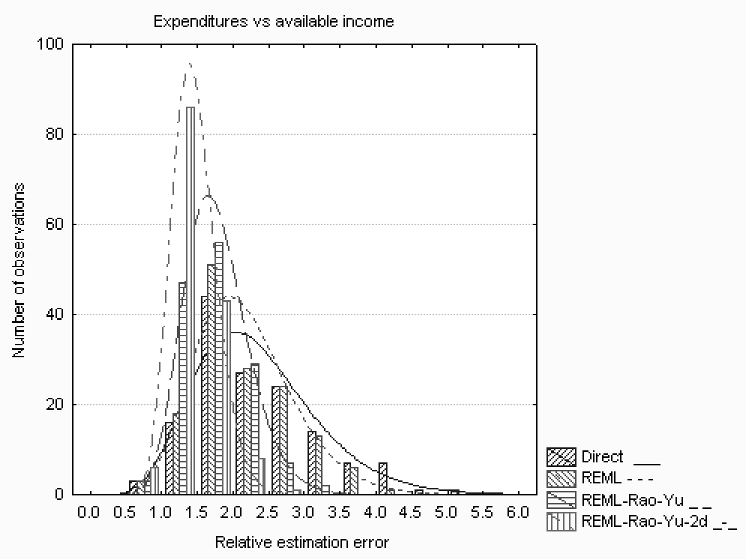

Figure 4. Distribution of REE for *expenditure* estimates in % in the years 2003-2011 (direct estimator and EBLUPs: ordinary and using Rao-Yu model – both 1 and 2-dimensional)

Source: Own calculations.

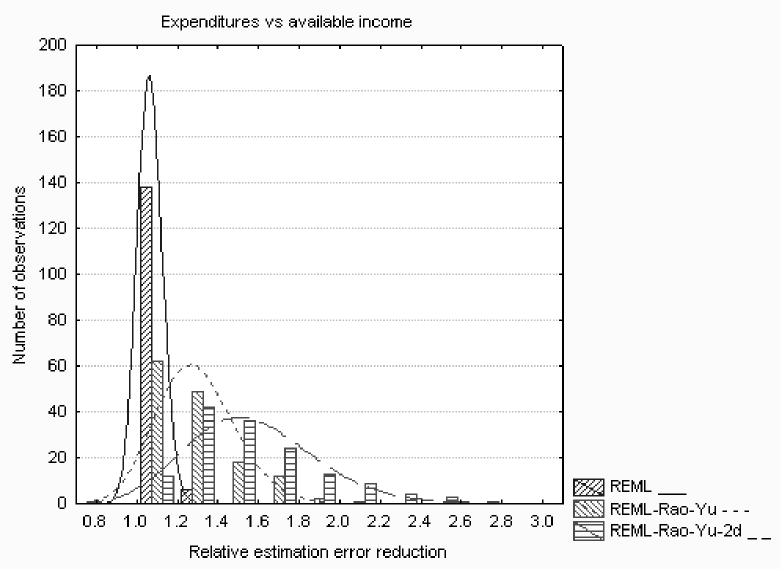

Figure 5. Distribution of REE reduction for *expenditure* estimates in the years 2003-2011 (direct estimator and EBLUPs: ordinary and using Rao-Yu model – both 1 and 2-dimensional)

Source: Own calculations.

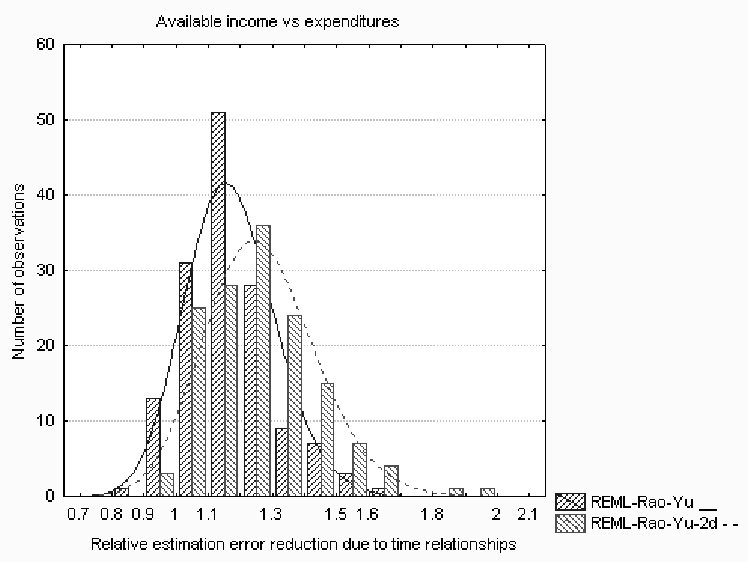

Figure 6. Distribution of REE reduction for *available income* using Rao-Yu EBLUP estimators due to time-related effects (referenced to the ordinary EBLUPs for one and two-dimensional models)

Source: Own calculations.

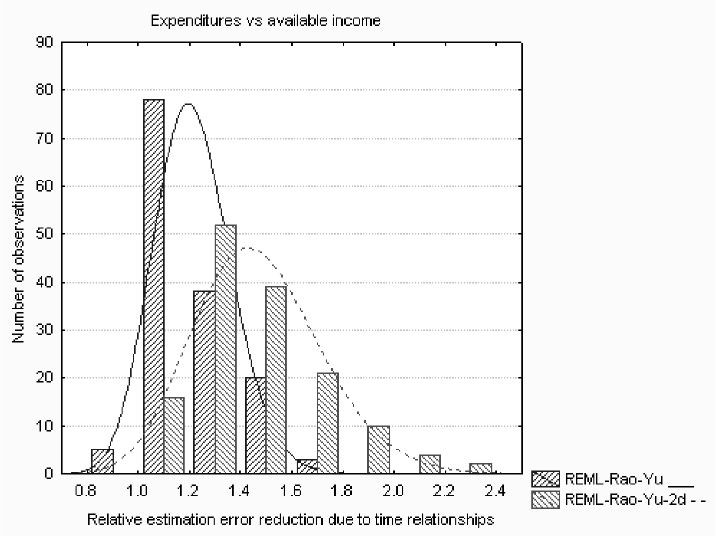

Figure 7. Distribution of REE reduction for *expenditure* using Rao-Yu EBLUP estimators due to time-related effects (referenced to the ordinary EBLUPs for one and two-dimensional models).

Source: Own calculations.

Table 4. Selected diagnostics for 2d Rao-Yu estimators referenced to the ordinary EBLUPs for different categories of income by region in the years 2003-2011

First dependent variable Y_1	Second dependent variable Y_2	$u_{c,(1,2)}$	$\rho_{(Y_1,Y_2)}$	$\frac{\mathrm{REE}_{EBLUP}}{\mathrm{REE}_{\mathrm{R-Y2d}}}$ for Y_1	$\frac{\mathrm{REE}_{EBLUP}}{\mathrm{REE}_{\mathrm{R-Y2d}}}$ for Y_2
Available	Expenditures	0.9464	0.9751	1.080	1.207
Available	Hired work	0.9800	0.9769	1.172	1.238
Available	Self-empl.	0.9321	0.8643	1.033	1.126
Available	Social benef.	0.6379	0.8067	1.001	1.098
Available	Retirm. pays	0.6261	0.8462	1.002	1.077
Available	Disabil. pens.	0.1912	-0.5435	0.999	1.048
Available	Family pens.	-0.0561	0.2464	0.996	1.057
Available	Other social	0.2896	-0.3227	0.997	1.032
Available	Unem.benef.	0.7253	-0.2659	1.010	1.092

Source: Own calculations.

Table 4 summarizes efficiency gains due to the application of two-dimensional models with respect to the classical Fay-Herriot one, which are especially visible for the cases of remarkable correlation between dependent variables Y_1 and Y_2. For the pairs presenting the Pearson correlation exceeding 0.9: *available income* and *expenditure* or *available income* and *income from hired work*, the relative estimation errors are significantly reduced. For example, the average REEs of EBLUPs of *income from hired work* are by 20% higher than the corresponding values obtained by means of the two-dimensional Rao-Yu model. It is worth noting that similar dependencies were observed for the univariate case of the Rao-Yu model (see e.g.: Jędrzejczak, Kubacki (2016)).

4. Conclusions

Multivariate small area models which make use of auxiliary information coming from repeated surveys can lead to significant quality improvements as they borrow information from time and space and additionally exploit the correlation between the considered parameters. In the paper, the advantages and limitations of bivariate small-area models for income distribution characteristics have been discussed. To assess the possible quality improvements, the multivariate Rao-Yu and Fay-Herriot models have been implemented and utilized to the estimation of income characteristics for the Polish households by region. Significant estimation error reductions have been observed for the variables that were evidently time-dependent and strictly correlated with each other and for the domains with relatively poor direct estimators. In the preliminary analysis of the models incorporating larger number of dependent variables also three- and fourdimensional Rao-Yu models have been specified but the gains from introducing additional dependent variables turned out to be rather ambiguous.

It would be advisable to check this method also for counties (poviats) and determine whether similar time-related relationships, which are observed for regions, could be observed for counties. The analysis presented here may also indicate that further comparisons between the Rao-Yu method and dynamic models, panel econometric models and nonlinear models should be conducted.

REFERENCES

BENAVENT, R., MORALES, D., (2015). Multivariate Fay–Herriot models for small area estimation, Computational Statistics & Data Analysis, Vol. 94, February 2016, pp. 372–390, http://www.sciencedirect.com/science/article/pii/S016794731500170X.

DATTA, G. S., FAY, R. E., GHOSH, M., (1991). Hierarchical and empirical Bayes multivariate analysis in small area estimation. In: Proceedings of Bureau of the Census 1991 Annual Research Conference, US Bureau of the Census, Washington, DC, pp. 63–79.

DATTA, G. S., GHOSH, M., NANGIA, N., NATARAJAN, K., (1996). Estimation of median income of four-person families: a Bayesian approach. In: Berry, D. A., Chaloner, K. M., Geweke, J. M. (Eds.), Bayesian Analysis in Statistics and Econometrics. Wiley, New York, pp. 129–140.

DATTA, G. S., DAY, B. MAITI, T., (1998). Multivariate Bayesian Small Area Estimation: An Application to Survey and Satellite Data, Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), Vol. 60, No. 3, Bayesian Analysis (Oct., 1998), pp. 344–362, http://sankhya.isical.ac.in/search/60a3/60a3ga.html.

DIALLO, M. S., (2014). Small Area Estimation under Skew-Normal Nested Error Models, 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.

FABRIZI, E., FERRANTE, M. R., PACEI, S., (2005). Estimation of poverty indicators at sub-national level using multivariate small area models, Statistics in Transition, December 2005, Vol. 7, No. 3, pp. 587–608.

FAY, R. E., (1987). Application of multivariate regression to small domain estimation, Small Area Statistics, Eds.: R. Platek, J. N. K., Rao, C. E., Sarndal, M. P., Singh. Wiley, New York, pp. 91–102.

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://ww2.amstat.org/sections/SRMS/Proceedings/y2012/Files/304438_73111.pdf.

FAY, R. E, DIALLO, M., PLANTY, M., (2013). Small Area Estimates from the National Crime Victimization Survey, [in:] Proc. Survey Research Methods Section of the American Statistical Association, pp. 1544–1557, http://ww2.amstat.org/sections/srms/Proceedings/y2013/Files/308383_80758.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.

FAY, R. 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://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/Fay_2012FCSM_I-B.pdf.

GERSHUNSKAYA, J., (2015). Combining Time Series and Cross-sectional Data for Current Employment Statistics Estimates, Proceedings of the Joint Statistical Meetings 2015 Survey Research Methods Section, Seattle, Washington, August 8 13, 2015, http://ww2.amstat.org/sections/srms/Proceedings/y2015/files/233962.pdf.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D., SANTAMARÍA, L., (2005). Analytic and bootstrap approximations of prediction errors under a multivariate Fay–Herriot model. Working Paper 05-49 (10), Statistics and Econometrics Series 061, Departamento de Estadística, Universidad Carlos III de Madrid, https://e-archivo.uc3m.es/bitstream/handle/10016/230/ws054910.pdf.

JANICKI, R., (2016). Estimation of the Difference of Small Area Parameters from Different Time Periods. Center for Statistical Research & Methodology Research Report Series (Statistics #RRS2016-01). U.S. Census Bureau, https://www.census.gov/srd/papers/pdf/RRS2016-01.pdf.

JĘDRZEJCZAK, A., KUBACKI, J., (2016). Estimation of Mean Income for Small Areas in Poland Using Rao-Yu Model, Acta Universitatis Lodziensis, Folia Oeconomica, 3 (322), pp. 37–53.

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://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/Li_2012FCSM_I-B.pdf.

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.

OECD, (2008). Growing Unequal? Income Distribution and Poverty in OECD Countries, http://www.oecd-ilibrary.org/social-issues-migration-health/growing-unequal_9789264044197-en.

OECD, (2011). Divided We Stand: Why Inequality Keeps Rising, OECD Publishing, http://dx.doi.org/10.1787/9789264119536-en.

PORTER, A. T., HOLAN, S. H., WIKLE, C. K., (2015). Multivariate spatial hierarchical Bayesian empirical likelihood methods for small area estimation. STAT, 4, 108–116, DOI: 10.1002/sta4.81, http://onlinelibrary.wiley.com/doi/10.1002/sta4.81/abstract.

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

RAO, J. N. K., MOLINA, I., (2015). Small Area Estimation (2nd edition). John Wiley & Sons, Inc., 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://ww2.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.

APPENDIX

The macro presented below describes simple calculations for 3-dimensional Rao-Yu model using sae2 package and eblupRY function.

```
library(sae2)
library(RODBC)
channel1 <- odbcConnectExcel("Input.xls")
command <- paste("select * from [Sheet1$]", sep="")
base <- sqlQuery(channel1, command)
data <- c(base$DOCHG_SD, base$D901_SD, base$D905_SD)
D <- 16
T <- 9
n_var <- 3
var_ptr <- vector(mode = "integer", length = D*T*n_var)
for(i in 1:D) {
  for(j in 1:n_var) {
    for(k in 1:T) {
      var_ptr[(i-1)*(T*n_var)+(j-1)*T+k] <- (j-1)*(D*T)+(i-1)*T+k
    }
  }
}
errmat <- diag((data[var_ptr])^2)
resultT.RY <- eblupRY(list(DOCHG_AVG ~ PKBPC_ABS, D901_AVG ~
PKBPC_ABS, D905_AVG ~ PKB_PC), D=D, T=T, vardir = errmat,data=base,
ids=base$WOJ, MAXITER = 500)
```