Tomasz Żądło
Uniwersytet Ekonomiczny w Katowicach

ON SOME PROBLEMS OF PREDICTION OF DOMAIN TOTAL IN LONGITUDINAL SURVEYS WHEN AUXILIARY INFORMATION IS AVAILABLE

Introduction

In the survey sampling the problem of estimation or prediction of subpopulations' (domains') characteristics has become very important issue. What is more, in the case of longitudinal surveys there is a possibility to increase the accuracy of the estimators or predictors by using information from other periods or even to estimate or predict subpopulation's characteristic for the period when the number of sampled domain elements equals zero. Domains with small or zero sample sizes are called small areas. In small area estimation empirical versions of Henderson's [1950] best linear unbiased predictors (BLUP) are widely used under different longitudinal area level models (see e.g. Rao [2003] chapter 8.3 and Rao and Yu [1994]). In the paper the class of unit level longitudinal models with auxiliary variables is proposed assuming that the population and the domains affiliation may change in time. In the paper the predictors which are empirical versions of Royall's [1976] BLUP under some special cases of the proposed model are derived. They can be used to predict the domain total based on any longitudinal data (including e.g. random and purposive samples, panel data and rotating samples) for any (including future) periods. Their mean squared errors (MSEs) and MSEs' estimators are also derived. In the Monte Carlo simulation study the problems of the accuracy of the predictor and biases of the MSE estimators are analyzed based on real data including several cases of model misspecification. The results of the simulation show that the proposed predictor and the proposed MSE estimator may perform very well even in some cases of model misspecification.

1. Basic notations

Let us introduce some notation presented earlier by Żądło [2009b]. In the paper longitudinal data for periods $t = 1,...,M$ are considered. In the period t the population of size N_t is denoted by Ω_t. The population in the period t is divided into D disjoint domains (subpopulations) Ω_{dt} of size N_{dt}, where $d = 1,...,D$. Let the set of population elements for which observations are available in the period t be denoted by s_t and its size by n_t. The set of domain elements for which observations are available in the period t is denoted by s_{dt} and its size by n_{dt}. Let: $\Omega_{rdt} = \Omega_{dt} - s_{dt}$, $N_{rdt} = N_{dt} - n_{dt}$.

Let M_{id} denotes the number of periods when the ith population element may be potentially observed in the dth domain (when the ith population element belongs to the dth domain). Let us denote the number of periods when the ith population element (which belongs to the dth domain) is observed by m_{id}. Let $m_{rid} = M_{id} - m_{id}$. We assume that the population may change in time and that one population element may change its domain affiliation in time (from technical point of view observations of some population element which change its domain affiliation are treated as observations of new population element). It means that i and t completely identify domain affiliation but additional subscript d will be needed as well. More about this assumptions will be written at the end of the next section.

The set of elements which belong at least in one of periods $t = 1,...,M$ to sets Ω_t is denoted by Ω and its size by N. Similarly, sets Ω_d, s, s_d, Ω_{rd} of sizes N_d, n, n_d, N_{rd} respectively are defined as sets of elements which belong at least in one of periods $t = 1,...,M$ to sets Ω_{dt}, s_t, s_{dt}, Ω_{rdt} respectively. The d*th domain of interest in the period of interest t^* will be additionally denoted by a symbol * in the subscript i.e. Ω_{d*t*}, and the set of elements which belong at least in one of periods $t = 1,...,M$ to sets Ω_{d*t*} will be denoted by Ω_{d*}.

Values of the variable of interest are realizations of random variables Y_{idj} for the ith population element which belongs to the dth domain in the period t_{ij}, where $i = 1,...,N$, $j = 1,...,M_{id}$, $d = 1,...,D$. The vector of size $M_{id} \times 1$ of random variables Y_{idj} for the ith population element which belongs to the dth domain will be denoted by $\mathbf{Y_{id}} = \left[Y_{idj} \right]$, where $j = 1,...,M_{id}$. Let us consider values of the variables of interest $Y_{i'd'j'}$ for the *i'*th population element which belongs to the *d'*th domain observed in periods $t_{i'j'}$, where *i'* = 1,...,*n*, *j'* = 1,...,$m_{i'd'}$, *d'* = 1,...,*D*. The vector of random variables $Y_{i'd'j'}$ (where *i'* = 1,...,*n*, *j'* = 1,...,$m_{i'd'}$, *d'* = 1,...,*D*) of size $m_{i'd'} \times 1$ will be denoted by $\mathbf{Y_{s\,i'd'}} = \left[Y_{i'd'j'} \right]$, where *j'* = 1,...,$m_{i'd'}$. The vector of random variables $Y_{i''d''j''}$ of size $m_{ri''d''} \times 1$ for the *i''*th population element which belongs to the *d''*th domain for observations which are not available in the sample is denoted by $\mathbf{Y_{r\,i''d''}} = \left[Y_{i''d''j''} \right]$, where *j''* = 1,...,$m_{ri''d''}$.

The proposed approach may be used to predict the domain total for any (past, current and future) periods. If the problem of prediction of the domain total for the future period is considered, the number of periods *M* includes future period or periods. What is more, in this case the division of the population into domains and values of the auxiliary variables in the future are assumed to be known.

2. Superpopulation model

We consider some class of superpopulation models (studied earlier by Żądło [2009b]) used for longitudinal data (compare Verbeke, Molenberghs, [2000]; Hedeker, Gibbons [2006]) which are – what is important for further considerations – special cases of the General Linear Model (GLM) and the General Linear Mixed Model (GLMM). The following two-stage model is assumed. Firstly:

$$\mathbf{Y_{id}} = \mathbf{Z_{id}}\boldsymbol{\beta}_{\mathbf{id}} + \mathbf{e_{id}}, \tag{1}$$

where *i* = 1,...,*N*; *d* = 1,...,*D*, $\mathbf{Y_{id}}$ is a random vector of size $M_{id} \times 1$, $\mathbf{Z_{id}}$ is known matrix of size $M_{id} \times q$, $\boldsymbol{\beta}_{\mathbf{id}}$ is a vector of unknown parameters of size $q \times 1$, $\mathbf{e_{id}}$ is a random component vector of size $M_{id} \times 1$. Vectors $\mathbf{e_{id}}$ (*i* = 1,...,*N*; *d* = 1,...,*D*) are independent with **0** vectors of expected values and variance-covariance matrices $\mathbf{R_{id}}$. Although $\mathbf{R_{id}}$ may depend on *i* it is often assumed that $\mathbf{R_{id}} = \sigma_e^2 \mathbf{I_{M_{id}}}$ where $\mathbf{I_{M_{id}}}$ is the identity matrix of rank M_{id}. Secondly, we assume that:

$$\boldsymbol{\beta}_{\mathbf{id}} = \mathbf{K}_{\mathbf{id}}\boldsymbol{\beta} + \mathbf{v}_{\mathbf{id}}, \tag{2}$$

where i = 1,...,N; d = 1,...,D, $\mathbf{K}_{\mathbf{id}}$ is known matrix of size $q \times p$, $\boldsymbol{\beta}$ is a vector of unknown parameters of size $p \times 1$, $\mathbf{v}_{\mathbf{id}}$ is a vector of random components of size $q \times 1$. It is assumed that vectors $\mathbf{v}_{\mathbf{id}}$ (i = 1,...,N; d = 1,...,D) are independent with **0** vectors of expected values and variance-covariance matrix $\mathbf{G}_{\mathbf{id}} = \mathbf{H}$ what means that $\mathbf{G}_{\mathbf{id}}$ does not depend on i.

Similar assumptions to (1) and (2) are presented by Verbeke, Molenberghs [2000, p. 20] but there are two differences. Firstly, in the book assumptions are made for profiles defined by elements. In this paper assumptions are made for profiles defined by elements and domains affiliation i.e. $\mathbf{Y}_{\mathbf{id}}$ (of size $M_{id} \times 1$) what allows to take the possibility of population changes in time into account. Secondly, in the book the assumptions are made only for the sampled elements (i.e. i = 1,...,n). In this paper they are made for all of population elements (i = 1,...,N).

Based on (1) and (2) it is obtained that:

$$\mathbf{Y}_{\mathbf{id}} = \mathbf{X}_{\mathbf{id}}\boldsymbol{\beta} + \mathbf{Z}_{\mathbf{id}}\mathbf{v}_{\mathbf{id}} + \mathbf{e}_{\mathbf{id}}, \tag{3}$$

where i = 1,...,N; d = 1,...,D, $\mathbf{X}_{\mathbf{id}} = \mathbf{Z}_{\mathbf{id}}\mathbf{K}_{\mathbf{id}}$ is known matrix of size $M_{id} \times p$. Let $\mathbf{V}_{\mathbf{id}} = D^2_{\xi}(\mathbf{Y}_{\mathbf{id}})$. Hence,

$$\mathbf{V}_{\mathbf{id}} = \mathbf{Z}_{\mathbf{id}}\mathbf{H}\mathbf{Z}^T_{\mathbf{id}} + \mathbf{R}_{\mathbf{id}}. \tag{4}$$

Let $\mathbf{A}_d$ be a column vector and $col_{1\le d\le D}(\mathbf{A}_d) = \begin{bmatrix} \mathbf{A}^T_{\mathbf{1}} & \ldots & \mathbf{A}^T_d & \ldots & \mathbf{A}^T_D \end{bmatrix}^T$ be a column vector obtained by stacking $\mathbf{A}_d$ vectors. Note that by stacking $\mathbf{Y}_{\mathbf{id}}$ vectors (i.e. $\mathbf{Y} = col_{1\le d\le D}(col_{1\le i\le N_d}(\mathbf{Y}_{\mathbf{id}}))$ from (3) we obtain the formula of the GLMM. Let $\mathbf{V} = D^2_{\xi}(\mathbf{Y})$. Hence,

$$\mathbf{V} = diag_{1\le d\le D}diag_{1\le i\le N_d}(\mathbf{V}_{\mathbf{id}}) \tag{5}$$

Unknown elements of **V** will be denoted by **δ**. Let, $\mathbf{Y}_{\mathbf{s}} = col_{1\le d\le D}col_{1\le i\le n_d}(\mathbf{Y}_{\mathbf{sid}})$, $\mathbf{V}_{\mathbf{ss}} = D^2_{\xi}(\mathbf{Y}_{\mathbf{s}})$, $\mathbf{V}_{\mathbf{ssid}} = D^2_{\xi}(\mathbf{Y}_{\mathbf{sid}})$. Hence,

$$\mathbf{V_{ss}} = diag_{1\le d\le D} diag_{1\le i\le n_d}(\mathbf{V_{ssid}}) = diag_{1\le d\le D} diag_{1\le i\le n_d}(\mathbf{Z_{sid}HZ_{sid}^T} + \mathbf{R_{sid}}) \quad (6)$$

where $\mathbf{Z_{sid}}$ is known matrix of size $m_{id} \times q$, $\mathbf{R_{sid}} = D_\xi^2(\mathbf{e_{sid}})$ and $\mathbf{e_{sid}}$ is $m_{id} \times 1$ random components vector.

3. EBLUP, its MSE AND MSE estimator

At the beginning let us compare BLUPs proposed by Henderson [1950] and Royall [1976]. Firstly, Royall derived the BLUP assuming the GLM which is generalization of the GLMM assumed by Henderson. Secondly, Royall predicts linear combination of **Y** given by $\theta = \boldsymbol{\gamma}^T \mathbf{Y}$ what is more general then linear combination of **β** and **v** given by $\theta_s = \mathbf{l^T}\boldsymbol{\beta} + \mathbf{m^T v}$ studied by Henderson. Thirdly, in both cases linear predictors are considered: $\hat{\theta} = \mathbf{g_s^T Y_s}$ by Royall [1976] and $\hat{\theta}_s = \mathbf{a^T Y_s} + b$ by Henderson, which forms are equivalent because $b =$ 0 under unbiasedness. Hence, Royall's BLUP may be treated as the generalization of Henderson's BLUP. In the paper the BLUP proposed by Royall is studied (and its empirical version – EBLUP) where the element *k* of the γ vector is given by:

$$\gamma_k = \begin{cases} 0 & \text{if} \quad i \notin \Omega_{d^*t^*} \\ 1 & \text{if} \quad i \in \Omega_{d^*t^*} \end{cases} \quad (7)$$

To obtain the BLUP of d^*th domain total in t^*th period and its MSE for model (3) general formulae proposed by Royall should be used with (7) and block-diagonal form of variance-covariance matrix (5). If the unknown parameters in the formula of the BLUP proposed by Royall are replaced by their estimates, two-stage predictor called the EBLUP is obtained. Kackar and Harville [1981] prove unbiasedness of empirical version of the BLUP proposed by Henderson under some weak assumptions. The proof of unbiasedness of empirical version of the BLUP proposed by Royall under similar weak assumptions (inter alia symmetric but not necessarily normal distribution of random components for the model assumed for the whole population), is presented in Żądło [2004]. The approximation of the MSE and its estimator for the empirical version of the BLUP proposed by Henderson are derived inter alia by Prasad and Rao [1990] and Datta and Lahiri [2000]. The approximation of the MSE and its estimator for the empirical version of the BLUP proposed by Royall are derived in Żądło [2009a] based on results presented in Datta and Lahiri [2000].

4. Special cases of superpopulation model

In the section we consider two special cases of the model (3). The first model is longitudinal random regression coefficient model similar to the one proposed in Dempster, Rubin and Tsutakawa [1981] and studied later e.g. in Moura and Holt [1999] and for one auxiliary variable in Prasad and Rao [1990]. Unlike the proposed longitudinal model, these authors only consider a model with domain-specific random effects (and for one period). We assume that:

$$Y_{idj} = (\beta_d + v_{id})x_{idj} + e_{idj} = \beta_d x_{idj} + v_{id} x_{idj} + e_{idj}, \tag{8}$$

where $i = 1,2,...,N$; $d = 1,2,...,D$, $j = 1,2,...,M_{id}$. Special case of (8) where

$$\forall_d \ \beta_d = \beta \tag{9}$$

will also be considered. What is more (similarly to Verbeke, Molenberghs [2000]), we assume that e_{idj} and v_{id} are mutually independent and $e_{idj} \sim (0, \sigma_e^2)$ and $v_{id} \sim (0, \sigma_v^2)$. Hence,

$$Cov_\xi(Y_{idj}, Y_{i'j'd'}) = \begin{cases} 0 & \text{if} \quad i \neq i' \vee d \neq d' \\ \sigma_e^2 + x_{idj}^2 \sigma_v^2 & \text{if} \quad i = i' \wedge j = j' \\ x_{idj} x_{i'j'd'} \sigma_v^2 & \text{if} \quad i = i' \wedge d = d' \wedge j \neq j' \end{cases}, \tag{10}$$

The second model is nested error regression models similar to the one proposed in Battese, Harter and Fuller [1988]. Unlike the proposed longitudinal model, these authors only consider a model with domain-specific random effects (and for one period). We assume that:

$$Y_{idj} = \mathbf{x_{idj}} \boldsymbol{\beta}_\mathbf{d} + v_{id} + e_{idj}, \tag{11}$$

where $\mathbf{x_{idj}} = \begin{bmatrix} x_{idj1} & x_{idj2} & \dots & x_{idjp} \end{bmatrix}$, e_{idj} and v_{id} are mutually independent and $e_{idj} \sim (0, \sigma_e^2)$ and $v_{id} \sim (0, \sigma_v^2)$. Special case of (11) where:

$$\forall_d \; \boldsymbol{\beta}_{\mathbf{d}} = \boldsymbol{\beta} \tag{12}$$

will also be considered. Hence,

$$Cov_{\xi}(Y_{idj}, Y_{i'j'd'}) = \begin{cases} 0 & \text{if} \quad i \neq i' \vee d \neq d' \\ \sigma_e^2 + \sigma_v^2 & \text{if} \quad i = i' \wedge j = j' \\ \sigma_v^2 & \text{if} \quad i = i' \wedge d = d' \wedge j \neq j' \end{cases} . \tag{13}$$

For all of the superpopulation models presented in this section the vector of unknown variance parameters will be denoted by $\boldsymbol{\delta} = \begin{bmatrix} \sigma_e^2 & \sigma_v^2 \end{bmatrix}^T$.

We have assumed that the population and the domain affiliation of population elements may change in time. Observations of new element of the population or observations of the population element after the change of its domain affiliation are treated as realizations of new profile (3). Hence, because of the covariance structure (5) where nonzero covariances are only within profiles, we assume the lack of correlation of observations for some population element before and after the change of the domain affiliation.

5. Prediction under a longitudinal random regression coefficient model

Based on Royall's theorem [1976], it is possible to derive the BLUP of the d^*th domain total in the t^*th (past, current or future) period and its MSE under longitudinal simple random regression coefficient model (8). They are given by:

$$\hat{\theta}_{BLU} = \sum_{i \in s_{d^*t^*}} Y_{id^*t^*} + \hat{\beta}_{d^*} \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*} + \sigma_v^2 \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*} b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} x_{id^*j} (Y_{id^*j} - x_{id^*j} \hat{\beta}_{d^*}) \tag{14}$$

where $\hat{\beta}_{d^*} = \left(\sum_{i=1}^{n_{d^*}} b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2 \right)^{-1} \left(\sum_{i=1}^{n_{d^*}} b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} Y_{id^*j} x_{id^*j} \right)$, $b_{id^*} = \sigma_e^2 + \sigma_v^2 \sum_{j=1}^{m_{id^*}} x_{id^*j}^2$

and

$$MSE_{\xi}(\hat{\theta}_{BLU}) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) \tag{15}$$

where:

$$g_1(\boldsymbol{\delta}) = N_{rd^*t^*}\sigma_e^2 + \sigma_v^2 \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*}^2 - \sigma_v^4 \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*}^2 b_{id*}^{-1} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2 \tag{16}$$

$$g_2(\boldsymbol{\delta}) = \left(\sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*} - \sigma_v^2 \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*} b_{id*}^{-1} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2 \right)^2 \left(\sum_{i=1}^{n_{d^*}} b_{id*}^{-1} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2 \right)^{-1} . \tag{17}$$

Let the unknown variance parameters in (14) be replaced by their maximum likelihood (ML) or restricted maximum likelihood (REML) estimates under normality. Hence, we obtain the two-stage predictor called EBLUP. Using general theorems proved in Żądło [2009a] it is possible to derive the formula of the MSE of the EBLUP and its estimators. Firstly, under assumptions presented in Żądło [2009a] (including the GLMM with block-diagonal variance-covariance matrix and normality of random components) the MSE in this case is given by:

$$MSE_{\xi}(\hat{\theta}_{EBLU}) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) + g_3^*(\boldsymbol{\delta}) + o(D^{-1}) \tag{18}$$

where $g_1(\boldsymbol{\delta})$ and $g_2(\boldsymbol{\delta})$ are given by (16) and (17) respectively and

$$g_3^*(\boldsymbol{\delta}) = \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*}^2 b_{id*}^{-3} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2 \left(I_{vv}^{(-1)}\sigma_v^4 - 2I_{ve}^{(-1)}\sigma_e^2\sigma_v^2 + I_{ee}^{(-1)}\sigma_e^4 \right) \tag{19}$$

and

$$I_{vv}^{(-1)} = 2b^{-1} \sum_{d=1}^{D} \sum_{i=1}^{n_d} b_{id}^{-2} \left(\sum_{j=1}^{m_{id}} x_{idj}^2 \right)^2 , \tag{20}$$

$$I_{ve}^{(-1)} = -2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{idj}^2\right), \tag{21}$$

$$I_{ee}^{(-1)} = 2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d}\left((m_{id}-1)\sigma_e^{-4} + b_{id}^{-2}\right), \tag{22}$$

and

$$b = \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}\left((m_{id}-1)\sigma_e^{-4} + b_{id}^{-2}\right)\right)\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{idj}^2\right)^2\right) +$$
$$-\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{idj}^2\right)\right)^2$$

Secondly, under general assumptions presented in Żądło [2009a] (including the GLMM with block-diagonal variance-covariance matrix and normality of random components) the approximately unbiased (its bias is $o(D^{-1})$) estimator of the MSE (18) for REML estimators of $\boldsymbol{\delta}$ in this case is given by:

$$M\hat{S}E_{\xi}(\hat{\theta}_{EBLU}) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) + 2g_3^*(\boldsymbol{\delta}) \tag{23}$$

and for ML estimators of $\boldsymbol{\delta}$ by

$$M\hat{S}E_{\xi}(\hat{\theta}_{EBLU}) = g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) + 2g_3^*(\hat{\boldsymbol{\delta}}) +$$
$$-\frac{1}{2}\left[\mathbf{I}_{\boldsymbol{\delta}}^{-1}(\hat{\boldsymbol{\delta}})col_{1\le k\le q}tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{\delta}})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}})\right]\right]^T\frac{\partial g_1(\hat{\boldsymbol{\delta}})}{\partial\boldsymbol{\delta}} \tag{24}$$

where $g_1(\hat{\boldsymbol{\delta}}), g_2(\hat{\boldsymbol{\delta}}), g_3^*(\hat{\boldsymbol{\delta}})$ are given by (16), (17), (19) respectively where $\boldsymbol{\delta}$ is replaced by $\hat{\boldsymbol{\delta}}$, $\mathbf{I}_{\boldsymbol{\delta}}^{-1} = \begin{bmatrix} I_{vv}^{(-1)} & I_{ve}^{(-1)} \\ I_{ve}^{(-1)} & I_{ee}^{(-1)} \end{bmatrix}$, where $I_{vv}^{(-1)}$, $I_{ve}^{(-1)}$, $I_{ee}^{(-1)}$ are given by (20), (21), (22) respectively, $col_{1\le k\le q}tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{\delta}})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}})\right]$ and $\frac{\partial g_1(\hat{\boldsymbol{\delta}})}{\partial\boldsymbol{\delta}}$ are given by

$$col_{1\le k\le q}tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right]=-\left[\frac{\sum_{d=1}^{D}\left(\sum_{i=1}^{n_d}b_{id}^{-1}\sum_{j=1}^{m_{id}}x_{idj}^2\right)^{-1}\left(\sum_{i=1}^{n_d}b_{id}^{-2}\sum_{j=1}^{m_{id}}x_{idj}^2\right)}{\sum_{d=1}^{D}\left(\sum_{i=1}^{n_d}b_{id}^{-1}\sum_{j=1}^{m_{id}}x_{idj}^2\right)^{-1}\left(\sum_{i=1}^{n_d}b_{id}^{-2}\left(\sum_{j=1}^{m_{id}}x_{idj}^2\right)^2\right)}\right] \quad (25)$$

and

$$\frac{\partial g_1(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}=\left[\frac{\partial g_1(\boldsymbol{\delta})}{\partial\sigma_e^2}\quad\frac{\partial g_1(\boldsymbol{\delta})}{\partial\sigma_v^2}\right]^T=$$

$$=\left[\frac{N_{rd^*t^*}-\sigma_v^4\sum_{i=1}^{N_{rd^*t^*}}x_{id^*t^*}^2b_{id*}^{-2}\sum_{j=1}^{m_{id^*}}x_{id^*j}^2}{\sum_{i=1}^{N_{rd^*t^*}}x_{id^*t^*}^2+\sigma_v^4\sum_{i=1}^{N_{rd^*t^*}}x_{id^*t^*}^2b_{id*}^{-1}\left(2\sum_{j=1}^{m_{id^*}}x_{id^*j}^2-b_{id*}^{-1}\left(\sum_{j=1}^{m_{id^*}}x_{id^*j}^2\right)^2\right)}\right] \quad (26)$$

respectively, where $\boldsymbol{\delta}$ is replaced by $\hat{\boldsymbol{\delta}}$.

Under assumptions (8) and (9) the equations presented above remain true but $\hat{\beta}_{d*}$ in (14) should be replaced by

$$\hat{\beta}=\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}b_{id}^{-1}\sum_{j=1}^{m_{id}}x_{idj}^2\right)^{-1}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}b_{id}^{-1}\sum_{j=1}^{m_{id}}Y_{idj}x_{idj}\right),$$

$g_2(\boldsymbol{\delta})$ given by (17) should by replaced by

$$g_2(\boldsymbol{\delta})=\left(\sum_{i=1}^{N_{rd^*t^*}}x_{id^*t^*}-\sigma_v^2\sum_{i=1}^{N_{rd^*t^*}}x_{id^*t^*}b_{id*}^{-1}\sum_{j=1}^{m_{id^*}}x_{id^*j}^2\right)^2\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}b_{id}^{-1}\sum_{j=1}^{m_{id}}x_{idj}^2\right)^{-1}$$

and $col_{1\le k\le q}tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right]$ given by (25) should be replaced by:

$$col_{1\le k\le q}tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right]=-\begin{bmatrix}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}b_{id}^{-1}\sum_{j=1}^{m_{id}}x_{idj}^2\right)^{-1}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}b_{id}^{-2}\sum_{j=1}^{m_{id}}x_{idj}^2\right)\\ \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}b_{id}^{-1}\sum_{j=1}^{m_{id}}x_{idj}^2\right)^{-1}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}b_{id}^{-2}\left(\sum_{j=1}^{m_{id}}x_{idj}^2\right)^2\right)\end{bmatrix}$$

6. Prediction under a longitudinal nested error regression model

Based on Royall's theorem [1976], it is possible to derive the BLUP of the d^*th domain total in t^*th (past, current or future) period and its MSE under longitudinal nested error regression coefficient model (11). The BLUP is given by:

$$\hat{\theta}_{BLU}=\sum_{i\in s_{d^*t^*}}Y_{id^*t^*}+\sum_{i=1}^{N_{rd^*t^*}}\mathbf{x}_{\mathbf{id^*t^*}}\hat{\boldsymbol{\beta}}_{\mathbf{d^*}}+\sigma_v^2\sum_{i=1}^{N_{rd^*t^*}}b_{id^*}^{-1}\sum_{j=1}^{m_{id^*}}(Y_{id^*j}-\mathbf{x}_{\mathbf{id^*j}}\hat{\boldsymbol{\beta}}_{\mathbf{d^*}}),\qquad(27)$$

where $b_{id^*}=\sigma_e^2+\sigma_v^2m_{id^*}$, $\hat{\boldsymbol{\beta}}_{\mathbf{d^*}}=\left(\sum_{i=1}^{n_{d^*}}b_{id^*}^{-1}\mathbf{X}_{\mathbf{sid^*}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid^*}}\right)^{-1}\left(\sum_{i=1}^{n_{d^*}}b_{id^*}^{-1}\mathbf{X}_{\mathbf{sid^*}}^{\mathbf{T}}\mathbf{Y}_{\mathbf{sid^*}}\right)$ and $\mathbf{X}_{\mathbf{sid^*}}$ is $m_{id^*}\times p$ known matrix of auxiliary variables. The MSE of the BLUP (27) is given by general formula (15), where

$$g_1(\boldsymbol{\delta})=N_{rd^*t^*}\left(\sigma_e^2+\sigma_v^2\right)-\sigma_v^4\sum_{i=1}^{N_{rd^*t^*}}b_{id^*}^{-1}m_{id^*},\qquad(28)$$

$$g_2(\boldsymbol{\delta})=\left(\sum_{i=1}^{N_{rd^*t^*}}\mathbf{x}_{\mathbf{id^*t^*}}-\sigma_v^2\sum_{i=1}^{N_{rd^*t^*}}b_{id^*}^{-1}\sum_{j=1}^{m_{id^*}}\mathbf{x}_{\mathbf{id^*j}}\right)\left(\sum_{i=1}^{n_{d^*}}b_{id^*}^{-1}\mathbf{X}_{\mathbf{sid^*}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid^*}}\right)^{-1}\times$$

$$\times\left(\sum_{i=1}^{N_{rd^*t^*}}\mathbf{x}_{\mathbf{id^*t^*}}-\sigma_v^2\sum_{i=1}^{N_{rd^*t^*}}b_{id^*}^{-1}\sum_{j=1}^{m_{id^*}}\mathbf{x}_{\mathbf{id^*j}}\right)^{\mathbf{T}}\qquad(29)$$

If the unknown variance parameters in (27) are replaced by their ML or REML estimates under normality we obtain the EBLUP with the MSE given by general formula (18), where $g_1(\boldsymbol{\delta})$ and $g_2(\boldsymbol{\delta})$ are given by (28) and (29) respectively and

$$g_3^*(\boldsymbol{\delta}) = \sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-3} m_{id^*} \left(I_{vv}^{(-1)} \sigma_v^4 - 2 I_{ve}^{(-1)} \sigma_e^2 \sigma_v^2 + I_{ee}^{(-1)} \sigma_e^4 \right) \quad (30)$$

where

$$I_{vv}^{(-1)} = 2b^{-1} \sum_{d=1}^{D} \sum_{i=1}^{n_d} b_{id}^{-2} m_{id}^2 , \quad (31)$$

$$I_{ve}^{(-1)} = -2b^{-1} \sum_{d=1}^{D} \sum_{i=1}^{n_d} b_{id}^{-2} m_{id} , \quad (32)$$

$$I_{ee}^{(-1)} = 2b^{-1} \sum_{d=1}^{D} \sum_{i=1}^{n_d} \left((m_{id} - 1) \sigma_e^{-4} + b_{id}^{-2} \right), \quad (33)$$

and $b = \left(\sum_{d=1}^{D} \sum_{i=1}^{n_d} \left((m_{id} - 1) \sigma_e^{-4} + b_{id}^{-2} \right) \right) \left(\sum_{d=1}^{D} \sum_{i=1}^{n_d} b_{id}^{-2} m_{id}^2 \right) - \left(\sum_{d=1}^{D} \sum_{i=1}^{n_d} b_{id}^{-2} m_{id} \right)^2$

The approximately unbiased (its bias is $o(D^{-1})$) estimator of the MSE of the EBLUP for the REML estimators of $\boldsymbol{\delta}$ is given by (23) and for the ML estimators of $\boldsymbol{\delta}$ by (24) where $g_1(\hat{\boldsymbol{\delta}})$, $g_2(\hat{\boldsymbol{\delta}})$, $g_3^*(\hat{\boldsymbol{\delta}})$ are given by (28), (29), (30) respecively where $\boldsymbol{\delta}$ is replaced by $\hat{\boldsymbol{\delta}}$, $\mathbf{I}_{\boldsymbol{\delta}}^{-1} = \begin{bmatrix} I_{vv}^{(-1)} & I_{ve}^{(-1)} \\ I_{ve}^{(-1)} & I_{ee}^{(-1)} \end{bmatrix}$, where $I_{vv}^{(-1)}$, $I_{ve}^{(-1)}$, $I_{ee}^{(-1)}$ are given by (31), (32), (33) respectively, $col_{1 \le k \le q} tr \left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{\delta}}) \frac{\partial}{\partial \delta_k} \mathbf{I}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}) \right]$ and $\frac{\partial g_1(\hat{\boldsymbol{\delta}})}{\partial \boldsymbol{\delta}}$ are given by

$$col_{1\le k\le q} tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right]=$$

$$=-\begin{bmatrix} \sum_{d=1}^{D} tr\left(\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\left(\sum_{i=1}^{n_d} b_{id}^{-2}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right) \\ \sum_{d=1}^{D} tr\left(\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\left(\sum_{i=1}^{n_d} b_{id}^{-2}m_{id}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right) \end{bmatrix} \tag{34}$$

and

$$\frac{\partial g_1(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}=\begin{bmatrix}\frac{\partial g_1(\boldsymbol{\delta})}{\partial\sigma_e^2} & \frac{\partial g_1(\boldsymbol{\delta})}{\partial\sigma_v^2}\end{bmatrix}^T=\begin{bmatrix} N_{rd^*t^*}-\sigma_v^4\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-2}m_{id^*} \\ N_{rd^*t^*}+\sigma_v^4\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1}m_{id^*}\left(2-b_{id^*}^{-1}m_{id^*}\right) \end{bmatrix} \tag{35}$$

respectively, where $\boldsymbol{\delta}$ is replaced by $\hat{\boldsymbol{\delta}}$.

Under assumptions (11) and (12) the equations presented above remain true but $\hat{\boldsymbol{\beta}}_{\mathbf{d}^*}$ in (27) should be replaced by

$$\hat{\boldsymbol{\beta}}=\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{Y}_{\mathbf{sid}}\right),$$

and $g_2(\boldsymbol{\delta})$ given by (29) should be replaced by

$$g_2(\boldsymbol{\delta})=\left(\sum_{i=1}^{N_{rd^*t^*}}\mathbf{x}_{\mathbf{id^*t^*}}-\sigma_v^2\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1}\sum_{j=1}^{m_{id^*}}\mathbf{x}_{\mathbf{id^*j}}\right)\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\times$$

$$\times\left(\sum_{i=1}^{N_{rd^*t^*}}\mathbf{x}_{\mathbf{id^*t^*}}-\sigma_v^2\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1}\sum_{j=1}^{m_{id^*}}\mathbf{x}_{\mathbf{id^*j}}\right)^{\mathbf{T}}$$

and $col_{1\le k\le q} tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right]$ given by (34) should be replaced by:

$$col_{1\le k\le q} tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right]=$$

$$=-\begin{bmatrix} tr\left(\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right) \\ tr\left(\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}m_{id}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right) \end{bmatrix}$$

7. Simulation analyses

The limited Monte Carlo simulation analyses are based on real data on N = 314 Polish poviats (what is NUTS 4 level) excluding cites with poviat's rights for M = 4 years 2005-2008. Data are available at the website of the Polish Central Staistical Office – www.stat.gov.pl. The problem is to estimate subpopulations (domains) totals for D = 6 regions (NTS 1 level) in 2008. The variable of interest is poviats' own incomes (in PLN) and the auxiliary variable is the population size in poviats (in persons). Two simulations are conducted using R (R Development Core Team [2011]). In the simulations the accuracy of the proposed predictor is compared with accuracies of two calibration estimators [Rao 2003, pp. 17-18] which will be denoted by GREG1 and GREG2. Both calibration estimators are of the form $\hat{\theta}_{d^*t^*}^{GREG}=\sum_{i\in s_{d^*t^*}} w_{sit^*}y_{i^*t^*}$ but weights w_{sit^*} are solutions for GREG1 of

$$\begin{cases}\sum_{i\in s_{d^*t^*}}\frac{(w_{sit^*}-1/\pi_{it^*})^2}{1/\pi_{it^*}}\to\min \\ \sum_{i\in s_{d^*t^*}} w_{sit^*}\mathbf{x}_{\mathbf{id}^*\mathbf{t}^*}=\sum_{i\in\Omega_{d^*t^*}}\mathbf{x}_{\mathbf{id}^*\mathbf{t}^*}\end{cases}$$

and for GREG2 are subsequence of weights obtained as a solution of

$$\begin{cases} \sum_{i \in s_{t^*}} \dfrac{(w_{sit^*} - 1/\pi_{it^*})^2}{1/\pi_{it^*}} \to \min \\ \sum_{i \in s_{t^*}} w_{sit^*} \mathbf{x}_{\mathbf{id^*t^*}} = \sum_{i \in \Omega_{t^*}} \mathbf{x}_{\mathbf{id^*t^*}} \end{cases}$$

where π_{it^*} are inclusion probabilities in the period t^*. These calibration estimators are classic model-assisted estimators which are known as good alternatives for model-based methods especially in the case of possible model misspecification.

The first simulation is design-based. In this case a sample of size $n = 79$ elements (c.a. 25% of population size) is balanced panel sample (each sampled element is observed in all of 4 periods), which is drawn at random in the first period with inclusion probabilities proportional to the value of the auxiliary variable in this period. For this sample size it was possible to estimate all of domain totals in each iteration even using direct estimators GREG1 and GREG2. The number of samples drawn in the simulation equals 10 000.

The second simulation is model based. In this case one sample is drawn using the sample design described above what gives the division of the population into the sampled and unsampled part. Then 10 000 populations are generated using model (11) – (with one auxiliary variable and constant) with parameters computed (REML) based on real, whole population data and random components with the following distributions: in the model denoted in the simulation as N case – normal distributions of both random components, U case – uniform distributions of both random components and E case – "shifted" exponential distribution of both random components. What is more, to study the problem of model misspecification, equations for linear model are used but 10 000 population are generated based on modified model (11) where instead of the auxiliary variable its natural logarithm is used. Both random components have the following distributions: Nm case – normal distribution, Um – uniform distribution and Em – "shifted" exponential distribution.

What is important, the predictor presented for the model (11) simplifies to the BLUP (i.e. does not depend on unknown variance parameters) for the balanced sample. Hence, in the equation of the MSE estimator the $g_3^*(\boldsymbol{\delta})$ element is omitted.

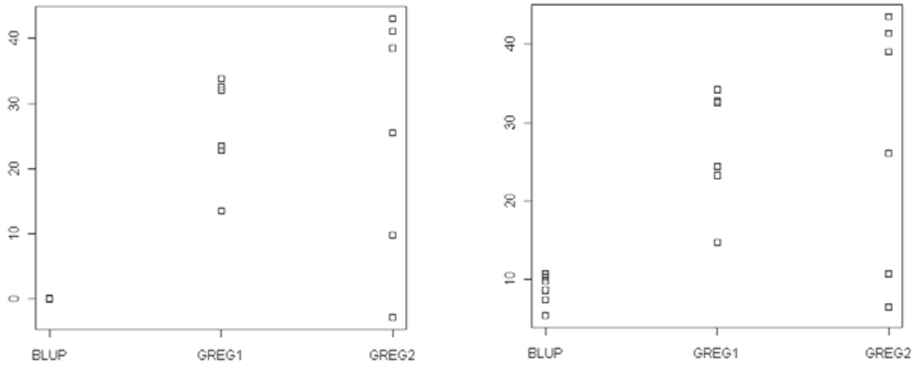

Fig. 1. Relative model-biases (on the left) and relative model RMSE (on the right) for N case (in %) for six domains

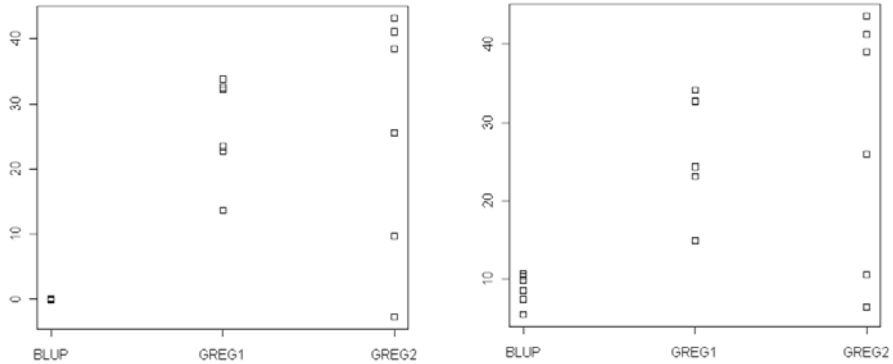

Fig. 2. Relative model-biases (on the left) and relative model RMSE (on the right) for U case (in %) for six domains

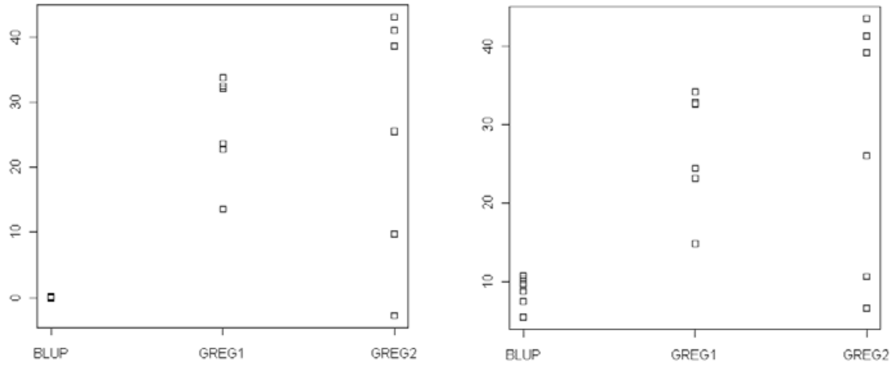

Fig. 3. Relative model-biases (on the left) and relative model RMSE (on the right) for E case (in %) for six domains

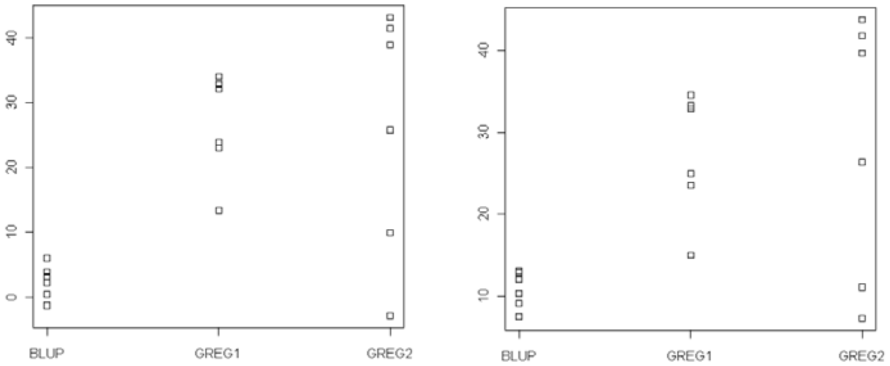

Fig. 4. Relative model-biases (on the left) and relative model RMSE (on the right) for Nm case (in %) for six domains

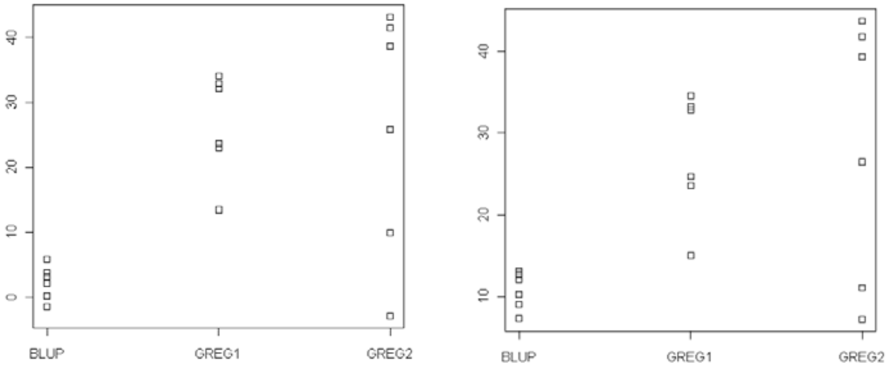

Fig. 5. Relative model-biases (on the left) and relative model RMSE (on the right) for Um case (in %) for six domains

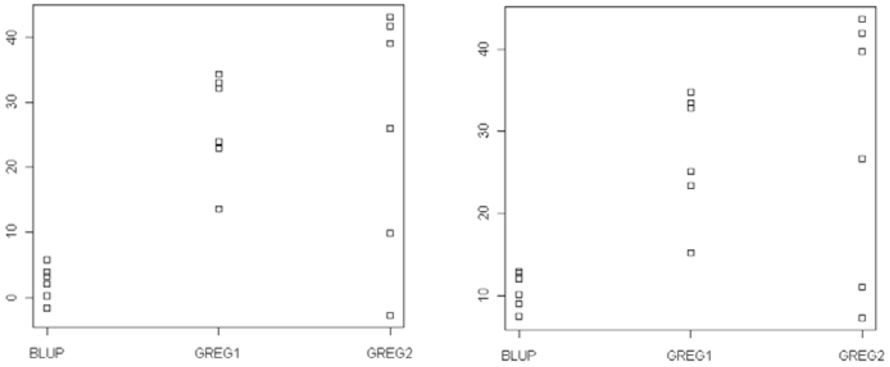

Fig. 6. Relative model-biases (on the left) and relative model RMSE (on the right) for Em case (in %) for six domains

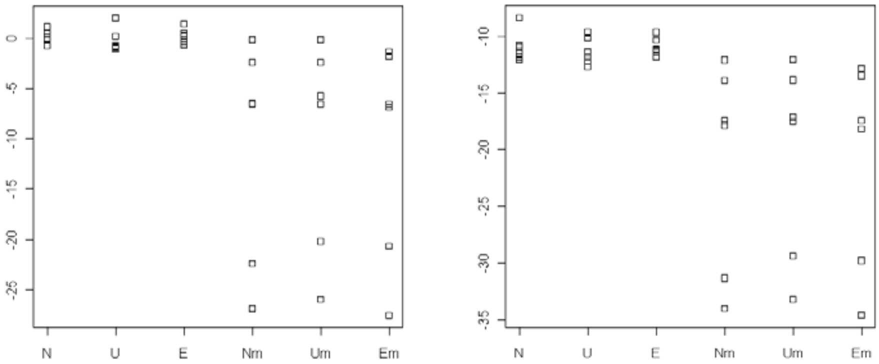

Fig. 7. Relative biases of MSE estimators of BLUP for REML (on the left) and ML (on the right) estimates of **δ** (in %) for six domains

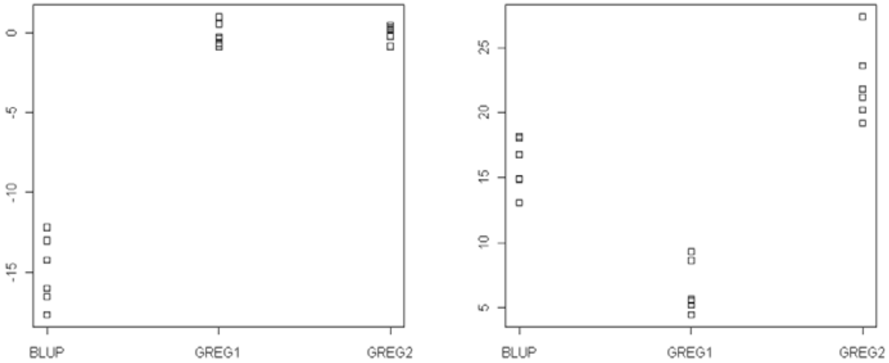

Fig. 8. Relative design-biases (on the left) and relative design RMSE (on the right) for six domains in %

Each point on the figures presents value of some statistic for one out of *D* = 6 domains. Comparing prediction accuracy of the BLUP and GREG1 and GREG2 (see Figures 1-6) it should be noted that the BLUP is better then both GREG estimators even in the cases of model misspecification. The absolute value of its relative bias does not exceed 10% for all of the considered cases of model misspecification (see Figures 4-6). The bias of the considered MSE estimator under normality is $o(D^{-1})$ (as proved in Żądło [2009a]) but for the data interesting case is studied – the number of domains is very small, it equals *D* = 6. It is known that for the big number of domains *D* the differences between biases of REML and ML MSE esimators (given by general equations (23) and (24) respectively) are small. In the simulation study the REML MSE estimator is (see Figure 7) less biased in all of the considered cases, what may have occured due to the relatively big (comparing with *D*) loss of degrees of freedom when ML method is used instead of REML to estimate $\boldsymbol{\delta}$. Let us limit further consideration to the REML MSE estimator. The absolute values of relative biases of MSE estimators (see Figure 7) are small not only under normality assumptions (N case), under which they were derived, but also for U and E cases. For the proof of robustness of some MSE estimators of the EBLUP of the form of Henderson's BLUP in some cases of model misspecification see Lahiri and Rao [1995]. The maximum absolute value of the relative biases of MSE estimators are less then 2,1% for these cases. When the true model is non-linear (the Nm, Um, Em cases) the biases obtained in the simulation are larger.

Results presented on the Figure 8 show that the design bias of the BLUP is larger than the design bias of both GREG1 and GREG2. Comparing the design accuracy for the data and large sample size (n = 79), the BLUP is better than GREG2 but worse than GREG1.

Conclusions

In the paper the EBLUP for longitudinal data is proposed. The predictor allows to predict the domain total for any (past, current, future) period assuming that population and domain affiliation of population elements may change in time. Its MSE is also derived and some MSE estimator is proposed. Its accuracy is analyzed for real data in the Monte Carlo simulation study.

Literature

Battese G.E., Harter R.M., and Fuller W.A. (1988): *An Error-components Model for Prediction of County Crop Areas Using Survey and Satellite Data*. "Journal of the American Statistical Association", No. 83.

Datta G.S., Lahiri P. (2000): *A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems*. "Statistica Sinica", No. 10.

Dempster A.P., Rubin D.B., Tsutakawa R.K. (1981): *Estimation in Covariance Components Models.* "Journal of the American Statistical Association", Vol. 76, No. 374.

Hedeker D., Gibbons R.D. (2006): *Longitudinal Data Analysis.* John Wiley and Sons, New Jersey.

Henderson C.R. (1950): *Estimation of Genetic Parameters (Abstract).* "Annals of Mathematical Statistics", No. 21.

Kackar R.N., Harville D.A. (1981): *Unbiasedness of Two-stage Estimation and Prediction Procedures for Mixed Linear Models*. "Communications in Statistics" Ser. A,10.

Lahiri P., Rao J.N.K. (1995): *Robust Estimation of Mean Squared Error of Small Area Estimators*. "Journal of the American Statistical Association", No. 90.

Moura F.A.S., Holt D. (1999): *Small Area Estimation Using Multilevel Models.* "Survey Methodology", No. 25.

Prasad N.G.N, Rao J.N.K. (1990): *The Estimation of Mean the Mean Squared Error of Small Area Estimators*. "Journal of the American Statistical Association", No. 85.

R Development Core Team (2011): *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna.

Rao J.N.K. (2003): *Small Area Estimation*. John Wiley and Sons, New York.

Rao J.N.K., Yu M. (1994): *Small-area Estimation by Combining Time-series and Cross-sectional Data*. "The Canadian Journal of Statistics", No. 22.

Royall R.M. (1976): *The Linear Least Squares Prediction Approach to Two-stage Sampling.* "Journal of the American Statistical Association", No. 71.

Verbeke G., Molenberghs G. (2000): *Linear Mixed Models for Longitudinal Data.* Springer, New York.

Żądło T. (2004): *On Unbiasedness of Some EBLU Predictor*. In: *Proceedings in Computational Satistics 2004.* Ed. J. Antoch. Physica-Verlag, Heidelberg-New York.

Żądło T. (2009a): *On MSE of EBLUP.* Statistical Papers. Springer, 50.

Żądło T. (2009b): *On Prediction of Domain Totals Based on Unbalanced Longitudinal Data*. In: *Survey Sampling in Economic and Social Research.* Eds. J. Wywiał, T. Żądło. Wydawnictwo AE, Katowice.

O PEWNYCH PROBLEMACH PREDYKCJI WARTOŚCI GLOBALNEJ W DOMENIE W BADANIACH WIELOOKRESOWYCH, GDY SĄ DOSTĘPNE INFORMACJE O ZMIENNYCH DODATKOWYCH

Streszczenie

W artykule wyprowadzono postacie najlepszych liniowych nieobciążonych predyktorów przy założeniu pewnych modeli będących uogólnieniami na przypadek danych

przekrojowo-czasowych modeli znanych z literatury statystyki małych obszarów. Ponadto wyprowadzono postacie błędów średniokwadratowych empirycznych wersji tych predyktorów oraz zaproponowano ich estymatory. W symulacji Monte Carlo porównywano dokładność zaproponowanego predyktora z dwoma ogólnymi estymatorami regresyjnymi po planie losowania i po modelu nadpopulacji (także w różnych przypadkach złej specyfikacji modelu). Ponadto analizowano obciążenia zaproponowanych estymatorów błędu średniokwadratowego.