ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

*Tomasz Żądło**

ON ACCURACY OF SOME EBLU PREDICTOR

ABSTRACT. In the paper we analyze the accuracy of the empirical best linear unbiased predictor (EBLUP) of the domain total (see Royall, 1976) assuming a special case of the general linear mixed model. To estimate the mean square error (MSE) of the EBLUP we use the results obtained by Datta and Lahiri (2000) for the predictor proposed by Henderson (1950) and adopt them for the predictor proposed by Royall (1976). In a simulation study we study real data on Polish farms from Dąbrowa Tarnowska region.

Key words: small area estimation, empirical best linear unbiased predictors, general mixed linear model.

I. BASIC NOTATIONS

The finite population Ω consists of N units, each of which has a value of a target variable y associated with it. The population vector of y's is $\mathbf{y}=\left[y_1,y_2,...,y_N\right]^T$ and it is treated as the realization of a random vector $\mathbf{Y}=\left[Y_1,Y_2,...,Y_N\right]^T$. The joint distribution of $\mathbf{Y}$ is denoted by ξ. From the population of N units, a sample s of n units is selected, and the y values of the sample units are observed. For any sample s we can reorder the population vector $\mathbf{y}$ so that the first n elements are those in the sample: $\mathbf{y}=\left[\mathbf{y}_s^T,\mathbf{y}_r^T\right]^T$ where $\mathbf{y}_s$ is the n-vector of observed values and $\mathbf{y}_r$ is the N_r-vector of unobserved values where $N_r=N-n$. The set of unsampled elements is denoted by $\Omega_r=\Omega-s$. Hence, the vector $\mathbf{Y}$ can be reordered as follows: $\mathbf{Y}=\left[\mathbf{Y}_s^T,\mathbf{Y}_r^T\right]^T$. The population is divided into D domains Ω_d (d=1,...,D), each of size N_d (d = 1,...,D). Let $s_d=\Omega_d\cap s$ consists of n_d elements (where

* Ph.D., Department of Statistics, University of Economics in Katowice.

n_d may be random), $\Omega_{rd} = \Omega_d - s_d$ and $N_{rd} = N_d - n_d$. For the domain of interest we a add star to the subscript d, for example the domain of interest is denoted by Ω_{d*} and its size by N_{d*}.

Let us introduce the general linear model (GLM). We assume that:

$$\begin{cases} E_\xi(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \\ D_\xi^2(\mathbf{Y}) = \mathbf{V} \end{cases}, \tag{1}$$

where $\mathbf{X}$ is a $N \times p$ matrix of values of p auxiliary variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters and $\mathbf{V}$ is a variance-covariance matrix depending on some parameters $\boldsymbol{\delta} = \left[\delta_1, ..., \delta_q\right]^T$. If the population elements are rearranged so that the first n elements of $\mathbf{Y}$ and the first n rows of $\mathbf{X}$ are for units in the sample, then $\mathbf{X} = \left[\mathbf{X}_s^T \quad \mathbf{X}_r^T\right]^T$, $\mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}$ where $\mathbf{X}_s$ is $n \times p$, $\mathbf{X}_r$ is $N_r \times p$, $\mathbf{V}_{ss}$ is $n \times n$, $\mathbf{V}_{rr}$ is $N_r \times N_r$, $\mathbf{V}_{sr}$ is $n \times N_r$ and $\mathbf{V}_{rs} = \mathbf{V}_{sr}^T$.

Introduce the general linear mixed model (GLMM) which is a special case of (1):

$$\begin{cases} \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e} \\ E_\xi(\mathbf{e}) = \mathbf{0} \wedge E_\xi(\mathbf{v}) = \mathbf{0} \\ D_\xi^2 \begin{bmatrix} \mathbf{v} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \end{cases} \tag{2}$$

where $\mathbf{Z}$ is known $N \times h$ matrix, and random vectors $\mathbf{v}$ and $\mathbf{e}$ are $h \times 1$ and $N \times 1$ respectively. If the population elements are rearranged so that the first n elements of $\mathbf{Y}$ are those in the sample, and the first n rows of $\mathbf{Z}$ are for units in the sample, then $\mathbf{e}$, $\mathbf{Z}$ and $\mathbf{R}$ can be expressed as: $\mathbf{e} = \begin{bmatrix} \mathbf{e}_s \\ \mathbf{e}_r \end{bmatrix}$, $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_s \\ \mathbf{Z}_r \end{bmatrix}$,

$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{ss} & \mathbf{R}_{sr} \\ \mathbf{R}_{rs} & \mathbf{R}_{rr} \end{bmatrix}$ where $\mathbf{e}_s$ is $n \times 1$, $\mathbf{e}_r$ is $N_r \times 1$, $\mathbf{Z}_s$ is $n \times h$, $\mathbf{Z}_r$ is $N_r \times h$, $\mathbf{R}_{ss}$ is $n \times n$, $\mathbf{R}_{rr}$ is $N_r \times N_r$, $\mathbf{R}_{sr}$ is $n \times N_r$ and $\mathbf{R}_{rs} = \mathbf{R}_{sr}^T$.

In the paper we will also discuss the GLMM with block-diagonal variance-covariance matrix which is a special case of (2) assuming that $Cov_\xi(Y_{id}, Y_{i'd'}) = 0$ for $d \neq d'$.

II. SUPERPOPULATION MODELS

In this section we introduce three special cases of the GLM and the GLMM.

Superpopulation model I. We assume that (Chambers and Ayoub, 2003, p.12):

$$Y_{id} = \mu + v_d + e_{id} \text{ (i=1,...N; d=1,...,D)}, \tag{3}$$

where μ is fixed, $v_d \overset{iid}{\sim} (0, \sigma_v^2)$, $e_{id} \overset{iid}{\sim} (0, \sigma_e^2)$ and v_d and e_{id} are independent. In our case additional normality assumption will be needed to derive MSE and its estimator.

What is interesting, from (3) we may obtain that (Valliant *et al.*, 2000, p. 256):

$$E_\xi(Y_{id}) = \mu$$

$$Cov_\xi(Y_{id}, Y_{i'd'}) = \begin{cases} \sigma_e^2 + \sigma_v^2 & \text{for} \quad i = i', d = d' \\ \sigma_v^2 & \text{for} \quad i \neq i', d = d' \\ 0 & \text{otherwise}. \end{cases} \tag{4}$$

Superpopulation model II. Let us assume (4) and that $\sigma_v^2 = 0$.

Superpopulation model III. Let us assume that random variables Y_{id} (i=1,...,N; d=1,...,D) are independent and

$$E_\xi(Y_{id}) = \mu_e \wedge D_\xi^2(Y_{id}) = \sigma_{ed}^2. \tag{5}$$

III. BLUPS AND THEIR MSES

In this paragraph we present the following theorem which gives the formulae of the BLU predictor and its MSE and their special cases for the superpopulation models presented in section II.

Theorem 1. (Royall (1976)). Assume that the population data obey the general linear model. Among the linear, model-unbiased predictors $\hat{\theta} = \mathbf{g}_s^T \mathbf{Y}_s$ of linear combination of random variables $\theta = \boldsymbol{\gamma}^T \mathbf{Y}$ (where $\boldsymbol{\gamma} = \left[\boldsymbol{\gamma}_s^T, \boldsymbol{\gamma}_r^T\right]^T$) the MSE is minimized by:

$$\hat{\theta}_{BLU} = \boldsymbol{\gamma}_s^T \mathbf{Y}_s + \boldsymbol{\gamma}_r^T \left[\mathbf{X}_r \hat{\boldsymbol{\beta}} + \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \left(\mathbf{Y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}} \right) \right], \tag{6}$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{X}_s \right)^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{Y}_s$.

The MSE of $\hat{\theta}_{BLU}$ is given by:

$$MSE_\xi(\hat{\theta}_{BLU}) = Var_\xi(\hat{\theta}_{BLU} - \theta) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}), \tag{7}$$

where

$$g_1(\boldsymbol{\delta}) = \boldsymbol{\gamma}_r^T \left(\mathbf{V}_{rr} - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{V}_{sr} \right) \boldsymbol{\gamma}_r, \tag{8}$$

$$g_2(\boldsymbol{\delta}) = \boldsymbol{\gamma}_r^T \left(\mathbf{X}_r - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{X}_s \right) \left(\mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{X}_s \right)^{-1} \left(\mathbf{X}_r - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{X}_s \right)^T \boldsymbol{\gamma}_r. \tag{9}$$

The proof of the theorem is presented in details for example by Valliant, Dorfman, Royall (2000) pp. 29–30. In the paper we consider the problem of prediction of the domain total, hence the i-th element of $\boldsymbol{\gamma}$ vector equals 1 when $i \in \Omega_{d*}$ and 0 otherwise.

BLUP and its MSE for superpopulation model I. The BLU predictor (6) of the domain total under the superpopulation model (3) simplifies to (Chambers and Ayoub, 2003, p. 13):

$$\hat{\theta}_{BLU} = \sum_{i \in s_{d*}} Y_i + N_{rd*} \hat{\beta} + N_{rd*} n_{d*} \sigma_v^2 (\sigma_e^2 + n_{d*} \sigma_v^2)^{-1} \left(\overline{Y}_{sd*} - \hat{\beta} \right). \tag{10}$$

where $\hat{\beta} = \left(\sum_{d=1}^{D} n_d \left(\sigma_e^2 + n_d \sigma_v^2 \right)^{-1} \right)^{-1} \sum_{d=1}^{D} \left(\sigma_e^2 + n_d \sigma_v^2 \right)^{-1} \sum_{i \in s_d} Y_i$.

The MSE of the BLUP of the domain total given by (7) may be written as follows:

$$MSE_{\xi}(\hat{\theta}_{BLU}) = E_{\xi}(\hat{\theta}_{BLU} - \theta)^2 = E_{\xi}(\hat{\theta}_{r\,BLU} - \theta_r)^2. \tag{11}$$

where

$$\hat{\theta}_{r\,BLU} = \boldsymbol{\gamma}_{\mathbf{r}}^{\mathrm{T}} \left[\mathbf{X}_{\mathbf{r}} \hat{\boldsymbol{\beta}} + \mathbf{V}_{\mathbf{rs}} \mathbf{V}_{\mathbf{ss}}^{-1} \left(\mathbf{Y}_{\mathbf{s}} - \mathbf{X}_{\mathbf{s}} \hat{\boldsymbol{\beta}} \right) \right] \text{ is the BLUP of } \theta_r = \sum_{i \in \Omega_{rd*}} Y_i$$

$$\theta_r = \theta - \sum_{i \in s_{d*}} Y_i = \sum_{i \in \Omega_{rd*}} Y_i = \boldsymbol{\gamma}_{\mathbf{r}}^{\mathrm{T}} \left(\mathbf{X}_r \boldsymbol{\beta} + \mathbf{Z}_r \mathbf{v} + \mathbf{e}_r \right). \tag{12}$$

Let $\theta_r^{\#} = \boldsymbol{\gamma}_{\mathbf{r}}^{\mathrm{T}} \left(\mathbf{X}_r \boldsymbol{\beta} + \mathbf{Z}_r \mathbf{v} \right) = \theta_r - \boldsymbol{\gamma}_{\mathbf{r}}^{\mathrm{T}} \mathbf{e}_{\mathbf{r}}$ Then from (12) and (11) we obtain that:

$$MSE_{\xi}(\hat{\theta}_{BLU}) = E_{\xi}(\hat{\theta}_{r\,BLU} - \theta_r^{\#})^2 + \boldsymbol{\gamma}_{\mathbf{r}}^{\mathrm{T}} \mathbf{V}_{\mathbf{rr}} \boldsymbol{\gamma}_{\mathbf{r}} - 2E_{\xi} \left(\boldsymbol{\gamma}_{\mathbf{r}}^{\mathrm{T}} \mathbf{e}_{\mathbf{r}} (\hat{\theta}_{r\,BLU} - \theta_r^{\#}) \right). \tag{13}$$

Chambers and Ayoub (2003) p.25 approximated the MSE of the BLUP by the first term on the right hand side of (13) given by $E_{\xi}(\hat{\theta}_{r\,BLU} - \theta_r^{\#})^2$. In this paper we use Royall's exact equation of the MSE given by (7). To derive (7) under the superpopulation model (3) we note that under (3) the following equalities hold:

$$\boldsymbol{\gamma}_{\mathbf{r}}^{\mathrm{T}} \mathbf{V}_{\mathbf{rs}} \mathbf{V}_{\mathbf{ss}}^{-1} = N_{rd*} \sigma_v^2 (\sigma_e^2 + n_{d*} \sigma_v^2)^{-1} \boldsymbol{\gamma}_s^T,$$

$$\boldsymbol{\gamma}_{\mathbf{r}}^{\mathrm{T}} \mathbf{V}_{\mathbf{rs}} \mathbf{V}_{\mathbf{ss}}^{-1} \mathbf{V}_{\mathbf{sr}} \boldsymbol{\gamma}_{\mathbf{r}} = n_{d*} \left(N_{rd*} \sigma_v^2 \right)^2 (\sigma_e^2 + n_{d*} \sigma_v^2)^{-1},$$

$$\boldsymbol{\gamma}_{\mathbf{r}}^{\mathrm{T}} \mathbf{V}_{\mathbf{rr}} \boldsymbol{\gamma}_{\mathbf{r}} = N_{rd*} (\sigma_e^2 + N_{rd*} \sigma_v^2),$$

$$\boldsymbol{\gamma}_{\mathbf{r}}^{\mathrm{T}} \mathbf{V}_{\mathbf{rs}} \mathbf{V}_{\mathbf{ss}}^{-1} \mathbf{X}_{\mathbf{s}} = n_{d*} N_{rd*} \sigma_v^2 (\sigma_e^2 + n_{d*} \sigma_v^2)^{-1},$$

$$\gamma_r^T\left(\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right) = N_{rd*}\sigma_e^2(\sigma_e^2 + n_{d*}\sigma_v^2)^{-1}.$$

Hence, the MSE of the BLUP under superpopulation model (3) simplifies to (7), where

$$g_1(\boldsymbol{\delta}) = N_{rd*}\sigma_e^2(\sigma_e^2 + N_{d*}\sigma_v^2)(\sigma_e^2 + n_{d*}\sigma_v^2)^{-1}, \tag{14}$$

$$g_2(\boldsymbol{\delta}) = \left(\sum_{d=1}^{D} n_d(\sigma_e^2 + n_d\sigma_v^2)^{-1}\right)^{-1}\left(N_{rd*}\sigma_e^2(\sigma_e^2 + n_{d*}\sigma_v^2)^{-1}\right)^2. \tag{15}$$

BLUP and its MSE for superpopulation model II. The BLUP and its MSE are given by (note that the following predictor is ξ-unbiased under (3)):

$$\hat{\theta}_{BLU} = \sum_{i\in s_{d*}} Y_i + N_{rd*}n^{-1}\sum_{i\in s} Y_i \text{ and } MSE_\xi(\hat{\theta}_{BLU}) = \left(N_{rd*} + N_{rd*}^2 n^{-1}\right)\sigma_e^2. \tag{16}$$

BLUP and its MSE for superpopulation model III. Under (5) BLUP and its MSE are given by (note that the following predictor is ξ-unbiased under (3)):

$$\hat{\theta}_{BLU} = N_{d*}n_{d*}^{-1}\sum_{i\in s_{d*}} Y_i \text{ and } MSE_\xi(\hat{\theta}_{BLU}) = \sigma_{ed*}^2 N_{d*}(N_{d*} - n_{d*})n_{d*}^{-1}. \tag{17}$$

IV. EBLUPS, THEIR MSES AND ESTIMATORS OF MSES

Note that the BLUPs for superpopulation models II and III do not depend on the unknown in practice parameters. In this cases we need only the following unbiased estimators of σ_e^2 and σ_{ed*}^2, given by $\hat{\sigma}_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_s\right)^2$ and $\hat{\sigma}_{ed*}^2 = \frac{1}{n_{d*}-1}\sum_{i=1}^{n_{d*}}\left(Y_i - \bar{Y}_{s_{d*}}\right)^2$, respectively to obtain unbiased estimators of MSEs presented in (16) and (17) respectively. Let discuss the problem of prediction of the domain total under the superpopulation model I. The BLU predictor (10) depends on the variance parameters $\boldsymbol{\delta} = \left[\sigma_e^2, \sigma_v^2\right]^T$ which are unknown in practical applications. Replacing $\boldsymbol{\delta}$ by an estimator $\hat{\boldsymbol{\delta}}$, we obtain two-stage

predictor called the empirical best linear unbiased predictor (the EBLU predictor). It is denoted by $\hat{\theta}_{EBLU}$ and it remains unbiased if (i) $E(\hat{\theta}_{EBLU})$ is finite; (ii) $\hat{\boldsymbol{\delta}}$ is any even, translation-invariant estimator of $\boldsymbol{\delta}$, that is $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s) = \hat{\boldsymbol{\delta}}(-\mathbf{Y}_s)$ and $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s - \mathbf{X}_s\mathbf{b}) = \hat{\boldsymbol{\delta}}(\mathbf{Y}_s)$ for all $\mathbf{Y}_s$ and $\mathbf{b}$; (iii) the distributions of $\mathbf{v}$ and $\mathbf{e}$ are both symmetric around $\mathbf{0}$ (not necessarily normal). This problem for Royall's predictors is discussed by Żądło (2004) and for Henderson's predictors by Kackar and Harville (1981). We should stress that many standard procedures for estimating $\boldsymbol{\delta}$ (including maximum likelihood - ML and restricted maximum likelihood - REML) yield even, translation-invariant estimators (Kackar and Harville (1981)).

To obtain the MSE of EBLUP for our case we adopt Datta and Lahiri (2000) results for Henderson's EBLUP. Under the general linear mixed model with the block diagonal variance-covariance matrix we assume that D is large and we neglect all terms of order $o(D^{-1})$. What is more the normality of random components and the following regularity conditions are assumed: (a) the elements of $\mathbf{X}_s$ and $\mathbf{Z}_s$ are uniformly bounded such that $\left\{\mathbf{X}_s^{\mathrm{T}}\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right\} = \left[O(D)\right]_{p\times p}$, (b) $\sup_{d\geq 1} n_d < \infty$ and $\sup_{d\geq 1} h_d < \infty$, (c) $\mathbf{X}_r^T\boldsymbol{\gamma}_r - \mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\boldsymbol{\gamma}_r = \left[O(1)\right]_{p\times 1}$, (d) $\frac{\partial}{\partial\delta_k}\mathbf{X}_s^{\mathrm{T}}\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\boldsymbol{\gamma}_r = \left[O(1)\right]_{p\times 1}$ for $k=1,\ldots,q$, (e) $\mathbf{R}_{sd}(\boldsymbol{\delta}) = \sum_{j=0}^{q}\delta_j\mathbf{C}_{dj}\mathbf{C}_{dj}^T$ and $\mathbf{G}_d(\boldsymbol{\delta}) = \sum_{j=0}^{q}\delta_j\mathbf{F}_{dj}\mathbf{F}_{dj}^T$, where $\mathbf{R}_{sd}$ and $\mathbf{G}_d$ are submatrices of $\mathbf{R}_s$ and $\mathbf{G}$ respectively for d-th domain, $\delta_0 = 1$, $\mathbf{C}_{dj}$ and $\mathbf{F}_{dj}$ $(d=1,\ldots,D,\ j=0,\ldots,q)$ are known matrices of order $n_d \times h_d$ and $h_d \times h_d$ respectively. The elements of the matrices $\mathbf{C}_{dj}$ and $\mathbf{F}_{dj}$ are uniformly bounded known constants such that $\mathbf{R}_{sd}$ and $\mathbf{G}_d$ $(d=1,\ldots,D)$ are all positive definite matrices. (In special cases, some of $\mathbf{C}_{dj}$ and $\mathbf{F}_{dj}$ may be null matrices.) (f) $\hat{\boldsymbol{\delta}}$ is an estimator of $\boldsymbol{\delta}$ which satisfies (i) $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta} = O_p(D^{-0,5})$, (ii) $\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}^{ML} = O_p(D^{-1})$ (iii) $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s) = \hat{\boldsymbol{\delta}}(-\mathbf{Y}_s)$, (iv) $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s - \mathbf{X}_s\mathbf{b}) = \hat{\boldsymbol{\delta}}(\mathbf{Y}_s)$ for any $\mathbf{b}$ and all $\mathbf{Y}_s$, where $\hat{\boldsymbol{\delta}}^{ML}$ is maximum likelihood (ML) estimator of $\boldsymbol{\delta}$. Conditions a), b), e) and f) are assumed by Datta and Lahiri (2000) who discussed the MSE of the Henderson's EBLUP. Conditions c) and d) may be treated as modifications of the assumptions c) and d) proposed by Datta and Lahiri (2000).

Under these assumptions and replacing $\mathbf{m}^T\mathbf{G}\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}$ in the proof presented by Datta and Lahiri (2000) by $\boldsymbol{\gamma}_r^T\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}$ we obtain that the MSE for Royall's EBLUP (i.e. the MSE of the predictor (6) where $\boldsymbol{\delta}$ is replaced by its estimator $\hat{\boldsymbol{\delta}}$), in the case when $\hat{\boldsymbol{\delta}}$ is maximum likelihood (ML) or restricted maximum likelihood (REML) estimator. Let $\mathbf{c}^T = \boldsymbol{\gamma}_r^T\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}$,

$$\frac{\partial \mathbf{c}^T}{\partial \boldsymbol{\delta}} = col_{1\le k\le q}\frac{\partial \mathbf{c}^T}{\partial \delta_k} = col_{1\le k\le q}\frac{\boldsymbol{\gamma}_r^T\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}}{\partial \delta_k}, col_{1\le k\le q}\mathbf{a}_k = \left[\mathbf{a}_1^T \quad \cdots \quad \mathbf{a}_q^T\right]^T,$$

$$\mathbf{I}_\delta = -E_\xi\left(\left[\frac{\partial^2 l}{\partial\delta_i\partial\delta_j}\right]_{q\times q}\right),$$

and l is log likelihood assuming multivariate normal distribution of $Y_1,...,Y_N$. Hence,

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

where

$$g_3^*(\boldsymbol{\delta}) = tr\left(\frac{\partial \mathbf{c}^T}{\partial \boldsymbol{\delta}}\mathbf{V}_{ss}\left(\frac{\partial \mathbf{c}^T}{\partial \boldsymbol{\delta}}\right)^T\mathbf{I}_\delta^{-1}\right). \tag{19}$$

Under superpopulation model I $g_1(\boldsymbol{\delta})$, $g_2(\boldsymbol{\delta})$ are given by (14) and (15) respectively and

$$g_3^*(\boldsymbol{\delta}) = N_{rd*}^2 n_{d*} a_{d*}^{-3}\left(\sigma_v^4 I_{vv} - 2\sigma_e^2\sigma_v^2 I_{ev} + \sigma_e^4 I_{ee}\right), \tag{20}$$

where

$$I_{vv} = 2a^{-1}\sum_{d=1}^{D} n_d^2 a_d^{-2},\ I_{ve} = -2a^{-1}\sum_{d=1}^{D} n_d a_d^{-2},\ I_{ee} = 2a^{-1}\sum_{d=1}^{D}\left((n_d-1)\sigma_e^{-4} + a_d^{-2}\right),$$

$$a_d = \sigma_e^2 + n_d\sigma_v^2,\ a = \left(\sum_{d=1}^{D}\left((n_d-1)\sigma_e^{-4} + a_d^{-2}\right)\right)\left(\sum_{d=1}^{D} n_d^2 a_d^{-2}\right) - \left(\sum_{d=1}^{D} n_d a_d^{-2}\right)^2.$$

Now we adopt the MSE estimator presented by Datta and Lahiri (2000) for our case. To estimate $\boldsymbol{\delta}$ we use REML because REML estimators are less biased than ML estimators. The bias of REML estimator is $o(D^{-1})$. What is important our MSE estimator is approximately unbiased in the sense that $E_\xi\left(\hat{MSE}_\xi\left(\hat{\theta}_{EBLU}(\hat{\boldsymbol{\delta}})\right)\right) = MSE_\xi\left(\hat{\theta}_{EBLU}(\hat{\boldsymbol{\delta}})\right) + o(D^{-1})$. Finally the estimator of (18) under superpopulation model I may be written as follows:

$$\hat{MSE}_\xi(\hat{\theta}_{EBLU}(\hat{\boldsymbol{\delta}})) = g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) + 2g_3^*(\hat{\boldsymbol{\delta}}), \tag{21}$$

where $g_1(\hat{\boldsymbol{\delta}})$, $g_2(\hat{\boldsymbol{\delta}})$, $g_3^*(\hat{\boldsymbol{\delta}})$ are given by (14), (15) and (20) respectively where $\boldsymbol{\delta} = \left[\sigma_e^2 \quad \sigma_v^2\right]^T$ is replaced by REML estimator $\hat{\boldsymbol{\delta}} = \left[\hat{\sigma}_e^2 \quad \hat{\sigma}_v^2\right]^T$.

V. SIMULATION STUDY

In the section we present the results of Monte Carlo simulation study prepared in R language (R Development Core Team, 2005). We analyze agricultural data on 8624 farms from Dąbrowa Tarnowska region in Poland obtained in 1996. The region is divided into D=79 villages and towns treated as domains of sizes between 20 and 610 farms. We draw one simple random sample without replacement of 862 farms from the population of 8624 farms which gives one division of the population into sampled and unsampled parts. Realizations of random sample sizes in domains are between 2 and 66 farms which means that the direct predictor presented in (16) gives estimates of total for each domain. We generate 5 000 sets of values of the variable of interest (sowing area in 100 square meters) both for sampled and unsampled part of the population based on superpopulation model (3) with σ_e^2 and σ_v^2 obtained from the entire population data and assuming normality of random components.

We study the accuracy of the following predictors in the simulation study: (a) the predictor (10) assuming that σ_e^2 and σ_v^2 are known, which is the BLUP under model (3) (it will be denoted by BLUP), (b) the predictor (10) where σ_e^2 and σ_v^2 are replaced by their estimates (based on the sample data using REML), which is the EBLUP under (3) (EBLUP), (c) the indirect predictor presented in (16) and direct predictor (17) (DP) presented in (17). We study accuracy of the predictors IP and DP under (3) to check their accuracy in the case of the model misspecification (the IP and DP are BLUPs under models which do not fulfil (3)).

Let us consider the simulation results obtained for 79 domains. What is important, all of predictors are model-unbiased under superpopulation model I (absolute simulation biases did not exceed 1,2%). Values of relative RMSE for 79 domains range for the BLUP from 8,22% to 30,76%, for the EBLUP from 8,24% to 31,01%, for the predictor DP from 8,49% to 51,01% and for the predictor IP from 29,37% to 35,74%. Notice that the increase of MSE due to the estimation of σ_e^2 and σ_v^2 (the difference between the MSE of the BLUP and the MSE of the EBLUP) for the considered real data is not high. Analyzing the values of the ratio of the MSE of the EBLUP and the MSE of the BLUP we note that its maximum value equals 1,0217 what means that the MSE of the EBLUP is higher than the MSE of the BLUP but not higher than only by 2,17% in all of 79 domains. What is more, the EBLUP has smaller MSE than the predictors IP and DP which are not functions of unknown parameters but are not BLUPs under the considered mixed model. It means that in our case the lost of the accuracy due to the estimation of variance components is smaller than the lost of the accuracy due to the model misspecification. What is important, the absolute value of the bias of the estimator of the MSE of the EBLUP is not high – it does not exceed 8,14017%. The MSE estimators of the IP and DP are not unbiased because they are derived under different superpopulation models and hence they are used in the case of model misspecification.

REFERENCES

Chambers R., Ayoub S. (2003), Small area estimation: A review of methods based on the application of mixed models, Southampton Staistical Sciences Research Institute Methodology Working Paper M03/16, University of Southampton

Datta, G. S., Lahiri, P. (2000), A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statistica Sinica*, 10, 613–627.

Henderson, C.R. (1950), Estimation of genetic parameters (Abstract), *Annals of Mathematical Statistics*, 21, 309–310.

Kackar, R.N., Harville, D.A. (1981), Unbiasedness of two-stage estimation and prediction procedures for mixed linear models, *Communications in Statistics, Series A*, 10, 1249–1261.

R Development Core Team (2005), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.

Royall, R.M. (1976), The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657–473.

Valliant, R., Dorfman, A.H., Royall, R.M. (2000), *Finite population sampling and inference. A prediction approach*, John Wiley & Sons, New York.

Żądło, T. (2004), On unbiasedness of some EBLU predictor. In: *Proceedings in Computational statistics 2004*, Antoch J. (red.), Physica-Verlag, Heidelberg, 2019–2026.

Tomasz Żądło

O DOKŁADNOŚCI PEWNEGO PREDYKTORA TYPU EBLU

W opracowaniu analizujemy dokładność empirycznych najlepszych liniowych nieobciążonych predyktorów wartości globalnej w domenie (ang. EBLUP – empirical best linear unbiased predictor) zakładając model nadpopulacji należący do klasy ogólnych mieszanych modeli liniowych. Do oceny błędu średniokwadratowego (ang. MSE – mean square error) predyktora typu EBLU wykorzystano rezultaty prezentowane przez Datta and Lahiri (2000) dla predyktora zaproponowanego przez Hendersona (1950) po zaadoptowaniu ich dla przypadku predyktora zaproponowanego przez Royalla (1976). W badaniu symulacyjnym wykorzystano rzeczywiste dane dotyczące gospodarstw rolnych w powiecie Dąbrowa Tarnowska uzyskane w spisie rolnym w 1996.