ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

Tomasz Żądło[*]

ON PREDICTION OF THE DOMAIN TOTAL UNDER SOME SPECIAL CASE OF TYPE A GENERAL LINEAR MIXED MODEL

Abstract. In the paper we present the best linear unbiased predictor (BLUP) and the empirical best linear unbiased predictor (EBLUP), their mean squared errors (MSE) and estimators of MSE of EBLUP under special case of Fay-Herriot model (Fay, Herriot (1979)). This is A type model what means that it is assumed for direct estimators of domain characteristics. What is more, it is assumed (even when EBLUP is studied) that variances of direct estimators are known. In the simulation based on real data, the influence of replacing the variances by their design-unbiased estimates or General Variance Function (GVF) estimates (Wolter (1985)) on predictor's biases and MSEs and on biases of MSE estimators is studied. The problem of non-normality of domain specific random components is also included.

Key words: BLUP, EBLUP, Fay-Herriot model, MSE estimators.

I. BASIC NOTATIONS AND SUPERPOPULATION MODEL

Population Ω, which consists of N – elements, is divided into D separate subpopulations Ω_d (d=1,..., D) called domains each of size N_d (d=1,..., D). From the population (random or purposive) sample s of size n is drawn. Let $s_d = s \cap \Omega_d$ is of size n_d. Domain mean and domain total are denoted by μ_d and $\theta_d = N_d \mu_d$ respectively. In the paper the special case (without auxiliary variables) of Fay-Herriot model (Fay, Herriot (1979)) is studied. It is assumed for direct estimators of domain means what means that it is type A model according to Rao (2003). Let us add, that models type B (Rao (2003)) are assumed for random variables which realizations are values of the variable of interest. Let us assumed for direct estimators of domain means (denoted by $\hat{\mu}_d$) that:

$$\hat{\mu}_d = \mu_d + e_d, \tag{1}$$

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

where $\mu_d = \mu + v_d$ is domain mean, μ is unknown parameter, e_d is design error, e_d and v_d (d=1,...,D) are independent, where $e_d \overset{iid}{\sim} N(0, W_d)$ and $v_d \overset{iid}{\sim} N(0, A)$, and it is assumed that W_d are known even in empirical case. This model is special case of the general linear model (GLM), the general linear mixed model (GLMM) and the GLMM with block-diagonal variance-covariance matrix where $\forall_{i,d} Z_{i,d} = 1$ and Y_i (I = 1,..., N) are replaced by $\hat{\mu}_d$ (d=1,..., D).

II. BLUP AND ITS MSE

For known A and W_d and without assumption of normality of random components the BLUP of mean in domain d and its MSE are given by (Datta, Rao, Smith (2005), Datta, Lahiri (2000), Lahiri, Rao (1995), Prasad, Rao (1990)):

$$\hat{\mu}_{d(s)}^{BLUP} = \hat{\mu}_d - B_d(A)(\hat{\mu}_d - \hat{\mu}) \tag{2}$$

where

$$B_d(A) = W_d(A + W_d)^{-1},\ \hat{\mu} = \left(\sum_{d=1}^{D}(A + W_d)^{-1}\right)^{-1}\left(\sum_{d=1}^{D}(A + W_d)^{-1}\hat{\mu}_d\right),$$

$$MSE_\xi(\hat{\mu}_{d(s)}^{BLUP}) = g_{1d(s)}(A) + g_{2d(s)}(A) \tag{3}$$

where

$$g_{1d(s)}(A) = AW_d(A + W_d)^{-1},\ g_{2d(s)}(A) = W_d^2(A + W_d)^{-2}\left(\sum_{u=1}^{D}(A + W_d)^{-1}\right)^{-1}.$$

Hence, for known domain sizes N_d, the BLUP of domain total and its MSE are given by: $\hat{\theta}_{d(s)}^{BLUP} = N_d\hat{\mu}_{d(s)}^{BLUP}$ and $MSE_\xi(\hat{\theta}_{d(s)}^{BLUP}) = N_d^2 MSE_\xi(\hat{\mu}_{d(s)}^{BLUP})$, where $\hat{\mu}_{d(s)}^{BLUP}$ and $MSE_\xi(\hat{\mu}_{d(s)}^{BLUP})$ are given by (2) and (3) respectively. Let us stress that these results are special case of Henderson' s theorem (Henderson (1950)).

III. ESTIMATION OF MODEL PARAMETERS

Deriving the form of BLUP it assumed that A and W_d are known. In practice it is usually assumed that W_d are known, but A is estimated based on the sample data. In further considerations we will discuss four estimators of A parameter: (i) maximum likelihood (ML) estimator $\hat{A}_{ML}$, (ii) restricted maximum likelihood (REML) estimator $\hat{A}_{RE}$ (iii) Fay-Herriot estimator (Fay, Herriot (1979) which is iteratively obtained solution of the following equality

$$\frac{1}{D-1}\mathbf{Y}^T Q(\hat{A}_{FH})\mathbf{Y} - 1 = 0 \tag{4}$$

where $\mathbf{Y}^T Q(\hat{A}_{FH})\mathbf{Y} = \sum_{d=1}^{D}(W_d + A)^{-1}(\hat{\mu}_d - \hat{\mu})^2$, and (iv) Prasad-Rao estimator (Prasad, Rao (1990)) given by $\max(0, \hat{A}_{PR})$, where

$$\hat{A}_{PR} = (D-1)^{-1}\sum_{d=1}^{D}(\hat{\mu}_d - \bar{\hat{\mu}}_d)^2 - D^{-1}\sum_{d=1}^{D}W_d \tag{5}$$

and $\bar{\hat{\mu}}_d = D^{-1}\sum_{d=1}^{D}\hat{\mu}_d$. In the first two cases normality of random variables is assumed.

Although it is assumed that variances of $\hat{\mu}_d$ (denoted by W_d) are known even in empirical cases (i.e. parameter A is replaced by its estimate), two types of estimators will be presented. Because of assumption of the independence of e_d we will consider in the simulation study stratified random sampling (without replacement) from domains. Hence the estimators of W_d (d=1,...,D) are given by:

$$\hat{W}_d = \frac{N_d - n_d}{N_d n_d}\frac{1}{n_d - 1}\sum_{i=1}^{n_d}(Y_i - \bar{Y}_{sd})^2 \tag{6}$$

What is more (e.g. Lahiri and Rao (1995)) values of (6) may also be smoothed using GVF (e.g. Wolter (1985)). Note that (Wolter (1985)) the choice of the function is based on empirical studies. In the paper we will use the following function (Wolter (1985) p. 203):

$$\log\left(W_d m_d^{-2}\right) = a - b\log(m_d) \tag{7}$$

where α and β are estimated based on (7) using least squared method, where W_d and μ_d are replaced by $\hat{W}_d$ and $\hat{\mu}_d$ respectively. Then, in (7) α, β and μ_d are replaced by their estimates and from (7) we obtain smoothed values of $\hat{W}_d$ denoted by $\hat{W}_{GVF\,d}$.

IV. EBLUP AND ITS MSE

EBLUPs of domain mean and domain total, denoted by $\hat{\mu}_{d(s)}^{EBLUP}$ and $\hat{\theta}_{d(s)}^{EBLUP}$, are given by (2) and $N_d\hat{\mu}_{d(s)}^{EBLUP}$ respectively, where A is replaced by one of discussed estimators. EBLUPs remain unbiased (for details see Kackar i Harville (1981)) inter alia because presented estimators of A are even, translation invariant functions of $\hat{\mu}_d$, i.e. $\hat{A}(-\hat{\mu}_d) = \hat{A}(\hat{\mu}_d)$ and $\hat{A}(\hat{\mu}_d + \mathbf{Xb}) = \hat{A}(\hat{\mu}_d)$ for any $\mathbf{b} \in R^p$. Note that normality of e_d and v_d is not required (only symmetry around 0).

Assuming that (Datta, Rao, Smith (2005) s.186) the elements of $\mathbf{X}$ are uniformly bounded and those of $\mathbf{X}^T\mathbf{V}_{ss}^{-k}(A)\mathbf{X}$ (k=1,2,3) are of order $O(D)$ and the W_d are bounded obove and bounded away from zero, MSE of EBLUP of domain mean for the discussed model is given by (Prasad, Rao (1990), Datta, Lahiri (2000)):

$$MSE_\xi(\hat{\mu}_{d(s)}^{EBLUP}(\hat{A})) = g_{1d(s)}(A) + g_{2d(s)}(A) + g_{3d(s)}(A) + o(D^{-1}) \tag{8}$$

where $g_{1d(s)}(A)$ and $g_{2d(s)}(A)$ are presented in (3). Form of $g_{3d(s)}(A)$ in equation (8) depends on the method of estimation of A. For A estimator proposed by Prasad and Rao (1990) it is given by:

$$g_{3d(s)}(A) = g_{3dPR(s)}(A) = 2W_d^2(A+W_d)^{-3}D^{-2}\sum_{u=1}^{D}(A+W_u)^2 \tag{9}$$

In the case of ML and REML estimators of A it is given by (Datta, Lahiri (2000)):

$$g_{3d(s)}(A) = g_{3dML(s)}(A) = g_{3dRE(s)}(A) = 2W_d^2(A+W_d)^{-3}\left(\sum_{u=1}^{D}(A+W_u)^{-2}\right)^{-1} \tag{10}$$

For A estimator proposed by Fay and Herriot it is as follows (Datta, Rao, Smith (2005)):

$$g_{3d(s)}(A) = g_{3dFH(s)}(A) = 2DW_d^2(A+W_d)^{-3}\left(\sum_{u=1}^{D}(A+W_d)^{-1}\right)^{-2} \tag{11}$$

MSE of EBLUP of domain total for the discussed model may be written as follows: $MSE_\xi(\hat{\theta}_{d(s)}^{EBLUP}(\hat{A})) = N_d^2 MSE_\xi(\hat{\mu}_{d(s)}^{EBLUP}(\hat{A}))$, where $MSE_\xi(\hat{\mu}_{d(s)}^{EBLUP}(\hat{A}))$ is given by (8).

V. MSE ESTIMATORS

In this section estimators of MSE (denoted by $\hat{MSE}_\xi(\hat{\mu}_{d(s)}^{EBLUP}(\hat{A}))$) will be presented which are approximately unbiased in the sense that $E_\xi\left(\hat{MSE}_\xi(\hat{\mu}_{d(s)}^{EBLUP}(\hat{A}))\right) - MSE_\xi(\hat{\mu}_{d(s)}^{EBLUP}(\hat{A})) = o(D^{-1})$. Datta and Lahiri (2000) proposed the following form of MSE estimator of EBLUP of domain mean:

$$\hat{MSE}_\xi(\hat{\mu}_{d(s)}^{EBLUP}(\hat{A})) = g_{1d(s)}(\hat{A}) + g_{2d(s)}(\hat{A}) + 2g_{3d(s)}(\hat{A}) - (B_d(\hat{A}))^2 b_{\hat{A}}(\hat{A}) \tag{12}$$

where $b_{\hat{A}}(A)$ is asymptotic (up to order $o(D^{-1})$) bias of $\hat{A}$, $B_d(A)$ is presented in (2).

For asymptotically unbiased $\hat{A}_{PR}$ and $\hat{A}_{RE}$ the last element in equation (12) is omitted.

For $\hat{A}_{ML}$ estimator from (12) we obtain (Datta, Lahiri (2000)):

$$\hat{MSE}_\xi(\hat{\mu}_{d(s)}^{EBLUP}(\hat{A}_{ML})) =$$

$$= g_{1d(s)}(\hat{A}_{ML}) + g_{2d(s)}(\hat{A}_{ML}) + 2g_{3dML(s)}(\hat{A}_{ML}) + (B_d(\hat{A}_{ML}))^2 \times$$

$$\times\left(\sum_{u=1}^{D}(A_{ML}+W_u)^{-2}\right)^{-1}\left[\left(\sum_{u=1}^{D}(A_{ML}+W_u)^{-1}\right)^{-1}\left(\sum_{u=1}^{D}(A_{ML}+W_u)^{-2}\right)\right] \tag{13}$$

For $\hat{A}_{FH}$ estimator (12) is given by (Datta, Rao, Smith (2005)):

$$\hat{MSE}_{\xi}(\hat{\mu}_{d(s)}^{EBLUP}(\hat{A}_{FH})) =$$

$$= g_{1d(s)}(\hat{A}_{FH}) + g_{2d(s)}(\hat{A}_{FH}) + 2g_{3dML(s)}(\hat{A}_{FH}) - 2(B_d(\hat{A}_{FH}))^2 \times$$

$$\times\left[\left(D\sum_{u=1}^{D}(A_{FH}+W_u)^{-2}\right)-\left(\sum_{u=1}^{D}(A_{FH}+W_u)^{-1}\right)^2\right]\left(\sum_{u=1}^{D}(A_{FH}+W_u)^{-1}\right)^{-3} \quad (14)$$

MSE estimators of EBLUPs of domain totals are given by: $\hat{MSE}_{\xi}(\hat{\theta}_{d(s)}^{EBLUP}(\hat{A})) = N_d^2 \hat{MSE}_{\xi}(\hat{\mu}_{d(s)}^{EBLUP}(\hat{A}))$. It is worth stressing (Lahiri, Rao (1995), Datta, Rao, Smith (2005)), that MSE estimators obtained for Prasad-Rao and Fay-Herriot estimators of A are robust on non-normality of random components.

VI. MONTE CARLO ANALYSIS

In the simulation prepared using R language (R Development Core Team, 2007), the population of 8624 farms in Dąbrowa Tarnowska region obtained during agricultural census in 1996 is studied. The population is divided into D = 79 domains (cities and villages). Domain sizes are between 20 and 610 farms. Because of assumption of independence of e_d, domains are treated as strata and stratified random sample (without replacement) with approximate proportional allocation (c.a. 10% of elements from each stratum) is drawn from the population. The problem of prediction of total sowing area in domains is studied, where $\hat{\mu}_d$ is sample mean of sowing area in the domain d.

Number of iterations equals is 5000. In each iteration values of $\hat{\mu}_d$ are generated based on the discussed model with W_d given by: $W_d = \frac{N_d - n_d}{N_d n_d}\frac{1}{N_d - 1}\sum_{i=1}^{N_d}\left(y_i - N_d^{-1}\sum_{i=1}^{N_d} y_i\right)^2$, where y_i are values of the variable of interest in the studied population data set. Random components e_d are generated independently based on $N(0,\sqrt{W_d})$. The value of A in simulation is based on REML algorithm (assuming normality) and the whole data set, where instead of $\hat{\mu}_d$ real values of domain means are used assuming zero sampling variances W_d. Random components v_d are generated independently based on normal, uniform and shifted exponential distribution (with expected value equal 0) with variance A.

If we compare values of $g_3(.)$ for different methods of estimation we will obtain quite large differences, but differences of approximate MSEs for different methods of estimation are smaller (the other components of approximate MSE are of higher order). The absolute relative biases of the considered predictors (different methods of estimation of A and W_d and different distributions of v_d are studied) do not exceed 1,8%. Hence, estimation of W_d do not have important influence on predictors' biases. For different distributions of v_d and different method of estimation of A parameter we obtain following values of relative root of MSE: for known W_d from 6,24% to 31,45%, for estimated W_d from 6,26% to 41,66% and for estimated and smoothed W_d from 6,65% to 45,69%. Although in the practice better model should be considered, it is important that the replacement W_d by their estimates has greater influence on MSE than the distribution of v_d and chosen method of estimation of A.

In the following table we present values: minimum (min), first quarter (Q_1), median (Me), mean, third quarter (Q_3) and maximum for 79 domains. In the columns we use following abbreviations for methods of estimation of A parameter: PR – Prasad-Rao, FH – Fay-Herriot, ML – maximum likelihood, RE – restricted maximum likelihood. The information of the distribution of v_d is as follows: N – normal, U – uniform and E – shifted exponential distribution.

Table 1. Relative biases of MSE estimators (in %)

Est. A	PR			ML			RE			FH		
	known W_d											
	N	U	E	N	U	E	N	U	E	N	U	E
min	–4.9	–3.1	–4.9	–4.4	–2.8	–5.4	–4.6	–3.1	–6.1	–4.8	–3.2	–5.4
Q_1	–1.9	–1.2	–0.7	–1.5	–0.8	–1.7	–2.0	–1.2	–2.4	–2.0	–1.4	–1.5
Me	–0.4	0.3	0.4	0.0	0.4	–0.3	–0.6	0.2	–0.7	–0.6	0.1	–0.0
mean	–0.6	0.2	0.5	0.0	0.8	–0.1	–0.6	0.2	–0.6	–0.6	0.1	–0.1
Q_3	0.7	1.3	2.0	1.3	2.1	1.3	0.7	1.5	0.8	0.6	1.2	1.3
max	4.6	5.3	7.6	5.2	5.8	6.2	4.7	5.1	5.3	4.7	5.2	6.4
	estimated W_d											
min	–96.0	–96.0	–96.0	–96.1	–96.0	–96.3	–96.1	–96.0	–96.3	–96.1	–96.0	–96.3
Q_1	–56.8	–54.6	–56.0	–58.4	–56.2	–57.9	–58.7	–56.6	–58.2	–58.6	–56.5	–58.2
Me	–27.0	–29.6	–26.0	–27.2	–30.5	–27.5	–27.9	–30.8	–28.1	–28.1	–30.8	–28.2
mean	–24.3	–24.2	–22.8	–20.0	–19.8	–19.8	–20.3	–20.2	–20.2	–21.1	–21.0	–20.7
Q_3	6.4	7.7	7.8	17.2	16.2	19.5	16.9	15.6	18.7	15.1	14.5	17.1
max	82.7	85.4	79.4	107.7	111.8	106.9	108.1	112.3	107.4	104.4	108.2	104.5
	smoothed. estimated W_d											
min	–90.3	–90.5	–90.1	–90.3	–90.6	–90.0	–90.3	–90.6	–90.1	–90.3	–90.5	–90.1
Q_1	–33.9	–34.7	–34.5	–32.8	–33.0	–33.4	–33.5	–34.0	–34.4	–33.6	–34.3	–34.5
Me	14.9	13.7	14.1	14.7	13.6	13.2	14.0	13.3	12.7	14.4	13.5	13.4
mean	19.4	20.2	20.5	18.6	19.4	19.2	18.4	19.1	19.0	18.9	19.6	19.8
Q_3	58.3	55.3	55.0	57.7	54.6	54.0	57.6	54.5	53.8	57.9	54.9	54.5
max	208.3	205.0	210.7	199.6	201.0	201.1	202.2	202.1	203.7	205.7	203.7	208.1

For known W_d obtained biases of MSE estimators are small – absolute relative biases do not exceed 7,6%. What is important, for asymmetric distribution (shifted exponential) the biases are higher and all of methods of estimation of A give similar results (although using ML and REML normality is assumed). When W_d are estimated (or estimated and smoothed) MSE estimators, which are derived under assumption of known W_d, have very large biases. In practice, new MSE estimators including this additional source of empirical predictors' variability should be proposed and used in this situation.

REFERENCES

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.

Datta G.S., Rao J.N.K, Smith D.D. (2005), *On measuring the variability of small area estimators under basic area level model*, Biometrika, 92, 1, 183–196.

Fay R.E., Herriot R.A. (1979), *Estimates of income for small places: An application of James-Stein procedures to census data*, Journal of the American Statistical Association, 74, 269–277.

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.

Lahiri Rao (1995), *Robust estimation of mean squared error of small area estimators*, Journal of the American Statistical Association, 90, 430, 758–766.

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, 85, 163–171.

R Development Core Team (2007). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org.

Rao J.N.K. (2003), *Small area estimation*, John Wiley & Sons, New York.

Wolter K.M. (1985), *Introduction to variance estimation*, Springer-Verlag, New York.

Tomasz Żądło

O PREDYKCJI WARTOŚCI GLOBALNEJ PRZY ZAŁOŻENIU PEWNEGO PRZYPADKU SZCZEGÓLNEGO OGÓLNEGO LINIOWEGO MODELU MIESZANEGO TYPU A

W pracy zaprezentowano najlepsze liniowe nieobciążone predyktory i empiryczne najlepsze liniowe nieobciążone predyktory ich błędy średniokwadratowe (MSE) oraz estymatory MSE dla przypadku szczególnego modelu Faya-Herriota (Fay, Herriot

(1979)). Model ten należy do klasy ogólnych mieszanych modeli liniowych typu A, co oznacza, że jest on zakładany dla wartości estymatorów bezpośrednich charakterystyk w domenach. Ponadto przyjmuje się, że wartości wariancji estymatorów bezpośrednich są znane. W artykule analizowano symulacyjnie z wykorzystaniem rzeczywistych danych wpływ zastąpienia nieznanych wariancji estymatorów bezpośrednich ich nieobciążonymi estymatorami i estymatorami otrzymanymi przy wykorzystaniu ogólnych funkcji wariancji na obciążenia predyktorów, wartość MSE oraz obciążenia estymatorów MSE.