Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663
www.czasopisma.uni.lodz.pl/foe/
5(331) 2017

DOI: http://dx.doi.org/10.18778/0208-6018.331.11

Małgorzata Karolina Krzciuk
University of Economics in Katowice. Department of Statistics, Econometrics and Mathematics,
malgorzata.krzciuk@uekat.pl

On the Simulation Study of Jackknife and Bootstrap MSE Estimators of a Domain Mean Predictor for Fay-Herriot Model

Abstract: We consider the problem of the estimation of the mean squared error (MSE) of some domain mean predictor for Fay-Herriot model. In the simulation study we analyze properties of eight MSE estimators including estimators based on the jackknife method (Jiang, Lahiri, Wan, 2002; Chen, Lahiri, 2002; 2003) and parametric bootstrap (Gonzalez-Manteiga et al., 2008; Buthar, Lahiri, 2003). In the standard Fay-Herriot model the independence of random effects is assumed, and the biases of the MSE estimators are small for large number of domains. The aim of the paper is the comparison of the properties of MSE estimators for different number of domains and the misspecification of the model due to the correlation of random effects in the simulation study.

Keywords: estimators of MSE, jackknife, parametric bootstrap, Empirical Best Linear Unbiased Predictor, Fay-Herriot model, simulation

JEL: C15, C83

1. Introduction

One of the main approaches in small area statistics is the model-based approach. In the paper we raise the issue of mean prediction for some domain under some model which belongs to the class of the linear mixed models.

The general linear mixed model is given by:

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

where X and Z are known matrices of auxiliary variables, β is the vector of the unknown parameters. The random effects v and stochastic disturbance e are independently distributed and have variance-covariance matrices denoted by G and R, respectively (cf. Jiang, 2007: 1–2; Rao, Molina, 2015: 98).

In the paper we analyze special case of (1). We consider the Fay-Herriot (1979) model which belongs to area level models, where the auxiliary information is available only on the area level. The model has the following form (cf. Prasad, Rao, 1990: 164; Lahiri, 2003: 206):

$$\hat{\theta}_d = \theta_d + e_d, \tag{2}$$

where:

$$\theta_d = \mathbf{x}_d^T\boldsymbol{\beta} + v_d \tag{3}$$

and $\hat{\theta}_d$ is the direct estimator of θ in the d-th domain ($d = 1, \ldots, D$). In (3) the vector of p values of auxiliary variables in d-th domain is denoted by $\mathbf{x}_d$ and $\boldsymbol{\beta}$ is the vector of p unknown parameters. The error associated with the sampling design e_d and random effects v_d are mutually independent and $e_d \overset{iid}{\sim} N(0, W_d)$, $v_d \overset{iid}{\sim} N(0, A)$ ($d = 1, \ldots, D$). When the assumptions (2) and (3) are met then $\mathbf{R} = diag_{1<d<D}(W_d)$, $\mathbf{G} = A\mathbf{I}_{D\times D}$ ($\mathbf{I}_{D\times D}$ – identity matrix of size $D\times D$). We assume that the variances W_d are known. In literature we find that (2) is the sampling model and (3) is the linking model (Jiang, Lahiri, 2006: 6).

The Fay-Herriot model allows to obtain reliable small area statistics by building the linking models for the direct estimators, the use of the auxiliary data, borrowing strength from other domains and elasticity in linking data from various sources (Datta, Rao, Smith, 2005: 184; Rueda, Mendez, Gomez, 2010: 571).

This model and its generalizations are applied in many areas, for example: estimating of income *per capita* for small areas in the United States (Fay, Herriot, 1979), estimating of p-variance for panel data from the study of natural resources of USA National Resources Inventory (Wang, Fuller, 2003), estimating of the av-

erage income of households and the kurtosis of income for the households (Jędrzejczak, 2011) and estimating unemployment rates in selected Canadian cities, (Rao, You, 1994). The Fay-Herriot model was also used by Bell (1997) to produce estimates of the number of school-aged children living in poverty per county, Lohr and Rao (2009) to compare the area-specific jackknife method with the naive estimators of MSE and the jackknife estimator proposed by Jiang, Lahiri and Wan (2002), Slud and Maiti (2006) for simulation studies of small area incomes and poverty estimation under transformed Fay-Herriot model.

2. BLU and EBLU predictor

The predictor which minimizes, in the class of linear model-unbiased predictors of θ, the MSE is called the Best Linear Unbiased Predictor (BLUP). Under the Henderson's theorem (1950) we consider the problem of prediction of the linear combination of vectors **v** and β given **v** by $\theta = \mathbf{l}^T\beta + \mathbf{m}^T\mathbf{v}$. The variance and covariance matrices $\mathbf{G} = \mathbf{G}(\delta)$ and $\mathbf{R} = \mathbf{R}(\delta)$, which are functions of the vector of parameters δ called variance components, are assumed to be known. For the general linear mixed model (1) the predictor is given by:

$$\hat{\theta}^{BLUP} = \mathbf{l}^T\hat{\boldsymbol{\beta}} + \mathbf{m}^T\hat{\mathbf{v}}, \tag{4}$$

where:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y} \tag{5}$$

and

$$\hat{\mathbf{v}} = \mathbf{G}\mathbf{Z}\mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right). \tag{6}$$

The variance-covariance matrix of the vector **Y** is given by $\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T$. Additionally, it should be noted that $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{v}}$ are functions of δ (Rao, 2003: 96–97).

For the analyzed Fay-Herriot model, where $\delta = A$, the BLUP is as follows:

$$\hat{\theta}_d^{BLUP} = \hat{\theta}_d - B_d(A)\left(\hat{\theta}_d - \mathbf{x}_d^T\hat{\boldsymbol{\beta}}\right), \tag{7}$$

where:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D}\frac{B_d(A)}{W_d}\mathbf{x}_d\mathbf{x}_d^T\right)^{-1}\left(\sum_{d=1}^{D}\frac{B_d(A)}{W_d}\mathbf{x}_d\hat{\theta}_d\right) \tag{8}$$

and

$$B_d\left(A\right)=W_d\left(A+W_d\right)^{-1}. \tag{9}$$

The MSE of (7) is given by:

$$MSE_{\xi}\left(\hat{\theta}_d^{BLUP}\right)=g_{1d}\left(A\right)+g_{2d}\left(A\right), \tag{10}$$

where:

$$g_{1d}\left(A\right)=AW_d\left(A+W_d\right)^{-1} \tag{11}$$

and

$$g_{2d}\left(A\right)=W_d^2\left(A+W_d\right)^{-2}\mathbf{x}_d^T\left(\sum_u^D\left(A+W_d\right)^{-1}\mathbf{x}_u\mathbf{x}_u^T\right)^{-1}\mathbf{x}_d. \tag{12}$$

In practical application the vector is unknown. The replacement of $\boldsymbol{\delta}$ by its estimator $\hat{\boldsymbol{\delta}}$ in (4) and (7) allows to obtain two stage predictor – the Empirical Best Linear Unbiased Predictor (EBLUP) (Rao, Molina, 2015: 101).

When the assumption (1) is fulfilled and furthermore: the expected value of the EBLUP is finite, $\hat{\boldsymbol{\delta}}$ is an even and translation invariant estimator, the distributions of stochastic disturbances and random effects are symmetric about zero, then $\hat{\theta}^{EBLUP}$ is model-unbiased (Kackar, Harville, 1981: 1258–1259).

For (7) the MSE has the general form (Prasad, Rao, 1990: 167; Data, Lahiri, 2000: 617–618):

$$MSE_{\xi}\left(\hat{\theta}_d^{EBLUP}\left(\hat{A}\right)\right)=g_{1d}\left(A\right)+g_{2d}\left(A\right)+g_{3d}\left(A\right)+o\left(D^{-1}\right), \tag{13}$$

where the last component, for A estimated using Restricted (Residual) Maximum Likelihood method is given by (Datta, Lahiri, 2000: 618):

$$g_{3d}\left(A\right)=2W_d^2\left(A+W_d\right)^{-3}\left(\sum_u^D\left(A+W_d\right)^{-2}\right)^{-1}. \tag{14}$$

Remaining elements in (13) are given by formulae (11) and (12), respectively.

3. Classic estimators of the MSE

In this section we present two MSE estimators, the naive one presented by Kackar and Harville (1984) and the estimator based on the Taylor expansion proposed by Datta and Lahiri (2000).

The first of them is given by (Kackar, Harville, 1984: 854–855):

$$\hat{MSE}_{\xi}^{N}\left(\hat{\theta}_{d}^{EBLUP}\left(\hat{A}\right)\right)=g_{1d}\left(\hat{A}\right)+g_{2d}\left(\hat{A}\right). \tag{15}$$

It should be noted that this estimator has the form of the MSE of BLUP (7), where we replace A by its estimator. The bias of the naive estimator is of $O(D^{-1})$ order. It is important that this estimator does not take into account the influence of estimating model parameters on the prediction accuracy.

The MSE estimator based on the Taylor expansion for REML estimates of A is given by (Datta, Lahiri, 2000: 618–619):

$$\hat{MSE}_{\xi}^{DL}\left(\hat{\mu}_{d}^{EBLUP}\left(\hat{A}\right)\right)=g_{1d}\left(\hat{A}\right)+g_{2d}\left(\hat{A}\right)+2g_{3d}\left(\hat{A}\right), \tag{16}$$

where $g_{3d}(A)$ is given by (14). The estimator takes into account the decrease of prediction accuracy resulting from the estimation of model parameters and its bias is of $o(D^{-1})$ order.

The properties of both estimators in case of some types of model misspecification are compared e.g. in Krzciuk (2015).

4. Jackknife method in estimation of MSE

In this section we present a special case of the jackknife estimator of the MSE, presented in Jiang, Lahiri, Wan (2002). These authors consider: wide class of mixed models and the problem of estimation of variance components using M-estimators and Empirical Best Predictor. In the article we analyze a special case of these assumptions: Fay-Herriot model, the estimator of A obtained using ML or REML method and the Empirical Best Unbiased Predictor.

The jackknife estimator considered by Jiang, Lahiri, Wan (2002) has the following form:

$$\begin{aligned}\hat{MSE}_{\xi}^{jack}\left(\hat{\theta}_{EBLUP}\right)&=g_{1d}(\hat{\boldsymbol{\delta}})-\frac{D-1}{D}\sum_{d=1}^{D}\left(g_{1d}(\hat{\boldsymbol{\delta}}_{-d})-g_{1d}(\hat{\boldsymbol{\delta}})\right)+\\&+\frac{D-1}{D}\sum_{d=1}^{D}\left(\hat{\theta}_{EBLUP}(\hat{\boldsymbol{\delta}}_{-d})-\hat{\theta}_{EBLUP}(\hat{\boldsymbol{\delta}})\right)^{2},\end{aligned} \tag{17}$$

where: D is the number of domains, $g_{1d}(\hat{\boldsymbol{\delta}})$ is given by (11) for $\hat{\boldsymbol{\delta}}$, $g_{1d}(\hat{\boldsymbol{\delta}}_{-d})$ can be written as (11) for $\hat{\boldsymbol{\delta}}_{-d}$ (Jiang, Lahiri, Wan, 2002: 1792–1793). Furthermore, $\hat{\boldsymbol{\delta}}_{-d}$ is calculated for data set $s - s_d$, where s_d is the set of elements of d-th domain in the sample. We should note that under some additional assumptions the estimator (17) is asymptotically unbiased and its bias is of $o(D^{-1-\varepsilon})$ order, where ε has value from the (0; 0.5) interval (Jiang, Lahiri, Wan, 2002: 1793).

Because of the possibility of obtaining negative values of (17) (Bell, 2001), we also consider the weighted jackknife MSE estimator studied by Chen, Lahiri, (2002; 2003):

$$\begin{aligned} \hat{MSE}_{\xi}^{wjack}\left(\hat{\theta}_{EBLUP}\right) &= g_{1d}(\hat{\boldsymbol{\delta}}) + g_{2d}(\hat{\boldsymbol{\delta}}) + \\ -\sum_{d=1}^{D} w_d \left(g_{1d}(\hat{\boldsymbol{\delta}}_{-d}) + g_{2d}(\hat{\boldsymbol{\delta}}_{-d}) - \left(g_{1d}(\hat{\boldsymbol{\delta}}) + g_{2d}(\hat{\boldsymbol{\delta}}) \right) \right) + \\ +\sum_{d=1}^{D} w_d \left(\hat{\theta}_{EBLUP}(\hat{\boldsymbol{\delta}}_{-d}) - \hat{\theta}_{EBLUP}(\hat{\boldsymbol{\delta}}) \right)^2 . \end{aligned} \tag{18}$$

Similarly to (17), $g_{1d}(\hat{\boldsymbol{\delta}}_{-d})$ and $g_{2d}(\hat{\boldsymbol{\delta}}_{-d})$ are given by (11) and (12), for $\hat{\boldsymbol{\delta}}_{-d}$. Chen, Lahiri (2003) give two proposals of weights for the considered Fay-Herriot model:

$$w_d = \frac{D-1}{D} \tag{19}$$

and

$$w_d = \mathbf{x}_d^T \left(\sum_{u=1}^{D} \mathbf{x}_u \mathbf{x}_u^T \right)^{-1} \mathbf{x}_d \,. \tag{20}$$

In the simulation study we will consider both of them.

5. Parametric bootstrap MSE estimators

In the paper we study two estimators of the MSE based on the parametric bootstrap method, proposed by Gonzales-Manteiga et. al. (2008) and Butar, Lahiri (2003). In both of them we generate the vector $\mathbf{Y}$ based on the following bootstrap model (cf. Chatterjee, Lahiri, Li, 2008: 1229–1230):

$$\mathbf{Y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{v}^* + \mathbf{e}^*, \tag{21}$$

where $\mathbf{v}^*$ and $\mathbf{e}^*$ are generated as follows: $\mathbf{v}^* \sim N\left(0, \mathbf{G}(\hat{\boldsymbol{\delta}})\right)$, $\mathbf{e}^* \sim N\left(0, \mathbf{R}(\hat{\boldsymbol{\delta}})\right)$, $\hat{\boldsymbol{\delta}}$ is the REML estimator of $\boldsymbol{\delta}$ and $\hat{\boldsymbol{\beta}}$ is the Least Squares (LS) estimator of β.

The first of the considered estimators, studied by Gonzales-Manteiga et. al. (2008), is given by:

$$\begin{aligned} \hat{MSE}_{\xi}^{boot}\left(\hat{\theta}_{EBLUP}\right) &= E_{*}\left(\hat{\theta}_{EBLUP}\left(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}^{*}),\hat{\boldsymbol{\delta}}^{*}\right)-\theta^{*}\right)^{2} = \\ &= B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_{EBLUP}\left(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}^{*(b)}),\hat{\boldsymbol{\delta}}^{*(b)}\right)-\theta^{*(b)}\right)^{2}, \end{aligned} \tag{22}$$

where $\hat{\boldsymbol{\delta}}^{*}$ is given by the same formula as $\hat{\boldsymbol{\delta}}$, where $\mathbf{Y}$ is replaced by $\mathbf{Y}^{*}$. Furthermore, $\theta^{*(b)}$ is the value of θ obtained in the b-th realization of the bootstrap model, where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$ are REML estimators. Additionally in the simulation study we will also consider the case where $\hat{\boldsymbol{\beta}}$ is an LS estimator of β according to Chatterjee, Lahiri, Li (2008). The expected value in bootstrap distribution is denoted by $E_{*}(.)$ (cf. Molina, Rao, 2010: 376–377).

The MSE estimator considered by Butar, Lahiri (2003) has the following form:

$$\begin{aligned} \hat{MSE}_{\xi}^{boot-BL}\left(\hat{\theta}_{EBLUP}\right) &= g_1(\hat{\boldsymbol{\delta}})+g_2(\hat{\boldsymbol{\delta}})-E_{*}\left(g_1(\hat{\boldsymbol{\delta}}^{*})+g_2(\hat{\boldsymbol{\delta}}^{*})\right)- \\ &-\left(g_1(\hat{\boldsymbol{\delta}})+g_2(\hat{\boldsymbol{\delta}})\right)+E_{*}\left(\hat{\theta}_{EBLUP}\left(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}^{*}),\hat{\boldsymbol{\delta}}^{*}\right)-\hat{\theta}_{EBLUP}(\hat{\boldsymbol{\delta}})\right)^{2}, \end{aligned} \tag{23}$$

where $g_1(\hat{\boldsymbol{\delta}}^{*})$ and $g_2(\hat{\boldsymbol{\delta}}^{*})$ are calculated based on (11) and (12) where $\hat{\boldsymbol{\delta}}$ is replaced by $\hat{\boldsymbol{\delta}}^{*}$. Butar, Lahiri (2003) prove that under some assumptions (23) is asymptotically unbiased in the following sense:

$$E_{\xi}\left(\hat{MSE}_{\xi}^{boot-BL}\left(\hat{\theta}_{EBLUP}\right)\right)-MSE\left(\hat{\theta}_{EBLUP}\right)=o\left(D^{-1}\right). \tag{24}$$

Among considered estimators, the classic jackknife estimator given by (17) and bootstrap estimator given by (23) are asymptotically unbiased under some assumptions. In the case of other estimators MSE bias is not known. We should note that the MSE of the estimators of MSE is not analyzed in small area estimation literature. Furthermore, properties of these estimators are not studied theoretically under misspecified models. We will study these problems in simulation analyses presented in the next two sections. Additionally, the classic estimator requires only to determine elements $g_{1d}(.)$ and $g_{2d}(.)$. These MSE components and the values of the EBLUP are needed to compute the MSE estimator based on the jackknife method, but its value can be negative (Bell, 2001). We can solve this problem using the weighted jackknife estimator. However, in this simulation, studies will show how important the formula of weights is. We should also pay attention to the estimator based on the bootstrap method, which has very simple form and where we only use values of EBLUP and

domain mean based on the parametric bootstrap model realizations. We need to specify the MSE components $g_{3d}(.)$ only for the MSE estimator based on the Taylor expansion.

6. Simulation study – biases of MSE estimators

The purpose of the simulation studies is the Monte Carlo analysis of the properties of the considered MSE estimators presented in sections 4–6, taking into account the increase of the number of domains and the problem of model misspecification.

In the simulation studies we use real data from the Local Data Bank (Polish Central Statistical Office). The considered population elements are Polish regions: poviats (NUTS 4), in the year 2013. The division of the population (of size $N = 379$) into $D = 16$ subpopulations is made according to larger regions – voivodships (NUTS 2). In the analyzed model (2) in the simulation θ_d is the average expenditure on health care in the domain. The average poviats' population (in thousands of people) in the domain in the previous year is the auxiliary variable in the model. Furthermore, the model includes the intercept. The sample, due to the assumption of independence of random components e_d, is drawn from the population as the stratified sample without replacement. We assume that the domains are strata and we assume the approximate proportional allocation of the sample from the strata (c.a. 15% elements from each strata). The relatively high fraction of elements drawn from the strata is due to the small number of elements in some of the strata.

In the simulation study, values of $\hat{\theta}_d$ are generated according to (2) where β is calculated based on the formula (5) for the whole population data set. Random effects are generated using normal distribution. In the simulation study we consider the cases where random effects are independent and where they are correlated. We assume simultaneous spatial autoregressive process (SAR process) for the $\mathbf{v}$ vector (Pratesi, Salvati, 2008: 115–116):

$$\mathbf{v} = \mathbf{G} = \left(\mathbf{I} - \rho\mathbf{W}^{-1}\right)^{-1}\mathbf{u}\,, \tag{25}$$

where $\mathbf{u}$ is D-element vector of independent random effect with variance σ_u^2 and ρ is the unknown parameter. Hence the variance-covariance matrix can by written as:

$$D_{\xi}^2\left(\mathbf{v}\right) = \mathbf{G} = \sigma_u^2\left[\left(\mathbf{I} - \rho\mathbf{W}\right)\left(\mathbf{I} - \rho\mathbf{W}^{-1}\right)\right]^{-1}. \tag{26}$$

The matrix $\mathbf{W}$ is the spatial weight matrix of size $D \times D$. So, we assume correlation between domains, not between the elements of the population. Usually, row

standardization is used to calculate elements of **W**. It should be noted that proximity of domains can be considered not only in geographical but also in economic sense. In geographical sense we can take into account whether objects have a common border (Karpuk, 2015) or the length of common border (Dacey, 1968). In economic sense we can use among others: the unemployment rate, the value of the investments and the number of entities (Pietrzak, 2010). The matrix of the weights can also be based on mutual trade relations, movement of capital and migration between the spatial units (Conley, 1999). In the simulation study, to determine the value of the weights, we use the value of the GDP per capita in the domain. This variable was also used by Kuc (2015). Additionally we assume the following four cases: $\rho = \{-0.8, -0.2, 0.2, 0.8\}$, to check the properties of the MSE estimators in case of this type of model misspecification. The fifth is the case when SAR process does not occur, which is consistent with the assumptions of the model. The problem of defining the weight matrix is presented widely in Suchecki (2010).

Stochastic disturbances are generated using normal distribution with expectation 0 and variances W_d. We calculate the values of W_d using the formula:

$$W_d = \frac{N_d - n_d}{N_d n_d} \frac{1}{N_d - 1} \sum_{i=1}^{N_d} \left(y_i - N_d^{-1} \sum_{i=1}^{N_d} y_i \right)^2 . \tag{27}$$

Often in practice these values are replaced by estimators (c.f. Żądło, 2009: 107) or by smoothed values of these estimators (Wolter, 1985). The parameter A is calculated using REML based on the real data.

In order to investigate the influence of the number of domains on biases of the analyzed MSE estimators we consider two cases where the number of domains equals 16 (original data) and 32. In the second case the original data are enlarged twice. This means that based on the original data we generate values of the average expenditure on health care in each domain twice, assuming the same vector of the auxiliary variables. We assume the number of Monte Carlo iterations at 5000 and the number of bootstrap iterations at 200. The simulation study was prepared using R language (R Development Core Team, 2016). The assumed number of Monte Carlo iterations, bootstrap iterations and size of the considered population can be expected to add up to very time-consuming computations. In each out of 5000 Monte Carlo iterations we have 200 bootstrap iterations (for three MSE estimators) and 16 or 32 jackknife iterations (for three MSE estimators), where iterative algorithms of restricted maximum likelihood methods were used to estimate model parameters.

In Figure 1 we present the values of the relative biases of the MSE estimators when the number of domains equals 16 and 32, for five cases of the value of the correlation coefficient and for the eight MSE estimators: NAIVE given by (15), DL given by (16), JACK given by (17), WJACKa given by (18) with weights calculat-

ed from (19), WJACKb given by (18) with weights calculated from (20), PBOOTa given by (22), PBOOTb given by (22) with the LS estimator of β and PBOOTc given by (23). For JACK and WJACKa estimators, we only present results within the assumed scale of values in Figure 1 and Figure 2. For other MSE estimators, all of the results are presented.

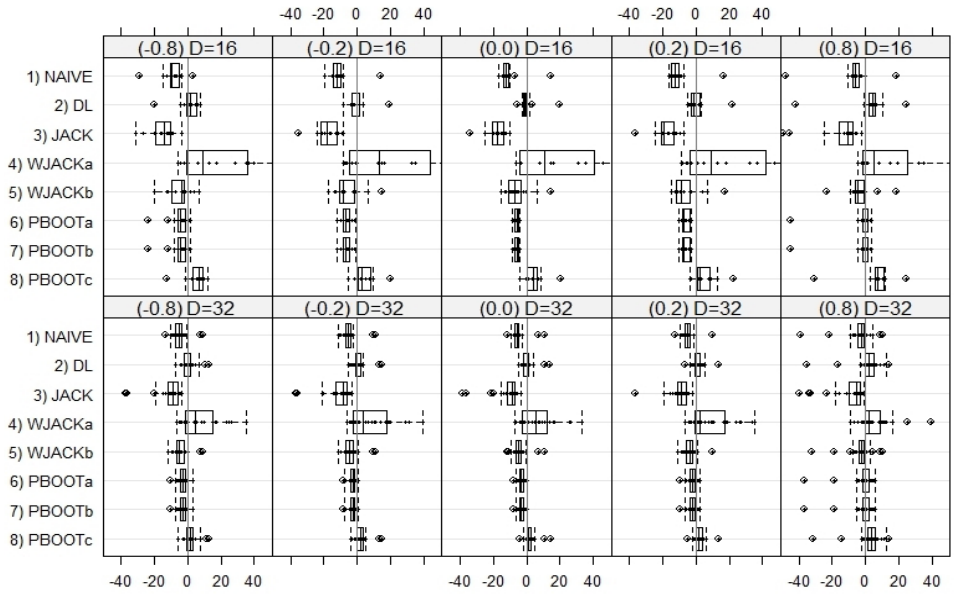

Figure 1. Values of relative biases of MSE estimators in %

Source: own elaboration

Each of the boxplots show 16 or 32 as the values of relative biases of individual MSE estimators in domain. The box represents values between the first and the third quartile and the points are the outliers – values not observed within 1.5 IQR (interquartile range). The line inside the box determines the value of the median of relative biases. In section 7 we present a similar figure for the relative RMSEs of the MSE estimators.

We should note that PBOOTa and PBOOTb estimators, in all cases, give very similar results, so the method estimation of β is irrelevant.

For $D = 16$ we obtain the value of the median of relative biases in domains closest to zero for DL estimator (except in the case of strong positive correlation). However, for weak correlation or the lack of correlation for PBOOTc estimator, we observe the absolute value of the median of relative biases in domains lower than 5%. For cases of strong correlation, similar results are obtained for WJACKa (for positive correlation), WJACKb, PBOOTa and PBOOTb estimators.

We can observe that, for $D = 32$, the values of the median of relative biases in domains for all estimators, except WJACKa, are quite stable regardless the

strength of the correlation. Only for strong positive correlation do we see bigger differences. For NAIVE, JACK, WJACKb, PBOOTa, PBOOTb estimators, we obtain the values of the median of relative biases in domains closer to zero than in other cases. It should be noted that for DL, PBOOTa, PBOOTb and PBOOTc, the absolute value of the median of relative biases in domains is smaller than 4%, in all of cases. For all of the estimators (except WJACKa) and for both of the considered numbers of domains, we see that the interquartile ranges of the values of relative biases are smallest in the case when the assumption of independence of random effects is met. Additionally, in the majority of cases, the lowest values of the interquartile ranges of the relative biases in domains (not higher than 4.2%) were obtained for PBOOTa and PBOOTb estimators.

If we analyze the case where the model is specified correctly, and compare cases when $D = 16$ and $D = 32$ we can see that for all of the estimators of the MSE the value of the median of relative biases in domains for increasing number of the domains is closer to zero.

In conclusion, the results suggest using the parametric bootstrap estimator given by (22) in practice.

7. Simulation study – RMSEs of MSE estimators

On Figure 2 we can see the values of the relative RMSEs of all estimators and all the cases considered in the previous section.

Comparing results for $D = 16$ and $D = 32$ we see that both the value of the medians of RMSEs of the estimators and their interquartile ranges decrease where there is an increasing number of domains.

For $D = 16$ the best results are obtained for NAIVE and DL estimators (except in the case of $\rho = 0.8$). However in some cases, similar results are observed for PBOOTa and PBOOTb estimators.

The values of medians of relative RMSEs for all estimators are stable for $D = 32$ and are not higher than 28%, except for JACK and WJACKa estimators. The lowest values of the median of relative RMSEs are obtained for NAIVE and DL estimators (not higher than 25%). However, for $D = 16$ we can also observe some stability for weak positive and negative correlation and the lack of correlation, though not for the WJACKa estimator.

To summarize, we show in the simulation study that the increasing number of domains has a strong influence on the properties of the considered MSE estimators. It causes a decrease in their relative biases and RMSEs. What is more, the results of the simulation study suggest that for a sufficiently large number of domains, all analyzed MSE estimators are robust on the studied model misspecification, resulting from the correlation of random effects. Furthermore, the results

of the simulation study show that very simple PBOOTa and PBOOTb estimators have good properties both for small numbers of domains and for the considered problem of model misspecification.

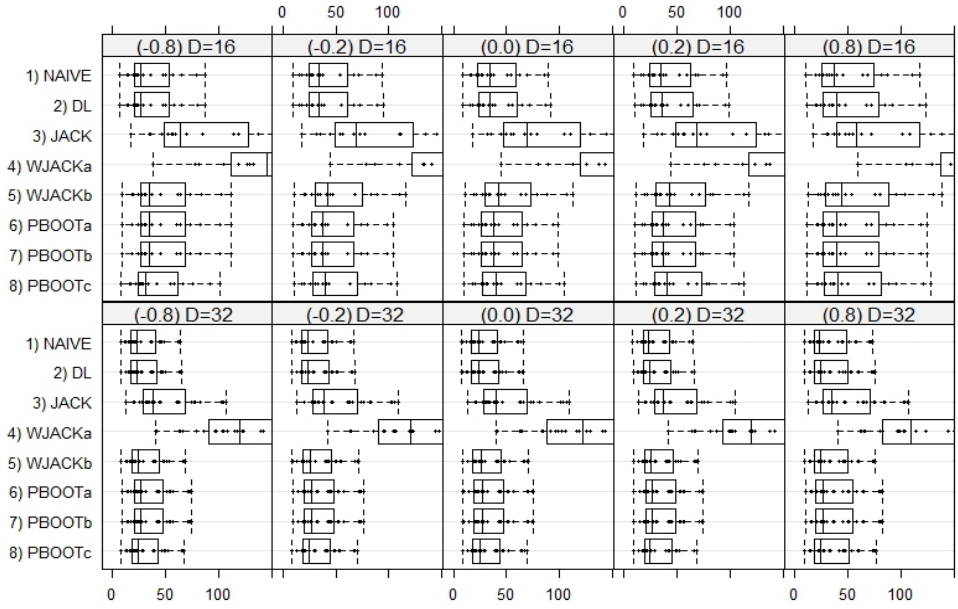

Figure 2. The values of relative RMSE the estimators in %

Source: own elaboration

As was the case with the relative biases of the MSE estimators, the results obtained for the relative RMSEs of the estimators suggest using the bootsrap estimator (22) in practice.

8. Conclusion

In the paper we study properties of MSE estimators of BLUPs of domains means. In the Monte Carlo simulation study we consider the influence of the model misspecification and the increase of the number of domains on their biases and RMSEs.

In the simulation study we show that for the considered real data and Fay-Herriot model, the values of relative biases and relative RMSEs of the estimators decrease due to the increase in the number of domains. Furthermore, it suggests that the considered estimators are quite robust on the considered type of model misspecification, which is presented in section 7 of the paper. We also show the advantages of the very simple parametric bootstrap estimator proposed by Gonzalez-Mantei-

ga et al. (2008). We obtain values of relative biases of the estimator close to zero even if there is a low number of domains and the model is misspecified. Results presented in the paper are an introduction to further, more detailed analyses.

References

Bell W. (1997), *Models for county and state poverty estimates. Preprint*, Statistical Research Division, U.S. Census Bureau.

Bell W. (2001), *Discussion with "Jackknife in the Fay-Herriot Model with An Example"*, "Proc. of the Seminar of Funding Opportunity in Survey Research", pp. 98–104.

Butar F.B., Lahiri P. (2003), *On Measures of Uncertainty of Empirical Bayes Small-Area Estimators*, "Journal of Statistical Planning and Inference", vol. 112, pp. 635–676.

Chatterjee S., Lahiri P., Li H. (2008), *Parametric Bootstrap Approximation to the Distribution of EBLUP and Related Prediction Intervals in Linear Mixed Models*, "The Annals of Statistics", vol. 36, no. 3, pp. 1221–1245.

Chen S., Lahiri P. (2002), *A Weighted Jackknife MSPE Estimator in Small-Area Estimation*, "Proceeding of the Section on Survey Research Methods", American Statistical Association, pp. 473–477.

Chen S., Lahiri P. (2003), *A Comparison of Different MSPE Estimators of EBLUP for the Fay-Herriot Model*, "Proceeding of the Section on Survey Research Methods", American Statistical Association, pp. 905–911.

Conley T.G. (1999), *GMM estimation with cross selection dependence*, "Journal of Econometrics", vol. 92(1), pp. 1–45.

Dacey M. (1968), *A review of measures of contiguity for two and k-color maps*, [in:] B. Berry, D. Marble (eds.), *Spatial analysis: A Reader in Statistical Geography*, Prentice Hall, Englewood Cliffs.

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

Datta G.S., Rao J.N.K., Smith D.D. (2005), *On Measuring the Variability of Small Area Estimators under a Basic Area Level Model*, "Biometrica", vol. 92, pp. 183–196.

Fay R.E. III, Herriot R.A. (1979), *Estimation of Incomes for Small Places: An Application of James-Stein Procedures to Census Data*, "Journal of the American Statistical Association", vol. 74, pp. 269–277, http://dx.doi.org/10.2307/2286322.

Gonzales-Manteiga W., Lombardia M., Molina I., Morales D., Santamaria L. (2008), *Bootstrap Mean Squared Error of Small-Area EBLUP*, "Journal of Statistical Computation and Simulation", vol. 78, pp. 433–462, http://dx.doi.org/10.1007/s00180-008-0138-4.

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

Jędrzejczak A. (2011), *Metody analizy rozkładów dochodów i ich koncentracji*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Jiang J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Science+Business Media, New York.

Jiang J., Lahiri P. (2006), *Mixed Model Prediction and Small Area Estimation*, "Test", vol. 15, no. 1, pp. 1–96, http://dx.doi.org/10.1007/BF02595419.

Jiang J., Lahiri P., Wan S.-M. (2002), *Unified Jackknife Theory for Empirical Best Prediction with M-estimation*, "The Annals of Statistics", vol. 30, pp. 1782–1810.

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

Kackar R.N., Harville D.A. (1984), *Approximations for Standard Errors of Estimators of Fixed and Random Effect in Mixed Linear Models*, "Journal of the American Statistical Association", vol. 79, pp. 853–862.

Karpuk M. (2015), *Wpływ czynników przestrzennych na ruch turystyczny w województwie zachodniopomorskim (2006–2012)*, "Zeszyty Naukowe Wydziału Nauk Ekonomicznych Politechniki Koszalińskiej", vol. 19, pp. 39–56.

Krzciuk M.K. (2015), *On the simulation study of the properties of MSE estimators in small area statistics*, Conference Proceedings. 33rd International Conference Mathematical Methods in Economics 2015, pp. 413–418.

Kuc M. (2015), *Wpływ sposobu definiowania macierzy wag przestrzennych na wynik porządkowania liniowego państw Unii Europejskiej pod względem poziomu życia ludności*, "Taksonomia 24", vol. 384, pp. 163–170.

Lahiri P. (2003), *On the Impact of Bootstrap in Survey Sampling and Small-Area Estimation*, "Statistical Science", vol. 18, no. 2, pp. 199–210.

Lohr S.L., Rao J.N.K. (2009), *Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models*, "Biometrika", vol. 96, pp. 457–468.

Molina I., Rao J. (2010), *Small Area Estimation of Powerty indicators*, "The Canadian Journal of Statistics", vol. 38, no. 3, pp. 369–385.

Prasad N.G.N., Rao J.N.K. (1990), *The Estimation of the Mean Squared Error of Small-Area Estimators*, "Journal of the American Statistical Association", vol. 85, no. 409, pp. 163–171, http://dx.doi.org/10.2307/2289539.

Pietrzak M.B. (2010), *Dwuetapowa procedura budowy przestrzennej macierzy wag z uwzględnieniem odległości ekonomicznej*, "Oeconomia Copernicana", vol. 1, pp. 65–78.

Pratesi M., Salvati N. (2008), *Small Area Estimation: The EBLUP Estimator Based on Spatially Correlated Random Area Effects*, "Statistical Methods and Applications", vol. 17, pp. 113–141, http://dx.doi.org/10.1007/s10260-007-0061-9.

Rao J.N.K. (2003), *Small Area Estimation*, John Wiley & Sons, Hoboken.

Rao J.N.K., You Y. (1994), *Small Area Estimation by Combining Time-Series and Cross-Sectional Data*, "Canadian Journal of Statistics", vol. 22, pp. 511–528.

R Development Core Team (2016), *A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna.

Rao J.N.K., Molina I. (2015), *Small Area Estimation*, John Wiley & Sons, Hoboken.

Rueda C., Mendez J.A., Gomez F. (2010), *Small Area Estimators on Restricted Mixed Models*, "Sociedad de Estadística e Investigación Operativa", vol. 16, pp. 558–579, http://dx.doi.org/10.1007/s11749-010-0186-2.

Slud E.V., Maiti T. (2006), *Mean-Squared Error Estimation in Transformed Fay-Herriot Models*, "Journal of the Royal Statistical Society. Series B (Statistical Methodology)", vol. 68, pp. 239–257.

Suchecki B. (2010), *Ekonometria przestrzenna. Metody i modele analizy danych przestrzennych*, C.H. Beck, Warszawa.

Wang J., Fuller W.A. (2003), *The Mean Squared Error of Small Area Predictors Constructed with Estimated Area Variances*, "Journal of the American Statistical Association", vol. 98, pp. 716–723.

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

Żądło T. (2009), *On prediction of the domain total under some special case of type A general Linear Mixed Models*, "Folia Oeconomica", vol. 228, pp. 105–112.

O badaniu symulacyjnym własności estymatorów MSE predyktora wartości średniej dla modelu Faya-Herriota, bazujących na metodzie jackknife oraz bootstrap

Streszczenie: W artykule rozważany jest problem estymacji błędu średniokwadratowego (MSE) w przypadku predykcji wartości średniej w domenie, w oparciu o model Faya-Herriota. W badaniu symulacyjnym analizowane są własności ośmiu estymatorów MSE, w tym bazujących na metodzie jackknife (Jiang, Lahiri, Wan, 2002; Chen, Lahiri, 2002; 2003) oraz parametrycznej metodzie bootstrap (Gonzalez-Manteiga et al., 2008; Buthar, Lahiri, 2003). W modelu Faya-Herriota zakładana jest niezależność składników losowych, a obciążenia estymatorów MSE są małe dla dużej liczby domen. Celem artykułu jest porównanie własności estymatorów MSE przy różnej liczbie domen i błędnej specyfikacji modelu, wynikającej z występowania korelacji efektów losowych w badaniu symulacyjnym.

Słowa kluczowe: estymatory MSE, metoda jackknife, parametryczna metoda bootstrap, empiryczny najlepszy liniowy nieobciążony predyktor, model Faya-Herriota, badanie symulacyjne

JEL: C15, C83

Received: 2017-01-14; verified: 2017-04-09. Accepted: 2017-10-25