STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 413–432, DOI 10. 21307

ON ASYMMETRY OF PREDICTION ERRORS IN SMALL AREA ESTIMATION

Tomasz Żądło[1]

ABSTRACT

The mean squared error reflects only the average prediction accuracy while the distribution of squared prediction error is positively skewed. Hence, assessing or comparing accuracy based on the MSE (which is the mean of squared errors) is insufficient and even inadequate because we should be interested not only in the average but in the whole distribution of prediction errors. This is the reason why we propose to use different than MSE measures of prediction accuracy in small area estimation. In the prediction accuracy comparisons we take into account our proposal for the empirical best predictor, which is a generalization of the predictor presented by Molina and Rao (2010). The generalization results from the assumption of a longitudinal model and possible changes of the population and subpopulations in time.

Key words: empirical best predictor, prediction errors, small area estimation.

1. Introduction

Nowadays, estimates of the population and large subpopulations characteristics are not sufficient for decision-makers. They require accurate estimates for subpopulations with small or even zero sample sizes. However, because of cost constraints, it is not possible to increase sample sizes continuously to make the estimation of smaller and smaller subpopulations possible using classical methods. The problem is solved using small area estimation methods "borrowing strength" from other subpopulations or time periods. There are three aims of our paper, with the first the main one and the second and the third the supplementary aims.

Firstly, our observation that the distribution of squared prediction errors has strong positive asymmetry (values of the third standardized moments obtained in the simulation study based on real data are presented in Table 1) has become a focus of our attention. It implies that their mean (known as MSE – Mean

[1] University of Economics in Katowice, Faculty of Management, Katowice, Poland.
E-mail: tomasz.zadlo@ue.katowice.pl.

Squared Error) does not have to be a good measure of prediction accuracy in terms of the average and, what is in our opinion even more important, their whole distribution should be studied. Hence, the main purpose of the paper is our proposal for assessing the prediction accuracy using new univariate and multivariate prediction measures based on quantiles of the distribution of absolute prediction errors. It will be shown that even if the accuracies of two predictors in terms of the average are similar, the accuracy comparison based on right tails of distributions of absolute prediction errors can give different results (which will be presented in e.g. Figure 4).

Secondly, in the accuracy comparisons resulting from the main aim we will include a generalization of the Empirical Best Predictor (EBP) proposed by Molina and Rao (2010). They proposed the predictor under a model assumed for data from surveys conducted in one period. We will propose a predictor assuming a longitudinal model. It means that in the case of longitudinal surveys we will be able to use information from previous periods to increase the prediction accuracy in the period of interest.

Thirdly, in the proposed longitudinal model we will take into account that the population and subpopulations may change in time. It will cover many longitudinal models known from small area estimation (which are special cases of the general linear mixed models) including models studied by:

- Saei and Chambers (2003), who assume mutually independent two random effects (domain-specific and time-specific) and random components (and the generalization, where AR(1) process is assumed for time-specific random effects),
- Saei and Chambers (2003), where domain-and-time-specific random effects with independent distributions in domains and AR(1) model in time are taken into account,
- Stukel and Rao (1999) and Nissinen (2009) p. 22 with mutually independent two random effects (domain-specific and element-specific) and random components,
- Nissinen (2009) p. 60, who assumes independent domain-specific random effects and autocorrelated (assuming AR(1)) random components in time,
- Molina, Morales, Pratesi, Tzavidis (2010) pp. 143-180 with independent domain-specific random effects, independent for domains and autoccorelated (assuming AR(1)) in time domain-and-time-specific random effects and heteroscedastic random components.

In the simulation study the properties of the proposed predictor (in terms of MSE and the proposed accuracy measures) will be studied under the proposed model taking into account the model misspecification as well.

2. Alternative prediction accuracy measures

In the case of positive asymmetry usually the mean is not the only measure used to describe the distribution. The MSE is the mean of squared errors (which have positive or even strong positive asymmetry) and it is usually used as the only accuracy measure. Moreover, a better predictor is usually defined as the one with smaller MSE. Żądło (2013) proposed a new measure of prediction accuracy Quantile of Absolute Prediction Error defined for the problem of prediction in the *d*th domain as follows:

$$QAPE(p)=\inf\left\{x:P\left(\left|U_d\right|\le x\right)\ge p\right\}, \tag{1}$$

where $U_d=\hat{\theta}_d-\theta_d$ is the prediction error of $\hat{\theta}_d$, which is the predictor of θ_d in the *d*th domain. It means that (1) is the quantile of order *p* of $|U_d|$. It means that at least $p100\%$ of realizations of absolute prediction errors in the *d*th domain are smaller or equal to $QAPE(p)$. In Żądło (2013) it was used to measure prediction accuracy of the empirical best linear unbiased predictor.

Żądło (2015) proposes multivariate versions of (1), which allow us to measure and compare accuracy in the case of simultaneous prediction in all of domains. It can be treated as the alternative to the average mean squared error studied, e.g. by Fabrizi and Trivisano (2010). Let prediction errors in D domains be denoted by $U_d=\hat{\theta}_d-\theta_d$, where $d=1,2,...,D$. Let us define the multivariate version of $QAPE$ as follows:

$$MQAPE(p)=\inf\left\{x:\sum_{d=1}^{D}P\left(\left|U_d\right|\le x\right)\ge Dp\right\}. \tag{2}$$

It means that it is the quantile of order *p* of a distribution of a mixture of random variables $|U_1|,...,|U_d|,...,|U_D|$ with equal weights. It means that at least $p100\%$ of realizations of absolute prediction errors in all domains are smaller or equal to $MQAPE(p)$.

Let relative prediction errors be denoted by $W_d=\frac{U_d}{\theta_d}=\frac{\hat{\theta}_d-\theta_d}{\theta_d}$, where $d=1,2,...,D$. Let us define relative $MQAPE$ as follows:

$$rMQAPE(p)=\inf\left\{x:\sum_{d=1}^{D}P\left(\left|W_d\right|\le x\right)\ge Dp\right\}. \tag{3}$$

It means that it is the quantile of order *p* of a distribution of a mixture of random variables $|W_1|,...,|W_d|,...,|W_D|$ with equal weights. It means that at least $p100\%$ of realizations of moduli of relative prediction errors in all domains are smaller or equal to $rMQAPE(p)$.

The estimation of (1), (2) and (3) is possible using a well-known parametric bootstrap method studied, e.g. by González-Manteiga et al. (2007, 2008) and

Molina and Rao (2010). Using the method, the estimator of the MSE is given by the mean of squared bootstrap realizations of prediction errors. Similarly, by computing quantiles of bootstrap realizations of:

- moduli of prediction errors in one of domains we can estimate (1),
- moduli of prediction errors in all of domains we can estimate (2) and
- moduli of relative prediction errors in all of domains we can estimate (3).

3. Model and predictor

We consider longitudinal data in periods $t=1,2,...,M$, where the population of size N_t in the period t is denoted by Ω_t. The population is divided into D disjoint subpopulations (domains) Ω_{dt} each of size N_{dt}, where $d=1,2,...,D$. A sample in the period t of size n_t is denoted by s_t. Let $s_{dt}=s_t\cap\Omega_{dt}$ and $\bar{\bar{s}}_{dt}=n_{dt}$. The d*th domain of interest in the period of interest t* will be denoted by Ω_{d*t*}. Let $\Omega_{rdt}=\Omega_{dt}-s_{dt}$, $N_{rdt}=N_{dt}-n_{dt}$, $\bigcup_{t=1}^{M}\Omega_t=\Omega$, $\bar{\bar{\Omega}}=N$, $\bigcup_{t=1}^{M}\Omega_{dt}=\Omega_d$, $\bar{\bar{\Omega}}_d=N_d$, $\bigcup_{t=1}^{M}\Omega_{rdt}=\Omega_{rd}$, $\bar{\bar{\Omega}}_{rd}=N_{rd}$, $\bigcup_{t=1}^{M}s_t=s$, $\bar{\bar{s}}=n$, $\bigcup_{t=1}^{M}s_{dt}=s_d$, $\bar{\bar{s}}_d=n_d$.

Let M_{id} be the number of periods when the ith population element belongs to the dth domain and m_{id} – the number of periods when the ith population element (which belongs to the dth domain) is observed. Let $M_{rid}=M_{id}-m_{id}$. It is assumed that the population may change in time and that one population element may change its domain affiliation in time. Hence, sets of population elements Ω_d (where $d=1,2,...,D$) may overlap.

The assumption that one population element may change its domain affiliation in time is very important in practice of longitudinal surveys. For example, let us consider the population of households and the division of the population into domains made according to the household size. In this case we should assume that some households can change their sizes in time, which causes the change of the domain affiliation. If a human population is under the study one may be interested in its characteristics for subpopulations defined according to some social or economic criteria. In the case of business surveys the population of firms may be divided into subpopulations according to some economic or financial criteria, what can imply even stronger changes of domains affiliations.

Values of the variable of interest (or the variable of interest after a transformation) are realizations of Y_{idj}'s for the ith population element, which belongs to the dth domain in the period t_{ij}, where $i=1,2,...,N$; $j=1,2,...,M_{id}$;

$d=1,2,...,D$. The vector $\mathbf{Y_{id}}=\left[Y_{idj}\right]_{M_{id}\times 1}$ will be called the profile and the vector $\mathbf{Y_{sid}}=\left[Y_{idj}\right]_{m_{id}\times 1}$ will be called the sample profile. Let the vector $\mathbf{Y_{rid}}=\left[Y_{idj}\right]_{M_{rid}\times 1}$ be the profile for non-observed realizations of random variables.

Let us introduce assumptions of the following longitudinal model, which is a special case of the general linear mixed model (e.g. Datta and Lahiri, 2000). The difference is introduced in the sizes of matrices, which allows us to take into account longitudinal data and possible changes in population and subpopulations in time. We assume that

$$\begin{cases}\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{v}+\mathbf{e}\\ D_{\xi}^{2}(\mathbf{v})=\mathbf{G}(\boldsymbol{\delta})\\ D_{\xi}^{2}(\mathbf{e})=\mathbf{R}(\boldsymbol{\delta})\\ Cov_{\xi}(\mathbf{v},\mathbf{e})=\mathbf{0}\end{cases}, \tag{4}$$

where ξ is the superpopulation model, $\mathbf{Y}=col_{1\le d\le D}col_{1\le i\le N_d}(\mathbf{Y_{id}})$, $\mathbf{e}=col_{1\le d\le D}col_{1\le i\le N_d}(\mathbf{e_{id}})$, where $\mathbf{e_{id}}$ is the $M_{id}\times 1$ vector of random components, $\mathbf{X}=col_{1\le d\le D}col_{1\le i\le N_d}(\mathbf{X_{id}})$, where $\mathbf{X_{id}}$ is the known matrix of size $M_{id}\times p$, $\boldsymbol{\beta}$ is the $p\times 1$ vector of unknown parameters, $\mathbf{Z}$ is the known matrix of size $\sum_{i=1}^{N}\sum_{d=1}^{D}M_{id}\times h$, $\mathbf{v}$ is the vector of random effects of size $h\times 1$, $\boldsymbol{\delta}$ is the vector of q unknown in practice parameters called variance components.

Let us consider the following decomposition of the vector $\mathbf{Y}$:

$$\mathbf{Y}=\begin{bmatrix}\mathbf{Y}_s^T & \mathbf{Y}_r^T\end{bmatrix}^T, \tag{5}$$

where $\mathbf{Y}_s$ is the vector of size $\sum_{i=1}^{N}\sum_{d=1}^{D}m_{id}\times 1$ of random variables, whose realizations are known, and $\mathbf{Y}_r$ is the vector of size $\sum_{i=1}^{N}\sum_{d=1}^{D}M_{rid}\times 1$ of random variables, which are not observed in the longitudinal survey. Then,

$$D_{\xi}^{2}(\mathbf{Y})=D_{\xi}^{2}\begin{bmatrix}\mathbf{Y}_s\\ \mathbf{Y}_r\end{bmatrix}=\mathbf{V}(\boldsymbol{\delta})=\begin{bmatrix}\mathbf{V}_{ss}(\boldsymbol{\delta}) & \mathbf{V}_{sr}(\boldsymbol{\delta})\\ \mathbf{V}_{rs}(\boldsymbol{\delta}) & \mathbf{V}_{rr}(\boldsymbol{\delta})\end{bmatrix}, \tag{6}$$

where under (4):

$$\mathbf{V}(\boldsymbol{\delta})=\mathbf{Z}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^T+\mathbf{R}(\boldsymbol{\delta}). \tag{7}$$

Let us consider the problem of predicting any given function of the random vector $\mathbf{Y}$ denoted by $\theta(\mathbf{Y})$ or shortly by θ. Among predictors $\hat{\theta}$ of θ, the Best Predictor (BP) is defined as the one, which minimizes (e.g. Molina and Rao 2010):

$$MSE_{\xi}(\hat{\theta}) = E_{\xi}(\hat{\theta} - \theta)^2 . \tag{8}$$

Hence, it is given by:

$$\hat{\theta}_{BP} = E_{\xi}(\theta \mid \mathbf{Y}_s), \tag{9}$$

which means that it may be obtained as a conditional expected value of θ assuming that the conditional distribution of $\mathbf{Y}_r \mid \mathbf{Y}_s$ is known.

We assume that the conditional distribution of $\mathbf{Y}_r \mid \mathbf{Y}_s$ can be derived (the example is presented in *Remark 1* in this section). In practice, it depends on the vector of unknown parameters, which will be denoted by $\boldsymbol{\tau}$. If we replace the parameters by their estimators, we obtain the Empirical Best Predictor (EBP) denoted by $\hat{\theta}_{EBP}$. Hence, the value of the EBP of $\theta(\mathbf{Y})$ can be obtained through the Monte Carlo approximation algorithm presented below (for prediction in surveys conducted in one period see Molina and Rao 2010).

(a) We estimate $\boldsymbol{\tau}$ based on the realization of $\mathbf{Y}_s$ and we obtain the value of the estimator denoted by $\hat{\boldsymbol{\tau}}$.

(b) Assuming that the distribution of $\mathbf{Y}_r \mid \mathbf{Y}_s$ can be derived, we generate L vectors $\mathbf{Y}_r$ (denoted by $\mathbf{Y}_r^{(l)}$, where $l = 1,2,...,L$) from the distribution of $\mathbf{Y}_r \mid \mathbf{Y}_s$, where the unknown vector $\boldsymbol{\tau}$ is replaced by $\hat{\boldsymbol{\tau}}$.

(c) We make L vectors denoted by $\mathbf{Y}^{(l)}$, where $\mathbf{Y}^{(l)} = \begin{bmatrix} \mathbf{Y}_s^T & \mathbf{Y}_r^{(l)T} \end{bmatrix}^T$ and $l = 1,2,...,L$, what means that L vectors $\mathbf{Y}^{(l)}$ include the same realization of $\mathbf{Y}_s$ and different realizations of $\mathbf{Y}_r$.

(d) The value of the EBP of $\theta(\mathbf{Y})$ is obtained as follows: $\hat{\theta}_{EBP} = L^{-1} \sum_{l=1}^{L} \theta(\mathbf{Y}^{(l)})$.

Due to the estimation of an unknown in practice vector of model parameters denoted by $\boldsymbol{\tau}$, the resulting predictor generally is not unbiased and it does not minimize the MSE (as the BP) but its value should be very close to the BP. Its MSE estimator, which takes into account the uncertainty resulting from the estimation of $\boldsymbol{\tau}$, can be obtained using parametric bootstrap method as in Molina and Rao (2010), where their model is replaced by (4).

Remark 1. If we additionally assume that the vector $\mathbf{Y}$ (which may be the vector of the variable of interest after a transformation) is normally distributed, which can be written as follows $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\delta}))$ (where under (4) $\mathbf{V}(\boldsymbol{\delta})$ is given by (7)), then $\boldsymbol{\tau} = \begin{bmatrix} \boldsymbol{\beta}^T & \boldsymbol{\delta}^T \end{bmatrix}^T$ and

$$\mathbf{Y}_r \mid \mathbf{Y}_s \sim N\left(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{V}_{rs}(\boldsymbol{\delta})\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right), \mathbf{V}_{rr}(\boldsymbol{\delta}) - \mathbf{V}_{rs}(\boldsymbol{\delta})\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{V}_{sr}(\boldsymbol{\delta})\right). \tag{10}$$

Hence, in the step (b) of the procedure presented above, vectors $\mathbf{Y}_r^{(l)}$, where $l=1,2,...,L$ are generated based on (10), where parameters are replaced by their estimates.

The idea of using EBPs was presented earlier by Molina and Rao (2010) but for studies conducted in one period. They study the general case assuming the general linear mixed model for studies conducted in one period. In the special case of their considerations $\mathbf{Y}$ is the vector of the variable of interest after the following transformation: $\mathbf{Y}=T(\ddot{\mathbf{Y}})$, where $\ddot{\mathbf{Y}}$ is the variable of interest, and

$$T(\ddot{\mathbf{Y}})=\ln(\ddot{\mathbf{Y}}+c), \tag{11}$$

where c is a constant. Then, they study the following model

$$Y_{id}=\mathbf{x}_{id}^T\boldsymbol{\beta}+v_d+e_{id}, \tag{12}$$

where $d=1,2,...,D$; $i=1,2,...,N$, $v_d\overset{iid}{\sim}N(0,\sigma_v^2)$, $e_{id}\overset{iid}{\sim}N(0,\sigma_e^2)$, e_{id} and v_d are independent, $\boldsymbol{\delta}=\begin{bmatrix}\sigma_e^2 & \sigma_v^2\end{bmatrix}^T$.

4. Simulation study – real data

We consider data on $N=378$ Polish poviats (NUTS-4 level) from years 2010-2012 (M=3). We have excluded one observation because of the lack of the data and one outlying observation (Warsaw). The problem of prediction of totals of the sold production of industry in $D=16$ domains (voivodships – NUTS-2 level) for companies with at least 10 employees is considered. The number of companies with at least 10 employees is the auxiliary variable. In the first period a sample of 38 poviats is drawn at random with probabilities proportional to the values of the auxiliary variable. Sample sizes in the domains are random and equal from 0 to 5 (with mean 2.375). The balanced panel survey is considered – elements sampled in the first period are observed until the end of the longitudinal survey (which gives 114 observations in 3 periods).

Because empirical best predictors are studied, the distribution of the variable of interest must be assumed. We consider the transformation of the variable of interest given by (11) and logarithmic transformation of the auxiliary variable. To test the distribution of the variable of interest, we use the transformation of residuals based on the Cholesky decomposition of the inverse of variance-covariance matrix (see, e.g. Jacqmin-Gadda et al. 2007). For the model chosen based on the AIC and BIC criteria (more details will be presented in the next paragraph) p-values for Shapiro-Wilk, Jarque-Bera and adjusted Jarque-Bera tests obtained for the sample equal 0.2297; 0.6046 and 0.446 respectively. For the considered model but without the transformations of both variables, p-values for the tests of normality were smaller than 10^{-12}. But if we test normality based on the whole population data (based on $M\times N=3\times 378=1134$ observations) in both cases (with and without transformations of variables) we should reject the null hypothesis on normality. That is why the problem of model misspecification will be taken into account in the simulation study as well.

To choose the appropriate model we consider different models: classic and mixed linear models with and without the auxiliary variable, with and without constant, nested-error models and models with random slopes. For mixed models with random slope we consider time-specific, domain-specific, time-and-domain-specific and finally profile-specific random effects. In models with nested errors we consider one random effect (time-specific, domain-specific, time-and-domain-specific and profile-specific) or two random effects (firstly: domain-specific and profile-specific; secondly: domain-specific and domain-and-time-specific). The model with the smallest both AIC and BIC criteria was the following model studied earlier by Stukel and Rao (1999) and Nissinen (2009) p. 22:

$$Y_{idt} = x_{idt}\beta_1 + \beta_0 + u_d + v_{id} + e_{idt}, \tag{13}$$

where Y_{idt} is the variable of interest after transformation (11), x_{idt} is the auxiliary variable after logarithmic transformation, $i = 1,2,...,N$, $d = 1,2,...,D$, $t = 1,2,...,M$, u_d, v_{id} and e_{idt} are mutually independent with zero expected values and variances given by σ_u^2, σ_v^2 and σ_e^2 respectively. Permutation test (with the test statistic given by the loglikelihood) was used to test the significance of the model parameters – at the significance level 0.05 tested parameters were significantly different from zero. Good properties of these tests are presented by Krzciuk and Żądło (2014a, 2014b).

The model-based simulation study was prepared using R software (R Core Team 2016). To mimic the real data, values of the variable of interest after transformation (11) are generated based on the model (13) with one auxiliary variable and the constant, where the parameters of the model are replaced by REML estimates obtained based on all of the observations (sampled and unsampled) of the real data. Hence, both random effects and random components are generated with zero expected values and variances σ_u^2, σ_v^2 and σ_e^2 equal REML estimates based on (13) and the whole population data. Random effects and random components u_d, v_{id} and e_{idt} are generated independently from:

- normal distributions,
- shifted exponential distributions (the third standardized moment is equal to 2),
- shifted gamma distributions (with the value of third standardized moment equal to 4) and
- shifted Pareto distributions (with the value of third standardized moment equal to 5).

It means that in the case of the normal and the shifted exponential distributions, the assumed values of the mean and the variance give explicitly values of the parameters of the distributions used in the simulation study. The case of the shifted gamma and the shifted Poisson distributions is more interesting because it is possible to set the values of the parameters of these distributions to obtain not only the assumed values of the mean and the variance but also the prespecified value of the third standardized moment (4 - for the shifted gamma and 5 - for the shifted Pareto distributions, as listed above).

In each iteration of the simulation study model parameters are estimated using restricted maximum likelihood, which gives consistent estimates even if the normality assumption is not met (Jiang 1996). The number of iterations equals 5000.

We study properties of the following predictors:

- the empirical best predictor based on the longitudinal model (13) under normality of random effects and random components (EBP),
- the empirical best predictor studied earlier by Molina and Rao (2010) based on the model (12) assumed for transformed data (EBP-MR),
- the empirical best linear unbiased predictor based on the Royall (1976) theorem for the longitudinal model with the smallest AIC and BIC criteria assumed for the untransformed data, i.e. for the mixed model with random regression coefficient with the profile-specific random effect (EBLUP),
- the synthetic regression estimator given by (SYNT-REG) given by (e.g. Bracha 1996, p. 260):

$$N_{dt}\left(\sum_{i\in s_t}\pi_{ti}^{-1}\right)^{-1}\sum_{i\in s_t}y_{ti}\pi_{ti}^{-1}+N_{dt}B\left(N_{dt}^{-1}\sum_{i\in\Omega_{dt}}x_{t*i}-\left(\sum_{i\in s_t}\pi_{ti}^{-1}\right)^{-1}\sum_{i\in s_t}x_{ti}\pi_{ti}^{-1}\right),$$

for $d=1,2,...,D$, where

$$B=\frac{\sum_{i\in s_t}\left(x_{ti}-\left(\sum_{i\in s_t}\pi_{t*i}^{-1}\right)^{-1}\sum_{i\in s_t}x_{ti}\pi_{ti}^{-1}\right)\left(y_{ti}-\left(\sum_{i\in s_t}\pi_{ti}^{-1}\right)^{-1}\sum_{i\in s_t}y_{ti}\pi_{ti}^{-1}\right)\pi_{ti}^{-1}}{\sum_{i\in s_t}\left(x_{ti}-\left(\sum_{i\in s_t}\pi_{ti}^{-1}\right)^{-1}\sum_{i\in s_t}x_{ti}\pi_{ti}^{-1}\right)^{2}\pi_{ti}^{-1}}$$, and π_{ti} is

the inclusion probability of the *i*th population element in the period *t*.

Because of small sample sizes in the domains (in some domains: 0) we study only indirect predictors and estimators. In each out of 5000 Monte Carlo iterations values of both empirical best predictors are computed based on $L=200$ generated population vectors.

Relative prediction biases for the considered estimators and predictors are presented in Figure 1. Each boxplot presents $D=16$ values of biases of a predictor of $D=16$ domains totals. For example, the values presented in the top-left boxplot are from ca 0.3% to ca 2.1%. The value 2.1% means that for one of the domains the relative bias of EBP predictor equals 2.1% (in this domain the value of the predictor is larger than the domain total on average by 2.1%). If the distributions of the random components and random effects for the transformed variables are normal, the biases of all predictors and estimators are small. EBP in this case is used under the correctly specified longitudinal model – the transformation of the variables, the assumed normal distribution and the assumed formula of the model (13) are correct. EBP-MR is used under the misspecified formula of the model (assumed for one period instead of the longitudinal data) but under the correct transformation of the variables and assuming correct (i.e. normal) distribution. Both EBLUP and SYNT-REG do not take into account the transformation of the variables. If the distribution is asymmetric, the biases are very large in many cases.

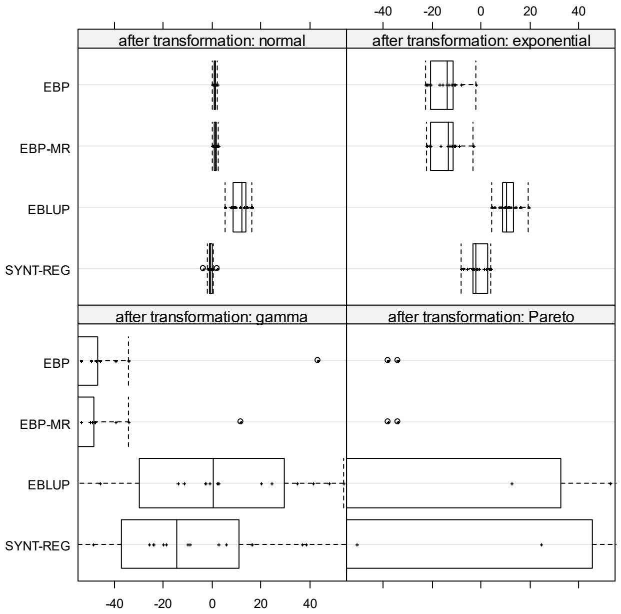

Figure 1. Relative prediction biases (in %) for different predictors and different distributions of random effects and random components (each boxplot presents values for D=16 domains)

Results of the comparisons of the accuracy between the proposed empirical best predictor EBP and other estimators and predictors based on the MSE are presented in Figure 2. Each boxplot presents $D=16$ values of ratios of the MSE of a predictor to the MSE of EBP for $D=16$ domains totals. For example, the values presented in the top-left boxplot are from ca 1.02 to ca 1.06. The value 1.06 means that for one of the domains the ratio of the MSE of EBP-MR predictor to the MSE of EBP predictor equals 1.06 (in this domain the value of the MSE of EBP-MR is higher than the MSE of EBP by 6%). If we compare the MSE of EBP-MR to the MSE of the EBP for other distributions, we see that the values of ratios are also very close to 1. In all of the cases the maximum gain in accuracy due to the usage of the proposed predictor measured by the MSE is smaller than 10%. The reasons of the results will be studied in the next section.

What is interesting, in the results presented in Figure 2 is the lack of stability comparing results for different distributions of random effects and random components. The reason of unstable results is strong positive asymmetry of the distribution of absolute prediction errors, especially if the distribution of random effects and random components is not normal (see values of the third standardized moments of absolute prediction errors presented in Table 1). Because the values of the prediction MSE (the values of the mean of squared errors) are strongly affected by outlying absolute prediction errors, results for alternative measures of prediction accuracy defined in section 2 will be presented as well.

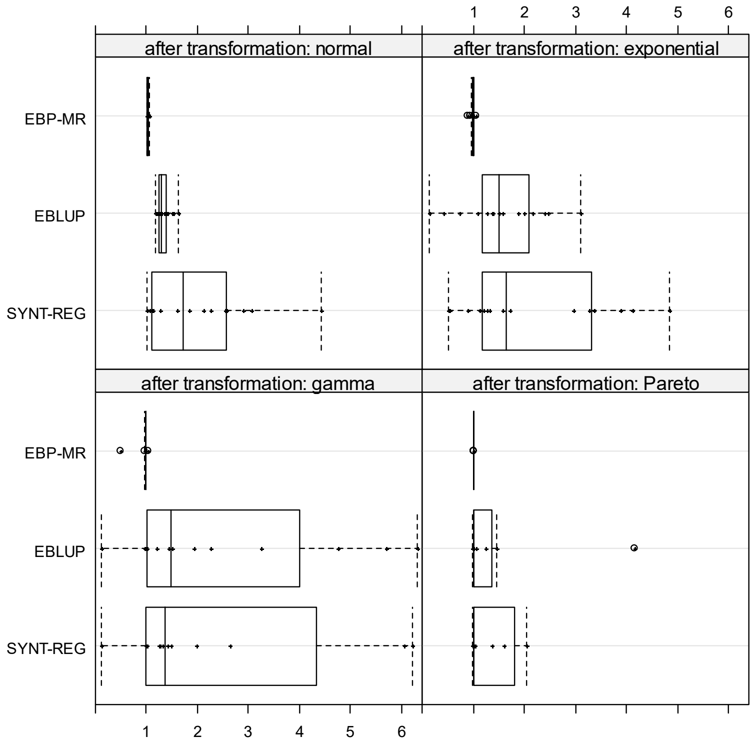

Figure 2. Values of MSE(.)/MSE(EBP) for different predictors and different distributions of random effects and random components (each boxplot presents values for *D*=16 domains)

Table 1. Third standardized moments of absolute prediction errors for different predictors and different distributions of random effects and random components (minimum and maximum for *D*=16 domains)

	After transformation:			
	normal	shifted exponential	shifted gamma	shifted Pareto
SYN-REG	1.3-4.0	8.3-52.7	24.4-70.6	33.1-70.7
EBLUP	1.2-3.6	12.4-55.4	28.7-70.6	34.3-70.7
EBP-MR	1.6-4.4	13.8-61.6	17.5-70.7	44.7-70.7
EBP	1.5-4.5	13.7-61.6	17.5-70.7	44.7-70.7

Firstly, we will compare the accuracy of the predictors based on the same real data. In this case we will use $QAPE(p)$ for $p=(0.5,0.75,0.9,0.95)$ for each domain. It is worth mentioning that the results presented in Figure 3 (and Figures 6-8 in Appendix) are more stable than the results presented in Figure 2. Each boxplot in Figure 3 presents $D=16$ values of ratios of the $QAPE(0.5)$ of a predictor to the $QAPE(0.5)$ of EBP for $D=16$ domains totals. As it was defined and discussed in the section 3, $QAPE(0.5)$ is the median of absolute prediction errors. For example, the values presented in the top-left boxplot are from ca 1 to ca 1.05. The value 1.05 means that for one of the domains the ratio of the $QAPE(0.5)$ of EBP-MR predictor to the $QAPE(0.5)$ of EBP predictor equals 1.05 (in this domain the value of the $QAPE(0.5)$ of EBP-MR is higher than the $QAPE(0.5)$ of EBP by 5%).

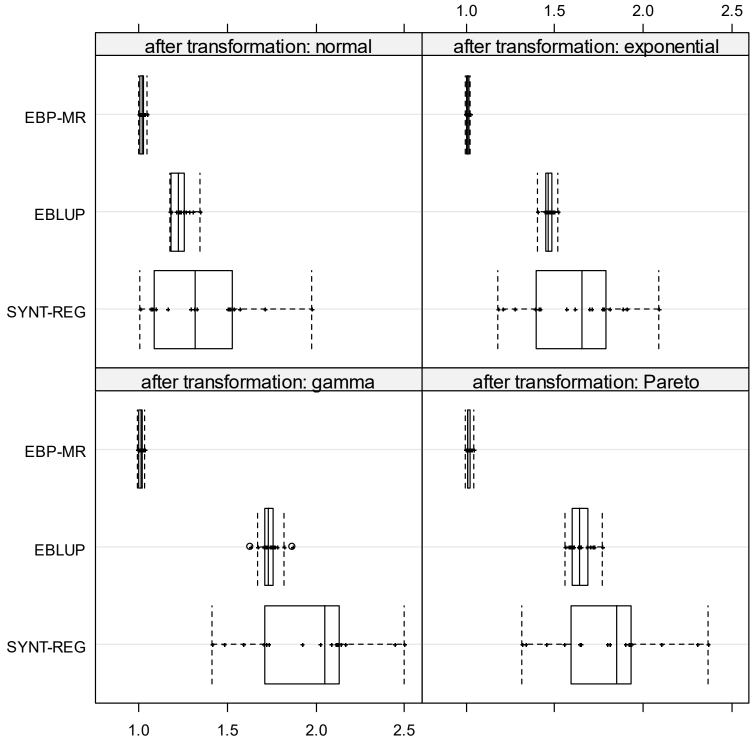

Figure 3. Values of QAPE0.50(.)/QAPE0.50(EBP) for different predictors and different distributions of random effects and random components (each boxplot presents values for *D*=16 domains)

Additionally, in Figure 4 the values of $rMQAPE(p)$ for $p=(0.5,0.75,0.9,0.95)$ are presented. As an example, we will interpret the value presented by the point in the top-left part of Figure 7 (for EBP under normal distribution of random effects and random components for the variable of interest after the transformation), which equals $rMQAPE(0.5)=18.2\%$. It means that at least 50% of moduli of relative prediction errors for all of the domains are smaller or equal to 18.2% and at least 50% of moduli of relative prediction errors for all of the domains are larger or equal to 18.2%. But $rMQAPE(0.5)$ informs only about the average (i.e. median) of absolute prediction errors. If we are interested in the right tail of the distribution of the absolute prediction errors, we can compute $rMQAPE(p)$ for $p>0,5$, e.g. $rMQAPE(0.75)=32.8\%$, $rMQAPE(0.9)=51.2\%$ and finally $rMQAPE(0.95)=67\%$ (see the top-left part of Figure 4 and top-left part of Figure 9 in Appendix).

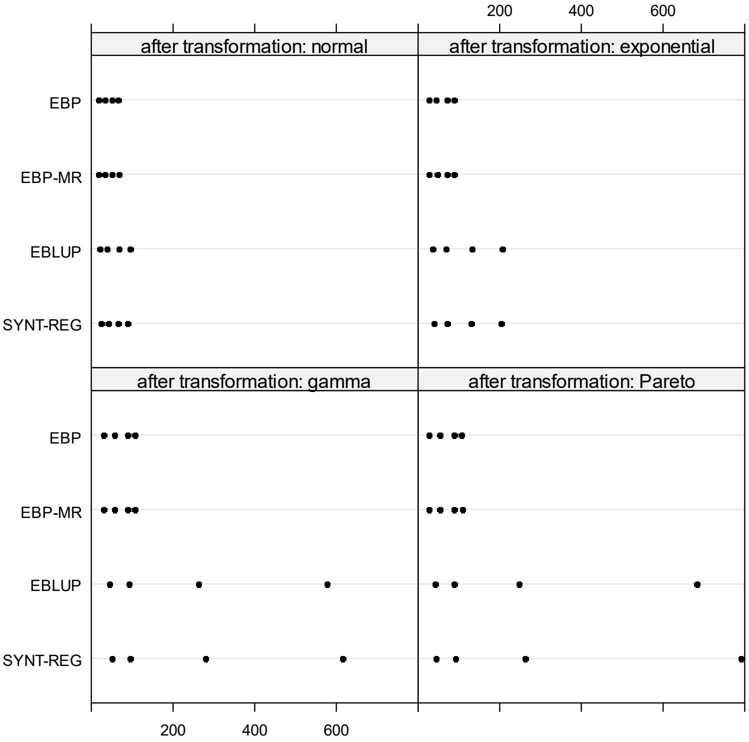

Figure 4. Values of rMQAPE(p) for p=(0.5, 0.75, 0.9, 0.95), different predictors and different distributions of random effects and random components - all results

It should be stressed that although values of $rMQAPE(0.5)$ for each predictor are quite similar even in the case of model misspecification (see the first point for each predictor in Figure 4 or Figure 9 in Appendix), the difference in the accuracy measured in right tails of the distribution of absolute prediction errors by $rMQAPE(0.95)$ can differ substantially especially in the case of model misspecification (see the last point for each predictor in Figure 4). For example in bottom-right panel in Figure 4, values of $rMQAPE(0.5)$ for EBP and SYNT-REG equal 27% and 45%, respectively, which means that $rMQAPE(0.5)$ for SYNT-REG is 1.67 times higher than for EBP. However, values of $rMQAPE(0.95)$ for EBP and SYNT-REG equal 102% and 776%, respectively, which means that $rMQAPE(0.95)$ for SYNT-REG is 7.6 times higher than for EBP. To sum up, the prediction accuracy measures presented in section 2 give us more detailed information on prediction accuracy, which is not limited to the average values (as in the case of the MSE). What is more, using *QAPE* we obtain more stable results of accuracy comparisons, especially in the case of model misspecification.

5. Simulation study – artificial data

In the previous section two problems were considered – the comparison of the accuracy and the choice of the appropriate measures of accuracy. One of the conclusions was the small difference in the accuracy (less than 10% in terms of the MSE for all considered distributions) between the proposed empirical best predictor for longitudinal surveys and the empirical best predictor proposed by Molina and Rao (2012) for surveys conducted in one period. To identify the reasons, we compare some results from the previous section (the column "real data" in Table 2) with two additional simulation scenarios, all of them under normality of random effects and random components for data after transformation (11) and assuming model (13) or its special case.

Table 2. Maximum values of ratios of accuracy measures for different simulation scenarios over D=16 domains under normality of random effects and random components

	Simulation scenario		
ratio of accuracy measures	real data	independent values of x	without x
MSE(EBP-MR)/MSE(EBP)	1.058	1.257	1.123
QAPE0.50(EBP-MR)/QAPE0.50(EBP)	1.046	1.119	1.040
QAPE0.75(EBP-MR)/QAPE0.75(EBP)	1.023	1.134	1.028
QAPE0.90(EBP-MR)/QAPE0.90(EBP)	1.029	1.117	1.042
QAPE0.95(EBP-MR)/QAPE0.95(EBP)	1.047	1.154	1.047

In the first scenario (results in Table 2 in the column "independent values of x"), we generate values of the variable of interest based on model (13) with values of all model parameters obtained for the real data (as in the previous section), but where the real auxiliary variable is replaced by the artificial one. Values of the auxiliary variable were generated independently from shifted gamma distribution assuming real values of the mean, variance and the third standardized moment for each year. In this case the maximum gain in accuracy of our EBP measured by MSE is 25.7% and measured by $QAPE$ is higher than 10%. It means that in the case of longitudinal surveys we should use auxiliary variable, which is weakly autocorrelated but even in this case the gain in accuracy will not be very large.

In the second scenario (results in Table 2 in the column "without x") we do not use the uxiliary variable both in the model and at the estimation stage. Hence, we compare prediction accuracy of empirical best predictors only under random parts of models (13) and (12). The accuracy measured by MSE of our EBP is higher by 12.3% compared with EBP-MR (by less than 5% in terms of $QAPE$). The reason is that model (13) chosen based on AIC and BIC for real longitudinal data is quite similar to the model (12) assumed by Molina and Rao (2010). In both models we have domain-specific random effects, although in the case of (13) it additionally implies non-zero covariances between observations within domains in different periods. The main difference between the models is the profile (element)-specific random effect in model (13), but results in the last column of Table 2 show that it does not imply a large gain in prediction accuracy. It means that the larger gain in accuracy can be obtained when the longitudinal model explains the variability of the variable of interest considerably better than the model assumed for one period.

To sum up, in this section based on the Monte Carlo analysis we have identified two reasons of the relatively small gain in accuracy, which was presented in the previous section, comparing our predictor with the predictor proposed by Molina and Rao (2010). Firstly, it has been autocorrelation in time of the auxiliary variable. Secondly, we have presented similarity of the proposed longitudinal model and the model studied by Molina and Rao (2010). Moreover, we have shown that in the studied cases the maximum gain in accuracy comparing these two predictors can be even higher than 25% in terms of MSE.

6. Real data application

In this section we consider values of the same predictors and estimators, the same data and the same sample as discussed in section 4. However, in this case their values are computed once based on the real data (they are not generated as in the simulation studies presented in section 4). Because the whole population data are available, we are able to compare estimates with real values of D=16 domains totals (see Figure 5). The largest differences between estimates and real values for the considered sample are observed for SYNT-REG and EBLUP, whereas the values of EBP-MR and the proposed EBP are very similar.

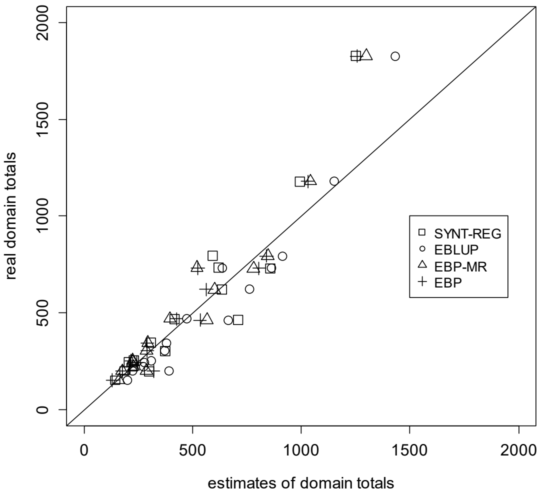

Figure 5. Values of estimates and real domain totals

7. Conclusions

In the paper the problem of assessing and comparing the prediction accuracy is studied. Because of strong positive asymmetry of absolute prediction error, it is shown that prediction accuracy measures alternative to the MSE should be used. These measures allow us to assess the prediction accuracy not limited to the average values and to obtain more stable results of accuracy comparisons, especially in the case of the model misspecification. In the accuracy comparisons based on the Monte Carlo simulation studies our proposal for the empirical best predictor is taken into account. Although its prediction accuracy was only slightly better for the considered data compared with the empirical best predictor proposed by Molina and Rao (2012), we present how to obtain a substantial gain in accuracy. The considerations are also supported by real data application.

REFERENCES

BRACHA, CZ., (1996). Teoretyczne podstawy metody reprezentacyjnej, PWN, Warszawa.

DATTA, G. S., 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.

FABRIZI, E., TRIVISANO, C., (2010). Robust linear mixed models for Small Area Estimation, Journal of Statistical Planning and Inference, Vol. 140, pp. 433–443.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M.J., MOLINA, I., MORALES, D., SANTAMARÍA, L., (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model, Computational Statistics and Data Analysis, Vol. 51, pp. 2720–2733.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D., SANTAMARÍA, L., (2008). Bootstrap mean squared error of small-area EBLUP, Journal of Statistical Computation and Simulation, Vol. 78, 443–462.

JACQMIN-GADDA, H., SIBILLOT, S., PROUST, C., MOLINA J.-M., THIÉBAUT, R., (2007). Robustness of the Linear Mixed Model to Misspecified Error Distribution, Computational Statistics & Data Analysis, Vol. 51, pp. 5142–5154.

JIANG, J., (1996). REML Estimation: Asymptotic Behavior and Related Topics, The Annals of Statistics, Vol. 24, pp. 255–286.

KRZCIUK M., ŻĄDŁO T., (2014a). On some tests of variance components for linear mixed models, Studia Ekonomiczne, Vol. 189, pp. 77–85.

KRZCIUK M., ŻĄDŁO T., (2014b). On some tests of fixed effects for linear mixed models, Studia Ekonomiczne, Vol. 189, pp. 49–57.

MOLINA, I., RAO, J. N. K., (2010). Small Area Estimation of Poverty Indicators, The Canadian Journal of Statistics, Vol. 38, pp. 369–385.

NISSINEN, K., (2009). Small Area Estimation With Linear Mixed Models For Unit-Level Panel and Rotating Panel Data, University of Jyväskylä Printing House, Jyväskylä.

R CORE TEAM, (2016). R: A Language and Environment For Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

ROYALL, R. M., (1976). The Linear Least Squares Prediction Approach to Two-Stage Sampling, Journal of the American Statistical Association, Vol. 71, pp. 657–473.

STUKEL, D. M, RAO, J. N. K., (1999). On Small-Area Estimation Under Two-Fold Nested Error Regression Models, Journal of Statistical Planning and Inference, Vol. 78, pp. 131–147.

ŻĄDŁO, T., (2013). On Parametric Bootstrap and Alternatives of MSE, Proceedings of 31st International Conference Mathematical Methods in Economics 2013, College of Polytechnics Jihlava, pp. 1081–1086.

ŻĄDŁO, T., (2015). Statystyka małych obszarów w badaniach ekonomicznych. Podejście modelowe i mieszane, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

APPENDIX

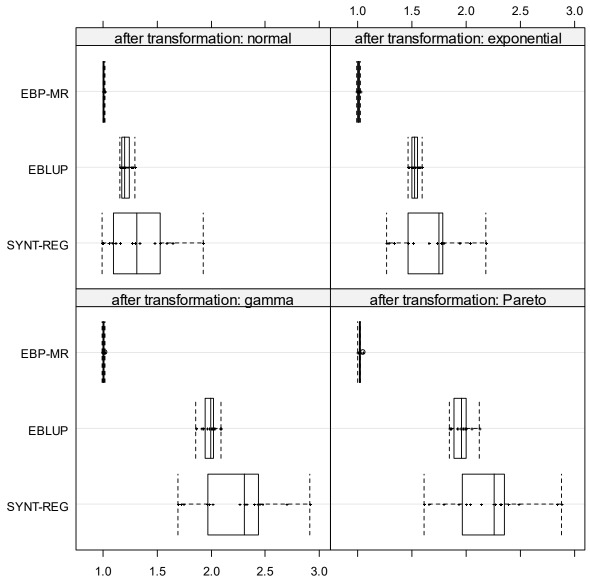

Figure 6. Values of QAPE0.75(.)/QAPE0.75(EBP) for different predictors and different distributions of random effects and random components (each boxplot presents values for *D*=16 domains)

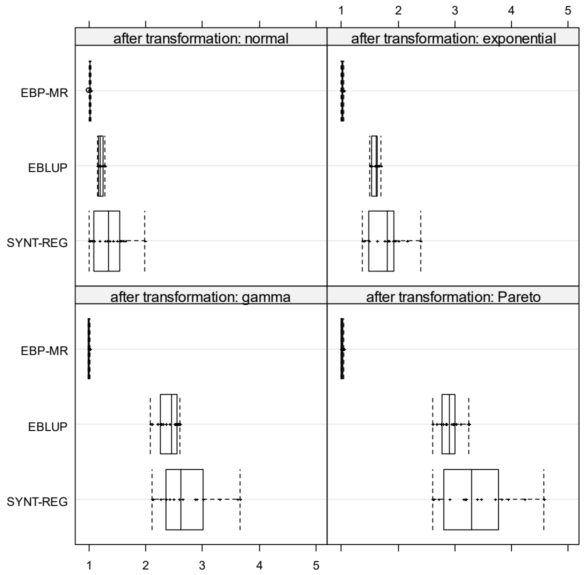

Figure 7. Values of QAPE0.90(.)/QAPE0.90(EBP) for different predictors and different distributions of random effects and random components (each boxplot presents values for *D*=16 domains)

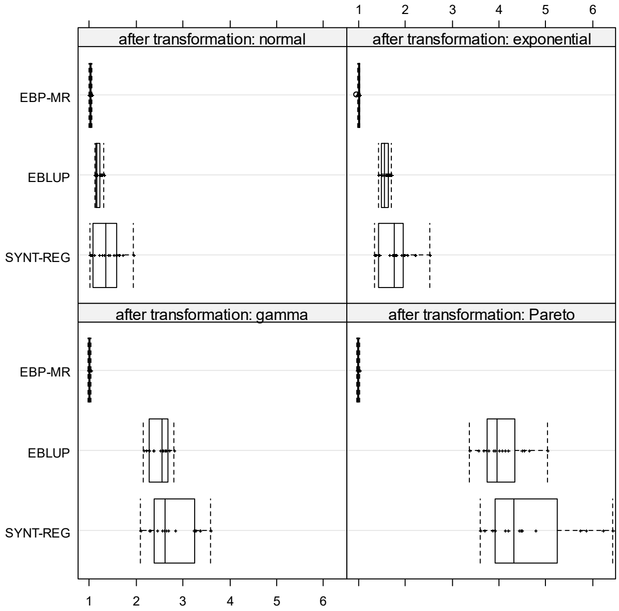

Figure 8. Values of QAPE0.95(.)/QAPE0.95(EBP) for different predictors and different distributions of random effects and random components (each boxplot presents values for *D*=16 domains)

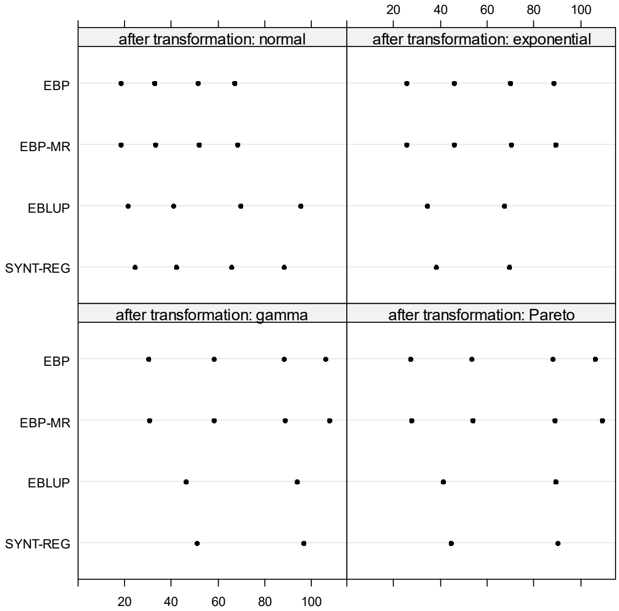

Figure 9. Values of rMQAPE(p) for p=(0.5, 0.75, 0.9, 0.95), different predictors and different distributions of random effects and random components – selected results