Wiadomości Statystyczne. The Polish Statistician, 2025, *70*(1), 1–18 https://doi.org/10.59139/ws.2025.01.1
Studia metodologiczne / Methodological studies
Zgłoszony/submitted: 18.04.2023, poprawiony/revised: 20.02.2024, zaakceptowany/accepted: 15.11.2024

Simulation comparison of the empirical properties of the best predictor under the assumption of linear mixed model with correlated and uncorrelated random effects

Małgorzata K. Krzciuk[a]

Abstract. In the paper, we are considering the problem of the prediction for small areas under the assumption of a linear mixed model with correlated random effects. The aim of the simulation study was to analyse the properties of the proposed empirical best predictors (EBP) under the assumption of the model with two correlated random effects specific for domains. Additionally, we addressed the problem of the reduced accuracy resulting from the estimation of model parameters and the influence of model misspecification in the case of the lack of correlation. Scripts for the Monte Carlo simulation analyses are prepared in R language and based on real data on entities registered in the REGON base in 2017 from the Local Data Bank of Statistics Poland.

The study demonstrates that the EBP for a model with correlation has good properties even if no correlation between random effects is assumed. In such a case, the mean loss of accuracy resulting from the model misspecification is no more than 2%. The results indicate that for larger sample sizes and more realisations of random effects, the model parameters, for example ρ, are estimated more accurately, which can have a significant impact on the properties of the proposed predictor, including its accuracy.

Keywords: small area estimation, prediction, empirical best predictor, EBP, linear mixed model, Monte Carlo, simulation study

JEL: C53, C51, C63, C83

Symulacyjne porównanie własności empirycznych najlepszych predyktorów przy założeniu liniowego modelu mieszanego ze skorelowanymi i nieskorelowanymi efektami losowymi

Streszczenie. W artykule rozważany jest problem predykcji charakterystyk dla małych obszarów przy założeniu liniowego modelu mieszanego ze skorelowanymi efektami losowymi. Przeprowadzone badanie symulacyjne ma na celu porównanie własności empirycznych najlepszych predyktorów (ang. *empirical best predictor* – EBP) przy założeniu modelu z dwoma skorelowanymi efektami losowymi specyficznymi dla domen. Ponadto uwzględniono problem zmniejszenia dokładności wynikający z szacowania parametrów modelu oraz wpływ błędnej specyfikacji modelu w przypadku braku korelacji. Skrypty analiz symulacyjnych Monte Carlo zostały opracowane w języku R z wykorzystaniem danych z Banku Danych Lokalnych GUS o podmiotach nowo zarejestrowanych w rejestrze REGON w 2017 r.

Z badania wynika, że EBP dla modelu z korelacją ma dobre własności nawet przy założeniu braku korelacji między efektami losowymi (przeciętny spadek dokładności wynikający z błędnej specyfikacji modelu nie przekroczył 2%). Uzyskane wyniki wskazują, że dla większych liczebności prób i większej liczby realizacji efektów losowych parametry modelu, w tym ρ, są

[a] Uniwersytet Ekonomiczny w Katowicach, Wydział Zarządzania, Polska / University of Economics, Faculty of Management, Poland. ORCID: https://orcid.org/0000-0002-5906-5744. E-mail: malgorzata.krzciuk@uekat.pl.

© Małgorzata K. Krzciuk
Artykuł udostępniony na licencji CC BY-SA 4.0 / Article available under the CC BY-SA 4.0 licence

szacowane dokładniej, co może mieć znaczący wpływ na własności proponowanego predyktora, w tym jego dokładność.
Słowa kluczowe: statystyka małych obszarów, predykcja, empiryczny najlepszy predyktor, EBP, liniowy model mieszany, Monte Carlo, badania symulacyjne

1. Introduction

In this paper, we are analysing the small area estimation problem. Small area estimation methods are very important in statistics. They make the estimation or prediction possible in the cases where classical estimation methods are inefficient or too costly, and allow the estimation based on very small, even zero-size samples in a subpopulation. The study adopts two main approaches to small area estimation, i.e. model-based and design-based. We analyse predictions using the empirical best predictor (EBP) of some characteristics in domains under a linear mixed model (LMM). The class of LMMs finds many applications in the prediction of economic data, which was widely described in the literature, e.g. by Pratesi and Salvati (2009) or Marhuenda et al. (2017). The issue of using small area estimation methods for economic data and official statistics is presented in e.g. Tzavidis et al. (2018). In this paper, we widen the former scope of research by taking into account the presence of a correlation between vector of random effects. More precisely, in our analyses, a model with two correlated vectors of random effects specific for domains is considered. This model was proposed to make predictions in the realm of small area estimation by means of the empirical best linear unbiased predictor (EBLUP) in Krzciuk (2020), but in this paper, we applied it to EBPs.

The aim of the simulation study was to analyse the properties of the proposed empirical best predictors (EBP) under the assumption of the model with two correlated random effects specific for domains. Additionally, we addressed the problem of the reduced accuracy resulting from the estimation of model parameters and the influence of model misspecification in the case of the lack of correlation. Particular attention was paid to the simulation bias and accuracy of the examined predictors. Parts of studies based on a model-based approach took into account the problem of model misspecification, and the last part was based on a randomisation approach.

2. Linear mixed model

In this paper, we assume a division of the population of size N denoted by Ω into D domains Ω_d, each of size N_d ($\sum_{d=1}^{D} N_d = N$), where $d = 1, 2, \dots, D$. We additionally assume the following decomposition of the vector of the study variable $\boldsymbol{Y}$: $\boldsymbol{Y} = [\boldsymbol{Y_s}^{\mathrm{T}} \quad \boldsymbol{Y_r}^{\mathrm{T}}]^{\mathrm{T}}$, where $\boldsymbol{Y_s}^{\mathrm{T}}$ is the vector of size n, for elements which were drawn to

sample s, and the vector $\boldsymbol{Y}_{\boldsymbol{r}}^{\mathrm{T}}$ of size $N_r = N - n$ corresponds to elements not drawn to the sample. Furthermore, the sample in the d-th domain of size n_d is denoted by s_d.

We are considering a model which belongs to the class of LMMs. An LMM has the following form:

$$\begin{cases} \boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{v} + \boldsymbol{e} \\ \mathrm{D}^2(\boldsymbol{v}) = \mathbf{G}(\boldsymbol{\delta}) \\ \mathrm{D}^2(\boldsymbol{e}) = \mathbf{R}(\boldsymbol{\delta}) \end{cases}, \tag{1}$$

where:
$\boldsymbol{Y}$ is the random vector of values of the dependent variable,
$\mathbf{X}$, $\mathbf{Z}$ are the known matrices of auxiliary variables,
$\boldsymbol{\beta}$ is the vector of unknown parameters,
$\boldsymbol{v}$ and $\boldsymbol{e}$ are independent vectors of random effects and stochastic disturbances with variance-covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively; additionally, $\boldsymbol{\delta}$ is the vector of unknown parameters called 'variance components' (see, e.g. Molina & Rao, 2015). For the LMM, vector $\boldsymbol{\delta}$ includes σ_v^2 and σ_e^2.

The variance-covariance matrix of $\boldsymbol{Y}$ has the following form:

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

where e.g. $\mathbf{V}_{ss}(\boldsymbol{\delta})$ and $\mathbf{V}_{rr}(\boldsymbol{\delta})$ are variance-covariance submatrix for elements which were drawn to the sample s and elements not drawn to the sample, respectively.

Moreover, for (1):

$$\mathbf{V}(\boldsymbol{\delta}) = \mathbf{Z}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z} + \mathbf{R}(\boldsymbol{\delta}). \tag{3}$$

Model (1) can also have the following form:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{v}_1 + \mathbf{Z}_2\boldsymbol{v}_2 + \cdots + \mathbf{Z}_h\boldsymbol{v}_h + \boldsymbol{e}, \tag{4}$$

where:

$$\mathrm{Var}\begin{bmatrix}\boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_h\end{bmatrix} = \begin{bmatrix}\mathbf{G}_{11} & \mathbf{G}_{12} & \cdots & \mathbf{G}_{1h} \\ \mathbf{G}_{21} & \mathbf{G}_{22} & \cdots & \mathbf{G}_{2h} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{G}_{h1} & \cdots & \cdots & \mathbf{G}_{hh}\end{bmatrix}, \tag{5}$$

$\boldsymbol{v}_i$ is a vector of i-th random effect, for $i = 1, \ldots, h$,
$\mathbf{Z}_i$ is the matrix of auxiliary variables, for $i = 1, \ldots, h$.

We should note that for $i \neq j$ it is possible that submatrix $\mathbf{G}_{ij} \neq \mathbf{0}$. We assume that in this case the covariance matrix of stochastic disturbances has the following form: $\mathbf{R}(\boldsymbol{\delta}) = \sigma_e^2 \mathrm{diag}(v_i)$ for $1 \leq i \leq N$.

In this and the next paragraph of this section, we are reviewing the models, having in mind the rest of the article. However, we mainly focus on models with correlated random effects. We consider some special cases of LMMs. The first group is comprised of LMMs with one random effect. Then, we look into the nested-error-regression models and models with a random slope. The nested-error-regression model was examined by Battese et al. (1988). If we assume that random effects are specific for domains, this model has the following form:

$$Y_{id} = \beta_1 x_{id} + \beta_0 + v_d + e_{id}, \tag{6}$$

where $d = 1, 2, \ldots, D$ (D is the total number of domains) and $i = 1, 2, \ldots, N$ (N is the population size).

Random effects v_d and stochastic disturbance e_{id} have variances σ_v^2 and σ_e^2, respectively and are mutually independent. Known matrix of auxiliary variables $\mathbf{Z}$ for this model is a block-diagonal matrix of dimension N by D:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1}_{N_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{N_2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \cdots & \cdots & \mathbf{1}_{N_D} \end{bmatrix}_{N \times D}, \tag{7}$$

where $\mathbf{1}_{N_1}, \mathbf{1}_{N_2}, \ldots, \mathbf{1}_{N_D}$ are unit matrices.

Variance-covariance matrices $\mathbf{G}$ and $\mathbf{R}$ in this case have the form of:

$$\mathbf{G}(\boldsymbol{\delta}) = \sigma_{v_d}^2 \mathbf{I}_{D \times D}, \tag{8}$$

$$\mathbf{R}(\boldsymbol{\delta}) = \sigma_e^2 \mathrm{diag}(v_i) \text{ for } 1 \leq i \leq N. \tag{9}$$

By inserting these matrices into the formula (3) we obtain the following covariance matrix of $\mathbf{Y}$:

$$\mathbf{V}(\boldsymbol{\delta}) = \mathrm{diag}_{1 \leq d \leq D} \mathbf{V}_{\boldsymbol{d}} = \mathrm{diag}_{1 \leq d \leq D} \left(\sigma_{v_d}^2 \mathbf{1}_{N_d} \mathbf{1}_{N_d}^{\mathrm{T}} + \sigma_e^2 \mathbf{I}_{N_d \times N_d} \right). \tag{10}$$

The second LMM with one random effect is a model with a random slope in the form of:

$$Y_{id} = (\beta_1 + v_d) x_{id} + \beta_0 + e_{id}. \tag{11}$$

Random effects v_d and stochastic disturbance e_{id} have variances σ_v^2 and σ_e^2, respectively, and are mutually independent. This model belongs to the class of models with random parameters, which are considered e.g. by Dempster et al. (1981). In this model, matrix $\mathbf{Z}$ has the following form:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{X_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X_2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \cdots & \cdots & \mathbf{X_D} \end{bmatrix}_{N \times D}, \tag{12}$$

where $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_D}$ are submatrices of auxiliary variables.

The variance-covariance matrix of random effects specific for domains in this case reads

$$\mathbf{G}(\boldsymbol{\delta}) = \sigma_{v_d}^2 \mathbf{I}_{D \times D}, \tag{13}$$

and for stochastic disturbance has the same form as in (6), so matrix $\mathbf{V}$ could be written as:

$$\mathbf{V}(\boldsymbol{\delta}) = \mathrm{diag}_{1 \le d \le D} \mathbf{V_d} = \mathrm{diag}_{1 \le d \le D} \left(\sigma_{v_d}^2 \mathbf{X}_d \mathbf{X}_d^{\mathrm{T}} + \sigma_e^2 \mathbf{I}_{N_d \times N_d} \right), \tag{14}$$

where $\mathbf{X}_d$ is submatrix of auxiliary variables.

The next two models, which are a special case of (4), include two random effects v_{1d} and v_{2d}. Both of them are specific for domains. In the first one, we assume that the vectors of random effects are uncorrelated, and the model has the following form (Krzciuk, 2020):

$$Y_{id} = (\beta_1 + v_{2d}) x_{id} + \beta_0 + v_{1d} + e_{id}, \tag{15}$$

where random effects with variances $\sigma_{v_{1d}}^2$ and $\sigma_{v_{2d}}^2$ and stochastic disturbance e_{id} with variance σ_e^2 are mutually independent.

In this case, following Krzciuk (2020), matrix of auxiliary variables $\mathbf{Z}$ is more complex:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1}_{N_1} & \mathbf{X_1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{N_2} & \mathbf{X_2} & \cdots & \mathbf{0} & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{N_D} & \mathbf{X_D} \end{bmatrix}_{N \times 2D}. \tag{16}$$

Matrix $\mathbf{G}(\boldsymbol{\delta})$ in this case is the block-diagonal matrix:

$$\mathbf{G}(\boldsymbol{\delta}) = \begin{bmatrix} \mathbf{G_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G_2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \cdots & \cdots & \mathbf{G_D} \end{bmatrix}_{2D\times 2D}, \tag{17}$$

where each of blocks has the form of $\mathbf{G_d} = \begin{bmatrix} \sigma^2_{v_{1d}} & 0 \\ 0 & \sigma^2_{v_{2d}} \end{bmatrix}$ with $d = 1, 2, \ldots, D$.

Taking into account (16) and (17), we obtain matrix $\mathbf{V}$ written as:

$$\mathbf{V}(\boldsymbol{\delta}) = \mathrm{diag}_{1\le d\le D}\mathbf{V_d} = \mathrm{diag}_{1\le d\le D}\left(\sigma^2_{v_d}\mathbf{1}_{N_d}\mathbf{1}^{\mathrm{T}}_{N_d} + \sigma^2_{v_{2d}}\mathbf{x}_d\mathbf{x}^{\mathrm{T}}_d + \sigma^2_e\mathbf{I}_{N_d\times N_d}\right). \tag{18}$$

We should note that (18) includes components of variance-covariance matrices (10) and (14) for both of the models with one random effect examined in the paper.

The last model is the LMM with two correlated vectors of random effects, v^{ρ}_{1d} and v^{ρ}_{2d} (Krzciuk, 2020):

$$Y_{id} = \left(\beta_1 + v^{\rho}_{2d}\right)x_{id} + \beta_0 + v^{\rho}_{1d} + e_{id}. \tag{19}$$

Random effects v^{ρ}_{1d} and v^{ρ}_{2d} have variances $\sigma^2_{v^{\rho}_{1d}}$ and $\sigma^2_{v^{\rho}_{2d}}$, respectively, and are correlated. Stochastic disturbance has variance σ^2_e. In this model, matrices $\mathbf{Z}$ and $\mathbf{G}$ have the same form as in the model with uncorrelated vectors of random effects, but block matrix $\mathbf{G_d}$ where d=1, 2, ..., D, has nonzero elements out of the diagonal:

$$\mathbf{G_d} = \begin{bmatrix} \sigma^2_{v^{\rho}_{1d}} & \rho\sigma_{v^{\rho}_{1d}}\sigma_{v^{\rho}_{2d}} \\ \rho\sigma_{v^{\rho}_{1d}}\sigma_{v^{\rho}_{2d}} & \sigma^2_{v^{\rho}_{2d}} \end{bmatrix}. \tag{20}$$

Therefore, it is a more general case than the model with uncorrelated vectors of random effects (14) and the matrix (17).

In variance-covariance matrix of $\mathbf{Y}$, we have additional elements, compared to (18), and it has the following form (Krzciuk, 2020):

$$\begin{aligned} \mathbf{V}(\boldsymbol{\delta}^{\rho}) = \mathrm{diag}_{1\le d\le D}\mathbf{V_d} = \mathrm{diag}_{1\le d\le D}(\sigma^2_{v^{\rho}_{1d}}\mathbf{1}_{N_d}\mathbf{1}^{\mathrm{T}}_{N_d} + \sigma^2_{v^{\rho}_{2d}}\mathbf{x}_d\mathbf{x}^{\mathrm{T}}_d + \\ +\rho\sigma_{v^{\rho}_{1d}}\sigma_{v^{\rho}_{2d}}\left(\mathbf{1}_{N_d}\mathbf{x}^{\mathrm{T}}_d + \mathbf{x}_d\mathbf{1}^{\mathrm{T}}_{N_d}\right) + \sigma^2_e\mathbf{I}_{N_d\times N_d}). \end{aligned} \tag{21}$$

We can also consider more complex models, for example LMMs with three random effects, where two are correlated, or an LMM with four random effects, where two pairs are correlated.

We should note that LMMs, which take into account the correlation between random effects, allow a wider analysis and thus can be applied to, for example, the estimation of plasma concentration of some drug by nonlinear mixed effects model (Dumont et al., 2014), analyses of the healthcare costs at the end of life (Menec et al., 2004), or analyses of the bias and the precision of the estimates for pharmacokinetics and pharmacodynamics (Ogungbenro et al., 2008).

Furthermore, in addition to the correlation between random effects, we can analyse the autocorrelation of a random effect. For example, Pratesi and Salvati (2008) use simultaneously-autoregressive process of random effects for the estimation of the annual per-capita mean income, Tiao and Ali (1971) consider normal mixed autoregressive moving average process of random effects, and Skoglund and Karlsson (2001) analyse serial correlation of random effects in the estimation of production of the Japanese chemical industry.

3. A model with correlated random effects vs. a two-level model

While considering the problem of the prediction under a model belonging to the class of LMMs, we should also pay attention to a two-level model. Let us compare model (4) with the two-level model examined by Moura and Holt (1999), Torabi and Rao (2008) and Molina and Rao (2015):

$$Y_{id} = \boldsymbol{x}_{di}^{\mathrm{T}}\boldsymbol{\beta}_d + e_{di}, \tag{22}$$

where:

$\boldsymbol{\beta}_d = \tilde{\mathbf{Z}}_d\boldsymbol{\beta} + \boldsymbol{v}_d$ and $\tilde{\mathbf{Z}}_d$ is a $p \times q$ matrix,

$\boldsymbol{\beta}$ is a $q \times 1$ vector of regression parameters,

$\boldsymbol{v}_d$ $(d = 1, 2, \dots, D)$ are – what is important in our case – assumed to be independent,

$\boldsymbol{v}_d \sim (\mathbf{0}, \boldsymbol{\Sigma})$ and e_{di} are independent and possibly heteroscedastic as in (4).

Model (4) considered in this paper is more general than the two-level model (22) because in (22):

- only one level of grouping is taken into account in the case of random effects, i.e. if d denotes domains, then only domain-specific random effects are taken into account ($\boldsymbol{v}_d$), while in the case of model (4), we can take into account e.g. four vectors of random effects: the domain-specific ($\boldsymbol{v}_d$), the group-specific ($\boldsymbol{v}_g$), the time-specific ($\boldsymbol{v}_t$) and the profile-specific ($\boldsymbol{v}_i$) one in one model;

- the same variance-covariance matrix $\boldsymbol{\Sigma}$ is assumed for all the vectors of random effects $\boldsymbol{v}_d$ $(d = 1, 2, \ldots, D)$, i.e. $\boldsymbol{v}_d \sim (\mathbf{0}, \boldsymbol{\Sigma})$, while in model (4), variance-covariance matrices of vectors of random effects can be defined in different ways, as shown in (5);
- vectors of random effects are assumed to be independent, while in (4), the form of the dependence can be defined (see covariance matrices $\mathbf{G}_{ij}$ in (5).

To sum up, the model with correlated random effects (4) can be simplified to the two-level model (22) if only one grouping variable is used to define random effects, if their variance-covariance matrices are identical (i.e. assuming in (5) that $\mathbf{G}_{ii} = \boldsymbol{\Sigma}$) and if all vectors of random effects in (4) are independent (and hence in (5), we assume that $\mathbf{G}_{ij} = \mathbf{0}$ for all $i \neq j$).

Another difference between the findings of our study and the results presented by Moura and Holt (1999) and Torabi and Rao (2008) is that those authors consider EBLUPs under their model, while we analyse the EBP in this paper.

Molina and Rao (2015) emphasise that the two-level model enables the effective use of unit level and area level covariates in one model. We believe that the above-mentioned generalisation of a model may be an even better choice for small area prediction problems in real surveys.

4. Empirical best predictor

In this section, we will be presenting predictors examined in our analyses. We are looking into the problem of the prediction of any function of the population vector $\boldsymbol{Y}$ denoted by θ. Among predictors $\hat{\theta}$, the best predictor is the one that minimises:

$$MSE_{\xi}(\hat{\theta}) = E_{\xi}(\hat{\theta} - \theta)^2, \tag{23}$$

in other words, the mean squared prediction error.

Hence, the best predictor has the form of (e.g. Molina & Rao, 2010):

$$\hat{\theta}_{BP} = E(\theta \,|\boldsymbol{Y}_s), \tag{24}$$

which means that it is the conditional expected value of the function of random variables θ, assuming that the form of the conditional distribution $\boldsymbol{Y}_r \,|\boldsymbol{Y}_s$ is known. In practice, the conditional distribution $\boldsymbol{Y}_r \,|\boldsymbol{Y}_s$ depends on the vector of unknown parameters $\boldsymbol{\tau}$ (i.e. $\boldsymbol{\tau} = [\boldsymbol{\beta}^{\mathrm{T}} \quad \boldsymbol{\delta}^{\mathrm{T}}]^{\mathrm{T}}$). If these parameters are replaced by their estimates, then we obtain the EBP, denoted by $\hat{\theta}_{EBP}$.

The value of the EBP of any function of random variables $\theta(\mathbf{Y})$ might be calculated using the Monte Carlo approximation (Molina & Rao, 2010). We can divide this procedure into four steps. The first one involves the estimation of vector $\boldsymbol{\tau}$ of parameters of the distribution of the examined random variables, using the realisation of vector $\boldsymbol{Y_s}$. In the second step we generate, taking into account the assumption on the known distribution $\boldsymbol{Y_r}|\boldsymbol{Y_s}$, L vectors $\boldsymbol{Y_r}$ ($\boldsymbol{Y_r^{(l)}}$ for $l = 1, 2, \dots, L$), where vector $\boldsymbol{\tau}$ is replaced by its estimate. The next step consists in creating L population vectors expressed $\boldsymbol{Y^{(l)}} = \left[\boldsymbol{Y_s^{\mathrm{T}}} \quad \boldsymbol{Y_r^{(l)\mathrm{T}}}\right]^{\mathrm{T}}$ ($l = 1, 2, \dots, L$). In the last step, we are calculating the EBP according to the formula: $\hat{\theta}_{EBP} = L^{-1}\sum_{l=1}^{L}\theta\left(T^{-1}(\boldsymbol{Y^{(l)}})\right)$, where $T(.)$ is the transformation function of the examined variable used to obtain the dependent variable in the linear mixed model (Molina & Rao, 2010).

If we assume model (19), the proposed best predictor has the following form:

$$\hat{\theta}_{BP}^{\rho} = E(\theta\,|\boldsymbol{Y_s}). \tag{25}$$

Analogically to predictor (24), it is the conditional expected value of the function of random variables θ, assuming that the form of the conditional distribution $\boldsymbol{Y_r}|\boldsymbol{Y_s}$ is known. In this case, and also in practice, the conditional distribution $\boldsymbol{Y_r}|\boldsymbol{Y_s}$ depends on the vector of the unknown parameters $\boldsymbol{\tau}^{\rho}$. This vector, however, must include the correlation coefficient, i.e. parameter ρ. If we replaced $\boldsymbol{\tau}^{\rho}$ by their estimate, we would also obtain the EBP, denoted by $\hat{\theta}_{EBP}^{\rho}$.

5. The dataset

In the application of the proposed predictor we are using real data from the Local Data Bank of Statistics Poland on municipalities in the southern and south-western macroregions ($N = 587$). We are therefore considering four voivodships (NUTS 2): Małopolskie, Śląskie, Opolskie and Dolnośląskie. The dependent variable is the number of newly registered entities in the REGON register in 2017. The population as of 2016 in thousands of people is the auxiliary variable. Additionally, we are dividing the population according to two grouping variables, namely the voivodships and types of municipalities (rural, urban and urban-rural). The total number of domains is therefore $D = 4 \times 2 = 8$.[1] We are using three iterative algorithms in one simulation study (Monte Carlo replications of the superpopulation model, Monte Carlo approximation of the EBP and the REML method of estimation of model parameters). It should be added that the Monte

[1] The main reason why we are considering the population of the size of 587 and the division of this population into eight domains is the time-consuming calculations presented in the next section.

Carlo approximation of the EBPs can be time-consuming even despite the parallel computations used in the simulation study. The complexity of the model due to the correlation of random effects also affects the computation time. For example, the time of the estimation of model parameters with two random effects is longer by approximately 50% for our sample data and by 70% for the studied population data than for the model with one random effect. Estimation of the model parameters with two correlated random effects is also more-time consuming than for the model without the correlation, by about 5%. Furthermore, the estimation of model parameters with twice as many domains (for $D = 16$ domains) is longer than for the model analysed in our study by approximately 20%.

Table 1. Selected descriptive statistics for the dependent variable – newly registered entities in the REGON register in 2017

Statistics	Domain							
	1	2	3	4	5	6	7	8
Minimum	17.00	8.00	43.00	14.00	25.00	13.00	24.00	10.00
1st quantile	64.75	32.25	91.00	41.00	49.50	24.00	76.50	33.75
Median	106.50	53.00	134.50	67.00	84.00	33.50	170.00	53.00
Mean	163.40	70.77	203.50	82.14	146.50	36.94	407.00	61.99
3rd quantile	191.00	79.00	279.80	104.00	131.50	46.25	439.50	80.50
Maximum	1,010.00	393.00	917.00	287.00	1,320.00	75.00	3,500.00	208.00
Coeficient of variation	1.04	0.91	0.59	0.88	1.54	0.48	1.46	0.58

Source: author's work.

In Tables 1 and 2 there are values for some descriptive statistics for dependent and auxiliary variables for all domains. We should note that the coefficient of variation for all domains is less than for the population ($CV = 1.80$). It could be a premise for using the linear mixed model with random effects specific for domains.

Table 2. Some descriptive statistics for auxiliary variable – population in 2016 (in thousands of people)

Statistics	Domain							
	1	2	3	4	5	6	7	8
Minimum	3.74	1.62	5.35	2.49	5.92	3.42	5.73	2.42
1st quantile	7.41	4.60	11.49	6.52	10.36	4.84	11.47	5.87
Median	12.69	5.53	16.87	8.57	13.69	6.72	22.43	7.86
Mean	18.89	7.26	23.43	10.01	21.41	6.77	51.97	9.05
3rd quantile	22.68	8.37	30.61	11.52	23.53	8.15	59.39	11.94
Maximum	114.60	29.51	110.10	28.00	118.70	14.60	298.10	23.58
Coeficient of variation	1.04	0.63	0.80	0.54	1.00	0.37	1.19	0.47

Source: author's work.

The sample is drawn using the Brewer sampling technique which involves the inclusion of the probabilities proportional to the auxiliary variable. This technique is presented in the paper by Brewer (1975). The sample size is $n = 59$ ($\sim 10\%$ of the population size). We are analysing this size of the population and the sample also due to the time-consuming calculations. In the study, we are considering the EBPs for two models:

- the linear mixed model with two correlated vectors of random effects (19);
- the linear mixed model with two uncorrelated vectors of random effects (15).

For the sample drawn from the population, we are calculating the values of predictors of total values in domains.

6. Simulation study

The main aim of the simulation study, presented in this section, is the comparison of the accuracy of the analysed predictors. In the study we are using the same data set and the same sample as introduced in the previous section. We are considering five models:

- the linear regression model with one dependent variable:

$$\log(Y_{id}) = \beta_1 \log(x_{id}) + \beta_0 + e_{id}; \tag{26}$$

- the nested error regression model (6);
- the model with a random slope (11);
- the linear mixed model with two uncorrelated vectors of random effects (15);
- the linear mixed model with two correlated vectors of random effects (19).

We should note that in all the studied models, the log-transformation of both variables is taken into account as in model (26).

We are taking into account models (15) and (19) in the model selection process to study how the inclusion of more than one random effect to the model and the correlation between random effects affects the analysed predictors for the economic dataset. The choice of the model is based on one of the most commonly used criteria for model selection, namely the Akaike Information Criterion (AIC):

$$\mathrm{AIC} = -2\ln(L(M)) + 2|M|, \tag{27}$$

where:

- L is the value of the likelihood function for model M;
- $|M|$ is number of model parameters (see, e.g. Biecek, 2012).

The chosen model is the last of the models we are examining in this paper: the LMM with two correlated random effects specific for domains. Due to the assumption of normality of stochastic disturbances and the distributions of random effects, the log-transformation of both variables was taken into account. To verify the model we use, for example, the permutation version of the conditional t test (Wolfinger, 1993), the permutation version of the test based on the likelihood function (Biecek, 2012) to test the significance of fixed effects, and the permutation version of the test based on the likelihood function (Biecek, 2012) to test the significance of variance components. Simulation analyses of the properties of these tests are presented, e.g., in Krzciuk and Żądło (2014a, 2014b).

The simulation analyses are divided into three phases. The first and the second are model-based simulation studies. We started these procedures by drawing a sample only once. Next, in the first phase, data after log transformation were generated on the basis of the model with correlated random effects, and in the second on the basis of the model with uncorrelated random effects. We should note that we were generating data using the estimated values of model parameters on the basis of the whole population data set and the information on auxiliary variable for all the elements of the population. To estimate model parameters, we were using the restricted maximum likelihood method and function lmer from the lme4 package in R (R development; R Core Team, 2022). Furthermore, in the first part of the simulation study, we assumed such values of ρ as in the original population dataset ($\rho = -0.97$). We used real values of the auxiliary variable, the grouping variable and the sample indicators, and generated values of the dependent variable.

In the last phase, we used design-based approach. In this part of the analysis, in each iteration of the simulation study a sample is drawn using the Brewer sampling technique.

In all of the phases of the simulation study we are assuming $L = 2{,}000$ of Monte Carlo iterations and 100 of EBP iterations. We are considering four predictors:

- EBP_1, the predictor based on the model with correlated vectors of random effects (19), where model parameters are replaced by their estimates;
- EBP_2, the predictor based on the model with uncorrelated vectors of random effects (15), where model parameters are replaced by their estimates;
- BP_1, the predictor (25) based on the model with correlated vectors of random effects, where model parameters are assumed to be known;
- BP_2, the predictor (24) based on the model with uncorrelated vectors of random effects, where model parameters are assumed to be known.

In model-based simulation studies, we are calculating simulation prediction relative biases of EBPs listed above $\mathrm{rB}_{\mathrm{sym}}(.)$:

$$\mathrm{rB}_{\mathrm{sym}}\left(\hat{\theta}_{EBP}\right) = \frac{\frac{1}{L}\sum_{k=1}^{L}\left(\hat{\theta}^{k}{}_{EBP} - \theta^{k}\right)}{\left|\frac{1}{L}\sum_{k=1}^{L}\theta^{k}\right|}, \tag{28}$$

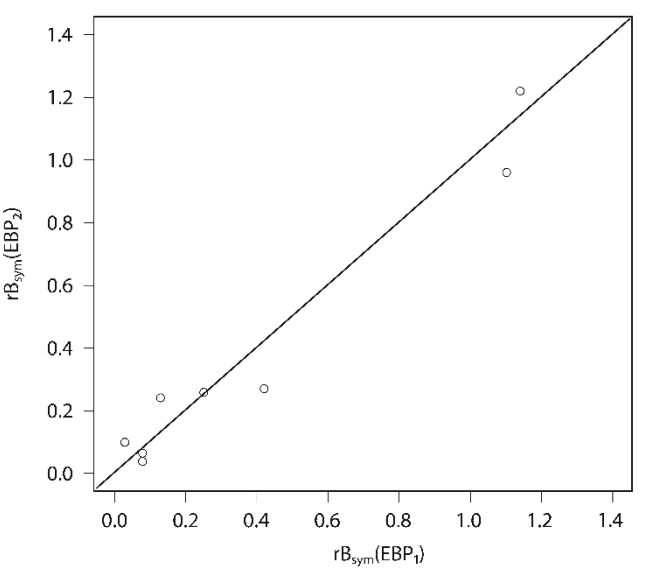
and adequate ratios of the simulation prediction MSEs. The MSE of the EBP here has the form of:

$$\mathrm{MSE}_{\mathrm{sym}}\left(\hat{\theta}_{EBP}\right) = \frac{1}{L}\sum_{k=1}^{L}\left(\hat{\theta}^{k}{}_{EBP} - \theta^{k}\right)^{2}, \tag{29}$$

where $\hat{\theta}^{k}{}_{EBP}$ and θ^{k} are values of the predictor and the value of the predicted characteristic in the k-th iteration, respectively. In the design-based experiment, the value of the predicted characteristic is fixed (non-random), and hence θ^{k} in (28) and (29) is the same for each iteration.

Figure 1 presents values of relative biases of the studied predictors, i.e. the proposed EBP_1 based on the model with two correlated vectors of random effects for domains and EBP_2, based on the model with uncorrelated vectors of random effects for the first phase of the analyses when $\rho \neq 0$.

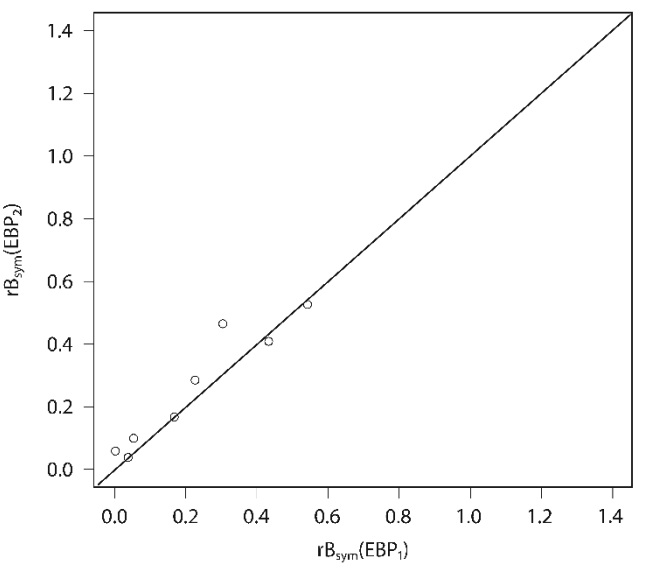
Figure 1. Values of relative simulation biases $\mathrm{rB}_{\mathrm{sym}}(.)$ of EBP_1 and EBP_2 (in %) for eight domains – model-based approach ($\rho \neq 0$)



Source: author's work.

In the first part of the analyses carried out according to model-based approach, we are generating data by means of the model with correlated vectors of random effects. The mean values of the relative simulation biases $rB_{sym}(.)$ over domains for the proposed EBP_1 are lower than for the EBP_2, under the assumption of the model with uncorrelated vectors of random effects. They total 0.129% and 0.161%, respectively. Surprisingly, we can observe a mean loss of accuracy for the proposed EBP compared to EBP_2 over domains $MSE_{sym}(EBP_1)/MSE_{sym}(EBP_2)$ by 1.9%, which probably results from the estimation of model parameters – the mean loss of accuracy over domains resulting from the estimation of model parameters (shown by ratio $MSE_{sym}(EBP_1)/MSE_{sym}(BP_1)$ is 16%. Hence, we can expect better results for larger sample sizes and higher number of domains, when ρ is estimated more accurately.

Figure 2. Values of relative simulation biases $rB_{sym}(.)$ of EBP_1 and EBP_2 (in %) for eight domains – model-based approach ($\rho = 0$)



Source: author's work.

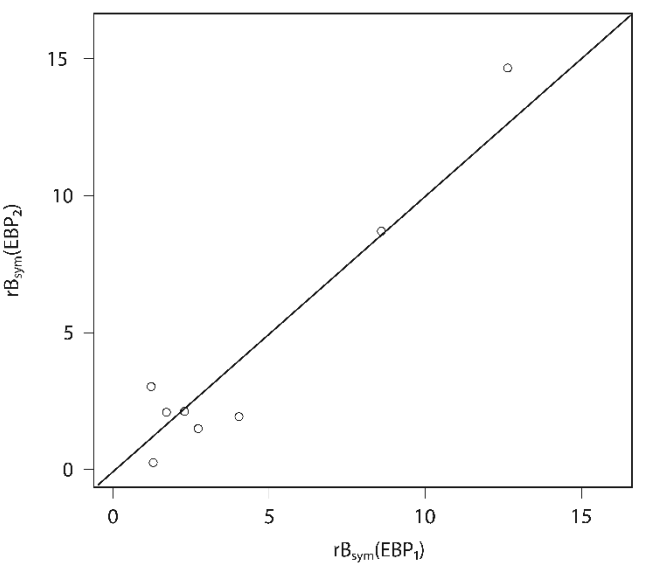
Figure 2 presents values of the relative biases of the studied predictors when data are generated by means of the model with uncorrelated vectors of random effects, so when ρ=0. We should note that even if the lack of correlation between random effects is assumed, the EBP for the model with correlation has a low value of relative biases. This can be seen in the lower scattering of points.

The second model-based part of the simulation study assumes that data are generated by means of the model with uncorrelated vectors of random effects. In this

case, mean rB_{sym} (.) over domains for the proposed EBP according to the model with correlation is quite similar to the EBP_2. They are close to 0.26%. The mean loss of the accuracy resulting from the model misspecification over domains, measured by the ratio $MSE_{sym}(EBP_1)/MSE_{sym}(EBP_2)$, equals 1.4%. The mean loss of accuracy resulting from the estimation of model parameters equals 10.6%. The results obtained under the assumption of a model based-approach suggest that the MSE of the proposed EBP should be lower for larger sample sizes and higher number of realisations of random effects (number of domains in this analysis), when ρ is estimated more accurately.

In Figure 3, we can see values of the relative design-biases of the analysed predictors in domains for the third phase of the simulation study conducted following the design-based approach.

Figure 3. Values of relative simulation biases $rB_{sym}(.)$ of EBP_1 and EBP_2 (in %) for eight domains – designed-based approach



Source: author's work.

In this approach, the proposed EBPs and EBP for models without correlation of random effects have similar properties for the analysed dataset. Means over domains simulation biases for EBP_1 and EBP_2 are close to 4.33%. In this case, the mean ratio $MSE_{sym}(EBP_1)/MSE_{sym}(EBP_2)$ over domains is 1.024.

7. Conclusions

In the paper we proposed the EBP for the LMM with correlated vectors of random effects. In the simulation study, we compared properties of this predictor with the EBP based on the model with uncorrelated vectors of random effects. In most cases, the absolute values of the relative bias for the proposed EBP are lower or relatively similar to the EBP based on the model with uncorrelated vectors of random effects. Additionally, we analysed the problem of the influence of the estimation of model parameters and model misspecification on the MSEs of the studied predictors. In the model-based approach, the mean loss of accuracy resulting from the model misspecification is lower or equal 2%. The mean loss of accuracy resulting from the estimation of model parameters is lower than 20%. The obtained results suggest that the MSE of the proposed EBP should be lower for larger sample sizes and higher number of realisations of random effects (here: the number of domains) when ρ is estimated more accurately. Even if the lack of correlation between random effects is assumed, the EBP for the model with correlation has good properties. In the design-based approach, the EBPs for models with and without correlation of random effects have similar properties.

These results constitute the basis for further research. According to Schreck et al. (2024), plasmode simulations might be a complement or the possible extension of simulation analyses. This alternative data-generation approach can be seen as a bridge between the parametric and nonparametric simulations, as the above authors point out. This approach, as mentioned by Gadbury et al. (2008), is based on resampling the covariate information from the original real dataset. Wyss et al. (2021) suggested that the advantage of plasmode simulations is that they can generate more realistic data, which could also allow an estimate with better properties.

References

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), 28–36. https://doi.org/10.1080/01621459.1988.10478561.

Biecek, P. (2012). *Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*. Wydawnictwo Naukowe PWN.

Brewer, K. E. W. (1975). A simple procedure for sampling Πpswor. *Australian Journal of Statistics*, *17*(3), 166–172. https://doi.org/10.1111/j.1467-842X.1975.tb00954.x.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in Covariance Components Models. *Journal of the American Statistical Association*, *76*(374), 341–353. https://doi.org/10.2307/2287835.

Dumont, C., Chenel, M., & Mentré, F. (2014). Influence of covariance between random effects in design for nonlinear mixed-effect models with an illustration in pediatric pharmacokinetics. *Journal of Biopharmaceutical Statistics*, *24*(3), 471–492. https://doi.org/10.1080/10543406.2014.888443.

Gadbury, G. L., Xiang, Q., Yang, L., Barnes, S., Page, G. P., & Allison, D. B. (2008). Evaluating Statistical Methods Using Plasmode Data Sets in the Age of Massive Public Databases: An Illustration Using False Discovery Rates. *PLoS Genetics*, *4*(6), 1–8. https://doi.org/10.1371/journal.pgen.1000098.

Krzciuk, M. K. (2020). On empirical best linear unbiased predictor under a Linear Mixed Model with correlated random effects. *Econometrics. Ekonometria. Advances in Applied Data Analysis*, *24*(2), 17–29. https://doi.org/10.15611/eada.2020.2.02.

Krzciuk, M. K., & Żądło, T. (2014a). On some tests of variance components for linear mixed models. *Studia Ekonomiczne. Zeszyty Naukowe Uniwesytetu Ekonomicznego w Katowicach*, 189, 77–85. https://www.ue.katowice.pl/fileadmin/_migrated/content_uploads/8_M.K.Krzciuk_T.Zadlo_On_some_tests_of_variance..._01.pdf.

Krzciuk, M. K., & Żądło, T. (2014b). On some tests of fixed effects for linear mixed models. *Studia Ekonomiczne. Zeszyty Naukowe Uniwesytetu Ekonomicznego w Katowicach*, 189, 49–57. https://www.ue.katowice.pl/fileadmin/_migrated/content_uploads/5_M.K.Krzciuk_T.Zadlo_On_some_tests_of_fixed..._01.pdf.

Marhuenda, Y., Molina, I., Morales, D., & Rao, J. N. K. (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *180*(4), 1111–1136. https://doi.org/10.1111/rssa.12306.

Menec, V., Lix, L., Steinbach, C., Ekuma, O., Sirski, M., Dahl, M., & Soodeen, R. A. (2004). *Patterns of Health Care Use and Cost at the End of Life*. Manitoba Centre for Health Policy. http://mchp-appserv.cpe.umanitoba.ca/reference/end_of_life.pdf.

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, *38*(3), 369–385. https://doi.org/10.1002/cjs.10051.

Molina, I., & Rao, J. N. K (2015). *Small Area Estimation*. John Wiley and Sons. https://doi.org/10.1002/9781118735855.

Moura, F. A. S., & Holt, D. (1999). Small area estimation using multilevel models. *Survey Methodology*, *25*(1), 73–80. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1999001/article/4714-eng.pdf?st=1HNkUIbE.

Nelsen, R. B. (1999). *An Introduction to Copulas*. Springer. http://dx.doi.org/10.1007/978-1-4757-3076-0.

Ogungbenro, K., Graham, G., Gueorguieva, I., & Aarons, L. (2008). Incorporating correlation in interindividual variability for the optimal design of multiresponse pharmacokinetic experiments. *Journal of Biopharmaceutical Statistics*, *18*(2), 342–358. https://doi.org/10.1080/10543400701697208.

Pratesi, M., & Salvati, N. (2008). Small Area Estimation: The EBLUP Estimator Based on Spatially Correlated Random Area Effects. *Statistical Methods & Applications*, *17*(1), 113–141. https://doi.org/10.1007/s10260-007-0061-9.

Pratesi, M., & Salvati, N. (2009). Small Area Estimation in the Presence of Correlated Random Area Effects. *Journal of Official Statistics*, *25*(1), 37–53.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. https://www.r-project.org/.

Schreck, N., Slynko, A., Saadati, M., & Benner, A. (2024). Statistical plasmode simulations-Potentials, challenges and recommendations. *Statistics in Medicine*, *43*(9), 1804–1825. https://doi.org/10.1002/sim.10012.

Skoglund, J., & Karlsson, S. (2001). Specification and estimation of random effects models with serial correlation of general form (SSE/EFI Working paper series in Economics and Finance, No. 433). https://swopec.hhs.se/hastef/papers/hastef0433.pdf.

Tiao, G. C., & Ali, M. M. (1971). Analysis of correlated random effects: linear model with two random components. *Biometrika*, *58*(1), 37–51. https://doi.org/10.2307/2334315.

Torabi, M., & Rao, J. N. K. (2008). Small area estimation under a two-level model. *Survey Methodology*, *34*(1), 11–17. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008001/article/10612-eng.pdf?st=ga0W1sNZ.

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *181*(4), 927–979. https://doi.org/10.1111/rssa.12364.

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics – Simulation and Computation*, *22*(4), 1079–1106. https://doi.org/10.1080/03610919308813143.

Wyss, R., Schneeweiss, S., van der Laan, M., Lendle, S. D., Ju, C., & Franklin, J. M. (2021). Using Super Learner Prediction Modeling to Improve High-dimensional Propensity Score Estimation. *Epidemiology*, *29*(1), 96–106. https://doi.org/10.1097/EDE.0000000000000762.