University of Economics in Katowice

your place

your space

your future

On estimating prediction accuracy when the model is misspecified

Małgorzata K. Krzciuk

The 4th Congress of Polish Statistics
July 2-4, 2024, Warsaw, Poland

The main aims of studies

Comparison in a simulation study of the properties of selected RMSE estimators of plug-in predictors, assuming the linear mixed model with correlated random effects, taking into account the model misspecification problem.

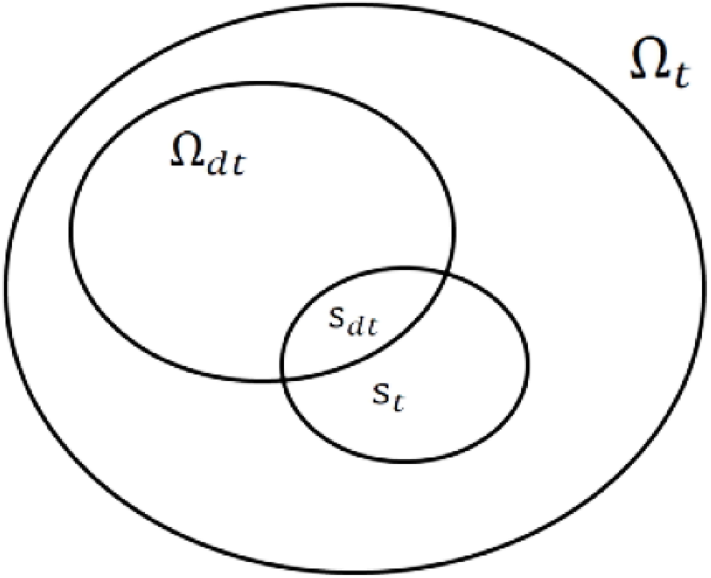
We analyse two types of model misspecification:

- lack of correlation,
- non-normality of the distribution.

The robustness of the methods, on which the analysed estimators are based, to non-normality of the distribution is discussed by Carpenter et al. (2003), Jelsema and Pedadda (2016) and Thai et al. (2013), among others.

Introduction - Small Area Estimation

Small area - domain for which we cannot obtain direct estimates with adequate precision (Rao and Molina 2015, p. 2).



Source: own elaboration

Linear mixed model

The general linear mixed model (cf. Jiang 2007, pp. 1-2):

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

where:

- $\mathbf{Y}$ – the random vector of values of the dependent variable;
- $\mathbf{X}$, $\mathbf{Z}$ – known matrices of auxiliary variables;
- $\boldsymbol{\beta}$ – the vector of unknown parameters;
- $\mathbf{v}$ and $\mathbf{e}$ – random effects and stochastic disturbance, independently distributed with variance-covariance matrices denoted by $\mathbf{G}(\boldsymbol{\delta})$ and $\mathbf{R}(\boldsymbol{\delta})$, where $\boldsymbol{\delta}$ is a vector of random components.

Linear mixed model

Variance-covariance matrix of $\mathbf{Y}$ (Littell et al. 2006, p. 736):

$$\mathbf{V}(\boldsymbol{\delta}) = \mathbf{Z}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^{\mathrm{T}} + \mathbf{R}(\boldsymbol{\delta}). \quad (2)$$

LMM with correlation of random effects specific for domains

This special case of LMM is given by (Krzciuk 2020, p. 20):

$$Y_{idt} = \left(\beta_1 + v_{2d}^{(\rho)}\right) x_{idt} + \beta_0 + v_{1d}^{(\rho)} + e_{idt}, \quad (3)$$

where:

- $v_{1d}^{(\rho)}$, $v_{2d}^{(\rho)}$ – random effects, $v_{1d}^{(\rho)} \sim iid\left(0, \sigma^2_{v_{1d}^{(\rho)}}\right)$, $v_{2d}^{(\rho)} \sim iid\left(0, \sigma^2_{v_{2d}^{(\rho)}}\right)$ and $cor\left(v_{1d}^{(\rho)}, v_{2d}^{(\rho)}\right) = \rho$, for $d = 1, 2, \ldots, D$;
- e_{idt} – stochastic disturbance with distribution $e_{idt} \sim iid\,(0, \sigma_e^2)$.

LMM with correlation of random effects specific for domains

Variance-covariance matrix of $\mathbf{Y}$ (Krzciuk 2023a, p. 38):

$$\mathbf{V}^{(\rho)}(\boldsymbol{\delta}) = \underset{1 \leq d \leq D}{diag} \mathbf{V}_d = \underset{1 \leq d \leq D}{diag} \Big(\sigma^2_{v_{1d}^{(\rho)}} \mathbf{1}_{N_d M} \mathbf{1}^T_{N_d M} + \sigma^2_{v_{2d}^{(\rho)}} \mathbf{x}_d \mathbf{x}_d^T$$

$$+ \rho \sigma_{v_{1d}^{(\rho)}} \sigma_{v_{2d}^{(\rho)}} \left(\mathbf{1}_{N_d M} \mathbf{x}_d^T + \mathbf{x}_d \mathbf{1}^T_{N_d M} \right) + \sigma_e^2 \mathbf{I}_{N_d M \times N_d M} \Big). \qquad (4)$$

LMM with correlation of random effects specific for domains

The matrix $\mathbf{G}$ is given by (Krzciuk 2023a, p. 37):

$$\mathbf{G}^{(\rho)} = \begin{bmatrix} \mathbf{G}_1^{(\rho)} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & \dots & \dots & \vdots \\ \vdots & \dots & \mathbf{G}_d^{(\rho)} & \dots & \vdots \\ \vdots & \dots & \dots & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \dots & \mathbf{G}_D^{(\rho)} \end{bmatrix}_{2D \times 2D}, \quad (5)$$

where submatrix for domain we can write as:

$$\mathbf{G}_d^{(\rho)} = \begin{bmatrix} \sigma^2_{v_{1d}^{(\rho)}} & \rho\sigma_{v_{1d}^{(\rho)}}\sigma_{v_{2d}^{(\rho)}} \\ \rho\sigma_{v_{1d}^{(\rho)}}\sigma_{v_{2d}^{(\rho)}} & \sigma^2_{v_{2d}^{(\rho)}} \end{bmatrix}. \quad (6)$$

Plug-in predictors

The plug-in predictor for:

$$\theta = \theta\left(K^{-1}(\mathbf{Y})\right) = \theta\left(K^{-1}\left(\begin{bmatrix}\mathbf{Y}_s^{\mathrm{T}} & \mathbf{Y}_{\mathrm{r}}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\right)\right) \tag{7}$$

can therefore be written as (cf. Chwila and Żądło, 2019, p. 20):

$$\hat{\theta}_{PLUG-IN} = \theta\left(K^{-1}\left(\begin{bmatrix}\mathbf{Y}_s^{\mathrm{T}} & \hat{\mathbf{Y}}_{\mathrm{r}}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\right)\right), \tag{8}$$

where $\hat{\mathbf{Y}}_{\mathrm{r}}^{\mathrm{T}}$ is a vector of fitted values obtained based on the model assumed for unobserved random variables, where the dependent variable is the back-transformed variable of interest.

Plug-in predictors and LMM with correlated vectors of random effects

The plug-in predictor, assuming model (3) can be denoted as (Krzciuk 2023a, p. 108):

$$\hat{\theta}^{\rho}_{PLUG-IN} = \theta\left(K^{-1}\left(\begin{bmatrix}\mathbf{Y}_{\mathrm{s}}^{\mathrm{T}} & \hat{\mathbf{Y}}_{\mathrm{r}(\rho)}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\right)\right), \qquad (9)$$

where $\hat{\mathbf{Y}}_{\mathrm{r}(\rho)}^{\mathrm{T}}$ is the vector of fitted values obtained based on the model (3), which was assumed for the unobserved variables.

Mean squared errors of plug-in predictors

The analyses addressed the problem of estimation of root of mean square errors $\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2$ i.e.:

$$\mathrm{RM\hat{S}E}(\hat{\theta}) = \sqrt{\mathrm{M\hat{S}E}(\hat{\theta})} \qquad (10)$$

Considered RMSE estimators of plug-in predictors:

- $\mathrm{RM\hat{S}E}_{\mathrm{PB}}$ – using the parametric bootstrap method;
- $\mathrm{RM\hat{S}E}_{\mathrm{R}}$ – using the residual bootstrap;
- $\mathrm{RM\hat{S}E}_{\mathrm{RC}}$ – using the residual bootstrap method with correction.

$\mathrm{RM\hat{S}E_{PB}}$ estimator

The estimator is calculated according to the algorithm (Rao and Molina 2015, pp. 183–186):

1. Estimation of model parameters, i.e. $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$ based on sample.
2. Generate B realizations $\mathbf{y}^{*(b)} = \left[\mathbf{y}_s^{*(b)} \quad \mathbf{y}_r^{*(b)}\right]$, where $b = 1,2,\dots,B$, according to the assumed model, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$.
3. B-times:
   - calculation $\theta^{*(b)} = \theta^{*(b)}\left(\mathbf{y}^{*(b)}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}\right)$,
   - estimation $\hat{\boldsymbol{\beta}}^{(b)}$ and $\hat{\boldsymbol{\delta}}^{(b)}$ based on $\mathbf{y}_s^{*(b)}$,
   - calculation $\hat{\theta}^{*(b)} = \hat{\theta}^{*(b)}\left(\mathbf{y}^{*(b)}, \hat{\boldsymbol{\beta}}^{(b)}, \hat{\boldsymbol{\delta}}^{(b)}\right)$;
4. Calculation :

$$\mathrm{RM\hat{S}E_{PB}}\left(\hat{\theta}\right) = \sqrt{B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}^{*(b)} - \theta^{*(b)}\right)^2} \quad (11)$$

Estimators $\mathrm{RM\hat{S}E_R}$ i $\mathrm{RM\hat{S}E_{RC}}$

The estimator is calculated according to the algorithm for $\mathrm{RM\hat{S}E_{PB}}$ however (cf. Żądło 2023, p. 11):

2. Generate B realizations:

$$\mathbf{y}^{*(b)} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}_1\mathbf{v}_1^{*(b)} + \ldots + \mathbf{Z}_l\mathbf{v}_l^{*(b)} + \ldots + \mathbf{Z}_L\mathbf{v}_L^{*(b)} + \mathbf{e}^{*(b)},$$

where $\mathbf{e}^{*(b)}$ is N-element vector defined as $srswr(col_{1\le i\le n}\hat{e}_i, N)$ and $\mathbf{v}_l^{*(b)}$ (where $l = 1,2,..,L$) is vector with dimensions $K_l J_l \times 1$ formed from the columns of the matrix: $srswr([\hat{\mathbf{v}}_{l1} \quad \ldots \quad \hat{\mathbf{v}}_{lk} \quad \ldots \quad \hat{\mathbf{v}}_{lK_l}], J_l)$ with dimensions $J_l \times K_l$.

In the analyses, we also include the correction more extensively discussed by the Carpenter et al. (2003).

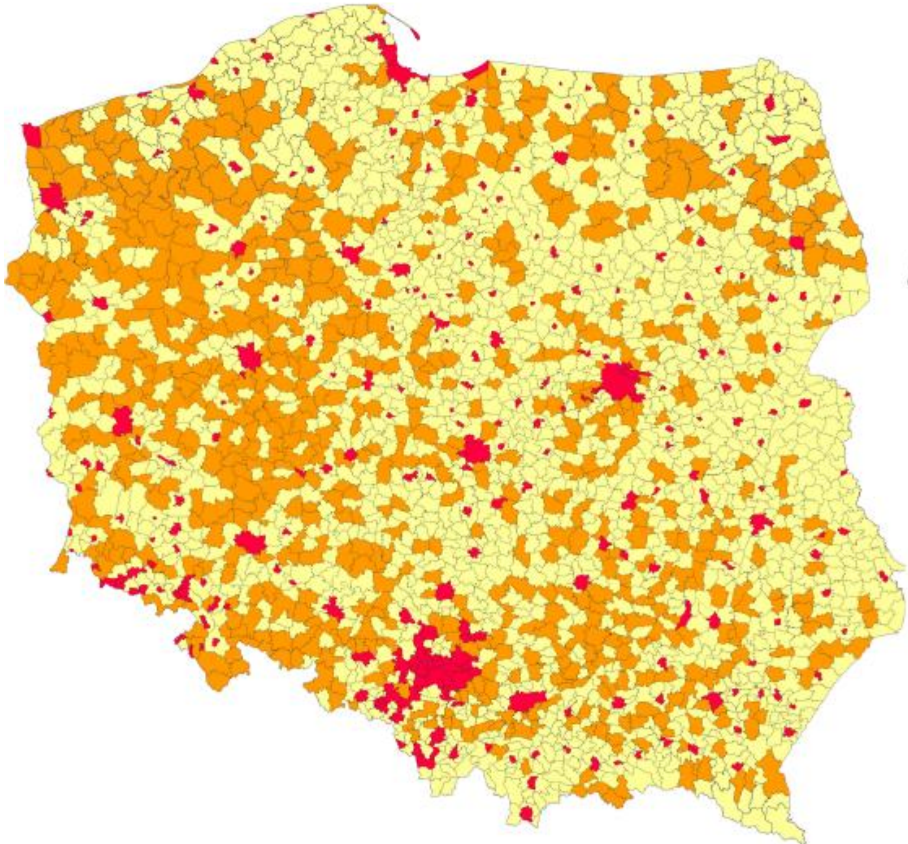
Simulation studies – dataset

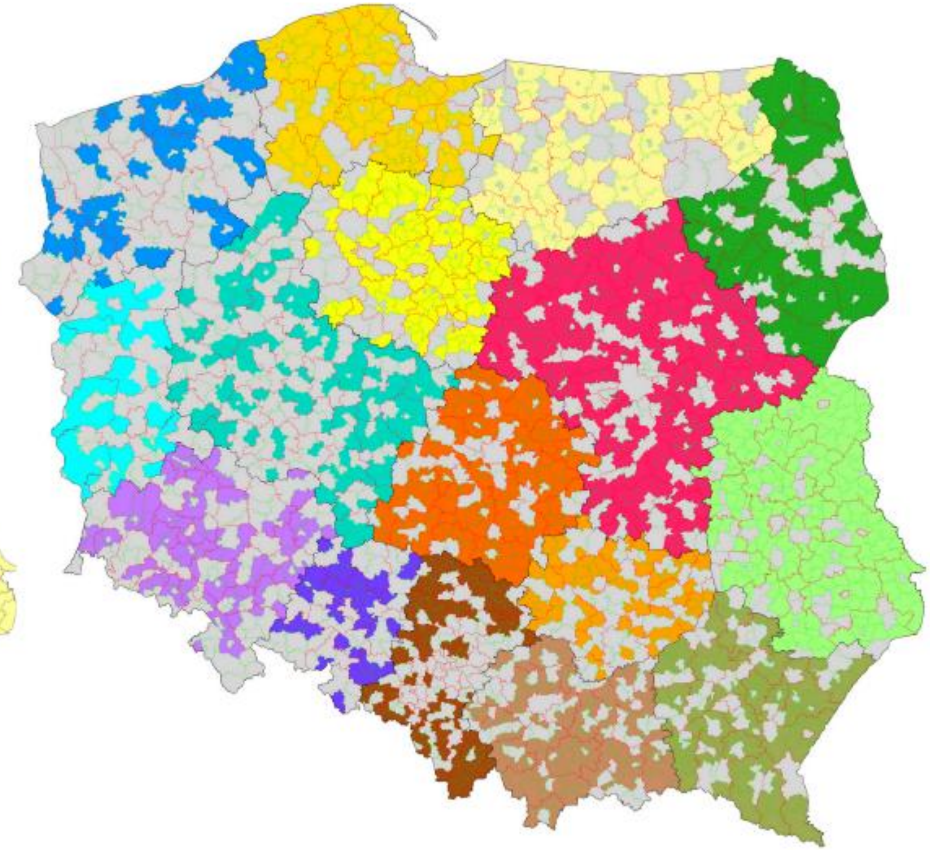
- **The study variable:** revenue of municipalities in million PLN in 2018–2020;
- **The auxiliary variable**: the total population in municipalities in thousands of people in 2017–2019;
- **The data comes from:** the Local Data Bank of Statistics Poland;
- **The size of the population**: N=7398 for 3 periods;
- **The size of the sample**: n=1503 (501 in one period);

Simulation studies – sample

- Sample in first period: **stratified sample** – strata are defined on the basis of the affiliation of municipalities to voivodeships;
- **Subpopulations**: **16 voivodeships** and **2 two types of municipalities** – rural and other (**16×2=32**);
- **Balanced panel**;
- Considered only rural municipalities **domains** (D=16);
- **Random** size of the sample in domains.

Simulation studies – division of municipalities into domains





Source: Krzciuk 2023a, p. 114-115

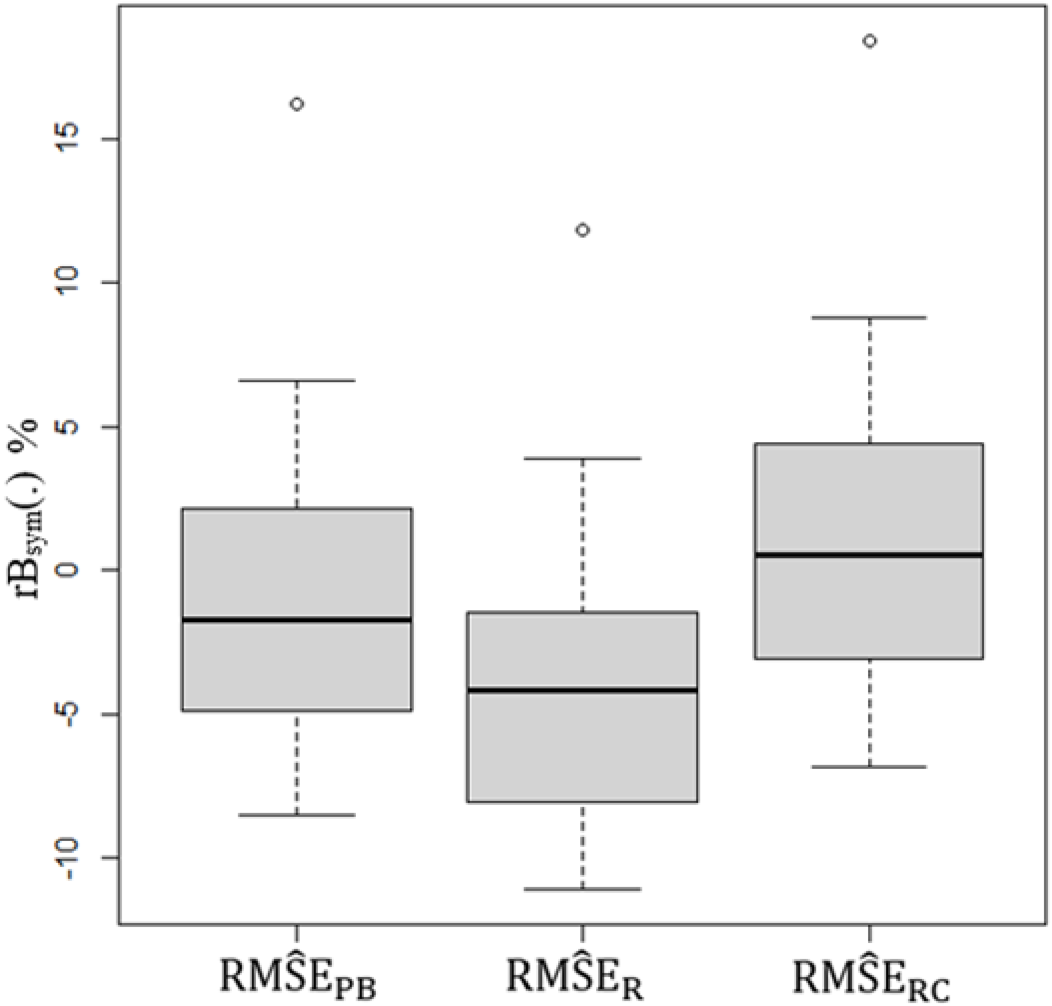
Simulation studies – assumptions

- **Model**: LMM with two correlated domain-specific random effects;
- **Characteristics**: total values in domain;
- **Predictor**: $\hat{\theta}^{\rho}_{PLUG-IN}$;
- **Estitators of RMSE**: $\mathrm{RM\hat{S}E_{PB}}$, $\mathrm{RM\hat{S}E_{R}}$, $\mathrm{RM\hat{S}E_{RC}}$;
- **Number of Monte Carlo iterations:** 1000;
- **Number of bootstrap iterations:** 200.

Simulation studies – assumptions

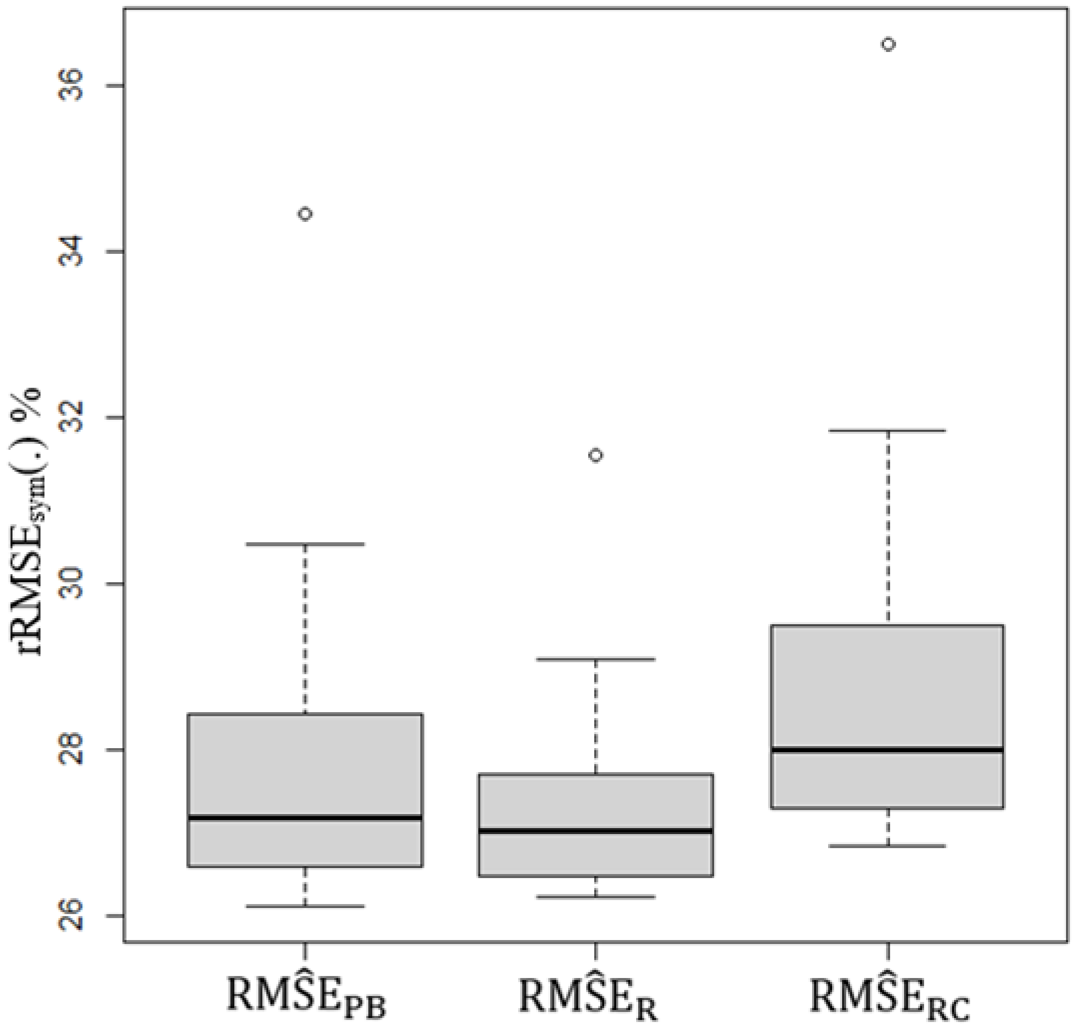
- multivariate normal distribution with expected values equal 0 and $\rho=-0.83$;
- multivariate normal distribution with expected values equal 0 and $\rho=0$;
- t copula $\rho=-0.83$, df = 3 with marginal distribution **shifted exponential** or **shifted gamma distribution**;
- normal copula $\rho=-0.83$ with marginal distribution: **shifted exponential** or **shifted gamma distribution**.

Simulation studies – $\mathrm{rB_{sym}}\left(\mathrm{RM\hat{S}E}\right)$ in % correct model specification



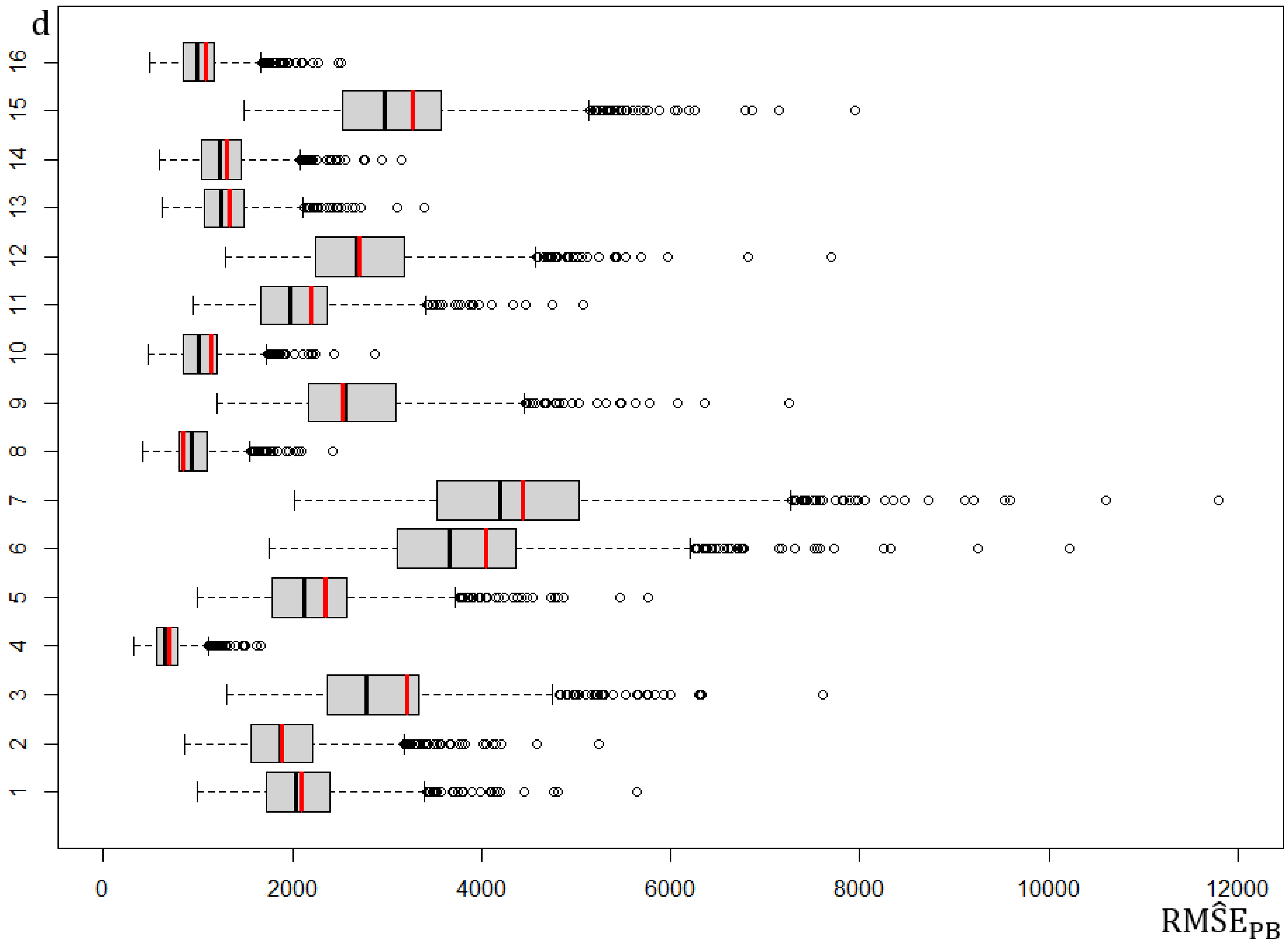
Source: Krzciuk (2023b)

Simulation studies – $\mathrm{rRMSE}_{\mathrm{sym}}(\mathrm{RM\hat{S}E})$ in % correct model specification



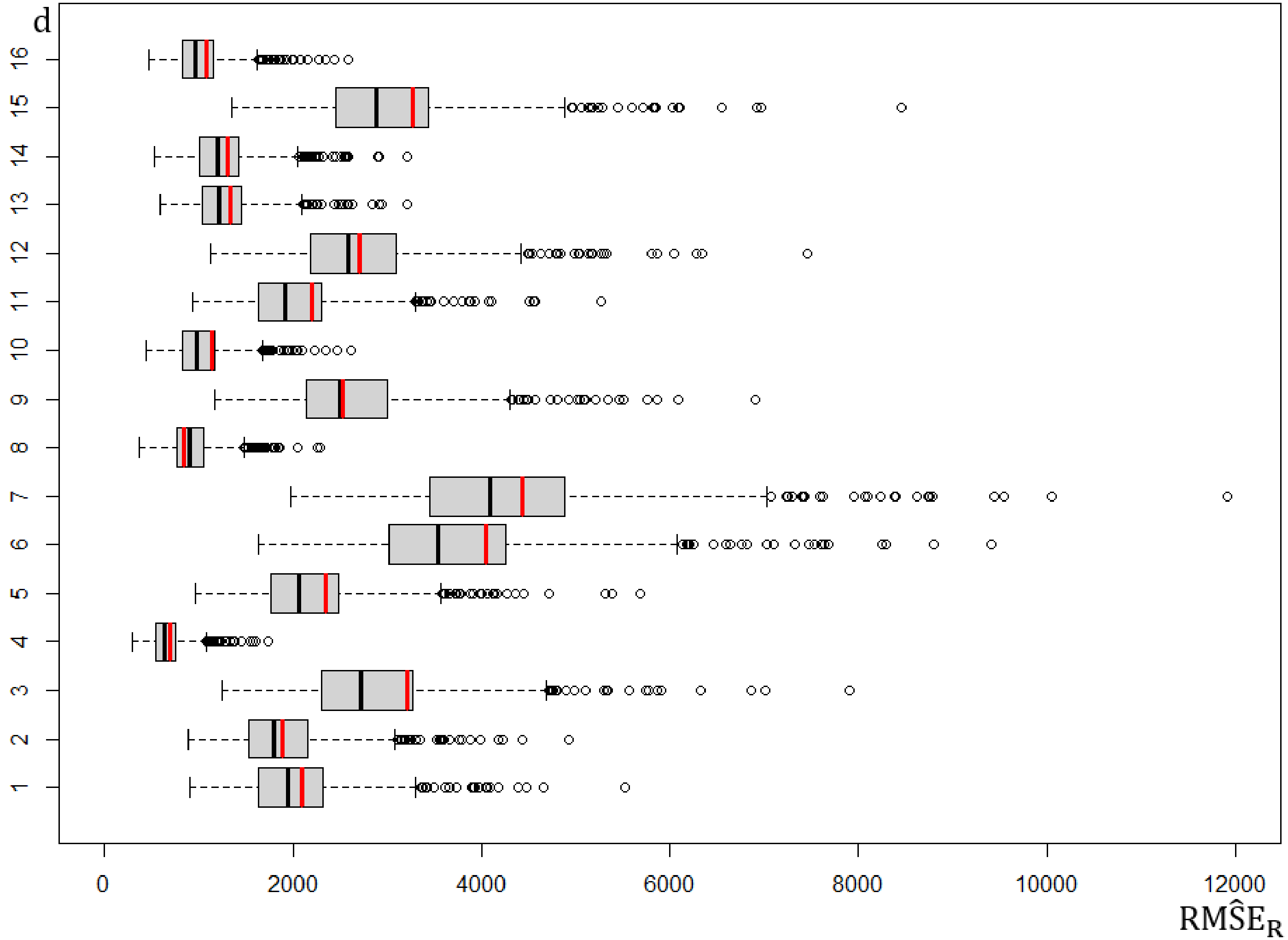
Source: Krzciuk (2023b)

Simulation studies – $\widehat{\mathrm{RMSE}}_{\mathrm{PB}}\left(\hat{\theta}^{\rho,\,\mathrm{wg}}_{PLUG-IN}\right)$ correct model specification



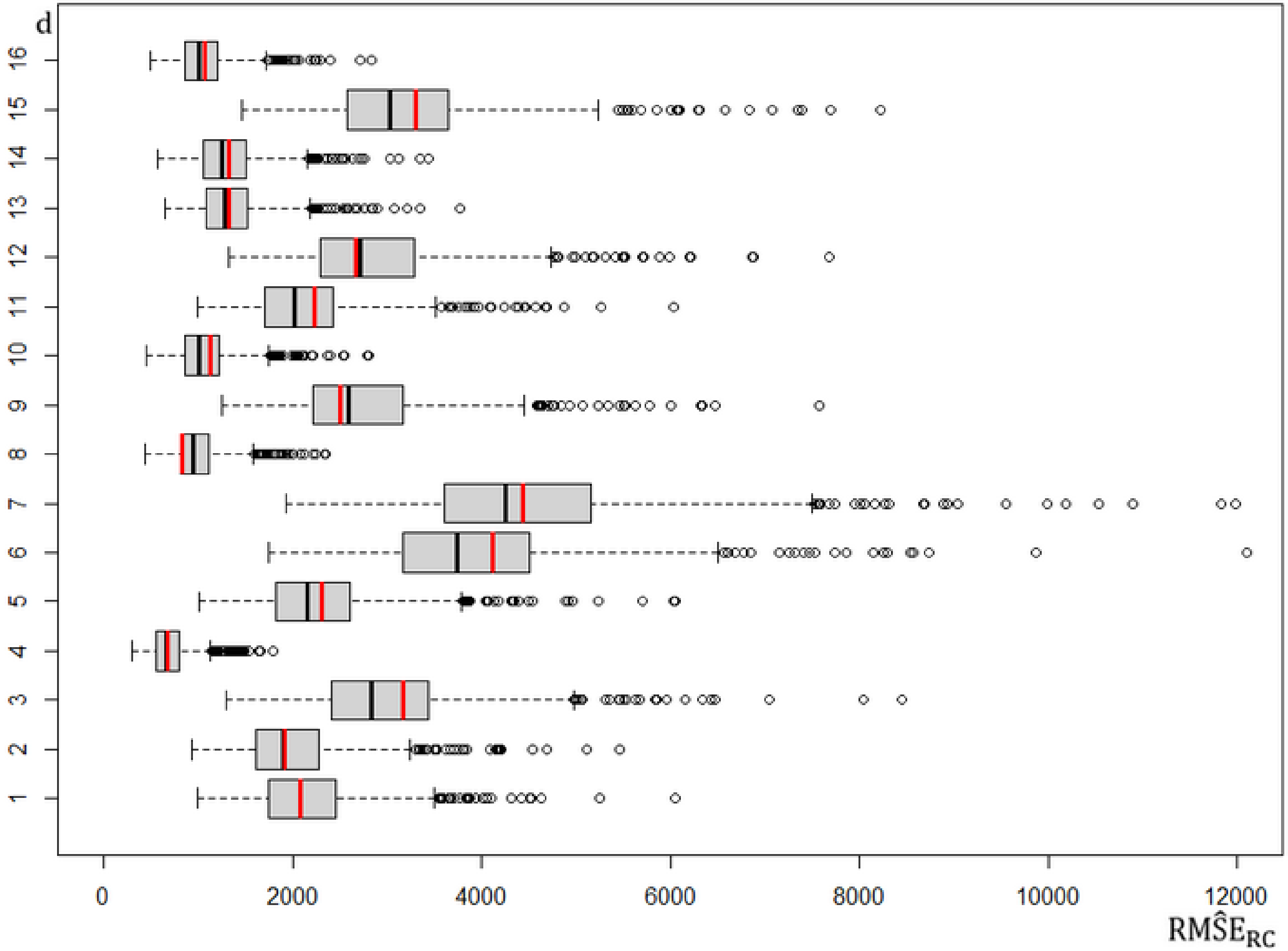
Source: Krzciuk (2023b)

Simulation studies – $\widehat{\mathrm{RMSE}}_{\mathrm{R}}\left(\hat{\theta}^{\rho,\,\mathrm{wg}}_{PLUG-IN}\right)$ correct model specification



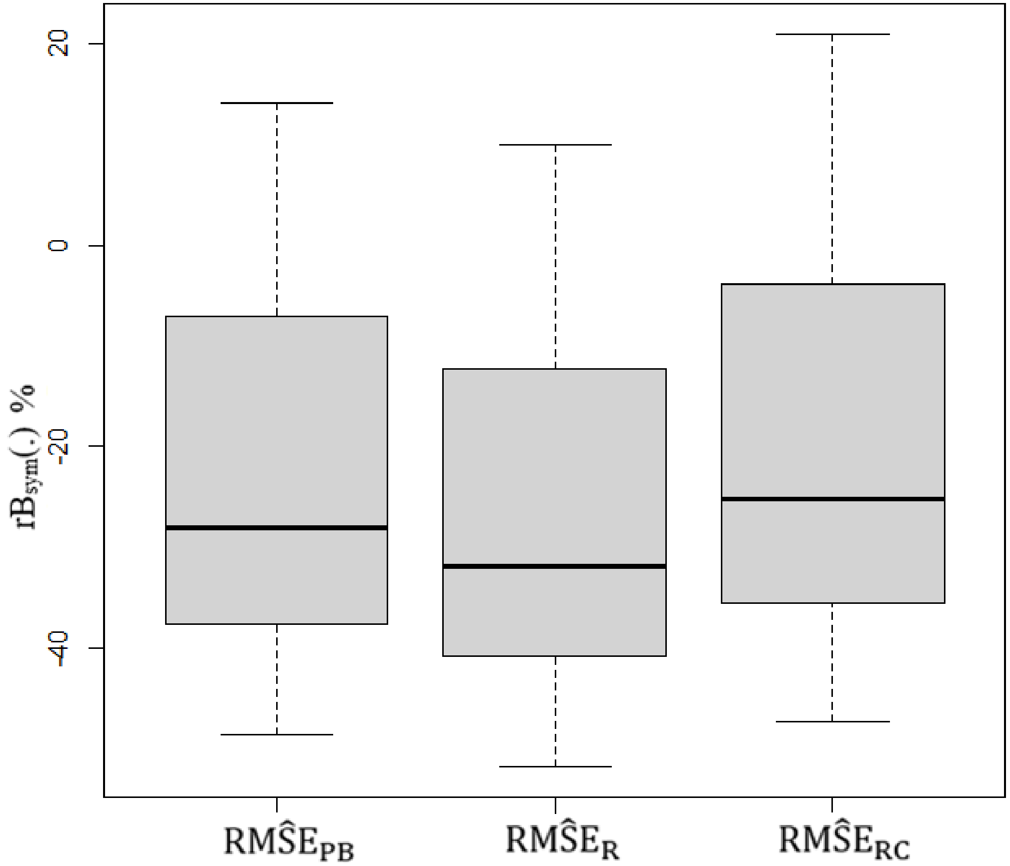
Source: Krzciuk (2023)

Simulation studies – $\widehat{\mathrm{RMSE}}_{\mathrm{RC}}\left(\hat{\theta}^{\rho,\,\mathrm{wg}}_{PLUG-IN}\right)$ correct model specification



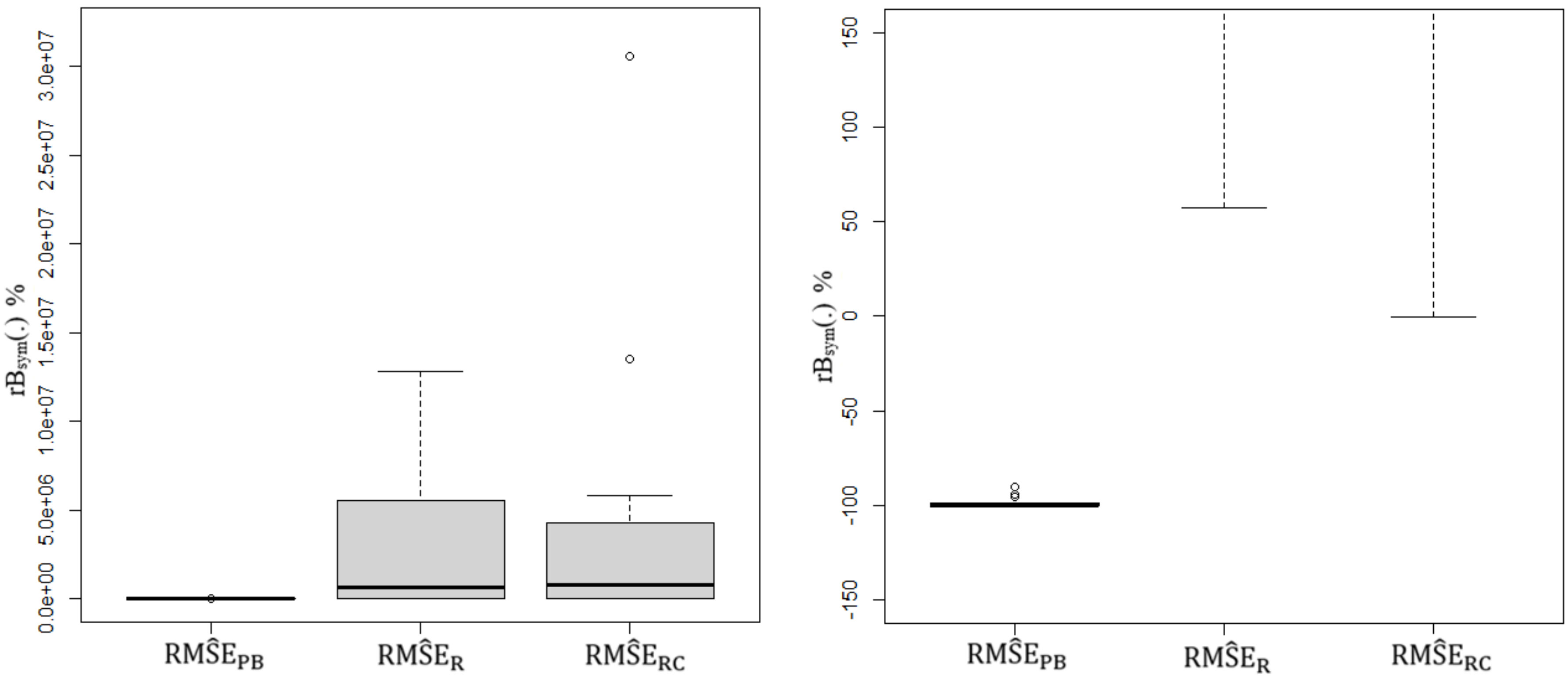
Source: Krzciuk (2023b)

Simulation studies – rB_{sym} $(\widehat{RMSE})$ in % model misspecification
(the lack of correlation)



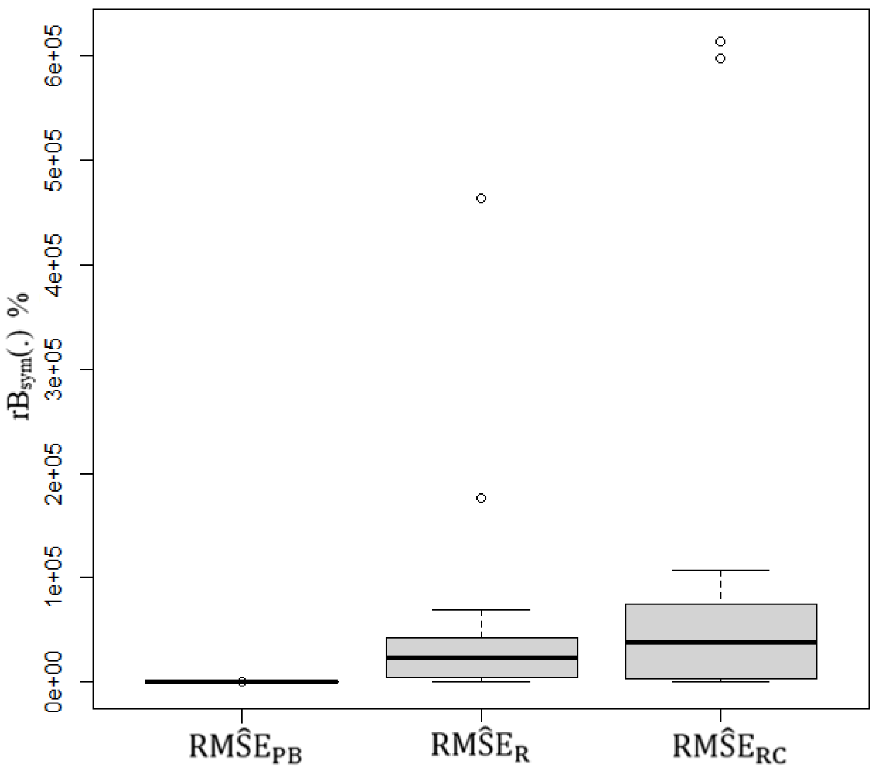
Source: own elaboration

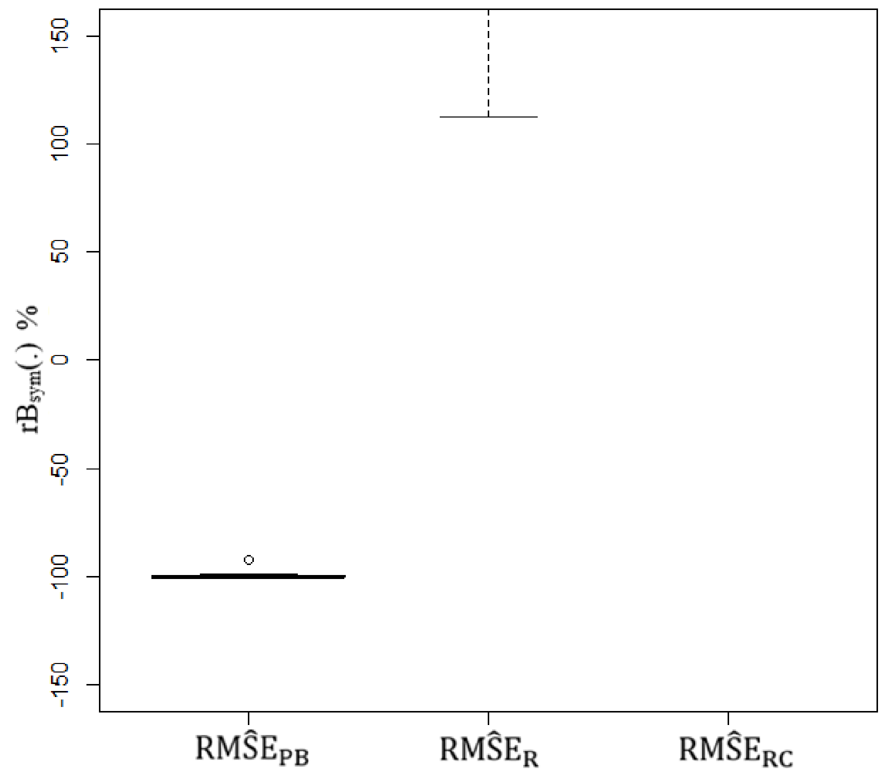
Simulation studies – rB_{sym} ($\widehat{RMSE}$) in % model misspecification (t copula, shifted gamma distribution)



Source: own elaboration

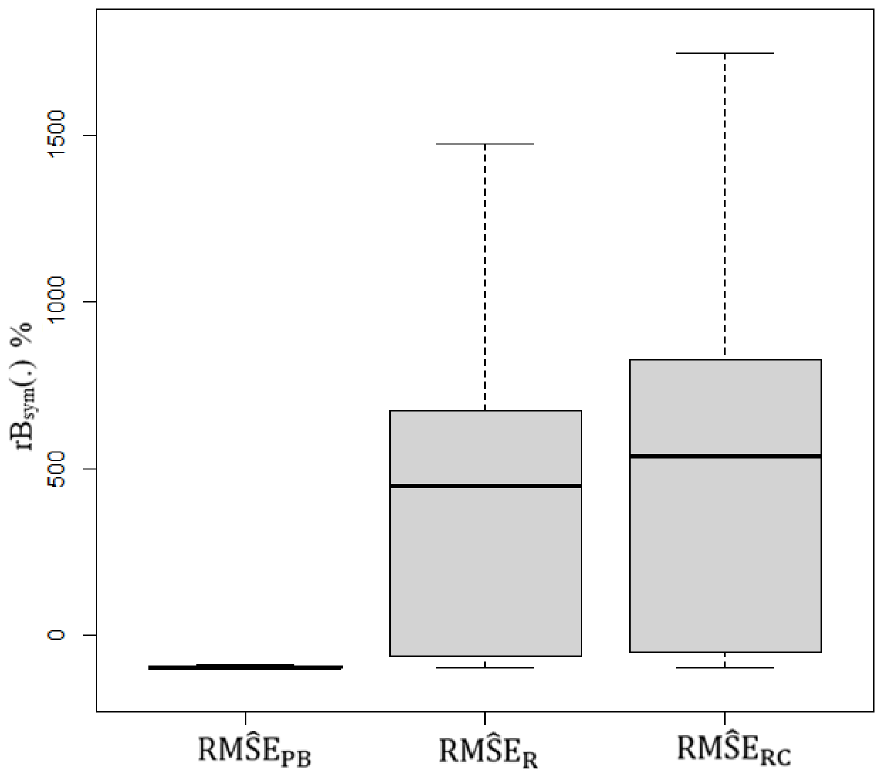
Simulation studies – rB_{sym} $(\widehat{RMSE})$ in % model misspecification (normal copula, shifted gamma distribution)

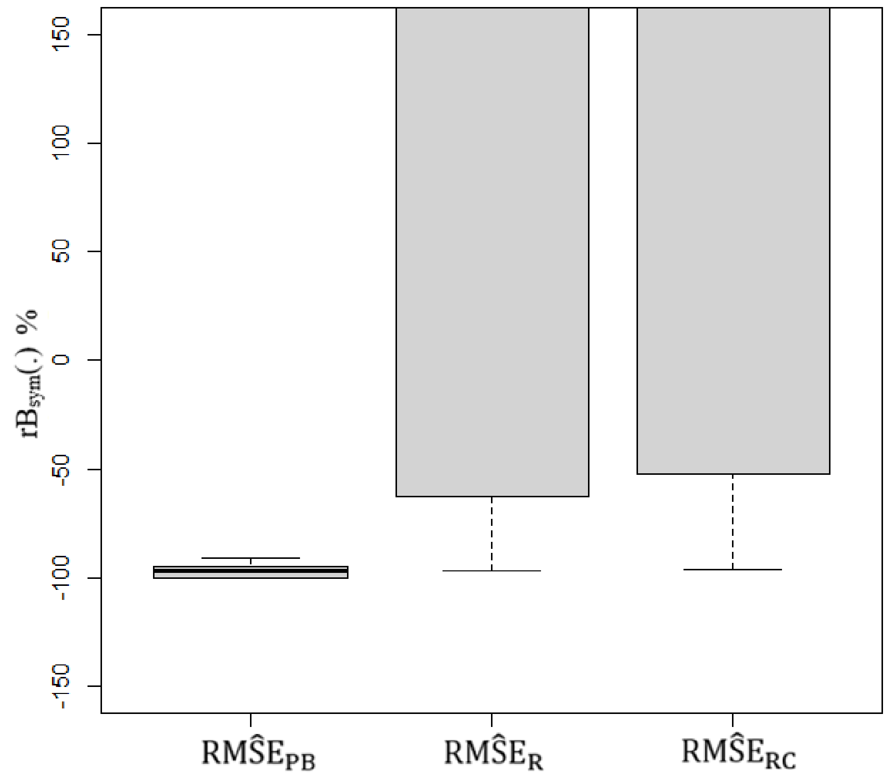




Source: own elaboration

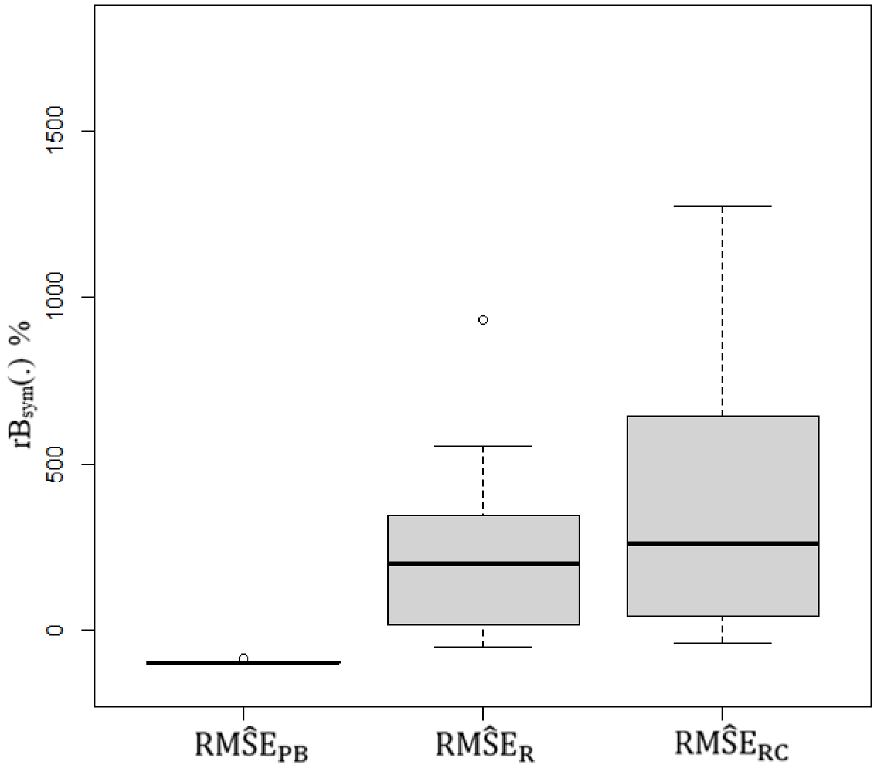
Simulation studies – $rB_{sym}(\widehat{RMSE})$ in % model misspecification (t copula, shifted exponential distribution)

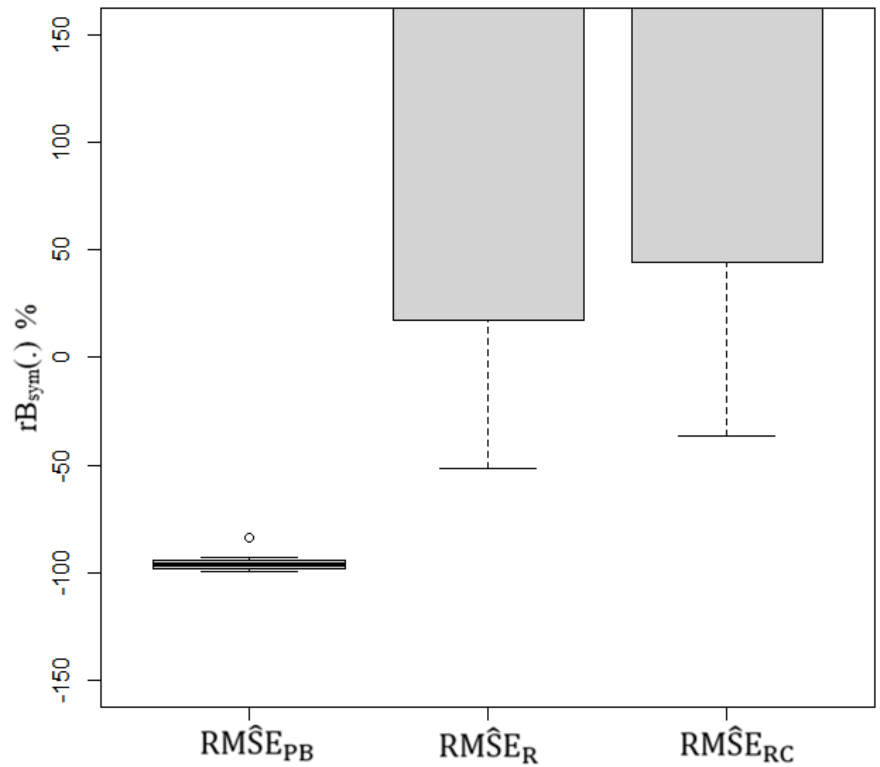




Source: own elaboration

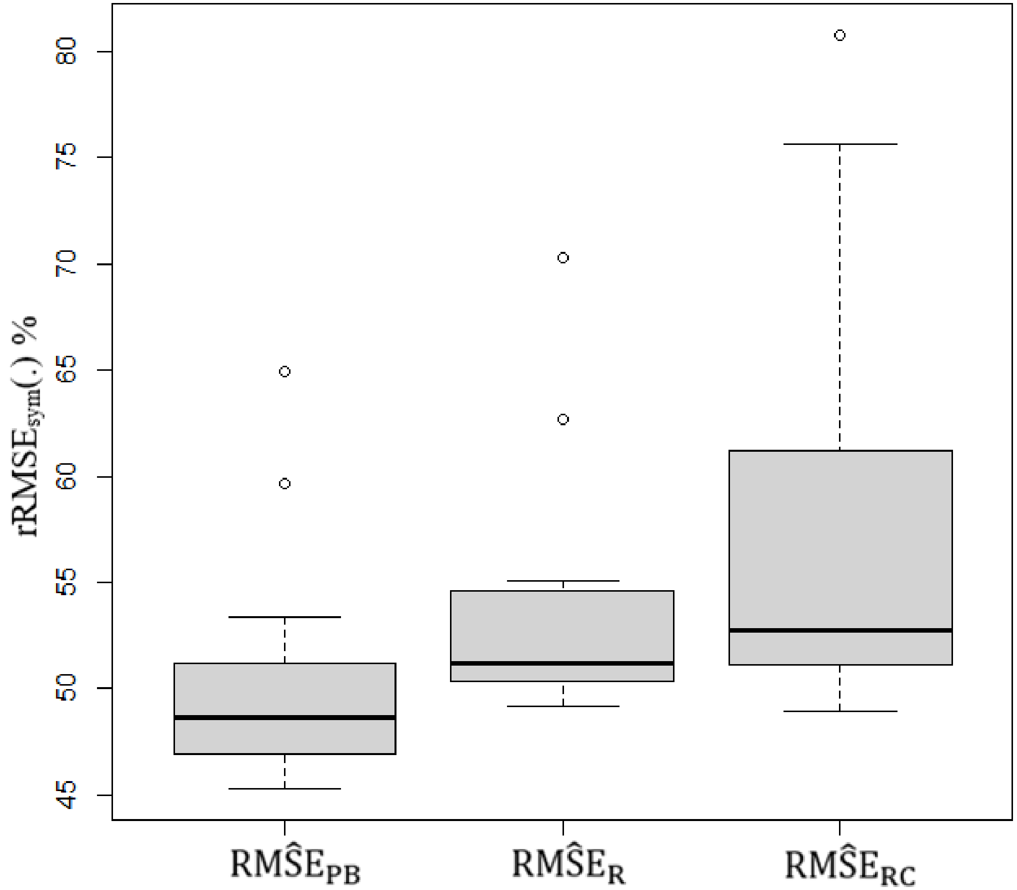
Simulation studies – $rB_{sym}(\widehat{RMSE})$ in % model misspecification (normal copula, shifted exponential distribution)





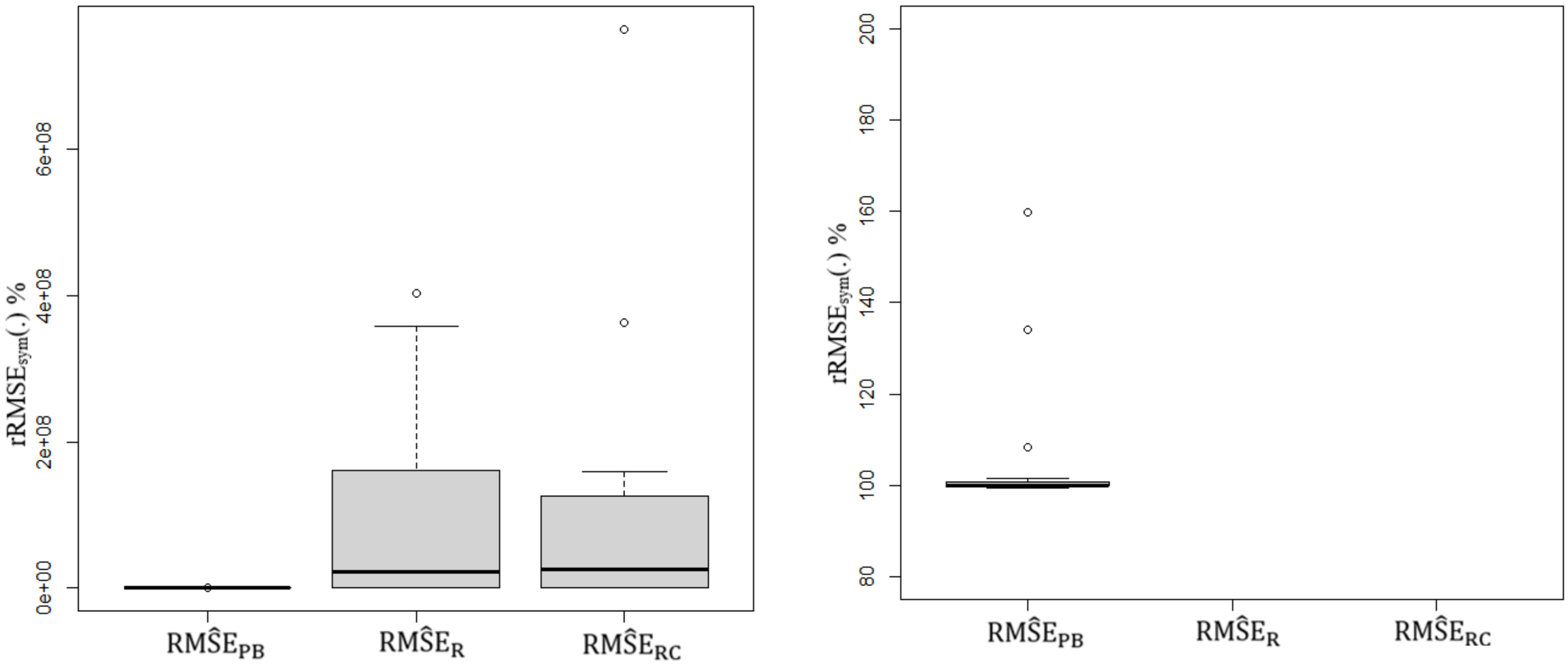
Source: own elaboration

Simulation studies – $\mathrm{rRMSE}_{\mathrm{sym}}$ $(\mathrm{RM\hat{S}E})$ in % model misspecification (the lack of correlation)



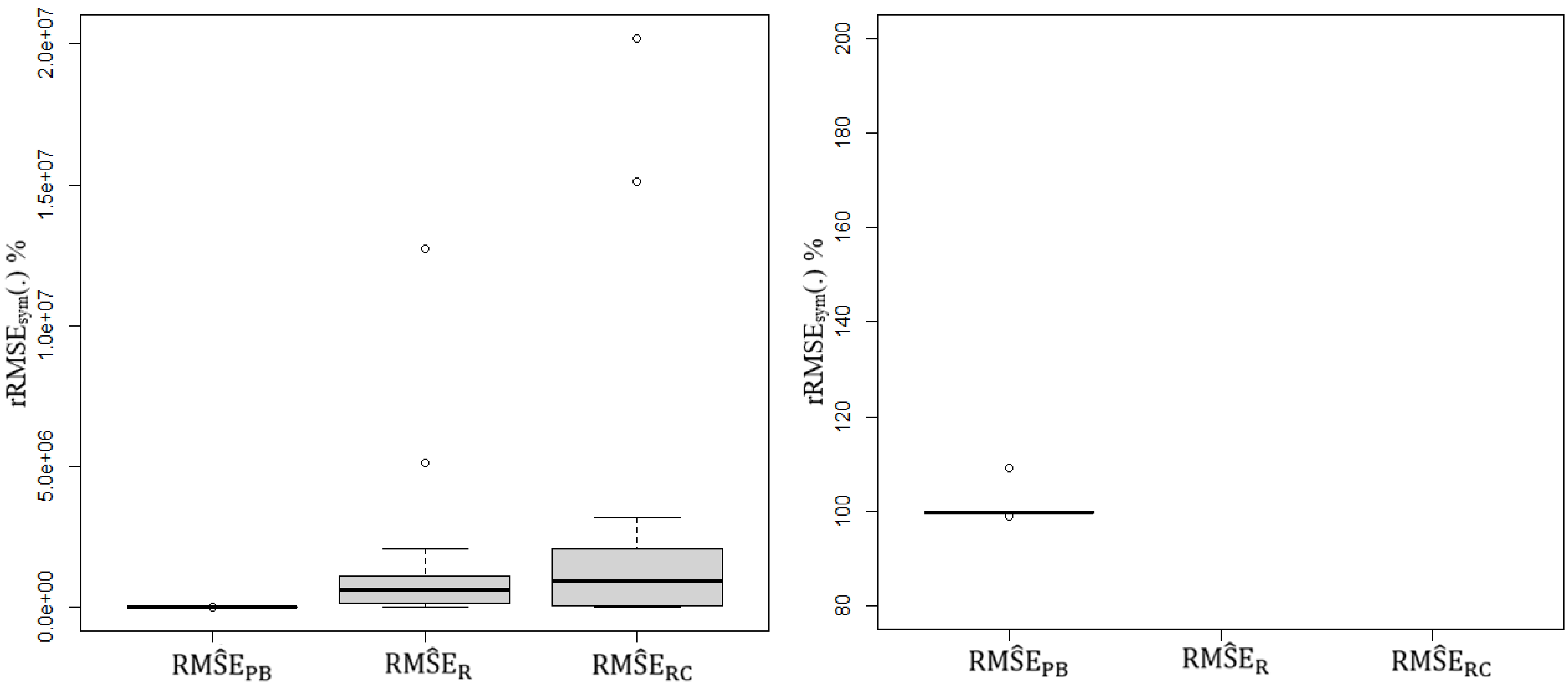
Source: own elaboration

Simulation studies – $rRMSE_{sym}$ ($R\hat{M}SE$) in % model misspecification (t copula, shifted gamma distribution)



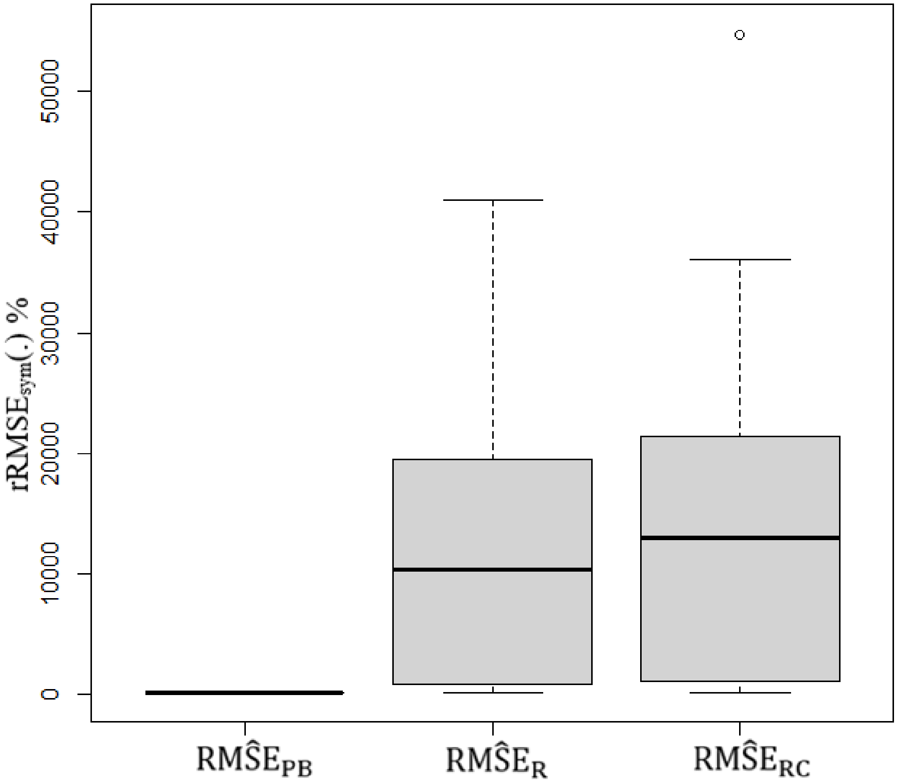
Source: own elaboration

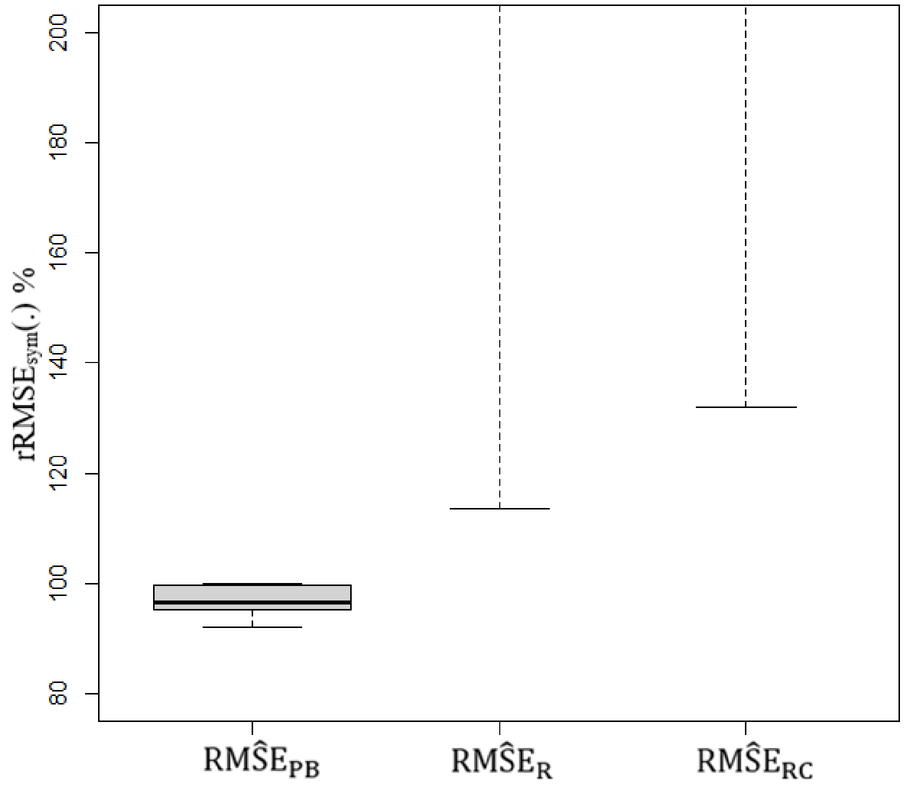
Simulation studies – rRMSE_{sym} ($\mathrm{RM\hat{S}E}$) in % model misspecification
(normal copula, shifted gamma distribution)



Source: own elaboration

Simulation studies – $rRMSE_{sym}$ ($R\hat{MS}E$) in % model misspecification
(t copula, shifted exponential distribution)

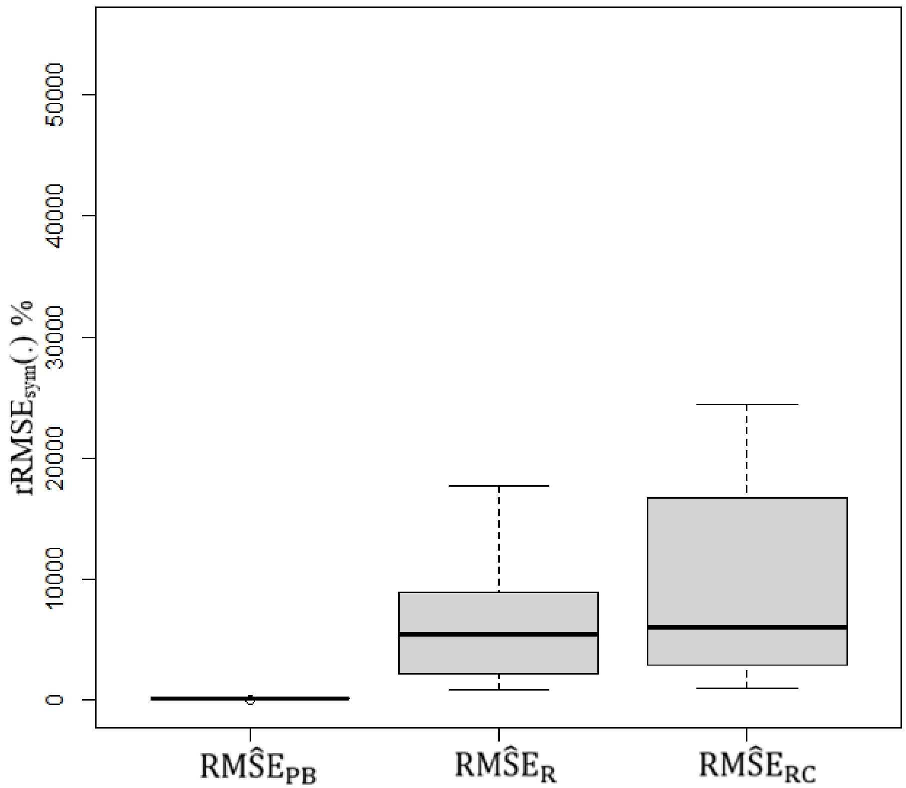


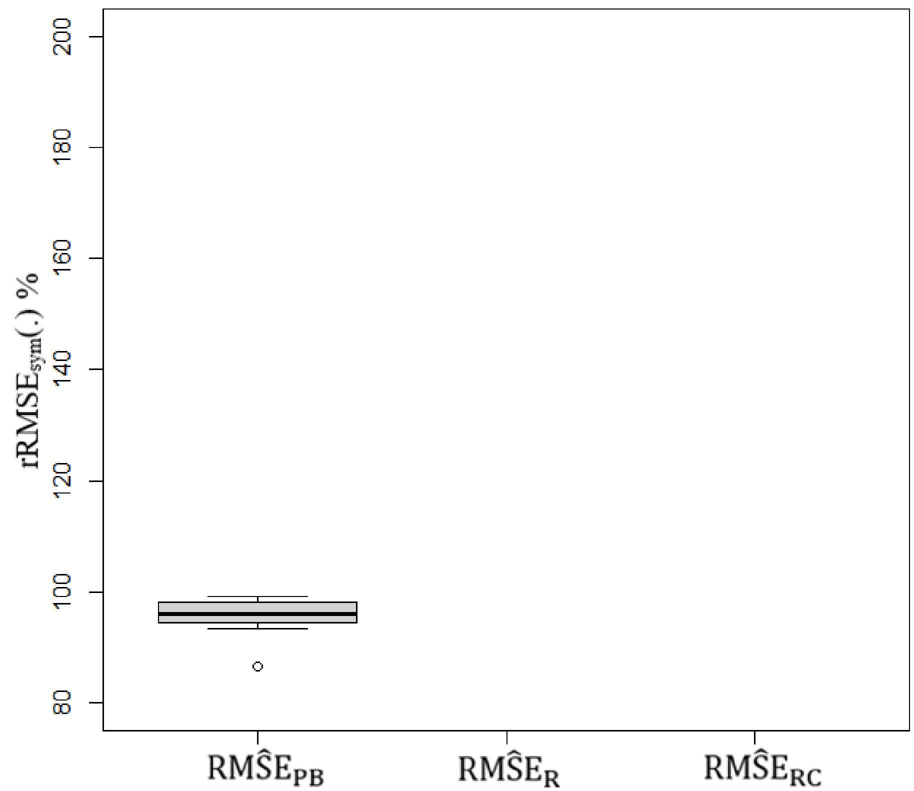


Source: own elaboration

Simulation studies – $\text{rRMSE}_{\text{sym}}$ ($\text{RM}\hat{\text{S}}\text{E}$) in % model misspecification

(normal copula, shifted exponential distribution)





Source: own elaboration

Conclusions – correct model specification

- For the $\hat{\theta}^{\rho,\,\mathrm{wg}}_{PLUG-IN}$ prediction, the medians of the considered $\mathrm{RM\hat{S}E}$ estimators were close to the RMSE value obtained from the simulation.
- The median of absolute relative bias of the analysed estimators did not exceed 5% and was close to 0 for the $\mathrm{RM\hat{S}E}_{RC}$ estimator.
- The lowest $\mathrm{rRMSE_{sym}}$ values were obtained for the estimator using the residual bootstrap method.

Conclusions – considered model misspecification

- The obtained results suggest greater robustness of the considered RMSE estimators to model misspecification due to lack of correlation.

- The results of simulation studies suggest greater robustness among the considered RMSE estimators of the estimator based on the parametric bootstrap method.

Bibliography

Carpenter, J.R., Goldstein, H. and Rasbash, J. (2003), A novel bootstrap procedure for assessing the relationship between class size and achievement, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52, 431-443.

Chatterjee S., Lahiri P. i Li H. (2008), Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models, *The Annals of Statistics*, 36, 1221-1245.

Chwila A., Żądło T. (2019), On properties of empirical best predictors, *Communications in Statistics - Simulation and Computation*, 1-34.

Jiang J. (2007), *Linear and generalized linear mixed models and their applications*, Springer, New York.

Bibliography

Jelsema C.M.D, Peddada S.D. (2016), CLME: An R Package for Linear Mixed Efects Models under Inequality Constraints, *Journal od Statistical Software*, 75, 1-32.

Krzciuk M. (2020), On empirical best linear unbiased predictor under a Linear Mixed Model with correlated random effects, *Econometrics*, 24, 2,17-29.

Krzciuk M.K. (2023a), Small area estimation – model-based approach in economic research, University of Economics in Katowice.

Krzciuk M.K. (2023b), O estymatorach MSE predyktorów typu plug-in dla liniowych modeli mieszanych ze skorelowanymi efektami losowymi, Paper presented at the 41st Conference Multivariate Statistical Analysis, 6-8.11.2023, Łódź.

Bibliography

Littell R.C., Milliken G.A., Stroup W.W., Wolfinger R.D., Schabenberger O. (2006), *SAS for Mixed Models, Second Edition*, Cary, NC: SAS Institute Inc.

Rao J.N., Molina I. (2015), Small area estimation, John Wiley & Sons.

Sklar, A. (1959), Fonctions de répartition à n dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris, 8, 229–231.

Thai H-T., Mentré F., Holford N. H. G., Veyrat-Follet Ch., Comets E. (2013), A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models, Pharmaceutics Statistics, 12, 129–140.

Żądło T. (2023), On bootstrap algorithms in survey sampling, *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie (po recenzji).*

Acknowledgement

The work has been co-financed by the Minister of Science under the "Regional Initiative of Excellence" programme.

Minister of Science
Republic of Poland

Regionalna
Inicjatywa
Doskonałości

Thank you
for Your attention

Mean squared errors of plug-in predictors

We consider the following bootstrap model
(cf. Chatterjee, Lahiri, Li 2008, pp. 1229-1230):

$$\mathbf{Y}^* = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{v}^* + \mathbf{e}^*$$

where:

- $\mathbf{v}^* \sim N\left(\mathbf{0}, \mathbf{G}(\widehat{\boldsymbol{\delta}})\right)$;
- $\mathbf{e}^* \sim N\left(\mathbf{0}, \mathbf{R}(\widehat{\boldsymbol{\delta}})\right)$;
- $\widehat{\boldsymbol{\beta}}$ is the LS estimator of $\boldsymbol{\beta}$;
- $\widehat{\boldsymbol{\delta}}$ is the REML or ML estimator of $\boldsymbol{\delta}$.

Simulation studies – copula functions

Let $\mathrm{H}(\mathrm{X}, \mathrm{Y})$ be a two-dimensional distribution functionwith boundary distributants $\mathrm{F}_1(X)$ and $\mathrm{F}_2(Y)$. Then there exists copula C satisfying the condition (Sklar, 1959):

$$\mathrm{H}(X, Y) = C(\mathrm{F}_1(X), \mathrm{F}_2(Y))$$

If $\mathrm{F}_1$and F_2 are continuous, then C is explicit.