Computational Statistics and Data Analysis 199 (2024) 108014

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bootstrap-based statistical inference for linear mixed effects under misspecifications

Katarzyna Reluga [a,*], María-José Lombardía [b], Stefan Sperlich [c]

[a] University of Bristol, School of Mathematics, Fry Building, Woodland Road, Bristol, BS81UG, United Kingdom
[b] University of A Coruña, CITIC, Facultad de Informatica, Campus de Elviña, A Coruña, 15071, Spain
[c] University of Geneva, Geneva School of Economics and Management, Uni Mail, 40 Boulevard du Pont-d'Arve, Geneva, 1211, Switzerland

ARTICLE INFO

Keywords:
Linear mixed models
Robust bootstrap inference
Small area estimation
Simultaneous inference

ABSTRACT

Linear mixed effects are considered excellent predictors of cluster-level parameters in various domains. However, previous research has demonstrated that their performance is affected by departures from model assumptions. Given the common occurrence of these departures in empirical studies, there is a need for inferential methods that are robust to misspecifications while remaining accessible and appealing to practitioners. Statistical tools have been developed for cluster-wise and simultaneous inference for mixed effects under distributional misspecifications, employing a user-friendly semiparametric random effect bootstrap. The merits and limitations of this approach are discussed in the general context of model misspecification. Theoretical analysis demonstrates the asymptotic consistency of the methods under general regularity conditions. Simulations show that the proposed intervals are robust to departures from modelling assumptions, including asymmetry and long tails in the distributions of errors and random effects, outperforming competitors in terms of empirical coverage probability. Finally, the methodology is applied to construct confidence intervals for household income across counties in the Spanish region of Galicia.

1. Introduction

Linear mixed models are frequently employed for modelling hierarchical and longitudinal data. Within this modelling framework, population parameters are represented by fixed regression parameters, while the additional between-cluster variation is captured by cluster-specific random effects. Bootstrap methods are considered for statistically valid inference for mixed effects, which are linear combinations of fixed and random effects. Mixed effects are recognized as excellent predictors of cluster-level parameters in various fields, such as small area estimation or medicine (cf. monographs of Verbeke and Molenberghs, 2000; Jiang, 2007; Rao and Molina, 2015).

Further inference on mixed parameters heavily relies on model and distributional assumptions. Consequently, numerous bootstrap methods have been introduced to partially alleviate this reliance and flexibly approximate distribution functions of estimators. While they can be derived analytically using model-dependent large sample theory, the application of the latter might lead to inaccurate results in finite samples, and it is rarely robust to misspecifications.

* Corresponding author.
E-mail address: katarzyna.reluga@bristol.ac.uk (K. Reluga).

https://doi.org/10.1016/j.csda.2024.108014
Received 24 January 2024; Received in revised form 20 June 2024; Accepted 21 June 2024

Available online 1 July 2024

The family of bootstrap methods for clustered data is rich. Field and Welsh (2007); Chambers and Chandra (2013) and more recently Flores-Agreda and Cantoni (2019) provide extensive reviews of bootstrap methods for clustered data. In fact, all essential procedures can be classified into three broad categories: bootstrapping by resampling clusters and observations within clusters (Davison and Hinkley, 1997; McCullagh, 2000), bootstrapping by random weighing of estimating equations (Field et al., 2010; Samanta and Welsh, 2013; O'Shaughnessy and Welsh, 2018) and bootstrapping by resampling predictors of random effects and/or residuals (Davison and Hinkley, 1997). The latter is referred to as a random effect bootstrap and can be further subcategorized into: parametric versions, which entails resampling from normal distributions with estimated variance (Butar and Lahiri, 2003; Hall and Maiti, 2006b; Chatterjee et al., 2008), and semiparametric versions in which stochastic components are obtained using estimated variance components, but without making distributional assumptions (Carpenter et al., 2003; Hall and Maiti, 2006a; Lombardía and Sperlich, 2008; Opsomer et al., 2008; Chambers and Chandra, 2013; Bertarelli et al., 2021). The main goal of all bootstrap schemes is to estimate the sample distribution or certain moments of a statistic using its bootstrap distribution. It is important to emphasize that the goal of this manuscript is *not* to compare the performance of all existing procedures or select the optimal scheme, even though such a comparison in the framework of mixed effects has not been attempted yet, and it might be an interesting future research direction.

There exist various criteria to assess the quality of bootstrap schemes for clustered data. In the context of inference for mixed parameters, the existing literature primarily concentrates on the bootstrap estimation of the mean squared error, which essentially involves an accurate estimation of the first few moments (see, e.g. Butar and Lahiri, 2003; Hall and Maiti, 2006a; Chatterjee et al., 2008). In contrast, proposed bootstrap methods are required to reproduce cumulative distribution functions of studentized and maximal statistics, which are core elements in cluster-wise and simultaneous inference for mixed parameters. While the former has been intensively studied in the literature, the latter has been neglected until recently despite its role in formulating valid comparative statements. These are of utmost importance in policy making or public health, where decision-makers compare the estimates across many clusters to decide on resource reallocation or the introduction of new policies (see Tzavidis et al., 2018; Reluga et al., 2023b; Kramlinger et al., 2022, for discussions on the need for comparative tools in the small area estimation which relies on mixed effects).

In this manuscript, statistical tools have been constructed for cluster-wise and simultaneous inference for mixed parameters under distributional model misspecifications, employing a semiparametric random effect bootstrap as in Carpenter et al. (2003) and Opsomer et al. (2008). This bootstrap scheme is proven to successfully reproduce cumulative distribution functions of studentized and maximal statistics. These results provide novel contributions that complement and extend the derivations of Carpenter et al. (2003), who, assuming normality, proved the consistency of bootstrap versions of fixed effects, as well as the variance of random effects and errors. Additionally, this work extends the methods developed by Reluga et al. (2023b) for cluster-wise and simultaneous inference for linear mixed effects under satisfied modelling assumptions.

The presented theory applies to the construction of intervals and corresponding testing procedures. The analysis demonstrates that the methods are asymptotically consistent under general regularity conditions, without assuming normality of stochastic components in the model. Simulations compare the performance of various bootstrap schemes to construct cluster-wise and simultaneous intervals for mixed parameters. The results indicate that the constructed intervals are robust to departures from modelling assumptions, including asymmetry and long tails in the distributions of errors and random effects, outperforming competitors in terms of empirical coverage probability. Additionally, the methodology is applied to construct simultaneous confidence intervals for household income across counties in the Spanish region of Galicia. Finally, this simple bootstrap-based inference can be considered complementary to other existing techniques that handle distributional model misspecifications and deal with outliers, such as robust inference methods (Chambers and Tzavidis, 2006b; Sinha and Rao, 2009) or estimation by data transformation (Rojas-Perilla et al., 2020).

The remainder of this paper is structured as follows. Section 2 contains a brief introduction to the inference for mixed parameters. Section 3 describes the methodology to construct cluster-wise and simultaneous inference tools under distributional model misspecification. Section 4 discusses potential extensions and limitations of proposed methodology in the broader context of constructing statistical methods under model misspecification. The simulation results and the data application are provided in Sections 5 and 6, respectively. Finally, Section 7 presents the conclusion. Proofs and regularity conditions can be found in Appendix A. The supplementary material (SM) contains further analytical details and simulations.

2. Inference for mixed parameters

Consider an n-dimensional response vector $y \in \mathbb{R}^n$ modelled by $y = X\beta + Zu + e$ where $X \in \mathbb{R}^{n\times(p+1)}$, $Z \in \mathbb{R}^{n\times q}$ are known full column rank design matrices for fixed and random effects, vector $\beta \in \mathbb{R}^{p+1}$ contains fixed effects, whereas random effects $u \in \mathbb{R}^q$ and errors $e \in \mathbb{R}^n$ are assumed to be mutually independent and identically distributed with $var(e) = G$ and $var(u) = R$. The model of Laird and Ware (1982) is used. It follows that for each cluster j one has

$$y_j = X_j\beta + Z_j u_j + e_j, \quad j = 1, \dots, m, \tag{1}$$

where $y_j \in \mathbb{R}^{n_j}$, $X_j \in \mathbb{R}^{n_j\times(p+1)}$, $Z_j \in \mathbb{R}^{n_j\times q_j}$, $e_j = (e_1^T, e_2^T, \dots, e_m^T)^T$, $u = (u_1, u_2, \dots, u_m)^T$. The total sample size is denoted with n, the number of clusters with m such that $n = \sum_{j=1}^m n_j$, where n_j is the number of observations in the j^{th} cluster. Furthermore, G and R are block-diagonal with blocks $G_j = G_j(\delta) \in \mathbb{R}^{q_j\times q_j}$ and $R_j = R_j(\delta) \in \mathbb{R}^{n_j\times n_j}$ which depend on variance parameters $\delta = (\delta_1, ..., \delta_h)^T$. Let $E(y) = X\beta$ and $var(y) = V = R + ZGZ^T$ where V is a block-diagonal with blocks $V_j = R_j + Z_j G Z_j^T$. Under normality of random effects and errors, $y_j \sim N(X_j\beta, V_j)$ and $y_j|u_j \sim N(X_j\beta + Z_j u, G_j)$. The methods of maximum likelihood and restricted maximum likelihood are often used to obtain an estimator $\hat{\delta} = (\hat{\delta}_1, \dots, \hat{\delta}_h)^T$ (see, for example, Verbeke and Molenberghs, 2000, Chapter 5). In contrast, β and u are estimated and predicted using two-stage techniques. In particular, after assuming that δ is

known, one can use maximum likelihood, estimating equations of Henderson (1950) or the h-likelihood of Lee and Nelder (1996) to obtain best unbiased linear estimator (BLUE) $\tilde{\beta} = \beta(\delta) = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ and the best unbiased linear predictors (BLUP) $\tilde{u}_j = u_j(\delta) = G_j Z_j^T V_j^{-1}(y_j - X_j \tilde{\beta})$. In the second stage, δ is replaced by $\hat{\delta}$ which results in empirical BLUE (EBLUE) $\hat{\beta} = \beta(\hat{\delta})$, and empirical BLUP (EBLUP) $\hat{u}_j = u_j(\hat{\delta})$. The goal is to develop valid inferential tools for general cluster-level parameters

$$\theta_j = k_j^T \beta + l_j^T u_j, \quad j = 1, \dots, m, \tag{2}$$

with known $k_j \in \mathbb{R}^{p+1}$ and $l_j \in \mathbb{R}^{q_j}$, i.e., θ_j is treated as random cluster-level parameter due to the randomness of u_j. The application of the two-stage approach leads to

$$\tilde{\theta}_j = \theta_j(\delta) = k_j^T \tilde{\beta} + l_j^T \tilde{u}_j, \quad \hat{\theta}_j = \theta_j(\hat{\delta}) = k_j^T \hat{\beta} + l_j^T \hat{u}_j \quad j = 1, \dots, m. \tag{3}$$

The focus is on developing methods to construct $1 - \alpha$ confidence (or prediction) intervals and perform hypothesis testing for mixed parameters θ_j, $j = 1, \dots$, following the ideas of Reluga et al. (2023b). Let $\sigma_j^2 = var(\hat{\theta}_j)$ and $\hat{\sigma}_j^2 = \widehat{var}(\hat{\theta}_j)$ be a corresponding estimator of the variability of the parameter estimate. A t-statistic and a maximal statistic are defined as follows:

$$t_j = \frac{\hat{\theta}_j - \theta_j}{\widehat{var}^{1/2}(\hat{\theta}_j)}, \quad j = 1, \dots, m, \quad M = \max_{j=1,\dots m} \left| t_j \right|. \tag{4}$$

An individual confidence interval $I_{j,1-\alpha}$ at $1-\alpha$-level for θ_j is a region which satisfies $P(\theta_j \in I_{j,1-\alpha}) = 1 - \alpha$. To construct symmetric intervals $I_{j,1-\alpha}$, it is enough to find a high quantile from the distribution of statistic t_j, that is $q_{j,1-\alpha} = \inf\{a \in \mathbb{R} : P(t_j \le a) \ge 1 - \alpha\}$. One thus have

$$I_{j,1-\alpha} : \left\{ \hat{\theta}_j + q_{j,\alpha/2} \times \hat{\sigma}_j, \hat{\theta}_j + q_{j,1-\alpha/2} \times \hat{\sigma}_j \right\}, \quad j = 1, \dots, m. \tag{5}$$

Due to the central limit theorem, $q_{j,\alpha/2}$ and $q_{j,1-\alpha/2}$ are often replaced by high quantiles from the standard normal distribution or the Student's t-distribution. Alternatively, one can consider constructing symmetric intervals with $\pm q_{j,1-\alpha/2}$. In simulations in Section 5, the performance of both symmetric and asymmetric intervals was tested, and it was concluded that the latter as defined in (5) performed better in terms of empirical coverage.

A similar strategy can be used to construct simultaneous confidence intervals $I_{1-\alpha}$ at $1-\alpha$-level which satisfy $P(\theta_j \in I_{1-\alpha}\ \forall j \in [m]) = 1 - \alpha$, where $[m] = \{1, \dots, m\}$. Let $q_{1-\alpha} = \inf\{a \in \mathbb{R} : P(M \le a) \ge 1 - \alpha\}$ be a high quantile from the distribution of statistic M. Then it follows that

$$I_{1-\alpha} = \bigtimes_{j=1}^{m} I_{j,1-\alpha}^s, \quad I_{j,1-\alpha}^s : \left\{ \hat{\theta}_j \pm q_{1-\alpha} \times \hat{\sigma}_j \right\}, \quad j = 1, \dots, m, \tag{6}$$

and it follows that $I_{1-\alpha}$ covers all clusters with probability $1 - \alpha$ (see Reluga et al., 2023a,b, for more details). The relation between confidence intervals and hypothesis testing allows us to define modified statistics t_j and M that can be used to carry out hypothesis testing. Due to the inherent similarity between intervals and tests, the derivation can be found in the SM.

The construction of the studentized statistics in (4) requires the estimation of σ_j^2. The most common measure to assess the variability of predictions is the mean squared error $\mathrm{MSE}(\hat{\theta}_j) = E[(\hat{\theta}_j - \theta_j)^2]$, where E denotes the expectation with respect to the model defined in (1). Nevertheless, following Chatterjee et al. (2008), a simpler choice of $\tilde{\sigma}_j^2 = l_j^T (G_j - G_j Z_j^T V_j^{-1} Z_j G_j) l_j$, which accounts for the variability of θ_j without accounting for the variability in the estimation of β or δ, leads to satisfactory numerical results. In Section 5 the empirical performance of cluster-wise and simultaneous intervals is tested using six different estimators of σ^2, and the application of $\tilde{\sigma}_j^2$ leads to the best results under many bootstrapping schemes.

3. Bootstrap-based inference for mixed effects

Bootstrap schemes to construct individual and simultaneous intervals are introduced, with the aim of ensuring robustness to model misspecifications, including asymmetry and long tails in the distribution of errors and random effects. Bootstrap-generated observations are denoted as:

$$y^* = X\hat{\beta} + Zu^* + e^*, \tag{7}$$

where e^* and u^* are bootstrap replica of random components in the model. The generation of e^* and u^* depends on the bootstrap scheme, which will be further elaborated upon. Setting $\delta^* = \hat{\delta}$, $V^* = \hat{V}$, $G^* = \hat{G}$, the following definitions are used $\tilde{\beta}^* = \beta(\delta^*) = (X^T V^{*-1} X)^{-1} X^T V^{*-1} y^*$, $\tilde{u}_j^* = u_j(\delta^*) = G_j^* Z_j^T V_j^{*-1}(y_j^* - X_j \tilde{\beta}^*)$. In addition, let $\hat{\delta}^*$ be an estimated version of δ^* obtained by regressing y^* on X. Then one obtains $\hat{\beta}^* = \beta(\hat{\delta}^*)$ and $\hat{u}_j^* = u_j(\hat{\delta}^*)$. Bootstrap mixed effects are thus defined as

$$\theta_j^* = k_j^T \hat{\beta} + l_j^T u_j^*, \qquad \hat{\theta}_j^* = \theta_j(\hat{\delta}^*) = k_j^T \hat{\beta}^* + l_j^T \hat{u}_j^*. \tag{8}$$

The bootstrap versions of the statistics of interest in (4) are given by

$$t_j^* = \frac{\hat{\theta}_j^* - \theta_j^*}{\widehat{var}^{1/2}(\hat{\theta}_j^*)}, \quad j = 1, \ldots, m, \quad M^* = \max_{j=1,\ldots m} \left| t_j^* \right|. \tag{9}$$

Statistics in (9) are used to construct bootstrap equivalents of intervals in (5) and (6), that is

$$I_{j,1-\alpha}^* : \left\{ \hat{\theta}_j + q_{j,\alpha/2}^* \times \hat{\sigma}_j, \hat{\theta}_j + q_{j,1-\alpha/2}^* \times \hat{\sigma}_j \right\}, \tag{10}$$

$$I_{1-\alpha}^* = \bigtimes_{j=1}^{m} I_{j,1-\alpha}^{*s}, \; I_{j,1-\alpha}^{*s} : \left\{ \hat{\theta}_j \pm q_{1-\alpha}^* \times \hat{\sigma}_j \right\}, \tag{11}$$

where $q_{j,1-\alpha}^* = \inf\{a \in \mathbb{R} : P(t_j^* \leq a) \geq 1-\alpha\}$ and $q_{1-\alpha}^* = \inf\{a \in \mathbb{R} : P(M^* \leq a) \geq 1-\alpha\}$ are quantiles from the distributions of t_j^* and M^*. Alternatively, one can consider a symmetric confidence interval in (10), that is $I_{j,1-\alpha}^* : \left\{ \hat{\theta}_j \pm q_{j,1-\alpha/2}^* \times \hat{\sigma}_j \right\}$. This choice has hardly any effect on the performance of $I_{j,1-\alpha}^*$ for large n_j, but it might make a difference once n_j is small. The most popular choice in the context of mixed models is to use a parametric bootstrap, drawing e^* and u^* from a postulated normal distribution with estimated variance parameters. In contrast, a semiparametric bootstrap method introduced by Carpenter et al. (2003) and generalised by Opsomer et al. (2008) is used. The empirical performance of this bootstrap scheme for fixed parameters and variance components has been studied by Chambers and Chandra (2013). The goal is to mimic the data generating process in model (1). Before explicitly writing down the bootstrap algorithm, it is important to provide some motivation behind it. Let $\tilde{y} = X\tilde{\beta} = X(X^T V^{-1} X)^{-1} X^T V^{-1} y = Hy$, $\tilde{e} = y - X\tilde{\beta} - Z\tilde{u} = (I - ZGZ^T V^{-1})(I - H)y = RV^{-1}(I - H)y$, $\hat{e} = y - X\hat{\beta} - Z\hat{u}$ and $\hat{u} = \hat{G}Z^T \hat{V}^{-1}(y - X\hat{\beta})$. Then, by some algebraic transformations one has $I - ZGZ^T V^{-1} = RV^{-1}$, and

$$var(\tilde{u}) = GZ^T\{V^{-1}(I-H)\}ZG, \quad var(\tilde{e}) = R\{V^{-1}(I-H)\}R.$$

Thus, $\hat{e}$ and $\hat{u}$ should be re-scaled before sampling with replacement to avoid the effects of shrinkage (Morris, 2002). Centring, that is subtracting the empirical mean, is also advisable to ensure that the empirical rescaled residuals have a mean zero. Thus one should consider sampling from

$$\hat{e}_{sc} = \hat{e}_s - \bar{\hat{e}}_s, \quad \bar{\hat{e}}_s = \sum_{i=1}^{n} \frac{\hat{e}_{si}}{n}, \quad \hat{e}_s = [R\{V^{-1}(I-H)\}]^{-1/2}\hat{e}, \tag{12}$$

$$\hat{u}_{sc} = \hat{u}_s - \bar{\hat{u}}_s, \quad \bar{\hat{u}}_s = \sum_{i=1}^{n} \frac{\hat{u}_{sj}}{m}, \quad \hat{u}_s = [GZ^T\{V^{-1}(I-H)\}Z]^{-1/2}\hat{u}. \tag{13}$$

Below is an algorithm to obtain bootstrap quantiles and construct intervals as in (10) and (11).

Algorithm 1. A semiparametric random bootstrap algorithm

1. Obtain consistent estimators $\hat{\beta}$ and $\hat{\delta}$.
2. For $b = 1, 2, \ldots, B$:
   (a) Obtain vectors $e^{*(b)} \in \mathbb{R}^n$, $u^{*(b)} \in \mathbb{R}^m$ by sampling independently with replacement from $\hat{e}_{sc}$ in (12) and $\hat{u}_{sc}$ in (13).
   (b) Generate sample $y^{*(b)} = X\hat{\beta} + Zu^{*(b)} + e^{*(b)}$ as in (7) and obtain $\theta_j^{*(b)} = k_j^T\hat{\beta} + l_j^T u_j^{*(b)}$, $j = 1, \ldots, m$.
   (c) Fit a LMM to the bootstrap sample of the previous step.
   (d) Obtain bootstrap estimates $\hat{\delta}^{*(b)}$, $\hat{\beta}^{*(b)}$, $\hat{\theta}_j^{*(b)}$, $t_j^{*(b)}$ and $M^{*(b)}$, $j = 1, \ldots, m$.
3. Estimate critical values $q_{j,\alpha/2}^*$, $q_{j,1-\alpha/2}^*$, $q_{1-\alpha}^*$ by the $[\{(\alpha/2)B\}+1]^{th}$ and $[\{(1-\alpha/2)B\}+1]^{th}$ order statistics of t_j^*, and $[\{(1-\alpha)B\}+1]^{th}$ order statistic of $M^{*(b)}$.
4. Construct bootstrap intervals as in (10) and (11).

Fisher consistency of $\hat{\delta}^*$ and $\hat{\beta}^*$ obtained using the semiparamteric bootstrap of Algorithm 1 has been proven by Carpenter et al. (2003). In Lemma 1 and 2 the consistency of statistics t_j^* and M^* is shown.

Lemma 1 *(Consistency of t_j^*). Let t_j and t_j^* be as defined in (4) and (9), $m \to \infty$, $n_j \to \infty$, $\lim_{n_j, n\to\infty} n_j/n = c$, $c > 0$. If the regularity conditions in Appendix A.1 are satisfied, then one has in probability*

$$\sup_{a \in \mathbb{R}} \left| F_{t_j}(a) - F_{t_j^*}(a) \right| \to 0.$$

Corollary 1 ensures the consistency of individual intervals introduced in (10).

Corollary 1 *(Consistency of $I_{j,1-\alpha}^*$). Lemma 1 implies that under the same regularity conditions one has*

$$P(\theta_j \in I_{j,1-\alpha}) \to 1 - \alpha, \quad \alpha \in (0,1).$$

Consistency of M^* does not follow from Lemma 1 by the delta method, because the max function is not differentiable. Lemma 2 provides a heuristic proof based on some results known from extreme value theory.

Lemma 2 *(Consistency of M^*). Let M and M^* be as defined in (4) and (9), $m \to \infty$, $n_j \to \infty$, $\lim_{n_j, n \to \infty} n_j/n = c$, $c > 0$. If the regularity conditions in Appendix A.1 are satisfied and Lemma 1 holds, then one has in probability*

$$\sup_{a \in \mathbb{R}} \left| F_M(a) - F_{M^*}(a) \right| \to 0.$$

Corollary 2 ensures the consistency of bootstrap simultaneous intervals of (11).

Corollary 2. *Lemma 2 implies that under the same assumptions one has*

$$P(\theta_j \in I^*_{1-\alpha} \ \forall j \in [m]) \to 1 - \alpha, \quad \alpha \in (0, 1).$$

4. Inference beyond moderate distributional model misspecifications

Misspecification of modelling assumptions is a very broad concept. The methodology presented here is readily applicable when dealing with distributional model misspecification, specifically when the true data generating process follows a linear mixed model in (1) but does not assume normality for the random effects or errors, a common situation in practice (cf., the data analysis in Section 6). However, one can consider at least two broader frameworks.

First, assume that the true data generating process G_j, i.e. $y_j \sim G_j$, is any distribution, which does not necessarily depend on θ_j. In this context, considering inference for θ_j becomes an ill-posed problem as the true θ_j might even not exist. Following White (1982), one approach is to consider cluster-wise and simultaneous inference for a pseudo-true parameter $\breve{\theta}_j$ which minimizes the Kullback–Leibler (KL) divergence between the true and the model density. The development of methodology to cover this case requires substantial theoretical extensions such as sup inf convergence of cumulative distribution functions of statistics which is an interesting future research direction. While these considerations are interesting, the primary advantage of this paper for practitioners lies in its provision of easy, yet theoretically well-grounded, solutions for anyone interested in conducting statistical inference on the cluster or area effects under common model misspecification.

Second, let us consider a misspecification in covariates, often referred to as an omitted variable problem. In this scenario, the true data-generating process follows a linear mixed model as shown in (1), but it is not modelled with the same linear mixed model, possibly due to a lack of access to all covariates or an erroneous omission of some polynomial terms. If the omitted terms are uncorrelated with the included covariates, the omission increases the variances of the random terms and may cause heteroscedasticity. However, nothing else changes, and proposed method remains unaffected. This differs from cases where the omitted variables are correlated with the included ones, leading to a correlation between the included covariates and random terms (u or e). Consequently, the estimates $\hat{\theta}_j$ can be systematically biased, resulting in the incorporation of bias in the estimation of the empirical cumulative distribution function F_{t_j} and its bootstrap version $F_{t_j^*}$, as discussed in Lombardía and Sperlich (2012). As a result, both the cluster-wise and simultaneous intervals are not guaranteed to provide good coverage. This was confirmed by the results of a simulation study, which, although not presented here, is available upon request from the authors. The problem might be alleviated by introducing a surrogate covariate highly predictive of the missing one.

Finally, this manuscript proposes simple tools for cluster-wise and simultaneous inference once analysed data suffers from moderate departures from modelling assumptions. Nevertheless, alternatives should be used if data are contaminated by outliers, a problem which is particularly severe in the small area estimation. In this situation, a common approach is to fit a robust working model to the sample data and then use it to predict non-sampled units in the population of interest, see Chambers and Tzavidis (2006a); Sinha and Rao (2009); Chambers et al. (2014). The construction of confidence intervals or the estimation of MSE in this challenging setup often requires bootstrapping. The bootstrapping scheme needs to be adapted such that outliers from the sample have a controlled effect on the bootstrap variability. Since the main goal is *not* to compare the performance of all existing bootstrap procedures in the presence of various types of outliers (see the literature review in Section 1), this discussion is limited to some recent bootstrap schemes whose performance was tested in the context of multi-level data and small area estimation. In the context of mixed and multi-level data, Modugno and Giannerini (2015) compared numerically several bootstrap schemes. In their simulation study, a variation of wild bootstrap performed the best in terms of the coverage of confidence intervals for fixed parameters and variance components. In the context of SAE, Bertarelli et al. (2021) proposed a bounded block bootstrap, an extension of REB that uses 'Huberized' cluster-level and individual-level residuals to restrict the effect of outliers. The latter proved to be effective for the construction of MSE for mixed effects. Nevertheless, neither Modugno and Giannerini (2015) nor Bertarelli et al. (2021) studied the performance of their schemes to construct cluster-wise or simultaneous intervals for mixed effects.

5. Simulation study

5.1. Setup

Numerical simulation studies are conducted to evaluate the finite sample properties of the bootstrap intervals. In all scenarios, outcomes are generated from a linear mixed effect model (1) with a fixed and a random intercept, and a uniformly distributed

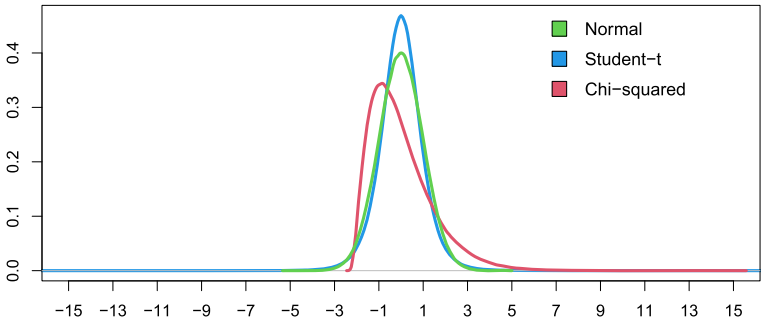

Fig. 1. Probability density functions used to generate errors and random effects in simulations.

covariate. Specifically, the settings are $x_{ij1} = 1$, $z_{ij} = 1$, $x_{ij2} \sim U(0,1)$. Three types of sample sizes are considered to mimic joint asymptotics: $n_j \in \{5, 10, 15\}$, $m \in \{25, 50, 75\}$. Furthermore, in each simulation run, errors and random effects are drawn from one of the following distributions: standard normal, Student's t with 6 degrees of freedom, or chi-squared with 5 degrees of freedom. Fig. 1 displays kernel density plots of 10^6 realisations of a random variable generated using one of the above mentioned distribution. The Student's t-distribution is symmetric, but it has very long tails in comparison to the normal distribution. On the other hand, chi-square is highly right skewed, with a long right tail. The simulation setting is deemed challenging enough to test the robustness of the methods to misspecifications of distributions of errors and/or random effects.

The distributions are centred to zero and re-scaled to variances $var(e_{ij}) = 1$ and $var(u_j) = 1$. The following scenarios are considered: (a) $e_{ij} \sim \chi^2_5$, $u_j \sim \chi^2_5$, (b) $e_{ij} \sim \chi^2_5$, $u_j \sim t_6$, (c) $e_{ij} \sim N(0,1)$, $u_j \sim N(0,1)$, (d) $e_{ij} \sim t_6$, $u_j \sim \chi^2_5$, (e) $e_{ij} \sim t_6$, $u_j \sim t_6$. The target parameter in each simulation run was set to $\theta_j = 1 + \bar{x}_{j2} + u_j$, where $\bar{x}_{j2} = \sum_{i=1}^{n_j} x_{ij2}/n_j$ and u_j is a generated random effect in cluster j which is different for each simulation run. The performance of cluster-wise intervals in (10) and simultaneous intervals in (11) at the $\alpha = 0.05$ level is compared. These intervals are obtained using the semiparametric bootstrap (SPB) presented in Section 3, a parametric bootstrap (PB) as in Chatterjee et al. (2008), a wild bootstrap as in Lombardía and Sperlich (2008) and a random effect block bootstrap as in Chambers and Chandra (2013). Different auxiliary distributions are considered while implementing a wild bootstrap: Rademacher distribution (WBR), Mammen's distribution (WBM, cf., Mammen, 1993), and a standard normal distribution (WBN). Two variations of REB, as originally proposed by Chambers and Chandra (2013), are implemented: with a random selection of clusters without scaling (REB) and with scaling (REBs). Additionally, the performance of the REB variants without the random selection of clusters is tested: without scaling (REBnc) and with scaling (REBsnc). Moreover, the performance of intervals constructed using large-sample asymptotic approximations (A) is evaluated with $(1-\alpha/2)$ and $(1-\alpha/(2\times m))$ quantiles from a normal distribution (the latter by Bonferroni correction).

Inference based on the asymptotic theory or parametric bootstrap is notorious for not being robust to model misspecifications. Wild bootstrap proved to be successful in alleviating the problem of heteroscedasticity (see, among other Sugasawa and Kubokawa, 2017, for a recent contribution), whereas the random effect bootstrap with scaling turned out to be successful in the inference for fixed effects and variance components in the presence of distributional misspecifications or autocorrelated errors (Chambers and Chandra, 2013). Therefore, they are included in the study for comparison. Finally, six different methods to retrieve an estimate of σ_j^2 are used: $\tilde{\sigma}_j^2$ defined in Section 2 (var), asymptotic version of the MSE estimator under normality (mse_a), and four bootstrap estimators of the MSE (mse_b, mse_{3t}, mse_{sp}, mse_{spa}). Since the choice of estimators of variability is of minor importance, the definition of these estimators and related simulation results are deferred to the SM. In total, the performance of 55 types of cluster-wise and simultaneous intervals is tested (9 bootstrap schemes × 6 variance estimators plus an asymptotic derivation).

The following criteria to assess the performance of intervals are employed: empirical coverage probability in (14) for individual and simultaneous intervals, average widths of the intervals in (15), and the coverage error in percentage (16), all of them over $N_s = 1000$ simulation runs:

$$\mathrm{Cov}_{ind} = \frac{1}{mN_s}\sum_{j=1}^{m}\sum_{n_s=1}^{N_s}\mathbf{1}\{\theta_j^{(n_s)} \in I^{*(n_s)}_{j,1-\alpha}\}, \quad \mathrm{Cov}_{sim} = \frac{1}{N_s}\sum_{n_s=1}^{N_s}\mathbf{1}\{\theta_j^{(n_s)} \in I^{*(n_s)}_{1-\alpha}\ \forall j \in [m]\}, \tag{14}$$

$$\mathrm{Len} = \frac{1}{mN_s}\sum_{j=1}^{m}\sum_{n_s=1}^{N_s}\rho_j^{(n_s)}, \quad \rho_j^{(n_s)} = 2q^{(n_s)}_{(\cdot)}\hat{\sigma}_j^{(n_s)}, \tag{15}$$

$$\%\mathrm{err}_{\mathrm{ind}} = \frac{|\mathrm{Cov}_{\mathrm{ind}} - (1-\alpha)|}{(1-\alpha)} \times 100, \quad \%\mathrm{err}_{\mathrm{sim}} = \frac{|\mathrm{Cov}_{\mathrm{sim}} - (1-\alpha)|}{(1-\alpha)} \times 100 \tag{16}$$

where $(\cdot)$ stands for subindices j, α for cluster-wise intervals, and index α for simultaneous intervals. In each simulation run, they are $B = 1000$ bootstrap samples generated.

Remark 1. In their paper, Chambers and Chandra (2013) introduced three variations of the REB bootstrap: "REB/0", "REB/1" and "REB/2". REB/0 and REB/1 correspond to REB and REBs in this manuscript. REB/2 involves scaling and bias adjustment after the

Table 1
Empirical coverage (Cov.), length (Len.) and error of coverage in percentage (%err_{ind}) of individual intervals at $1-\alpha=0.95$ level, $e_{ij}\sim\chi^2_5$, $u_j\sim\chi^2_5$.

Method	$\hat{\sigma}^2$	$m=25, n_j=5$			$m=50, n_j=10$			$m=75, n_j=15$		
		%err_{ind}	Cov.	Len.	%err_{ind}	Cov.	Len.	%err_{ind}	Cov.	Len.
A	mse_a	0.22	94.79	1.67	0.04	95.04	1.21	0.02	94.98	1.00
SPB	var	0.24	94.77	1.69	0.54	94.49	1.21	0.50	94.53	0.99
PB	var	0.17	94.84	1.69	0.06	94.95	1.21	0.13	94.88	0.99
REB	var	0.70	95.67	1.77	0.25	95.23	1.24	0.08	95.08	1.01
REBnc	var	8.20	87.21	1.56	4.89	90.36	1.15	3.54	91.64	0.96
REBs	mse_b	0.15	95.14	1.72	0.07	95.07	1.23	0.00	95.00	1.01
REBsnc	mse_b	0.15	95.14	1.72	0.07	95.07	1.23	0.00	95.00	1.01
WBM	var	7.02	88.33	3.58	4.82	90.42	1.13	3.71	91.48	0.94
WBN	var	6.55	88.78	3.69	4.81	90.43	1.15	3.74	91.45	0.96
WBR	var	8.93	86.52	1.47	5.86	89.44	1.11	4.41	90.82	0.93

Table 2
Empirical coverage (Cov.), length (Len.) and error of coverage in percentage (%err_{ind}) of simultaneous intervals at $1-\alpha=0.95$ level, $e_{ij}\sim\chi^2_5$, $u_j\sim\chi^2_5$.

Method	$m=25, n_j=5$				$m=50, n_j=10$				$m=75, n_j=15$			
	$\hat{\sigma}^2$	%err_{ind}	Cov.	Len.	$\hat{\sigma}^2$	%err_{ind}	Cov.	Len.	$\hat{\sigma}^2$	%err_{ind}	Cov.	Len.
A	mse_a	4.84	90.40	2.64	mse_a	5.68	89.60	2.03	var	6.21	89.10	1.73
SPB	var	1.47	93.60	2.82	mse_{sp}	1.58	93.50	2.17	mse_{spa}	2.11	93.00	1.85
PB	var	3.16	92.00	2.70	var	5.26	90.00	2.04	var	6.00	89.30	1.73
REB	mse_{sp}	1.37	96.30	3.08	mse_{bc}	2.95	97.80	2.46	mse_b	3.16	98.00	2.08
REBnc	mse_{sp}	2.00	96.90	3.10	mse_{sp}	3.47	98.30	2.48	mse_{sp}	3.58	98.40	2.10
REBs	mse_{bc}	3.58	98.40	3.38	mse_{3t}	3.79	98.60	2.58	mse_{bc}	3.58	98.40	2.16
REBsnc	mse_{bc}	4.00	98.80	3.37	mse_{bc}	3.68	98.50	2.58	mse_b	3.68	98.50	2.15
WBM	mse_{sp}	1.05	96.00	2.76	mse_{spa}	2.84	97.70	2.21	mse_{spa}	2.32	97.20	1.88
WBN	var	2.84	97.70	3.33	mse_{sp}	0.00	95.00	2.39	mse_{sp}	0.74	94.30	2.01
WBR	var	0.11	94.90	3.00	var	0.11	95.10	2.19	var	2.21	92.90	1.83

REB bootstrapping. The performance of the latter variation is not considered, as its construction is tailored to provide valid confidence intervals only for parameters β, σ_u^2, and σ_e^2. It remains unclear how one can generalize their proposal to be applicable for the inference of statistics t_j and M.

The variations REBnc and REBsnc, tested in this manuscript, were not proposed by Chambers and Chandra (2013), and as far as is known, their theoretical properties have not been studied. Initially, their numerical performance was tested out of thoroughness and scientific curiosity. These variations were included in this manuscript because, surprisingly, their numerical performance was found to be comparable to variants REB/0 and REB/1 in Chambers and Chandra (2013) (see results in lines 5 and 7 of Tables 1 and 2). This finding warrants additional scrutiny and could be explored in future research.

5.2. Results of simulations

5.2.1. Cluster-wise inference

Table 1 displays the numerical performance of individual intervals for the mixed effects θ_j under scenario (a): $e_{ij}\sim\chi^2_5$, $u_j\sim\chi^2_5$. For the bootstrap-based intervals, for each sample size, only the performance of the interval with the lowest %err_{ind} is reported. For example, SPB and var in the second row refers to the performance of the cluster-wise interval constructed using the semiparametric bootstrap and $\tilde{\sigma}_j^2$ defined in Section 2. The remaining results can be found in the SM. In this case, the performance of all methods apart from REBnc, WBM, WBN and WBR is similar – the coverage error is smaller than 1% which means that the distribution of errors and random effects seem to hardly affect the empirical coverage, even for the intervals based on the asymptotic theory (cf. results in the first line of Table 1). The cluster-wise individual intervals seem to be robust to distributional misspecification. Furthermore, the PB, SPM and REB have similar performance which partially explains their popularity in small area estimation based on linear mixed models. Since the analysis of the performance of cluster-wise intervals under other simulation scenarios leads to the same conclusions, those results are deferred to the SM.

5.2.2. Simultaneous inference

The situation changes dramatically in Table 2 which shows the results for simultaneous intervals for mixed effects θ_j under scenario (a): $e_{ij}\sim\chi^2_5$, $u_j\sim\chi^2_5$. Intervals obtained using Bonferroni correction (A) and parametric bootstrap (PB) suffer from a significant undercoverage. The SPB-based intervals suffer from a minor undercoverage, the application of the REB leads to overcoverage, whereas the intervals constructed by WBN are closest to the nominal coverage.

In Section 3 it was proved that asymptotically, the cumulative distribution function of statistic M^* should approach the cumulative distribution of statistic M. To understand better the results in Table 2, the shapes of the probability density functions of the bootstrap-

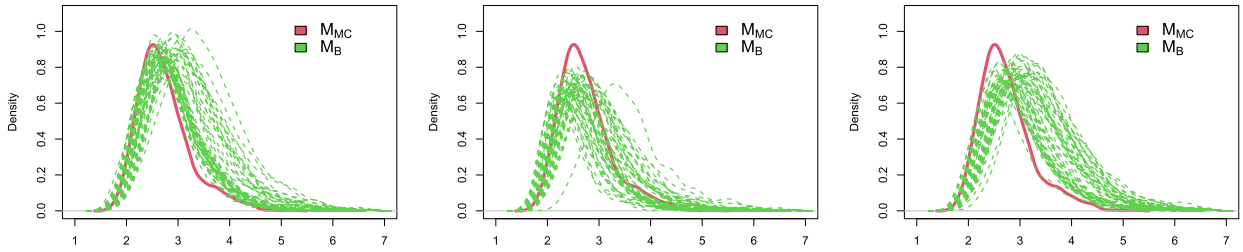

Fig. 2. Monte Carlo approximation (red) of the true density function of statistic M and bootstrap approximations (green): (left) SPB, (middle) REB, (right) WBN. (For interpretation of the colours in the figures, the reader is referred to the web version of this article.)

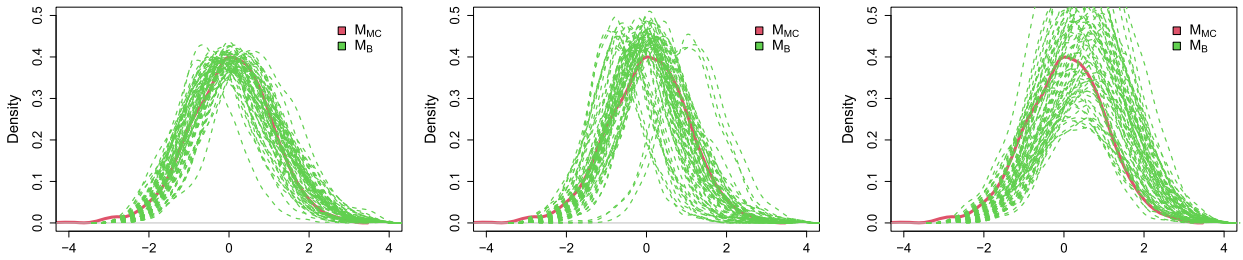

Fig. 3. Monte Carlo approximation (red) of the true density function of statistic t_1 and bootstrap approximations (green): (left) SPB, (middle) REB, (right) WBN. (For interpretation of the colours in the figures, the reader is referred to the web version of this article.)

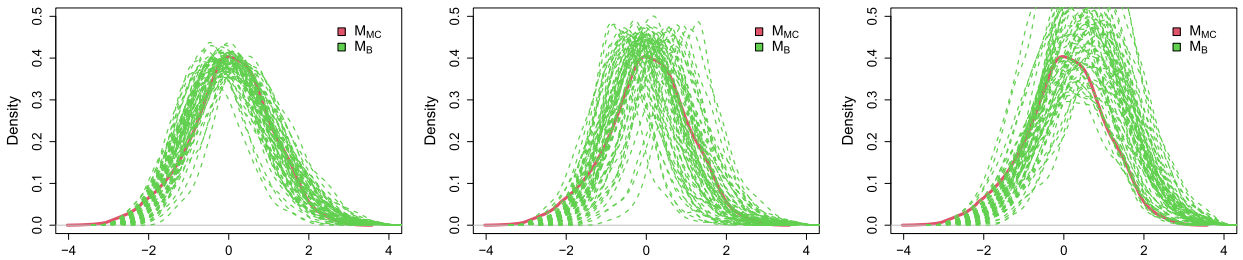

Fig. 4. Monte Carlo approximation of the true density function of statistic t_2 and: (left) the SPB approximations, (middle) the REB approximations, (right) WBN approximations. (For interpretation of the colours in the figures, the reader is referred to the web version of this article.)

based approximations in the simulation scenario with $m = 75, n_j = 15$ are visually assessed (the results obtained from simulation studies with the biggest sample size should be the most representative as one refers to the asymptotic theory to analyse them).

Fig. 2 displays the Monte Carlo approximation of the true cumulative distribution function of statistic M, and 50 out of 1000 bootstrap-based approximations in the simulation scenario with $m = 75, n_j = 15$. The left panel refers to the SPB-based simultaneous intervals with $\hat{\sigma}^2 = mse_{spa}$ (the second row of Table 2). As one can see, some of the bootstrap approximations suffer from some bias which may explain the small undercoverage. Nonetheless, the shape of the distribution is well maintained in each of the simulation runs in contrast to the middle and right panel which represent the REB and WBN approximations respectively. Both of them seem to be skewed to the right with fewer observations from the bulk of the data and more observations with extreme values which explains their overcoverage. This is also an indication that it is unclear if the REB or the WBN can be used to obtain results similar to those in Lemma 2.

The shape of the distributions in Fig. 2, and consequently the empirical coverage of simultaneous intervals, depend crucially on the ability of bootstrap schemes to reproduce the distribution of statistic M as defined in (4). Since the latter is taken as a maximum over studentized statistics t_j for which all three bootstrap schemes, SPB, REB and WBN, perform well, it might come as a surprise that only SPB performs well as a tool for simultaneous inference. To understand this, Figs. 3-4 present the Monte Carlo approximations of the true cumulative distribution function of two statistics t_j, along with 50 out of 1000 bootstrap-based approximations in the same simulation scenario as in Fig. 2. Similarly to Fig. 2, only the SPB-based bootstrap-based approximations capture the shape of the distributions of the true statistics in Figs. 3-4. In contrast, there is a lot of variability between bootstrap-based approximations under the REB; some of them are biased (moved to the right or the left from the truth), and for others, a bigger amount of data is located around the centre of the distribution – the bootstrap-based distributions are "taller" than the truth. High variability between the REB approximations explains the overcoverage of the REB for the simultaneous intervals. As for the WBN, the bootstrap-based approximations are once again quite variable, being sometimes flatter with longer tails, and sometimes taller with shorter tails. This also explains why the WBN performed poorly for individual intervals but well for simultaneous intervals; for some clusters, the quantiles were too low to give good individual coverage, but at least one of them was high enough to boost the maximum value which drives the distribution of statistics M, and affects the shape of the distribution in Fig. 2.

Table 3
The computation time (in seconds) to estimate high quantiles $q_{j,1-\alpha/2}$, $q_{j,\alpha/2}$, $q_{1-\alpha}$.

Method	PB	SPB	WRB	WNB	WMB	REB	REBs	REBnc	REBsnc
$m = 25$	17.187	17.111	17.095	16.961	17.049	18.544	18.517	18.764	18.918
$m = 50$	18.745	18.254	17.916	18.360	18.225	20.106	20.500	20.361	20.658
$m = 75$	21.565	21.535	21.349	21.924	21.654	24.748	24.029	23.917	24.103

Apart from the numerical results presented in this manuscript, there is a lack of theoretical or computational results regarding the robustness of bootstrap-based cluster-wise or simultaneous inference for mixed effects using wild or random effect block bootstrap. Therefore, making general statements about their performance is challenging, and the discussion was confined to commenting on numerical results in this section. Upon reviewing the literature on bootstrapping for linear mixed models, it has been observed that the majority of scholars test the performance of bootstrap methods in terms of empirical coverage of intervals for fixed effects and variance parameters. In this case, similar to the bootstrap estimation of the mean squared error for mixed effects (see discussion in Section 1), bootstrap schemes are required to reproduce the first few moments of the underlying data-generating process. As shown in Table 1 and Figs. 3-4, these quantities appear to be easier to reproduce by bootstrapping than the entire distribution of the statistic M. The validity of bootstrapping to accurately estimate moments can be assessed by analyzing either the bootstrap version of the likelihood or estimating equations, both converging to some sum-stable distribution.

5.2.3. Computational time

Finally, the computational time to construct different intervals depends on (a) the method to estimate σ_j^2, (b) the method to estimate high quantiles $q_{j,1-\alpha/2}$, $q_{j,\alpha/2}$, $q_{1-\alpha}$. For example, when using $\hat{\hat{\sigma}}_j^2$ as an estimator of σ_j^2, the asymptotic method (A) to construct intervals does not require any computational time, as the high quantiles from a standard normal distribution are tabularized. On the other hand, the average computation time (in seconds) to estimate quantiles for bootstrap-based methods is given in Table 3. The computations were performed using a personal laptop with an Intel(R) Core(TM) processor running at 2.40GHz and 32GB of RAM. The computation time for all considered sample sizes and methods does not exceed 25 seconds, making it very affordable.

5.2.4. Summary of results

In summary, SPB-based cluster-wise and simultaneous intervals seem to be a viable option when addressing distributional model misspecifications. Other bootstrap schemes provide satisfactory coverage only for either the cluster-wise or the simultaneous intervals, but not for both. As the findings in other simulation settings lead to the same conclusions, they were deferred to the SM.

5.3. On the comparison with the results of Chambers and Chandra (2013)

It is important to exercise caution when attempting direct comparisons between the results of Chambers and Chandra (2013) and the simulations in this manuscript. Several reasons underpin the decision to avoid such comparisons: (a) this paper explores different simulation scenarios, including varied data-generating processes and departures from model assumptions; (b) the focus is on different parameters of interest; (c) different assumptions and methods are employed to establish Fisher consistency; (d) access to the code used to generate their results is limited; and (e) two out of three bootstrap schemes proposed by Chambers and Chandra (2013) are implemented. Nevertheless, given the poor performance of SPB reported by Chambers and Chandra (2013) in some simulations, the above results might come as a surprise for certain readers, particularly within the small area estimation community, prompting comments on some of them.

Chambers and Chandra (2013) considered the construction of confidence intervals for fixed effects and variances of random effects using, among others, the SPB and the REB. They considered four simulation setups: without departures from normality (scenario A), with scaled and centred random effects and errors following χ^2-distribution with 1 degree of freedom (scenario B), with first-order autocorrelated errors within each cluster (scenario C), with first-order autocorrelated errors in the entire sample of size n (scenario D). Poor results for SPB were reported by the authors only for: (i) the intercept under scenarios A and B for $n_i = 20$ and under scenarios C and D regardless of n_i, (ii) the variance of random effects under scenario B regardless of n_i, (iii) the variance of errors under scenarios B, C, D regardless of n_i.

Regarding case (i), Chambers and Chandra (2013) acknowledge the lack of significant differences between the coverage rates for regression coefficients under different bootstrap schemes. They admit that the poor coverage of SPB for the intercept might stem from "a potential bias problem with our implementation of this method." Therefore, they decided to assess the quality of different bootstrap methods by comparing the coverage of their intervals for the variances of random effects and errors, which brings us to the analysis of the SPB in cases (ii) and (iii). The simulation setups B to D violate the assumptions imposed to prove the consistency of the SPB (i.e., the lack of correlation between errors and the existence of the first $4 + b, b > 0$ moments of the outcome variable, errors, and random effects, as outlined in regularity condition 3 in Appendix A.1). Therefore, the better performance of SPB in the simulation study can likely be attributed to (a) a meticulous implementation of the method proposed by Carpenter et al. (2003), which includes both centering and rescaling, as suggested by the authors; and (b) considering robustness in scenarios with moderate departures from modelling assumptions, such that the consistency of bootstrap-based equivalents of statistics of interest t_j and M can still be ensured.

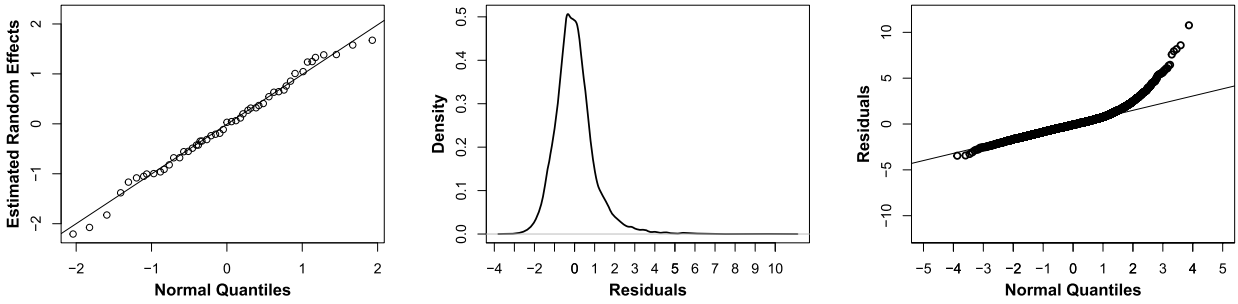

Fig. 5. REML empirical Bayes estimates of random effects: (left) QQ plot; Cholesky REML residuals: (middle) kernel density estimation and (right) QQ plot.

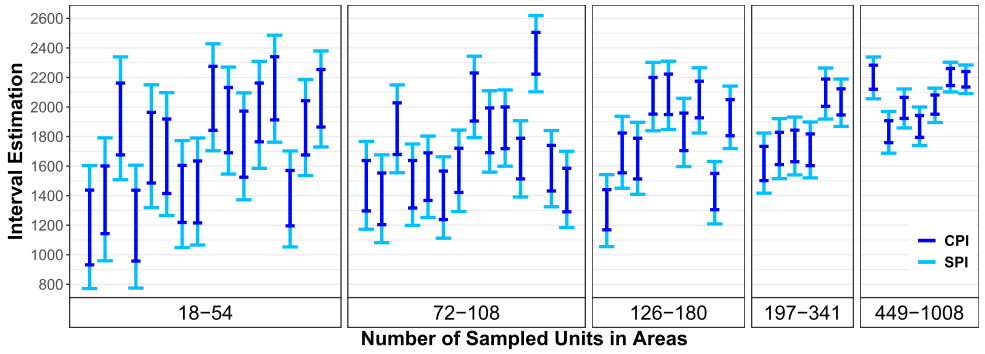

Fig. 6. 95% bootstrap CPI and SPI for the county-level averages of the monthly household income in Galicia.

6. Simultaneous inference for household income in Galicia

The method is applied to construct simultaneous confidence intervals for household income across $m = 52$ counties (*comarcas*) in the Spanish region of Galicia and $n = 9203$ households. Descriptive statistics of households across counties are as follows: Min = 18, $Q_1 = 50$, $Q_2 = 108$, Mean = 177, $Q_3 = 184$, Max = 1008. In some counties, the data were collected from only 18 households, indicating a small area estimation problem. Similarly to Reluga et al. (2023b), household income data from the Structural Survey of Homes of Galicia is considered. In their analysis, Reluga et al. (2023b) compared the performance of direct estimates of household income data and model-based estimates, and constructed cluster-wise and simultaneous intervals using parametric bootstrap. In the current analysis, the same linear mixed model and target parameter are considered: $\theta_j = k_j^T \beta + l_j^T u_j$. The parameter is estimated with EBLUP $\hat{\theta}_j = \hat{\bar{X}}_j^{dir} \hat{\beta} + \hat{u}_j$, where $\hat{\bar{X}}_d^{dir}$ is an estimate of the county-level means of covariates including age, education level, type of household, and variables indicating financial difficulties of the household at the end of a month (see Reluga et al., 2023b, for the list of covariates and model selection criterion).

Income data are well known to be right-skewed. Fig. 5 displays the REML empirical Bayes estimates of random effects together with the density and QQ-plot of Cholesky REML residuals. While the random effects (left panel) appear to follow a normal distribution, the skewness of the residuals (middle and right panel) is apparent. Parametric bootstrap is not appropriate to construct simultaneous intervals in settings with skewed residuals and/or random errors (cf., results in Section 5). The standard trick is to apply a log-transform to the income. However, the challenge lies in obtaining an unbiased back-transformation of the predicted log-income to real income. As demonstrated in Section 5, confidence intervals constructed using the semiparametric bootstrap are robust to this type of model misspecification, allowing us to avoid the log transformation.

Fig. 6 presents bootstrap-based cluster-wise and simultaneous confidence intervals. Clearly, Fig. 6 closely resembles Figure 5 in Reluga et al. (2023b) as the same covariates and model to obtain the EBLUP were used. The estimate of the high quantile $q_{1-\alpha}$ obtained using SPB is 1.95% greater than the one obtained by PB, resulting in a corresponding increase in the widths of the simultaneous intervals across all counties. However, this increase is not detrimental – one can still distinguish many counties with significantly different levels of household income. On the other hand, the change in widths of individual intervals when switching from PB to SPB ranges from -6% to 8.30%, with a mean of 1.07%. In summary, the application of the semiparametric bootstrap, which is robust to distributional model misspecifications, leads to a minor increase in interval widths compared to those of the parametric bootstrap, which is not robust and model-dependent. Fig. 7 displays maps of the lower and upper boundaries of the simultaneous intervals for the county-level averages of household income. Certainly, one can again observe similarities with former results, but also note that some of the poorest counties are now assigned to different categories. In other words, the semiparametric bootstrap, which is robust to distributional assumptions, leads to different conclusions for the most precarious areas.

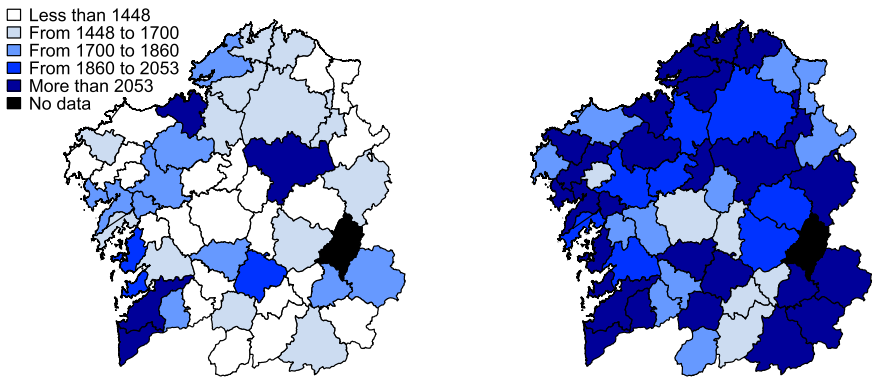

Fig. 7. 95% bootstrap SCI for the county-level averages of the average household income in Galicia: (left) lower boundary, (middle) upper boundary.

7. Discussion

Linear mixed effects are still popular for predicting cluster-level parameters in various domains. Yet, the underlying assumptions that should guarantee satisfactory numerical performance are often violated in practice. It has been shown that the application of a relatively simple bootstrapping scheme can remedy issues arising from distributional model misspecifications without the need to reach for more complex techniques such as robust estimation or data transformation. Furthermore, compared to methods based on parametric bootstrap, which is still the most commonly used technique in small area estimation, proposed bootstrap-based methodology offers at least three advantages: (a) robustness to distributional model misspecification, (b) no additional computational cost, and (c) simplicity of implementation. In addition, the numerical study confirms that mixed effects are fairly robust to distributional model misspecification, including asymmetry and long tails of distributions of errors and random effects, unless they undergo complex, nonlinear transformations such as max transformation. This is particularly important in the context of small area estimation in which mixed effects are often used for nonlinear indicators (Rojas-Perilla et al., 2020). For future research, it would be interesting to study how the bootstrap-based methodology could be extended to applications such as poverty analysis under model misspecification.

Acknowledgements

The authors gratefully acknowledge support from the Swiss National Science Foundation, projects 200021-192345 and P2GEP2-195898, as well as from the Instituto Galego de Estatística who provided us with the data set. In addition, this research has been supported by MICINN grant PID2020-113578RB-I00, and by Xunta de Galicia (Grupos de Referencia Competitiva ED431C 2020/14), GAIN (Galician Innovation Agency) and the Regional Ministry of Economy, Employment and Industry grant COV20/00604 and Centro de investigación del Sistema universitario de Galicia ED431G 2019/01, all of them through ERDF. The computations were performed at the University of Geneva using Baobab and Yggdrasil HPC Service and using the computational facilities of the Advanced Computing Research Centre, University of Bristol.

Appendix A. Regularity conditions and proofs

A.1. Regularity conditions

The regularity conditions from Shao et al. (2000) and Reluga et al. (2023b) are adopted. Let $\vartheta = (\beta, \delta)$, $\hat{\vartheta} = (\hat{\beta}, \hat{\delta})$ and $\vartheta_0 \in \Theta \subset \mathbb{R}^{p+h+1}$ be the true parameter value. The following assumptions are made:

1. Score equation $s_n(\vartheta) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \psi(y_{ij}, \vartheta)$ is well defined if: (a) $s_n(\vartheta)$ is continuous and differentiable for each fixed y, (b) $E\{s_n(\vartheta)\} = 0$ at ϑ_0, (c) ϑ_0 is an interior point of Θ and the estimator $\hat{\vartheta}$ is an interior point of the neighbourhood of ϑ_0.
2. $\liminf_n \lambda[n^{-1} \mathrm{var}\{s_n(\vartheta)\}] > 0$ and $\liminf_n \lambda[-n^{-1} E\{\nabla s_n(\vartheta)\}] > 0$ where $\nabla s_n(\vartheta) = \frac{\partial \psi(\vartheta)}{\partial \vartheta}$ and $\lambda[A]$ indicates the smallest eigenvalue of matrix A.
3. There exists a $b > 0$ such that $E\left(\left\|\psi(y_{ij}, \vartheta)\right\|^{4+b}\right) < \infty$, and $E\left[\{h_N(y_{ij})\}^{2+b}\right]$ in a compact neighbourhood N, where $h_C(y_{ij}) = \sup_{\vartheta \in N} \left\|\nabla s_n(\vartheta)\right\|$.
4. $V_j(\delta)$ has a linear structure in δ, $j = 1, \ldots, m$.

Conditions $1-3$ ensure that one can use the score equation s_n to estimate fixed parameters ϑ up to a vanishing term. Condition 4 implies that the second derivatives of R_j and G_j are 0.

A.2. Proofs

A.2.1. Proof of Lemma 1

Without loss of generality, it is assumed that the sequence of estimators t_j converges to a continuous distribution function F. A standard way of proving the consistency of bootstrap procedures (see, for example, Van der Vaart, 2000, Chapter 23) is to show that, for every a $F_{t_j}(a) \to F(a)$ in distribution and $F_{t_j^*}(a) \to F(a)$ given the original sample size in probability. Let $\hat{\vartheta}^* = (\hat{\beta}^*, \hat{\delta}^*)$ and E^* be a bootstrap operator of the expected value. Then t_j and t_j^* can be written as $t_j = f(\vartheta, \hat{\vartheta}, u_j)$ and $t_j^* = f(\hat{\vartheta}, \hat{\vartheta}*, u_j^*)$, respectively for a continuous and a differentiable function f. Consider the score equation $s_n(\vartheta)$ in A.1 and its bootstrap equivalent $s_n^*(\vartheta) = \sum_i^n \psi(y_i, \vartheta)$ with y replaced by y^*. It follows that $E^*\{s_n^*(\vartheta)\} = 0$ at $\vartheta = \hat{\vartheta}$ which yields the consistency of the sequence of bootstrap estimators $\hat{\vartheta}^*$. The consistency of random effects under random effect bootstrap was proven by Field and Welsh (2007) under $m \to \infty$, $n_j \to \infty$, $\lim_{n_j, n\to\infty} n_j/n = c$, $c > 0$, which is in alignment with results of Jiang (1998). One thus has that $\sqrt{n}(\hat{\theta}_j^* - \theta_j^*)$ and $\sqrt{n}(\hat{\theta}_j - \theta_j)$ converge to the same distribution. Finally, the consistency result follows by Slutsky's lemma.

A.2.2. Proof of Corollary 1

The proof follows along the same lines as in Lemma 23.3 of Van der Vaart (2000). By Lemma 1, the sequences of distribution functions F_{t_j} and $F_{t_j^*}$ converge weakly to F, which implies the convergence of their quantile functions $F_{t_j}^{-1}$ and $F_{t_j^*}^{-1}$ at every continuity point. One thus concludes that $q_{j,1-\alpha}^* = F_{t_j^*}^{-1}(1-\alpha) \to F^{-1}(1-\alpha)$ almost surely, and

$$P\{\theta_j \geq \hat{\theta}_j - \hat{\sigma}_j q_{j,1-\alpha}^*\} = P\left\{\frac{\hat{\theta}_j - \theta_j}{\hat{\sigma}_j} \leq q_{j,1-\alpha}^*\right\} \to P\left\{t_j \leq F^{-1}(1-\alpha)\right\} = 1-\alpha,$$

which completes the proof.

A.2.3. Proof of Lemma 2

Observe that $F_M(a) = P(M < a) = P(t_1 \leq a, \ldots, t_m \leq a, -t_1 \leq a, \ldots, -t_m \leq a)$. Since t_j, $j = 1, \ldots, m$ are asymptotically independent and identically distributed. The following approximation is used: $F_M(a) \approx \prod_{j=1}^{2m} F_j(a)$ where $F_j(a)$ is a proper, non-degenerate distribution. By classical results of extreme value theorem (or Fisher–Tippett–Gnedenko theorem, see Beirlant et al., 2004; Embrechts et al., 2013, for more details), one can assume that there exist sequences of re-normalizing constants $\{b_j > 0\}$, $\{c_j\}$ such that $P\{(M_\theta - c_j)/b_j \leq a\}$ converges to a non-degenerate distribution function $H(a)$ as $j \to \infty$, i.e., $F_j(a)$ belongs to the max-domain of attraction of some non-degenerate, continuous distribution $H(a)$. The consistency of $F_{M^*}(a)$ follows by evoking the properties of the random effect bootstrap and the arguments used in the proof of Lemma 1.

A.2.4. Proof of Corollary 2

The proof follows along the same lines as in Corollary 1 with statistic t_j replaced by M.

Appendix B. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2024.108014.

References

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.L., 2004. Statistics of Extremes: Theory and Applications, vol. 558. John Wiley & Sons.

Bertarelli, G., Chambers, R., Salvati, N., 2021. Outlier robust small domain estimation via bias correction and robust bootstrapping. Stat. Methods Appl. 30, 331–357.

Butar, F.B., Lahiri, P., 2003. On measures of uncertainty of empirical Bayes small-area estimators. J. Stat. Plan. Inference 112, 63–76.

Carpenter, J.R., Goldstein, H., Rasbash, J., 2003. A novel bootstrap procedure for assessing the relationship between class size and achievement. J. R. Stat. Soc., Ser. C 52, 431–443.

Chambers, R., Chandra, H., 2013. A random effect block bootstrap for clustered data. J. Comput. Graph. Stat. 22, 452–470.

Chambers, R., Tzavidis, N., 2006a. M-quantile models for small area estimation. Biometrika 93, 255–268.

Chambers, R.L., Tzavidis, N., 2006b. M-quantile models for small area estimation. Biometrika 93, 255–268.

Chambers, R., Chandra, H., Salvati, N., Tzavidis, N., 2014. Outlier robust small area estimation. J. R. Stat. Soc., Ser. B, Stat. Methodol. 76, 47–69.

Chatterjee, S., Lahiri, P., Li, H., 2008. Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. Ann. Stat. 36, 1221–1245.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application. 1. Cambridge University Press.

Embrechts, P., Klüppelberg, C., Mikosch, T., 2013. Modelling Extremal Events: for Insurance and Finance, vol. 33. Springer Science & Business Media.

Field, C., Pang, Z., Welsh, A.H., 2010. Bootstrapping robust estimates for clustered data. J. Am. Stat. Assoc. 105, 1606–1616.

Field, C.A., Welsh, A.H., 2007. Bootstrapping clustered data. J. Aust. Math. Soc. B 69, 369–390.

Flores-Agreda, D., Cantoni, E., 2019. Bootstrap estimation of uncertainty in prediction for generalized linear mixed models. Comput. Stat. Data Anal. 130, 1–17.

Hall, P., Maiti, T., 2006a. Nonparametric estimation of mean-squared prediction error in nested-error regression models. Ann. Stat. 34, 1733–1750.

Hall, P., Maiti, T., 2006b. On parametric bootstrap methods for small area prediction. J. Aust. Math. Soc. B 68, 221–238.

Henderson, C.R., 1950. Estimation of genetic parameters. Biometrics 6, 186–187.

Jiang, J., 1998. Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. Stat. Sin. 8, 861–885.

Jiang, J., 2007. Linear and Generalized Linear Mixed Models and Their Applications. Springer Series in Statistics.

Kramlinger, P., Krivobokova, T., Sperlich, S., 2022. Marginal and conditional multiple inference in linear mixed models. J. Am. Stat. Assoc. 1, 1–12.

Laird, N.M., Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics, 963–974.
Lee, Y., Nelder, J.A., 1996. Hierarchical generalized linear models. J. Aust. Math. Soc. B 58, 619–656.
Lombardía, M.J., Sperlich, S., 2008. Semiparametric inference in generalized mixed effects models. J. Aust. Math. Soc. B 70, 913–930.
Lombardía, M.J., Sperlich, S., 2012. A new class of semi-mixed effects models and its application in small area estimation. Comput. Stat. Data Anal. 56, 2903–2917.
Mammen, E., 1993. Bootstrap and wild bootstrap for high dimensional linear models. Ann. Stat., 255–285.
McCullagh, P., 2000. Resampling and exchangeable arrays. Bernoulli, 285–301.
Modugno, L., Giannerini, S., 2015. The wild bootstrap for multilevel models. Commun. Stat., Theory Methods 44, 4812–4825.
Morris, J.S., 2002. The BLUPs are not "best" when it comes to bootstrapping. Stat. Probab. Lett. 56, 425–430.
Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G., Breidt, F.J., 2008. Nonparametric small area estimation using penalized spline regression. J. Aust. Math. Soc. B 70, 265–286.
O'Shaughnessy, P., Welsh, A.H., 2018. Bootstrapping longitudinal data with multiple levels of variation. Comput. Stat. Data Anal. 124, 117–131.
Rao, J.N.K., Molina, I., 2015. Small Area Estimation. John Wiley & Sons.
Reluga, K., Lombardía, M.J., Sperlich, S., 2023a. Simultaneous inference for empirical best predictors with a poverty study in small areas. J. Am. Stat. Assoc. 118, 583–595.
Reluga, K., Lombardía, M.J., Sperlich, S., 2023b. Simultaneous inference for linear mixed model parameters with an application to small area estimation. Int. Stat. Rev. 91, 193–217.
Rojas-Perilla, N., Pannier, S., Schmid, T., Tzavidis, N., 2020. Data-driven transformations in small area estimation. J. R. Stat. Soc. A 183, 121–148.
Samanta, M., Welsh, A.H., 2013. Bootstrapping for highly unbalanced clustered data. Comput. Stat. Data Anal. 59, 70–81.
Shao, J., Kübler, J., Pigeot, I., 2000. Consistency of the bootstrap procedure in individual bioequivalence. Biometrika 87, 573–585.
Sinha, S.K., Rao, J., 2009. Robust small area estimation. Can. J. Stat. 37, 381–399.
Sugasawa, S., Kubokawa, T., 2017. Heteroscedastic nested error regression models with variance functions. Stat. Sin. 27, 1101–1123.
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. J. R. Stat. Soc. A 181, 927–979.
Van der Vaart, A.W., 2000. Asymptotic Statistics, vol. 3. Cambridge University Press.
Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer.
White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.