International Statistical Review

International Statistical Review (2022) doi: 10.1111/insr.12519

Simultaneous inference for linear mixed model parameters with an application to small area estimation

Katarzyna Reluga[1], **María-José Lombardía**[2] and **Stefan Sperlich**[3]

[1]*Division of Biostatistics, School of Public Health, University of California, Berkeley, California, USA*
[2]*CITIC, University of A Coruña, A Coruña, Spain*
[3]*Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland*
Correspondence *Katarzyna Reluga, Division of Biostatistics, School of Public Health, University of California, Berkeley, California, USA. Email: katarzyna.reluga@berkeley.edu*

Summary

Over the past decades, linear mixed models have attracted considerable attention in various fields of applied statistics. They are popular whenever clustered, hierarchical or longitudinal data are investigated. Nonetheless, statistical tools for valid simultaneous inference for mixed parameters are rare. This is surprising because one often faces inferential problems beyond the pointwise examination of fixed or mixed parameters. For example, there is an interest in a comparative analysis of cluster-level parameters or subject-specific estimates in studies with repeated measurements. We discuss methods for simultaneous inference assuming a linear mixed model. Specifically, we develop simultaneous prediction intervals as well as multiple testing procedures for mixed parameters. They are useful for joint considerations or comparisons of cluster-level parameters. We employ a consistent bootstrap approximation of the distribution of max-type statistic to construct our tools. The numerical performance of the developed methodology is studied in simulation experiments and illustrated in a data example on household incomes in small areas.

Key words: Max-type statistic; mixed parameter; multiple testing; small area estimation; simultaneous confidence interval.

1 Introduction

The family of linear mixed effects models (LMMs) developed by Henderson (1950) has been extensively applied in the statistical analysis of clustered and longitudinal data (Jiang, 2007; Verbeke & Molenberghs, 2000) as well as for the treatment-level analysis in medicine (Francq *et al.*, 2019). This modelling framework arises naturally in many fields such as environmental sciences, economics, medicine and so on. Under LMM one supposes that the extra between-cluster variation (or between-subject variation in longitudinal studies) is captured by cluster-specific random effects. Cluster-level parameters might be the most relevant part of the statistical analysis. In particular, they can be modelled by random effects themselves, or more frequently, by mixed effects which are often linear combinations of fixed and random effects. Mixed parameters are particularly appealing in, among others, animal husbandry, ecology and small area estimation (SAE) (see, e.g. the monograph of Rao & Molina, 2015, and the review of Tzavidis *et al.* 2018). The latter critically observed that although the resulting mixed parameter estimates 'are a set of numbers of identical definition and simultaneous interest' and that one should thus consider a simultaneous rather than a point estimation problem, the topic of 'ensemble properties of small area estimates (…) has been largely overlooked'. They mention benchmarking and rank estimation as examples in this direction. This is in line with our observation regarding related literature on Bayesian hierarchical models in which the authors examine constrained Bayes and triple-goal estimation (Gosh, 1992; Shen & Louis, 1998). However, most of this literature hardly considers LMM. To the best of our knowledge, beyond such constrained and rank estimation, simultaneous inference for mixed parameters is still missing. This is surprising given the utility of such inference in applied domains, for example, within public health centres carrying out studies on demographic groups, or when statistical offices report to policy makers for resource distribution. The existing pointwise inference or joint estimation of mixed parameters is not less relevant or useful; nevertheless, simultaneous inference would provide a framework for formulating statistically valid statements about a set of mixed parameters.

Numerous national and regional governments as well as international organisations conduct studies on socio-economic conditions in order to implement targeted policy interventions. The European Union, World Bank and statistical institutes regularly draft reports on economic development and poverty across countries, regions and provinces. When looking at such regional estimates, practitioners often aim to simultaneously assess and compare them. Nevertheless, existing methods are often not suitable to carry out such assessments or comparisons on a sound statistical basis. As soon as one begins to formulate a comparative statement about the situation in several areas simultaneously, the area-wise (or cluster-wise) analysis is rendered statistically invalid by an additional variability arising from the joint consideration. Consider cluster-wise prediction intervals (CPI) for mixed parameters; the coverage probabilities of $100(1-\alpha)$ intervals refer to the mean across all clusters. This implies that, by construction, about 100α per cent of the provided intervals (sometimes more) do not contain the true parameter. In other words, each time a statistical institute publishes its estimates for all areas with prediction intervals, the latter fail in at least $100\alpha\%$ to contain the true value. The same holds true for multiple comparisons via testing. The aim of this paper is to develop statistical tools that fill this gap. The investigation of such methods is not only theoretically appealing, but also relevant for practitioners.

We develop simultaneous prediction intervals (SPIs) and multiple testing (MT) procedures to disprove or support simultaneous hypotheses about certain characteristics. More specifically, our main proposal is to use a max-type statistic for a set of mixed parameters. We then employ a bootstrap procedure to consistently approximate the distribution of this statistic. The latter permits us to recover a critical value to construct an operational SPI or conduct MT procedures. Despite the unquestionable utility of such tools in the context of LMM, to the best of our knowledge, we are the first who investigate their theoretical and empirical properties. Furthermore, we compare the performance of our method with alternative simultaneous inference techniques that we adapted from regression and nonparametric curve estimation, namely a Monte Carlo procedure, a Bonferroni adjustment, and Beran's method. We also analytically derive simultaneous intervals based on the volume-of-tube formula of Weyl (1939) to approximate the tail probabilities. In spite of being conservative, the volume-of-tube method works well under some specific examples (Sun & Loader, 1994; Sun *et al.*, 1999). Our mathematical derivations confirm the former statement, but also demonstrate that this method is not operational in our context. We therefore defer its derivation to the supporting information.

Our methods are different from those considered by Sun *et al.* (1999) or Maringwa *et al.* (2008) within the framework of longitudinal studies. They propose to apply, respectively, the volume-of-tube formula and Monte Carlo (MC) sampling to construct simultaneous bands for linear combinations of fixed effects only. In contrast, we investigate a more complex problem of examining mixed effects. Our proposal also differs from the derivation of Krivobokova *et al.* (2010) who employ a mixed model representation for penalised splines to construct uniform bands for one-dimensional regression curves. Contrary to us, the authors can use a simplified version of the volume-of-tube formula. Our results are distinct from those of Ganesh (2009) who constructs simultaneous Bayesian credible intervals for a linear combination of area-level parameters under the model of Fay & Herriot (1979). We consider a more general inferential problem in a broader class of LMM within the frequentist framework. Furthermore, our contribution to the area of MT is a practical methodology used under LMM for the first time. The employment of the max-type statistic might be considered as a complement to the study of Kramlinger *et al.* (2018) who examined chi-square statistics for constructing MT and confidence sets for mixed parameters. In the classical linear regression literature, max-type and chi-square statistics have been considered as complements, and are both well established in the practitioners' toolbox. We believe that this is equally valid for mixed parameters. The former are more popular for SPI, whereas chi-square statistics are widely recognised for MT. Finally, Reluga *et al.* (2021) consider simultaneous inference for empirical best predictors under generalised linear mixed models, whereas our paper seeks to address simultaneous inference under LMM. Our study examines the statistical properties of SPI in contrast to the literature that investigates CPI. Starting from the work of Cox (1975) and Morris (1983), researchers proposed numerous methods based on analytical derivations (e.g. Basu *et al.*, 2003; Kubokawa, 2010; Yoshimori & Lahiri, 2014) and resampling (e.g. Hall & Maiti, 2006; Chatterjee *et al.*, 2008). However, CPI and SPI serve different purposes and are not alternatives to each other.

The remainder of the paper is organised as follows. In Section 2 we introduce the modelling framework and the parameter of interest. The construction of SPI and the MT procedure by making use of a max-type statistic is outlined in Section 3. In Section 4 we introduce bootstrap-based SPI and MT and prove their consistency. Section 5 contains potential alternatives which we adapted to our setting. We investigate the finite sample performance of our method in Section 6, and apply it to study the household income in Galicia in Section 7. Section 8 contains final remarks and conclusions. Technical details are deferred to Appendix A1 and the supporting information. The latter also includes the discussion of extensions of our method.

2 Linear Mixed Model Inference

Consider a classical LMM formulation $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$, where $\boldsymbol{X}, \boldsymbol{Z}$ are known, full column rank matrices for a fixed and a random part, $\boldsymbol{\beta}$ is a vector of fixed effects, $\boldsymbol{u}$ is a vector of random effects, and $\boldsymbol{e}$ denotes stochastic errors. It is common to assume $\boldsymbol{u}$ and $\boldsymbol{e}$ to be mutually independent with $\boldsymbol{u} \overset{ind}{\sim} N_q(\boldsymbol{0}, \boldsymbol{G})$ and $\boldsymbol{e} \overset{ind}{\sim} N_n(\boldsymbol{0}, \boldsymbol{R})$. More specifically, consider a LMM with a block diagonal covariance matrix (LMMb):

$$\boldsymbol{y}_d = \boldsymbol{X}_d\boldsymbol{\beta} + \boldsymbol{Z}_d\boldsymbol{u}_d + \boldsymbol{e}_d, \;\; d = 1, \ldots, D, \tag{1}$$

where n_d is the number of units in the d^{th} cluster (or area), $\boldsymbol{y}_d \in \mathbb{R}^{n_d}$, $\boldsymbol{X}_d \in \mathbb{R}^{n_d \times (p+1)}$ and $\boldsymbol{Z}_d \in \mathbb{R}^{n_d \times q_d}$. Here, D is the number of clusters, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ an unknown vector of regression coefficients, $\boldsymbol{u}_d \overset{ind}{\sim} N_{q_d}(\boldsymbol{0}, \boldsymbol{G}_d)$, $\boldsymbol{e}_d \overset{ind}{\sim} {}_{n_d}(\boldsymbol{0}, \boldsymbol{R}_d)$, and $n = \sum_{d=1}^{D} n_d$. We assume that $\boldsymbol{G}_d = \boldsymbol{G}_d(\boldsymbol{\theta}) \in \mathbb{R}^{q_d \times q_d}$ and $\boldsymbol{R}_d = \boldsymbol{R}_d(\boldsymbol{\theta}) \in \mathbb{R}^{n_d \times n_d}$ depend on variance parameters $\boldsymbol{\theta} =$

$(\theta_1, \ldots, \theta_h)^t$. LMM can be easily retrieved applying the notation of Prasad & Rao (1990). Under this setup, suppose that the variance-covariance $\boldsymbol{V}$ is nonsingular $\forall\theta_i,\ i = 1, \ldots, h$ with $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X\beta}$ and $\mathbb{V}\text{ar}(\boldsymbol{y}) = \boldsymbol{R} + \boldsymbol{ZGZ}^t = \boldsymbol{V}(\boldsymbol{\theta}) = :\boldsymbol{V}$. Two important examples of LMM that are extensively used, especially in SAE, are the *nested error regression model* (NERM) of Battese *et al.* (1988), and the *Fay–Herriot model* (FHM) of Fay & Herriot (1979). The former is defined as

$$y_{dj} = \boldsymbol{x}_{dj}^t\boldsymbol{\beta} + u_d + e_{dj},\ \ d = 1, \ldots, D,\ \ j = 1, \ldots, n_d, \tag{2}$$

where y_{dj} is the quantity of interest for the j^{th} unit in the d^{th} cluster, $\boldsymbol{x}_{dj} = (1, x_{dj1}, \ldots, x_{djp})^t$, $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$ and $e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2)$ for $d = 1, \ldots, D$, $j = 1, \ldots, n_d$. Here $\boldsymbol{y}_d = (y_{d1}, \ldots, y_{dn_d})$, $\boldsymbol{X}_d = \text{col}_{1\leqslant j\leqslant n_d}\boldsymbol{x}_{dj}^t$, $q_d = 1$, $Z_d = \boldsymbol{1}_{n_d}$ with $\boldsymbol{1}_{n_d}$ a n_d vector of ones, $\boldsymbol{e}_d = (e_{d1}, \ldots, e_{dn_d})^t$, $\boldsymbol{\theta} = (\sigma_e^2, \sigma_u^2)^t$, $\boldsymbol{R}_d(\boldsymbol{\theta}) = \sigma_e^2\boldsymbol{I}_{n_d}$ with $\boldsymbol{I}_{n_d}$ the $n_d \times n_d$ identity matrix and $\boldsymbol{G}_d(\boldsymbol{\theta}) = \sigma_u^2$. In contrast, the FHM is often referred to as an area-level model and consists of two levels. The model at level 1, called the sampling model, assumes that the direct estimators y_d of a cluster mean μ_d^F are design unbiased, and satisfy $y_d = \mu_d^F + e_d,\ e_d \overset{iid}{\sim} N(0, \sigma_{e_d}^2),\ d = 1, \ldots, D$. Under FHM, the sampling variance $\sigma_{e_d}^2 = \mathbb{V}ar(y_d|\mu_d^F)$ is supposed to be *known* for each cluster *d*. On the other hand, the linking model at level 2 is $\mu_d^F = \boldsymbol{x}_d^t\boldsymbol{\beta} + u_d,\ u_d \overset{iid}{\sim} N(0, \sigma_u^2),\ d = 1, \ldots, D$, where $\boldsymbol{x}_d = (1, x_{d1}, \ldots, x_{dp})^t$ is a $(p+1)$-vector of cluster-level auxiliary variables. Observe that the FHM can be rewritten as a LMM with $n_d = q_d = 1,\ Z_d = 1,\ \boldsymbol{\theta} = \sigma_u^2,\ \boldsymbol{R}_d(\sigma_u^2) = \sigma_{e_d}^2$, that is

$$y_d = \boldsymbol{x}_d^t\boldsymbol{\beta} + u_d + e_d,\ \ d = 1, \ldots, D. \tag{3}$$

Due to the data availability, the FHM is more frequently used in practice. While cluster-level information can be easily obtained (for example, using open access internet repositories), this is clearly not the case for unit-level information.

Assuming LMM, one is often interested in a simultaneous or comparative inference for general mixed parameters

$$\mu_d = \boldsymbol{k}_d^t\boldsymbol{\beta} + \boldsymbol{m}_d^t\boldsymbol{u}_d,\ \ d = 1, \ldots, D, \tag{4}$$

with $\boldsymbol{k}_d \in \mathbb{R}^{p+1}$ and $\boldsymbol{m}_d \in \mathbb{R}^{q_d}$ known. μ_d is a cluster conditional mean, but other parameters can be explored as well. Henderson (1975) developed the best linear unbiased predictor (BLUP) of a linear combination of random and fixed effects when $\boldsymbol{V}$ is completely known. Applying their idea one obtains the BLUP estimator for (4), that is $\tilde{\mu}_d: = \tilde{\mu}_d(\boldsymbol{\theta}) = \boldsymbol{k}_d^t\tilde{\boldsymbol{\beta}} + \boldsymbol{m}_d^t\tilde{\boldsymbol{u}}_d$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_h)^t$, $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y}$, and $\tilde{\boldsymbol{u}}_d = \tilde{\boldsymbol{u}}_d(\boldsymbol{\theta}) = \boldsymbol{G}_d\boldsymbol{Z}_d^t\boldsymbol{V}_d^{-1}(\boldsymbol{y}_d - \boldsymbol{X}_d\tilde{\boldsymbol{\beta}})$. In practice $\boldsymbol{\theta}$ is usually unknown, hence one uses $\widehat{\boldsymbol{\theta}}: = \widehat{\boldsymbol{\theta}}(\boldsymbol{y})$ which yields the EBLUP

$$\widehat{\mu}_d: = \widehat{\mu}_d(\widehat{\boldsymbol{\theta}}) = \boldsymbol{k}_d^t\widehat{\boldsymbol{\beta}} + \boldsymbol{m}_d^t\widehat{\boldsymbol{u}}_d,\ \ d = 1, \ldots, D, \tag{5}$$

with $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}),\ \widehat{\boldsymbol{u}} = \widehat{\boldsymbol{u}}(\widehat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_h)^t$. Having assumed certain conditions on the distributions of random effects and errors, as well as the variance components $\boldsymbol{\theta}$ (see Appendix A.1), Kackar & Harville (1981) proved that the two-stage procedure provides an unbiased estimator for μ_d.

To construct a studentised max-type statistic, it is important to assess the variability of prediction. The most common measure of uncertainty is the mean squared error $\text{MSE}(\widehat{\mu}_d) =$

$\mathbb{E}(\widehat{\mu}_d - \mu_d)^2$. Here, $\mathbb{E}$ denotes the expectation with respect to model (1). We can decompose the MSE into

$$\mathrm{MSE}(\widehat{\mu}_d) = \mathrm{MSE}(\tilde{\mu}_d)+\mathbb{E}(\widehat{\mu}_d - \tilde{\mu}_d)^2 + 2\mathbb{E}\{(\tilde{\mu}_d - \mu_d)(\widehat{\mu}_d - \tilde{\mu}_d)\}, \tag{6}$$

where $\mathrm{MSE}(\tilde{\mu}_d)$ accounts for the variability when the variance components $\boldsymbol{\theta}$ are known. Assuming LMMb and $\boldsymbol{b}_d^t = \boldsymbol{k}_d^t - \boldsymbol{o}_d^t\boldsymbol{X}_d$ with $\boldsymbol{o}_d^t = \boldsymbol{m}_d^t\boldsymbol{G}\boldsymbol{Z}_d^t\boldsymbol{V}_d^{-1}$, the $\mathrm{MSE}(\tilde{\mu}_d)$ reduces to

$$\boldsymbol{m}_d^t(\boldsymbol{G}_d - \boldsymbol{G}_d\boldsymbol{Z}_d^t\boldsymbol{V}_d^{-1}\boldsymbol{Z}_d\boldsymbol{G}_d)\boldsymbol{m}_d + \boldsymbol{b}_d^t\left(\sum_{d=1}^{D}\boldsymbol{X}_d^t\boldsymbol{V}_d^{-1}\boldsymbol{X}_d\right)^{-1}\boldsymbol{b}_d = : g_{1d}(\boldsymbol{\theta})+g_{2d}(\boldsymbol{\theta}), \tag{7}$$

where g_{1d} accounts for the variability of $\tilde{\mu}_d$ once $\boldsymbol{\beta}$ is known, and g_{2d} for the estimation of $\tilde{\boldsymbol{\beta}}$. The second term in (6) is intractable, but there exists a vast literature which deals with its estimation (see Rao & Molina, 2015, for a review). The third term disappears under normality of errors and random effects; it is rarely considered. Following Chatterjee *et al.* (2008), we suggest to construct SPIs using only $\boldsymbol{g}_1(\widehat{\boldsymbol{\theta}}) = (g_{11}(\widehat{\boldsymbol{\theta}}), \ldots, g_{1D}(\widehat{\boldsymbol{\theta}}))^t$, where $g_{1d}(\widehat{\boldsymbol{\theta}})$ is defined in (7) with $\boldsymbol{\theta}$ replaced by a consistent estimate. In fact, simulations studies in Reluga (2020) indicate that alternative measurements of variability do not improve the performance of SPI based on the max-type statistic.

3 Simultaneous Prediction Interval and Multiple Testing Procedure for Mixed Parameters Using Max-type Statistics

We concentrate on the construction of SPI and MT procedures for the mixed parameter in (4). In particular, we consider a confidence region $\mathcal{I}_{1-\alpha} = \times_{d=1}^{D}\mathcal{I}_{d,\,1-\alpha}$ such that $P(\mu_d \in \mathcal{I}_{1-\alpha}\ \forall d \in [D]) = 1 - \alpha$, $[D] = \{1, \ldots, D\}$. This is equivalent to finding a critical value $c_{S_0}(1-\alpha)$ which satisfies

$$\alpha = P\left(\left|\frac{\widehat{\mu}_d - \mu_d}{\widehat{\sigma}(\widehat{\mu}_d)}\right| \geqslant c_{S_0}(1-\alpha)\ \text{for some}\ d \in [D]\right) = P\left(\max_{d=1,\ldots,D}\left|\frac{\widehat{\mu}_d - \mu_d}{\widehat{\sigma}(\widehat{\mu}_d)}\right| \geqslant c_{S_0}(1-\alpha)\right),$$

where we denote by $\widehat{\sigma}(\widehat{\mu}_d)$ the estimated variability of $\widehat{\mu}_d$ (for example, the estimated square root of $\mathrm{MSE}(\widehat{\mu}_d)$). The critical value $c_{S_0}(1-\alpha)$ is in fact the $(1-\alpha)^{th}$-quantile of the studentised statistic

$$S_0 : = \max_{d=1,\ldots,D}|S_{0d}|,\ \text{where}\ S_{0d} = \frac{\widehat{\mu}_d - \mu_d}{\widehat{\sigma}(\widehat{\mu}_d)},\ \ c_{S_0}(1-\alpha) : = \inf\{t \in \mathbb{R} : P(S_0 \leqslant t) \geqslant 1 - \alpha\}. \tag{8}$$

It follows that with probability $1-\alpha$, a region defined as

$$\mathcal{I}_{1-\alpha}^{S} = \mathop{\times}_{d=1}^{D}\mathcal{I}_{d,\,1-\alpha}^{S},\ \ \text{where}\ \mathcal{I}_{d,\,1-\alpha}^{S} = \{\widehat{\mu}_d \pm c_{S_0}(1-\alpha)\widehat{\sigma}(\widehat{\mu}_d)\},$$

covers all mixed parameters. Because the probability density function (pdf) of S_0 is right skewed, we suggest to consider its upper quantile and construct symmetric $\mathcal{I}_{d,\,1-\alpha^S}$, $d \in [D]$. This approach can be regarded as a variation of the studentised maximum modulus method of Tukey (1953). At this point, we formally define CPI to circumvent all possible doubts concerning its relation to SPI. Let $c_d(1-\alpha) : = \inf\{t \in \mathbb{R} : P(S_{0d} \leqslant t) \geqslant 1 - \alpha\}$. CPI is defined as

$$\mathcal{I}_{d,\,1-\alpha}^{CPI} = \{\widehat{\mu}_d \pm c_d(1-\alpha) \times \widehat{\sigma}(\widehat{\mu}_d)\}\ \forall d \in [D],$$

which covers μ_d with probability $1 - \alpha$. Due to the central limit theorem, the most common choice is $c_d(1 - \alpha) = \Phi^{-1}(1 - \alpha/2)$, that is, a quantile from a normal distribution. Thanks to the correspondence between interval estimation and hypothesis testing, our methodology is applicable for the latter. Consider a following pair of hypotheses:

$$H_0 : \boldsymbol{A\mu} = \boldsymbol{h} \text{ versus } H_1 : \boldsymbol{A\mu} \neq \boldsymbol{h}, \tag{9}$$

where $\boldsymbol{A} \in \mathbb{R}^{D\prime \times D}$ with $D' \leqslant D$ and $\boldsymbol{h} \in \mathbb{R}^{D\prime}$ is a vector of constants. A test based on a max-type statistic t_H rejects H_0 at the α -level if $t_H \geqslant c_{H_0}(1 - \alpha)$ with $c_{H_0}(1 - \alpha) := \inf\{t \in \mathbb{R} : P(S_{H_0} \leqslant t) \geqslant 1 - \alpha\}$,

$$t_H := \max_{d=1,\ldots,D} |t_{H_d}|,\ S_{H_0} := \max_{d=1,\ldots,D} |S_{H_0 d}|,\ t_{H_d} = \frac{\widehat{\mu}^H_d - h_d}{\widehat{\sigma}(\widehat{\mu}^H_d)} \text{ and } S_{H_0 d} = \frac{\widehat{\mu}^H_d - \mu^H_d}{\widehat{\sigma}(\widehat{\mu}^H_d)}, \tag{10}$$

where $\boldsymbol{\mu}^H = (\mu^H_1, \ldots, \mu^H_{D'})^t = \boldsymbol{A\mu} \in \mathbb{R}^{D'}$ and $\widehat{\boldsymbol{\mu}}^H$ its estimated counterpart. In other words, $\boldsymbol{h} \notin \mathcal{I}^{H_0}_{1-\alpha}$ with $\mathcal{I}^{H_0}_{1-\alpha} = \times^D_{d=1} \mathcal{I}^{H_0}_{d,1-\alpha}$, where $\mathcal{I}^{H_0}_{d,1-\alpha} = \left\{\widehat{\mu}^H_d \pm c_{H_0}(1 - \alpha)\widehat{\sigma}(\widehat{\mu}^H_d)\right]$. In practice, a standard problem is to test for statistical differences between various clusters with respect to some characteristic. Our test is based on a single step procedure and exhibits a weak control of a family-wise error (FWER). If one aims at testing multiple hypotheses with a strong control of FWER, the step-down technique of Romano & Wolf (2005) could be implemented. As this is beyond the scope of this paper, details and related simulation results are deferred to the supporting information.

Remark *In this article, we consider the construction of SPIs for parameters which can be written as a linear combination of fixed regression parameters* $\boldsymbol{\beta}$ *and random effects* $\boldsymbol{u}_d$*, see Equation 4. Nevertheless, the methodology based on the max-type statistic can be applied to more general parameters such as non-linear mixed parameters estimated by the best predictors (Reluga* et al.*, 2021) or even head count ratios and Gini coefficients. The estimation of the latter necessitates the transformation of data, therefore it might be necessary to apply further adjustments in the estimation procedure (see, e.g. Rojas-Perilla* et al.*, 2020).*

4 Bootstrap-based Simultaneous Prediction Interval and Multiple Testing Procedure

It is challenging to estimate the distribution of S_0 in (8) and to recover critical values because, among others, mixed effects μ_d are unknown, $d = 1, \ldots, D$. Nevertheless, an almost straightforward way to approximate critical value $c_{S_0}(1 - \alpha)$ is to use a parametric bootstrap procedure which circumvents a direct application of the normal asymptotic distribution (González-Manteiga *et al.*, 2008). It can also provide faster convergence (Hall & Maiti, 2006; Chatterjee *et al.*, 2008). Let B be the number of bootstrap samples $(\boldsymbol{y}^{*(b)}, \boldsymbol{X}, \boldsymbol{Z})$. The bootstrap analogue of expression in (8) is

$$S^{*(b)}_B := \max_{d=1,\ldots,D} \left|S^{*(b)}_{Bd}\right|,\ S^{*(b)}_{Bd} = \frac{\widehat{\mu}^{*(b)}_d - \mu^{*(b)}_d}{\widehat{\sigma}^*(\widehat{\mu}^{*(b)}_d)},\ b = 1, \ldots, B. \tag{11}$$

The critical value can be consistently approximated by the $(1 - \alpha)^{th}$-quantile of (11), that is, $c_{BS}(1 - \alpha) := \inf\{t^* \in \mathbb{R} : P(S^*_B \leqslant t^*) \geqslant 1 - \alpha\}$. Consequently, the bootstrap SPI is defined as

$$\mathcal{I}^{BS}_{1-\alpha} = \mathop{\times}_{d=1}^{D} \mathcal{I}^{BS}_{d,1-\alpha},\ \text{ where } \mathcal{I}^{BS}_{d,1-\alpha} = \{\widehat{\mu}_d \pm c_{BS}(1 - \alpha)\widehat{\sigma}(\widehat{\mu}_d)\}. \tag{12}$$

Our choice of $\widehat{\sigma}(\widehat{\mu}_d) = \sqrt{g_{1d}(\widehat{\boldsymbol{\theta}})}$ is motivated by the asymptotic analysis of Chatterjee *et al.* (2008). Validity of the above bootstrap method is shown by adapting Theorem 3.1 of these authors (henceforth Theorem CLL, provided in the supporting information) and combining it with some results from the extreme value theory.

Proposition 1 Suppose that the assumptions of Theorem CLL and regularity conditions R.1-R.7 from Section A.1 hold. Then

$$\begin{gathered}\sup\\ q \in \mathbb{R}\left|P^*(S^*_B \leqslant q) - P(S_0 \leqslant q)\right| = o_P(1).\end{gathered}$$

An important implication of Proposition 1 is the coverage probability of $\mathcal{I}_{1-\alpha}{}^{BS}$.

Corollary 1. *Under Proposition 1 it holds that*

$$P\left(\mu_d \in \mathcal{I}_{1-\alpha}{}^{BS}\ \forall d \in [D]\right) \overset{D\to\infty}{\to} 1 - \alpha.$$

According to the theoretical developments for max-type statistics, the Kolmogorov distance defined in Proposition 1 converges to 0 at best at polynomial rate $(\log(\,\cdot\,))^{c_2}/n^{c_3}$, where $(\,\cdot\,)$ is the number of parameters for which we wish to obtain the maximum (in our case D), and c_2, c_3 some constants, see, for example, Chernozhukov *et al.* (2013). We are not aware of results for max-type statistics that one could employ to obtain second order correctness without such $\log(\,\cdot\,)$ term. Observe that in our setting $n \to \infty$ is equivalent to $D \to \infty$, because we assumed that n_d is bounded (see Appendix A.1). Certainly, our result would still hold if both, n_d and D grow. In case of a fixed D, we would replace Proposition 1 using explicitly a variation of studentised maximum modulus distribution (cf. Stoline & Ury, 1979).

Due to the relation between interval estimation and tests, the methodology developed for SPI can be used to find a critical value for MT procedure in (9). In particular, consider slightly modified bootstrap statistics

$$S^{*(b)}_{BH_0} := \max_{d=1,\ldots,D}\left|S^{*(b)}_{BH_0 d}\right|,\quad S^{*(b)}_{BH_0 d} = \frac{\widehat{\mu}_d^{*H(b)} - \mu_d^{*H(b)}}{\widehat{\sigma}^*\left(\widehat{\mu}_d^{*H(b)}\right)},$$

with $\boldsymbol{\mu}^{*H(b)} = \left(\mu_1^{*H(b)}, \ldots, \mu_{D'}^{*H(b)}\right)^t := \boldsymbol{A}\boldsymbol{\mu}^{*(b)} \in \mathbb{R}^{D'}$, and its estimated versions

$$\widehat{\boldsymbol{\mu}}^{*H(b)} = \left(\boldsymbol{a}_1^t\left(\boldsymbol{k}_1^t\widehat{\boldsymbol{\beta}}^{*(b)} + \boldsymbol{m}_1^t\widehat{\boldsymbol{u}}_1^{*(b)}\right), \ldots, \boldsymbol{a}_D^t\left(\boldsymbol{k}_D^t\widehat{\boldsymbol{\beta}}^{*(b)} + \boldsymbol{m}_D^t\widehat{\boldsymbol{u}}_D^{*(b)}\right)\right)^t := \boldsymbol{A}\widehat{\boldsymbol{\mu}}^{*(b)},$$

where $\boldsymbol{a}_d \in \mathbb{R}^D$ are the rows of $\boldsymbol{A}$. These are applied to find a bootstrap approximation for the critical value $c_{H_0}(1 - \alpha)$ of our test, namely $c_{BH_0}(1 - \alpha) := \inf\{t \in \mathbb{R} : P(S^*_{BH_0} \leqslant t) \geqslant 1 - \alpha\}$. It is worth mentioning that we do not need to generate bootstrap samples under H_0 to obtain the critical values of our test.

In Section 2, we defined NERM and FHM as popular examples of LMM. We describe a parametric bootstrap procedure that yields promising results when constructing SPI under these models. Under NERM and FHM, we use simplified versions of g_{1d} in (7) derived by Prasad & Rao (1990) as the estimators of $\widehat{\sigma}^2(\widehat{\mu}_d)$: $g^N_{1d}(\widehat{\boldsymbol{\theta}}) = \widehat{\sigma}^2_u/(\widehat{\sigma}^2_u + \widehat{\sigma}^2_e/n_d)(\widehat{\sigma}^2_e/n_d)$ for NERM and $g^F_{1d}(\widehat{\boldsymbol{\theta}}) = \widehat{\sigma}^2_u\sigma^2_{e_d}/(\widehat{\sigma}^2_u + \sigma^2_{e_d})$ for FHM. Under NERM the bootstrap algorithm is

1 From the original sample, obtain consistent estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}} = (\widehat{\sigma}_e^2, \widehat{\sigma}_u^2)$.
2 Generate D independent copies of $W_1 \sim N(0, 1)$. Construct $\boldsymbol{u}^* = (u_1^*, u_2^*, \dots, u_D^*)$ with $u_d^* = \widehat{\sigma}_u W_1,\ d = [D]$.
3 Generate n independent copies of $W_2 \sim N(0, 1)$. Construct $\boldsymbol{e}^* = (e_1^*, e_2^*, \dots, e_n^*)$ with $e_j^* = \widehat{\sigma}_e W_2,\ j = [n]$.
4 Create a bootstrap sample $\boldsymbol{y}^* = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{u}^* + \boldsymbol{e}^*$.
5 Fit the model to the bootstrap sample and obtain bootstrap estimates $\widehat{\boldsymbol{\beta}}^*$, $\widehat{\boldsymbol{\theta}}^* = (\widehat{\sigma}_e^{2*}, \widehat{\sigma}_u^{2*})$, μ_{dj}^* and $\widehat{\mu}_{dj}^*$.
6 Repeat Steps 2–5 B times. Calculate $S_B^{*(b)}$, $b = 1, \dots, B$, using $g_{1d}^{N*(b)}(\widehat{\boldsymbol{\theta}}^{*(b)})$ to obtain $c_{BS}(1 - \alpha)$ and $\mathcal{I}_{1\,-\,\alpha}{}^{BS}$.

Here, $g_{1d}^{N*(b)}(\widehat{\boldsymbol{\theta}}^{*(b)}) = \widehat{\sigma}_u^{2*(b)}/(\widehat{\sigma}_u^{2*(b)} + \widehat{\sigma}_e^{2*(b)}/n_d)(\widehat{\sigma}_e^{2*(b)}/n_d)$ is the bootstrap equivalent of $g_{1d}^N(\widehat{\boldsymbol{\theta}})$. To implement the analogous bootstrap under FHM, we need to modify step 1 and define $\widehat{\boldsymbol{\theta}} = \widehat{\sigma}_u^2$ as well as $g_{1d}^{F*(b)}(\widehat{\boldsymbol{\theta}}^{*(b)}) = \widehat{\sigma}_u^{2*(b)}\sigma_{e_d}^2/(\widehat{\sigma}_u^{2*(b)} + \sigma_{e_d}^2)$. Additionally, we need to replace step 3 by:

3.' Generate D independent copies of a variable $W_2 \sim N(0, 1)$. Construct vector $\boldsymbol{e}^* = (e_1^*, e_2^*, \dots, e_D^*)$ with elements $e_d^* = \sigma_{e_d} W_2,\ d = [D]$.

The parametric bootstrap algorithm can be modified to accommodate more complex models, for example, with spatial or temporal correlation, by adapting accordingly the process of generating errors and random effects. An indisputable advantage of the bootstrap approach is its generality. As soon as we can mimic a data generating process for the assumed model, it can be implemented and applied to construct SPI and carry out MT for any kind of estimator. In addition, bootstrap SPI are relatively robust to model misspecifications (see results in Table 3), in particular when the number of units in each cluster grows. This is in alignment with related remarks of Jiang (1998). On the other hand, bootstrap is computationally more expensive than an analytical derivation.

We conclude this section with a practical extension of our results. In the testing problem (9) we have already allowed for a scenario where only $D' < D$ hypotheses were considered, even though all data were used to estimate fixed parameters and predict random effects. Similarly, one might be interested in the construction of SPI with a joint coverage probability for a subset of $D' < D$ clusters. Without loss of generality, we assume that our goal is to construct SPI for the first D' cluster-level mixed parameters. Then, in the definition of S_0 in (8) and S_B^* (11) one replaces $\max_{d=1,\dots,D}$ by $\max_{d=1,\dots,D'}$ and proceeds along the same lines as for D areas. If $D' = O(D)$, we can evoke the same results from the extreme value theory as in case of Proposition 1 to prove a result similar to Corollary 1, that is

Corollary 2. *Let $D' < D,\ d \in [D'],\ D' = O(D)$. Consider*

$$S_{B'}^{*(b)} := \max_{d=1,\dots,D'} \left|S_{Bd}^{*(b)}\right|, \quad c_{B'S}(1 - \alpha) := \inf\{t^* \in \mathbb{R} : P(S_{B'}^* \leqslant t^*) \geqslant 1 - \alpha\},$$

$$\mathcal{I}_{1-\alpha}^{B'S} = \mathop{\times}_{d=1}^{D'} \mathcal{I}_{d,1-\alpha}^{B'S}, \quad \mathcal{I}_{d,1-\alpha}^{B'S} = \{\widehat{\mu}_d \pm c_{B'S}(1 - \alpha)\widehat{\sigma}(\widehat{\mu}_d)\},$$

where $S_{Bd}^{*(b)}$ as defined in (11). Then, under Proposition 1, it holds that

$$P\big(\mu_d \in \mathcal{I}_{1-\alpha}{}^{B}\ \forall d \in [D']\big) \xrightarrow{D' \to \infty} 1 - \alpha.$$

5 Alternative Methods for Simultaneous Prediction Interval and Multiple Testing Procedure

Although, to the best of our knowledge, we are the first who introduce SPI and MT procedures for mixed parameters, various approaches have been put forward to tackle the problem of simultaneous confidence bands for linear regression surfaces. For example, Bonferroni t-statistics are a straightforward tool to compare a set of fixed parameters. Furthermore, other authors, such as Working & Hotelling (1929), Scheffé (1953), Sun & Loader (1994) or Beran (1988), to mention a few, developed equally important methodologies for the simultaneous inference of fixed parameters. In the rest of this section, we adapt some of the methods from the linear regression and nonparametric curve estimation to our setting. One could thus treat them as alternative approaches to our proposal.

5.1 Monte Carlo Procedure

Consider LMMb defined in (1). One way to obtain BLUP estimates for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ is to solve the mixed model equations of Henderson (1950):

$$\begin{bmatrix} \boldsymbol{X}^t\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^t\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^t\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^t\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^t\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^t\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}, \tag{13}$$

which can be re-expressed in the following simplified form:

$$\boldsymbol{K}\tilde{\boldsymbol{\phi}} = \boldsymbol{C}^t\boldsymbol{R}^{-1}\boldsymbol{y}, \text{ where } \boldsymbol{K} = \boldsymbol{C}^t\boldsymbol{R}^{-1}\boldsymbol{C}+\boldsymbol{G}^{+}, \;\; \boldsymbol{G}^{+} = \begin{bmatrix} \boldsymbol{0}_{(p+1)\times(p+1)} & \boldsymbol{0}_{(p+1)\times D} \\ \boldsymbol{0}_{D\times(p+1)} & \boldsymbol{G}^{-1}_{D\times D} \end{bmatrix}, \tag{14}$$

with $\tilde{\boldsymbol{\phi}} = \left(\tilde{\boldsymbol{\beta}}^t, \tilde{\boldsymbol{u}}^t\right)^t$, $\boldsymbol{C} = [\boldsymbol{X}\,\boldsymbol{Z}]$. For some $\boldsymbol{x} = \left(1, x_1, \ldots, x_p\right)^t$ with $x_1, \ldots, x_p \in \mathcal{X} \subset \mathbb{R}^p$, $\boldsymbol{z} = \left(z_1, \ldots, z_q\right)^t \in \mathcal{Z} \subset \mathbb{R}^q$ and $\boldsymbol{c} = \left(\boldsymbol{x}^t, \boldsymbol{z}^t\right)^t \in \mathcal{X} \times \mathcal{Z} = :\mathcal{C}$ one has $\boldsymbol{x}^t\tilde{\boldsymbol{\beta}} + \boldsymbol{z}^t\tilde{\boldsymbol{u}} = \boldsymbol{c}^t\tilde{\boldsymbol{\phi}}$. Having reformulated the LMMb, and assuming normality for errors and random effects one obtains

$$Z = \frac{\boldsymbol{c}^t\left(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}\right)}{\sqrt{\mathbb{V}\text{ar}\left\{\boldsymbol{c}^t\left(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}\right)\right\}}} = \frac{\boldsymbol{c}^t\left(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}\right)}{\sqrt{\boldsymbol{c}^t\left(\boldsymbol{C}^t\boldsymbol{R}^{-1}\boldsymbol{C}+\boldsymbol{G}^{+}\right)^{-1}\boldsymbol{c}}} \sim \text{N}(0,1). \tag{15}$$

The asymptotic result in (15) is a building block of an analytical derivation based on the volume-of-tube formula of Weyl (1939) to construct simultaneous inference tools for nonparametric curves (see, e.g. Sun *et al.*, 1999; Krivobokova *et al.*, 2010). When using LMM for spline regression, Ruppert *et al.* (2003) proposed a simple numerical approach to construct confidence bands of one-dimensional nonparametric curves by the empirical approximation of (15), that is

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\boldsymbol{u}} - \boldsymbol{u} \end{bmatrix} \approx N\left\{\boldsymbol{0}, \left(\boldsymbol{C}^t\widehat{\boldsymbol{R}}^{-1}\boldsymbol{C}+\widehat{\boldsymbol{G}}^{+}\right)^{-1}\right\}. \tag{16}$$

Because of the unknown variance parameter $\boldsymbol{\theta}$, the result in (16) holds only approximately. We apply expression (16) to simulate the distribution of S_0 in (8), and set

$$S_0 = \max_{d=1,\,\ldots,\,D} |S_{0d}| \approx \max_{d=1,\,\ldots,\,D} \frac{\left| \overline{\boldsymbol{c}}_d^{\,t} \begin{bmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\boldsymbol{u}}_d - \boldsymbol{u}_d \end{bmatrix} \right|}{\widehat{\sigma}(\widehat{\mu}_d)} =: \max_{d=1,\,\ldots,\,D} |S_{MCd}| = S_{MC},$$

where $\overline{\boldsymbol{c}}^d = (\boldsymbol{k}_d^t, \boldsymbol{m}_d^t)^t$. Afterwards, we draw K realisations from normal distribution in (16), estimate the critical value $c_{S_0}(1 - \alpha)$ by the $([(1 - \alpha)K]+1)^{th}$ order statistic of S_{MC} and construct MC SPI as follows

$$\mathcal{I}^{MC}_{1-\alpha} = \bigtimes_{d=1}^{\mathrm{D}} \mathcal{I}^{MC}_{d,\,1-\alpha} \text{ where } \mathcal{I}^{MC}_{d,\,1-\alpha} = \{\widehat{\mu}_d \pm c_{MC}(1 - \alpha)\widehat{\sigma}(\widehat{\mu}_d)\}. \tag{17}$$

We can similarly obtain a critical value for MT. The consistency of ${\mathcal{I}_{1-\alpha}}^{MC}$ follows from Equation 15 which is a standard result for mixed models. The same results from the extreme value theory as in the proof of Proposition 1 might be invoked to prove the consistency for the maxima. Monte Carlo SPI are easy to implement and less computer intensive than bootstrap. Yet, they are less robust to departures from the normality of errors and random effects (cf. Section 6).

5.2 Bonferroni Procedure

Classical simultaneous inference has been considered via Bonferroni correction. If all statistics $(\tilde{\mu}_d - \mu_d)/\sigma(\tilde{\mu}_d)$, $d = 1, \ldots, D$, were independent Gaussian pivots, the critical value to construct SPI or MT could be selected as $c_{BO}(1 - \alpha) = \Phi^{-1}(1 - \alpha/2D)$. One may use quantiles from the normal instead of the t-distribution, because the number of mixed parameters is allowed to grow to infinity such that the latter distribution converges to the former, cf. the high-dimensional regression setting in Chernozhukov *et al.* (2013). Having retrieved the value of interest, a Bonferroni SPI is defined as

$$\mathcal{I}^{BO}_{1-\alpha} = \bigtimes_{d=1}^{D} \mathcal{I}^{BO}_{d,\,1-\alpha}, \text{ where } \mathcal{I}^{BO}_{d,\,1-\alpha} = \{\widehat{\mu}_d \pm c_{BO}(1 - \alpha)\widehat{\sigma}(\widehat{\mu}_d)\}. \tag{18}$$

While the same critical value might be used in MT procedure (9), it provides a weak control of FWER. Using Bonferroni's methodology, we do not try to approximate the distribution of statistic S_0 in Equation 8. In this context, it is recommended to use a more accurate estimate of variability. Hence, we suggest setting $\widehat{\sigma}(\widehat{\mu}_d) = \sqrt{\mathrm{mse}(\widehat{\mu}_d)}$ which is an estimated version of MSE defined in (6). An application of this procedure is simple and does not require much computational effort; it is going to be our benchmark under asymptotic independence of parameters. However, the results of Romano & Wolf (2005) confirm that the method of Bonferroni performs poorly for correlated random variables, a problem that is even aggravated when allowing for spatiotemporal and/or temporal dependencies, see our discussion in Section 8. Similarly to bootstrap SPI, Bonferroni bands are fairly robust to the distributional departures from normality of errors and random effect if the number of units in each cluster is large.

5.3 Volume-of-Tube Procedure

The Monte Carlo procedure in Section 5.1 was introduced to deal with the shortcomings of the analytical derivation based on the volume-of-tube formula. In our supporting information, we derive a critical value to construct the volume-of-tube SPI $\mathcal{I}_{1-\alpha}{}^{VT}$. As expected, the approximation is conservative, that is, the coverage probability of $\mathcal{I}_{1-\alpha}{}^{VT}$ is higher than the nominal level $1-\alpha$. On the top of that, the formula to estimate the critical value contains several constants, and it is not clear how one should proceed to estimate them within our modelling context. Some ideas were derived for simpler one-dimensional models. Sun *et al.* (1999) proposed so-called a derivative and a perturbation methods, while Sun & Loader (1994) suggested nonparametric estimation. It is unclear, though, how to extend their implementations to the LMM setting. Bootstrap approximation can be regarded as an alternative. However, in this case, it would be easier to use bootstrap directly as described in Section 4. Finally, the application of the volume-of-tube formula results in two sources of errors; from the approximation itself and from the estimation of the constants, making the approximation less reliable. Owing to the above-mentioned shortcomings of the volume-of-tube SPIs and practical limitations in their implementation, we deferred their derivation to the supporting information.

5.4 Beran Procedure

Beran (1988) developed a procedure to obtain balanced simultaneous intervals with an overall coverage probability $1-\alpha$ within the context of models without random effects. His technique is based on so-called roots and bootstrapping to approximate their cumulative distribution functions (cdfs). Beran's method is as computer intensive as bootstrap SPI, but in comparison with the former it might provide a poorer coverage rate as its convergence in sup-norm is not guaranteed, cf. simulation results in our supporting information. Last but not least, it is not necessarily robust to the distributional departures from normality of errors and random effects. For the sake of comparison, we followed Beran's methodology and implemented it under LMM. Nevertheless, due to the inferior performance, further discussion together with the numerical results are deferred to the supporting information.

6 Simulation Experiments

We carry out simulations to examine finite sample properties of bootstrap (BS), Monte Carlo (MC), Bonferroni (BO) and Beran (BE) SPIs as well as to evaluate the empirical power of MT procedures under various scenarios. Due to their unsatisfactory performance, the results for Beran's SPI are deferred to the supporting information. We analyse all methods under NERM and FHM. As far as the former is concerned, we set $x_{dj1}=1,\ x_{dj2}\sim U(0,1)\ \forall\ d\in[D]$ and $j\in[n_d]$, whereas under the FHM we set $x_{d1}=1,\ x_{d2}\sim U(0,1)\ \forall d\in[D]$ with $\boldsymbol{\beta}=(1,1)^t$ in both models. The number of simulation runs is $I=2500$, each with $B=1000$ bootstrap samples. The covariates are fixed in all simulation runs. We consider small to medium numbers of clusters with $D\in\{15,30,\ 60,90\}$.

When NERM is considered, we first set $n_d=5\ \ \forall d\in[D],\ e_{dj}\sim N(0,\ \sigma_e^2),\ u_d\sim N(0,\ \sigma_u^2)$ such that the intraclass correlation coefficient $\text{ICC}=\sigma_u^2/(\sigma_u^2+\sigma_e^2)$ equals 1/3, 1/2 or 2/3 (see the first column of Tables 1,2). Then we relax the modelling assumptions by allowing e_{dj} and u_d to deviate from normality to become heavy-tailed or asymmetric. Namely, we draw them from centred chi-square distribution with 5 degrees of freedom, student-t distribution with 6 degrees of freedom and skewed student-t distribution with 5 degrees of freedom and the skewness parameter equal to 1.25. We always rescale them to variances σ_e^2 and σ_u^2 as indicated in parentheses in Tables 1-3. Furthermore, we allow the number of units to grow with the number of clusters, cf. Jiang (1998). Our choice of the skewed t-distribution is motivated by the data example in

Table 1. *ECP (in %), WS and VS under NERM with normal errors and random effects*

		ECP (in %)			*WS (VS)*		
	$D{:}n_d$	*BS*	*MC*	*BO*	*BS*	*MC*	*BO*
	15:5	95.4	92.9	93.8	1.876 (0.031)	1.754 (0.022)	1.794 (0.024)
ICC= $2/3$	30:5	95.2	93.9	94.4	1.947 (0.015)	1.890 (0.013)	1.910 (0.013)
$(\sigma_e^2, \sigma_u^2) = (0.5, 1)$	60:5	94.9	93.7	94.2	2.041 (0.008)	2.011 (0.007)	2.023 (0.007)
	90:5	95.2	94.4	94.9	2.101 (0.006)	2.079 (0.005)	2.088 (0.005)
	15:5	96.7	91.2	94.4	2.695 (0.113)	2.358 (0.046)	2.488 (0.049)
ICC= $1/2$	30:5	95.5	92.8	94.4	2.671 (0.027)	2.552 (0.024)	2.608 (0.024)
$(\sigma_e^2, \sigma_u^2) = (1, 1)$	60:5	95.0	93.7	94.5	2.774 (0.014)	2.719 (0.012)	2.750 (0.012)
	90:5	95.2	94.2	94.8	2.850 (0.010)	2.811 (0.009)	2.833 (0.009)
	15:5	98.3	87.3	96.5	2.816 (0.205)	2.156 (0.065)	2.488 (0.087)
ICC= $1/3$	30:5	97.3	90.6	94.8	2.641 (0.050)	2.346 (0.032)	2.485 (0.022)
$(\sigma_e^2, \sigma_u^2) = (1, 0.5)$	60:5	95.3	92.7	94.5	2.616 (0.012)	2.513 (0.015)	2.577 (0.012)
	90:5	95.0	93.0	94.6	2.663 (0.010)	2.597 (0.010)	2.643 (0.009)

Note: The nominal coverage probability is 95 %.

Table 2. *ECP (in %), WS and VS for a subset of D' areas under the NERM with normal errors and random effects*

			ECP (in %)			*WS (VS)*		
	$D{:}n_d$	D'	*BS*	*MC*	*BO*	*BS*	*MC*	*BO*
	15:5	3	95.3	94.9	95.1	2.006 (0.029)	2.006 (0.029)	2.033 (0.033)
ICC $= 1/2$	30:5	6	95.9	95.6	95.7	2.175 (0.017)	2.175 (0.017)	2.187 (0.017)
$(\sigma_e^2, \sigma_u^2) = (1, 1)$	60:5	12	96.5	96.1	96.1	2.350 (0.009)	2.350 (0.009)	2.358 (0.009)
	90:5	18	95.0	94.6	94.6	2.455 (0.007)	2.455 (0.007)	2.457 (0.007)

Note: The nominal coverage probability is 95%.

Table 3. *ECP (in %), WS and VS under NERM with chi-square, t-distributed and skewed t-distributed departures from normality, centred and rescaled to variances given in parentheses*

		ECP (in %)			*WS (VS)*		
	$D{:}n_d$	*BS*	*MC*	*BO*	*BS*	*MC*	*BO*
	15:5	92.8	88.0	92.4	2.322 (0.086)	2.322 (0.086)	2.476 (0.103)
ICC= $1/3$	30:10	91.4	90.2	91.2	1.878 (0.012)	1.876 (0.012)	1.899 (0.012)
$e_{dj} \sim \chi_5(1)$, $u_d \sim \chi_5(0.5)$	60:20	91.3	90.5	91.1	1.455 (0.002)	1.455 (0.002)	1.460 (0.002)
	90:30	92.3	92.5	92.8	1.238 (0.001)	1.238 (0.001)	1.241 (0.001)
	15:5	93.1	89.9	91.6	1.742 (0.046)	1.742 (0.046)	1.783 (0.050)
ICC= $2/3$	30:10	90.8	90.4	90.4	1.370 (0.007)	1.370 (0.007)	1.378 (0.007)
$e_{dj} \sim \chi_5(0.5)$, $u_d \sim N(1)$	60:20	91.5	90.9	91.3	1.043 (0.001)	1.043 (0.001)	1.046 (0.001)
	90:30	92.9	92.3	92.6	0.883 (0.000)	0.883 (0.000)	0.885 (0.000)
	15:5	90.5	83.5	95.1	2.111 (0.105)	2.111 (0.105)	2.492 (0.164)
ICC= $1/3$	30:10	92.5	89.7	91.9	1.794 (0.014)	1.794 (0.014)	1.843 (0.014)
$e_{dj} \sim t_6(1)$, $u_d \sim t_6(0.5)$	60:20	91.6	91.2	91.9	1.419 (0.002)	1.419 (0.002)	1.428 (0.002)
	90:30	92.1	91.9	92.2	1.217 (0.001)	1.217 (0.001)	1.222 (0.001)
	15:5	92.7	89.0	91.1	1.750 (0.043)	1.750 (0.043)	1.791 (0.049)
ICC= $2/3$	30:10	92.4	91.4	91.9	1.365 (0.007)	1.365 (0.007)	1.373 (0.007)
$e_{dj} \sim t_6(0.5)$, $u_d \sim N(1)$	60:20	93.5	93.7	93.9	1.041 (0.001)	1.041 (0.001)	1.044 (0.001)
	90:30	94.2	93.8	94.0	0.882 (0.000)	0.882 (0.000)	0.884 (0.000)
ICC $= 1/9$	26:50	95.5	93.9	96.5	1.257 (0.002)	1.257 (0.002)	1.328 (0.001)
$e_{dj} \sim st_{5,1.25}(2)$, $u_d \sim N(0.25)$	52:100	94.3	93.8	94.6	0.992 (0.000)	0.992 (0.000)	1.002 (0.000)

Note: The nominal coverage probability is 95%.

Section 7. In particular, it aims to mimic the pdf of estimated errors in our application, see the middle panel of Figure 4. We consider a scenario with $D = 52$ and $n_d = 100$, that is, the number of areas in the data example with $n_d = 100$ being close to the median of the number of units across counties. We also evaluate the performance of our method for a smaller sample size with $D = 26$ and $n_d = 50$. Because the results hardly differ when estimating $\boldsymbol{\theta}$ using restricted maximum likelihood (REML) or the method of moments, we only present the former.

In the simulation study with FHM, we apply a similar setting as in Datta *et al.* (2005). Random effects and errors are independent, centred and normally distributed with unknown variance $\sigma_u^2 = 1$ and known $\sigma_{e_d}^2$. Each fifth part of the total number of clusters is assigned to a different value of $\sigma_{e_d}^2$; in Scenario 1, we have 0.7, 0.6, 0.5, 0.4, 0.3, whereas in Scenario 2, 2.0, 0.6, 0.5, 0.4, 0.2. That is, we consider the case of known heteroscedasticity for errors. Variance σ_u^2 is estimated using REML, Henderson's method (Prasad & Rao, 1990) and the method of Fay & Herriot (1979). We present results only for the former as the other methods perform similarly for our SPI and MT. All simulated scenarios are almost optimal settings for the Bonferroni procedure as the mixed parameter estimates are asymptotically independent. Therefore we can take it as a benchmark (cf. comments in Section 5.2).

We use three criteria to evaluate the performance of different methods to construct SPI: the empirical coverage probability (ECP), the average width (WS), and the average variance of widths (VS):

$$\text{ECP} = \frac{1}{I}\sum_{k=1}^{I} \boldsymbol{1}\{\mu_d^{(k)} \in \mathcal{I}_{1-\alpha}{}^{P}\, \forall d \in [D]\}, \ \text{where P} = \text{BS}, \ \text{MC}, \ \text{BE or BO},$$

$$\text{WS} = \frac{1}{DI}\sum_{d=1}^{D}\sum_{k=1}^{I} \rho_d^{(k)}, \ \rho_d^{(k)} = 2c_P^{(k)}(1-\alpha)\widehat{\sigma}^{(k)}(\widehat{\mu}_d), \ \text{where P} = \text{BS}, \ \text{MC}, \ \text{BE or BO},$$

$$\text{VS} = \frac{1}{D(I-1)}\sum_{d=1}^{D}\sum_{k=1}^{I}\left(\rho_d^{(k)} - \overline{\rho}^d\right)^2, \ \overline{\rho}^d = \sum_{k=1}^{I}\rho_d^{(k)}/I.$$

ECP is the percentage of times all cluster-level parameters are inside their SPI. On the other hand, WS is calculated for each cluster over the widths of the intervals from I simulations, and averaged over all clusters to obtain an aggregated indicator. Lower values of WS are preferable. Finally, for assessing their variability, we compute the variance of widths over the simulations, and average them over all clusters (VS). We prefer lower values of VS, as they would indicate that the length of intervals is stable.

Last but not least, in practice, $c_{BS}(1-\alpha)$ is approximated by $[\{(1-\alpha)B\}+1]^{th}$ order statistics of the empirical bootstrap distribution. In addition, to construct $\mathcal{I}_{1-\alpha}{}^{MC}$ in (17) we can use $\boldsymbol{g}_1$ or the variance expression from the denominator in (15). Since the resulting numerical differences were negligible, we present results only for the latter.

Table 1 shows the numerical results of our criteria to compare the performance of different methods when errors and random effects are normally distributed. Under these scenarios, BS attains the nominal level of 95% even for a small number of clusters ($D = 15$). Yet, due to the overestimation of variability of the cluster parameters, this method suffers from an overcoverage when ICC$= 1/3$ for $D = 15$ and $D = 30$. Furthermore, although our simulations constitute a nearly optimal design for the Bonferroni method, BO exhibits almost always undercoverage. MC has worse performance, the convergence to the nominal level is slower with ECP oscillating around 94% only for $D = 60$ and $D = 90$ when ICC$= 2/3$ or ICC$= 1/2$. It does not attain the nominal coverage under the third scenario. The second part of Table 1 summarises results for WS and VS. As expected, the width increases with growing D. Nonetheless, the speed of this increase is moderate; with a growing number of areas, the SPI has to cover more parameters, but the estimate of variability decreases (for more details, see Reluga *et al.*, 2021). When we consider VS, we conclude that BS is more variable than other methods for $D = 15$, but this difference decreases for increasing D. In SAE, undercoverage is often considered a more severe type of error than overcoverage, partly due to the difficulties to detect and alleviate it (Yoshimori, 2015). On the other hand, overcoverage is often a result of an excessive variability in small samples which is illustrated in Table 1. Having this in mind, we conclude that BS is the most satisfactory method.

Table 2 shows the performance of SPI constructed for a subset of $D' = D/5$ clusters. It illustrates finite sample performance of Corollary 2. Because the simulations under other scenarios led to the same conclusions, we only consider $\sigma_u^2 = \sigma_e^2 = 1$. In each simulation run we construct SPI for D' areas, but use all data to compute variances, fixed and random effects. The ECP is similar as in Table 1. In contrast, the widths of SPI are narrower than those in Table 1, because they are constructed to cover simultaneously only $D' < D$ mixed parameters. This empirical study confirms a practical relevance of our proposal. In fact, it shows that one can construct reliable SPI or conduct MT for an arbitrary subset of cluster-level parameters.

While the vast majority of the SAE literature heavily relies on the normality of random effects and errors, especially regarding MSE estimation and CPI construction, this assumption may be violated in practice. Thus, we conduct a robustness study regarding departures from normality of errors and/or random effects. The first part of the empirical results of our criteria under this setting is presented in Table 3. Further simulation results are deferred to the supporting information. First, but not surprisingly, the overall performance of all methods is worse than in Table 1, especially for asymmetric χ^2 distributed errors. Second, BS and BO are still superior to all other methods. In addition, in case of chi-square distributed departures, the coverage is higher for ICC$= 1/3$ with $D = 15$ due to overestimated variability, then it drops for $D = 30$ and $D = 60$, and increases for $D = 90$ in accordance with our asymptotic theory. Importantly, under the scenario which mimics the data application, that is, with skewed t-distributed errors and normal random effects, the ECP is close to the nominal level. However, we must conclude that the considered SPIs do not attain the nominal coverage probability if errors exhibit more severe deviations from normality than those observed in our application, irrespective of the presence of deviations from normality of random effects. The issue of undercoverage might be alleviated by the use of a more sophisticated bootstrapping scheme but requiring different theoretical derivations and therefore beyond the scope of this paper (cf. Reluga, 2020). Simulations therein and in our supporting information confirm that the deviation from normality of random effects hardly affects the coverage of SPI. A similar conclusion was drawn by McCulloch & Neuhaus (2011) in the study of the bias of estimated fixed effects and EBLUPs. When it comes to the right hand side of Table 3, WS decreases with a growing sample size due to the increase of n_d. Even though the critical values increase with the growing number of clusters, $\widehat{\sigma}^2(\widehat{\mu}_d)$ decreases at a faster rate. For this reason the average width of intervals is decreasing too.

Let us revisit the differences between CPI and SPI. Figure 1 displays 95% bootstrap SPI in light blue and CPIs of Chatterjee *et al.* (2008) in dark blue. The critical values for CPIs have been calculated using parametric bootstrap (cf. Chatterjee *et al.*, 2008, for details). In comparison with CPI, SPI covers all clusters with a certain probability. Black dots represent the true mixed parameters μ_d. Out of thirty, three cluster-level parameters (eighth, twenty-second and twenty-third) are clearly outside of their CPI, and another four (first, seventh, eleventh and twenty-fifth) are on their boundary. It does not happen by chance or by the simulation design, but by the construction of CPIs: for $100(1 - \alpha)\%$ CPI, about $100\alpha\%$ of the true mixed parameters are not covered by their intervals. Figure 1 illustrates even a more severe case with 10% of

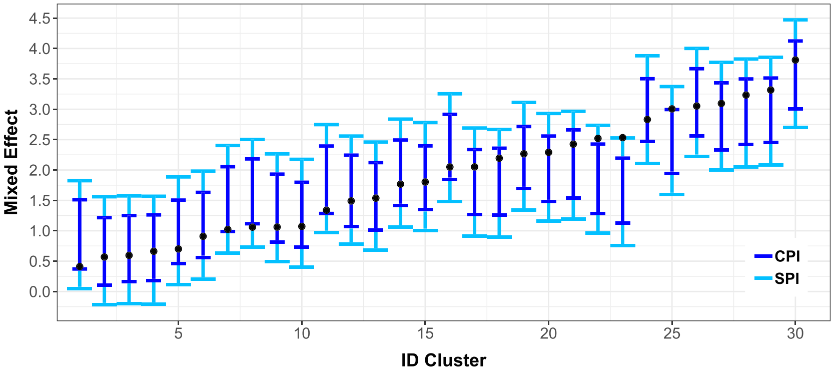

FIGURE 1. *95% CPI and bootstrap SPI for mixed effect means, $e_{dj} \sim N(0.5)$, $u_d \sim N(1)$, ICC $= 2/3$ and $D = 30$. Black dots are true mixed parameters*

the true parameters not covered by CPIs. In contrast, SPI contains all of the true mixed parameters. Moreover, SPI is not excessively wide compared with CPI. In fact, SPI is just as wide as necessary; the twenty-third cluster-level mean is right at the boundary. Undoubtedly, CPI and SPI are methodologically different and constructed to cover distinct sets with a certain probability. One can thus argue that their direct comparison is flawed and should not be investigated. We do not claim otherwise; rather, Figure 1 serves as an illustration of the practical relevance of SPI as a valid tool for comparing mixed parameters across clusters. Moreover, Figure 1 demonstrates that the cluster-wise inference can lead to erroneous conclusions once applied to perform joint statements or comparisons.

Regarding our multiple testing procedure, Figure 2 displays the empirical power of bootstrap and MC based max-type tests for $H_0 : \boldsymbol{\mu} = \boldsymbol{h}$ versus $H_1 : \boldsymbol{\mu} = \boldsymbol{h} + \boldsymbol{1}_D\delta$. For this simulation, we set $\boldsymbol{h} := \boldsymbol{\mu}$ under H_0, whereas under H_1 we add a constant $\delta \in [-2, 2]$ to each element of $\boldsymbol{h}$. As expected, the power curves get steeper with a growing sample size for each simulation scenario (by columns from top to bottom). For larger D, the curves almost coincide under all three scenarios. Moreover, ICC influences the Type II error – the curves are the steepest and almost not distinguishable for ICC$= 2/3$. Even though the MC based test achieves a higher power under H_0 when ICC$= 1/2$ and ICC$= 1/3$ for small and medium D (four plots on the top right of Figure 2), the bootstrap test performs significantly better in terms of attaining the nominal level. In contrast, we can conclude that MC test does not reach a nominal level of the test at the true value. In fact, its seemingly stronger power is a consequence of rejecting too often at the null hypothesis.

We finally turn to the analysis under FHM. Because the performance of our MT method under FHM leads to similar conclusions as under NERM, we restrict the presentation to ECP, WS and VS for different SPIs. Table 4 displays the obtained results. Bootstrap SPI suffers from overcoverage for small D, similarly to Bonferroni's SPI. The overcoverage is mostly likely caused by the same reasons as for NERM. Surprisingly, Bonferroni's intervals fail to achieve the nominal level for larger numbers of clusters.

7 Application to the Household Income Data of Galicia

We consider the household income data of the Structural Survey of Homes of Galicia (SSHG) which contains many potentially correlated covariates. It is of great interest for the Galician Institute of Statistics (IGS), and the regional government alike, to study the household income across counties (*comarcas*), for example, to adjust regional policies and resource

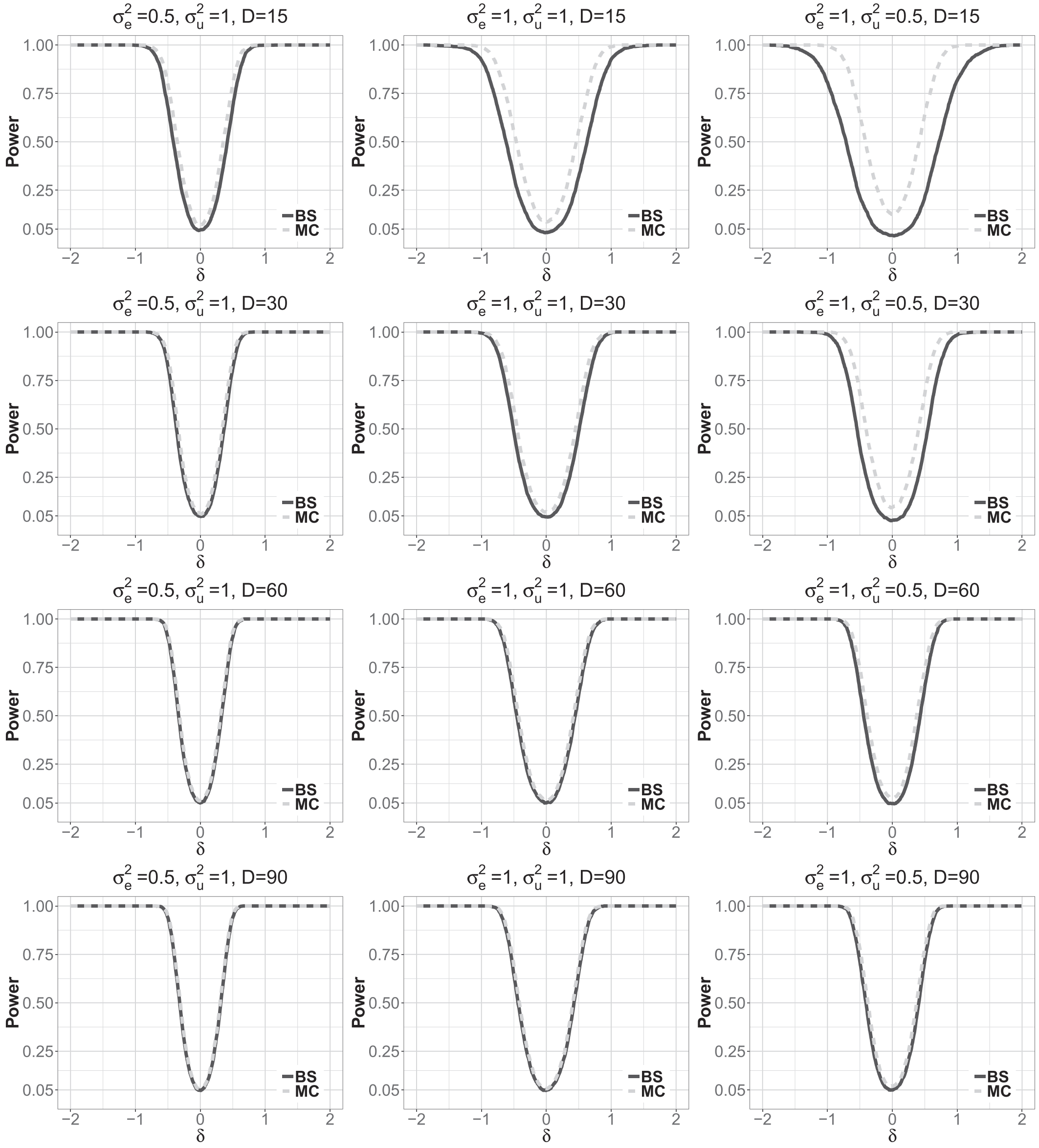

FIGURE 2. *Power of MT $H_0 : \boldsymbol{\mu} = \boldsymbol{h}$ versus $H_1 : \boldsymbol{\mu} = \boldsymbol{h} + \boldsymbol{1}_D \delta$ for BS-based and MC-based multiple tests (MT) under simulation scenarios with different values of ICC: 2/3 (left column), 1/2 (middle column) and 1/3 (right column)*

allocations. The IGS provides direct design-based estimates and/or EBLUP of the average household income accompanied by their variability measures or area-specific confidence intervals. However, the joint consideration of county-level parameters or comparisons between them is often important too. We start from the classical design-based and model-based area-wise analysis. Afterwards we complete it with the simultaneous inference for counties in Galicia.

The SSHG contains data on 23 628 individuals within 9 203 households which were collected in 2014 and published in 2015. It comprises information about the total income as well as different characteristics on individual and household level. The variable of interest is the monthly household income. This variable was obtained by taking the twelfth of the total yearly

Table 4. *ECP (in %), WS and VS under FHM with normal errors and random effects*

			ECP (in %)			*WS (VS)*	
	D	*BS*	*MC*	*BO*	*BS*	*MC*	*BO*
S.1	15	97.3	95.6	96.5	3.728 (0.016)	3.516 (0.024)	3.691 (0.019)
	30	96.6	95.2	96.6	3.792 (0.017)	3.664 (0.023)	3.818 (0.013)
	60	95.7	92.6	93.9	3.973 (0.014)	3.804 (0.031)	3.873 (0.027)
	90	95.2	93.3	94.4	4.024 (0.016)	3.920 (0.025)	3.970 (0.022)
S.2	15	98.0	95.7	95.9	4.073 (0.034)	3.749 (0.061)	3.962 (0.096)
	30	97.1	95.6	96.1	3.795 (0.017)	3.667 (0.023)	4.028 (0.040)
	60	97.4	93.4	94.9	4.198 (0.035)	3.956 (0.067)	4.029 (0.064)
	90	96.6	93.9	94.6	4.218 (0.037)	4.006 (0.054)	4.119 (0.053)

Note: The nominal coverage probability is 95%.

Table 5. *Descriptive statistics of the number of units across comarcas in provinces of Galicia*

A Coruña								*Lugo*			
Min	Q_1	Q_2	Q_3	Max	Total	Min	Q_1	Q_2	Q_3	Max	Total
18	36	90	197	930	3231	18	76.5	90	193.5	449	1619
Ourense								Pontevedra			
Min	Q_1	Q_2	Q_3	Max	Total	Min	Q_1	Q_2	Q_3	Max	Total
18	22.5	90	158	683	1637	36	94.5	162	368	1008	2716

Note: Statistics: Min - minimum, Q1 - first quartile, Q2 - median, Q3 - third quartile, Max - maximum.

income which consists of paid work, own professional activity and miscellaneous benefits. Following Lombardía *et al.* (2018), we consider following covariates: age, education level, type of household, and variables indicating financial difficulties of the household at the end of a month. Galicia is divided into four provinces (A Coruña, Lugo, Ourense and Pontevedra) which are further divided into 53 counties, the small areas that constitute our clusters. There are eighteen counties in A Coruña, thirteen in Lugo, twelve in Ourense and ten in Pontevedra. As the SSGH does not contain data from the county Quiroga in Lugo, we limit the study to the remaining 52 counties. Table 5 displays descriptive statistics of the number of units across the counties of each province. The model based approach is motivated by the scarcity of data (in some counties fewer than 20 observations were collected).

Even though the SSHG does not produce official estimates of totals Y_d^{dir}, X_{di}^{dir} and means $\overline{Y}_d^{dir}, \overline{X}_{di}^{dir}$ at the county level, we calculated them using

$$\widehat{Y}_d^{dir} = \sum_{j \in \mathcal{R}_d} w_j y_j,\ \ \widehat{\overline{Y}}_d^{dir} = \widehat{Y}_d^{dir} / \widehat{N}_d^{dir},\ \ \widehat{X}_{di}^{dir} = \sum_{j \in \mathcal{R}_d} w_j x_{ji},\ \ \widehat{\overline{X}}_{di}^{dir} = \widehat{X}_{di}^{dir} / \widehat{N}_d^{dir} \text{ and } \widehat{N}_d^{dir} = \sum_{j \in \mathcal{R}_d} w_j, \tag{19}$$

where $\widehat{N}_d^{dir}$ stands for the estimate of the county size N_d^{dir}, $\mathcal{R}_d$ is the sample in county d and w_j is an official calibrated sample weight. In addition, we have $w_j = 1/\pi_j$ where $\pi_j \neq 0$ is the first-order inclusion probability. We used the same design-based direct variance estimator as Lombardía *et al.* (2018), that is

$$\widehat{var}\left(\widehat{\overline{Y}}_d^{dir}\right) = \frac{1}{\left(\widehat{N}_d^{dir}\right)^2} \sum_{j \in \mathcal{R}_d} w_j (1 - w_j) \left(y_j - \widehat{\overline{Y}}_d^{dir}\right)^2. \tag{20}$$

Furthermore, we calculated the coefficient of variation (CV) of direct estimates at the county level. In 12 counties CV >10%, and in three of them CV >15%. A direct estimate is considered official, and thus publishable, if its CV is lower than a certain threshold set by a statistical office. For example, the Office for National Statistics in the UK sets this threshold to 20% for the labour force statistics (Lombardía *et al.*, 2018). Although the CV of our estimates does not exceed this threshold, it is still high enough to consider a model-based framework.

We construct design and model-based point and CPI estimates of monthly incomes. For the design-based estimation, we employed $\overline{\widehat{Y}}^{dir}_{d}$ and $\widehat{var}(\overline{\widehat{Y}}^{dir}_{d})$ defined in (19) and (20). Within the model-based framework, we need to estimate the county-level means of covariates $\overline{X}^{di}$ by $\overline{\widehat{X}}^{dir}_{di}$ in (19) due to the lack of access to register information. Afterwards, we consider $\mu_d = \boldsymbol{k}^t_d\boldsymbol{\beta} + \boldsymbol{m}^t_d\boldsymbol{u}_d$ in (4) as our target parameter with $\boldsymbol{k}_d = \overline{\widehat{\boldsymbol{X}}}^{dir}_{d}, \boldsymbol{m}_d{=}1$ and calculate EBLUP in (5) by $\widehat{\mu}_d = \overline{\widehat{\boldsymbol{X}}}^{dir}_{d}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}_d$ $\forall d \in [D]$. Because SSHG contains information on the household level, we can fit NERM to these data. Figure 3 shows design-and model-based point estimates of monthly household incomes together with 95% CPIs. Model-based CPIs were constructed using the parametric bootstrap (cf. Chatterjee *et al.*, 2008). We can use CPIs to compare different methods for the same cluster-level parameter, but not to make comparisons across different counties. For a better presentation, we divided the plot into five panels based on the number of units in each county. First, we can see that the widths of both direct-based and model-based intervals decrease with increasing sample size. Second, the widths of direct CPIs are much wider that their model-based counterparts. In fact, direct estimates for certain areas (for example, the fourth and sixth area in the first panel) are too wide to make any informative conclusion. This confirms the necessity of a model-based framework.

Remark *Due to the lack of access to administrative registers, we replaced population means of auxiliary variables* $\overline{\boldsymbol{X}}^{dir}_{d}$ *by their SSHG estimates* $\overline{\widehat{\boldsymbol{X}}}^{dir}_{d}$ *which is a common approach in the SAE literature (see, e.g. Chandra* et al.*, 2015; Lombardía* et al.*, 2017; 2018). As pointed out correctly by one of our referees, this step increases the total uncertainty of the predictor (in our case EBLUP* $\widehat{\mu}_d = \overline{\widehat{\boldsymbol{X}}}^{dir}_{d}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{u}}_d$*, but this applies more generally to EBP and other predictors) of the small area parameter of interest. Yet, following a widely accepted practice in the SAE research using LMM (see, e.g. Lombardía* et al.*, 2017; 2018), we did not account for this additional variability and treated* $\overline{\widehat{\boldsymbol{X}}}^{dir}_{d}$ *as a non-random quantity in the subsequent steps of the statistical inference (the estimation of MSE and the construction of CPIs and SPIs). In the context of LMM-based prediction of population means, Chandra* et al. *(2015) quantified exactly the additional variability of EBLUP* $\widehat{\mu}_d$*, and it reduces to adding* $\widehat{\boldsymbol{\beta}}^t\widehat{var}(\overline{\widehat{X}}^{dir}_{d})\widehat{\boldsymbol{\beta}}$ *to MSE in (6). Nevertheless, this additional term is often disregarded in large surveys such as SSHG due to its relative small contribution to the total*

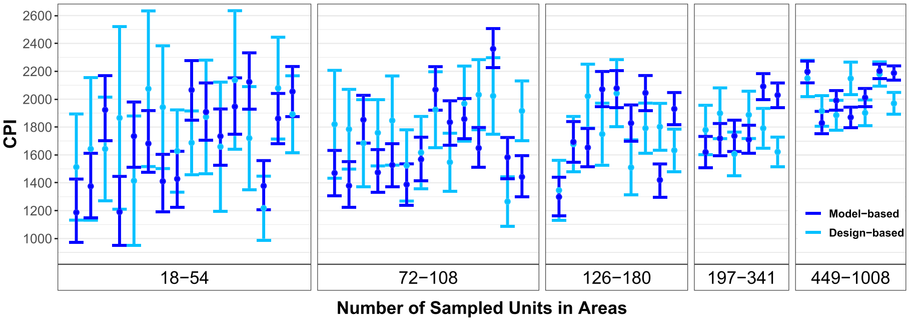

FIGURE 3. *Design and model-based 95% CPI.*

variability of $\widehat{\mu}_d$ with respect to the variability of $\widehat{\beta}$ and $\widehat{u}_d$ (in fact, the correction of Chandra et al. *(2015) is hardly ever used in applied survey-based studies in spite of its relevance). A comprehensive study measuring the relative increase of the variability of $\widehat{\mu}_d$ by replacing $\overline{\boldsymbol{X}}^{dir}_d$ by $\hat{\overline{\boldsymbol{X}}}^{dir}_d$ as well as the construction of the optimal MSE estimator in this context could be interesting topics of further research.*

Table 6 displays the covariates with their standard deviations as well as the estimated coefficients with standard errors and *p*-values. We performed a variable selection in two stages. First, we selected a subset of covariates that exhibited the highest Spearman's rank correlation with the household income. Afterwards, we applied a generalised AIC which uses a quasi-likelihood with generalised degrees of freedom, see Lombardía *et al.* (2017). The procedure selected covariates describing characteristics of the household and characteristics of the head of household. The estimates of variance parameters are $(\widehat{\sigma}_e^2, \widehat{\sigma}_u^2) = (758558.60, 19746.24)$.

It is well known that income data are right skewed. Unsurprisingly, our dependent variable exhibits this feature too. It is therefore popular to consider log-income or more sophisticated transformations. Because the naive back-transformation of the dependent variable could cause a serious bias due to the Jensen inequality, different estimation and inference methods were recently suggested by Rojas-Perilla *et al.* (2020) and references therein. Specifically, in our data example log transformation of household income did not help to overcome the problem of skewness. We thus decided to proceed with the household income on the original scale, and assess the sensitivity of our method to the departures from normality of errors and/or random effects after fitting a LMM. As we shall see, the undertaken inference is not compromised by

Table 6. *Descriptive statistics and coefficient estimates with standard errors and p-values*

Dependent variable		*Dir Mean*		*Dir Stdev*		
Inc	*Monthly household income*	*1914.884*		*13.766*		
Characteristics of the household		*Mean*	*Stdev*	$\hat{\beta}$	*S.E.*	*p-value*
Type1	= 1 if households consists of 1 person	0.208	0.406	– 1616.177	33.191	0.000
Type2	= 1 if households consists of more than 1 person	0.023	0.149	−933.004	65.622	0.000
Type3	= 1 if households consists of a couple with children	0.304	0.460	−601.977	31.416	0.000
Type4	= 1 if households consists of a couple without children	0.246	0.431	−983.258	31.972	0.000
Type5	= 1 if households consists of a single parent	0.093	0.290	– 1056.531	39.714	0.000
Type67	= 1 if households consists of one or several centres or other	Benchmark variable: dropped in fitting				
Dif1	= 1 if a lot of difficulties coming to the end of a month	0.123	0.328	−982.786	31.132	0.000
Dif2	= 1 if some difficulties coming to the end of a month	0.445	0.497	−514.372	20.266	0.000
Dif3	= 1 if no difficulties coming to the end of a month	Benchmark variable: dropped in fitting				
Ten1	=1 if property without mortgage	0.663	0.473	301.229	26.218	0.000
Ten2	=1 if property with mortgage	0.168	0.374	417.126	32.072	0.000
Ten34	=1 if ceded, rental or another type of property	Benchmark variable: dropped in fitting				
Characteristics of the household head						
Educ1	= 1 if primary education	0.232	0.422	−902.255	30.763	0.000
Educ2	= 1 if secondary education	0.515	0.500	−731.925	23.455	0.000
Educ3	= 1 if higher education	Benchmark variable: dropped in fitting				
Age1	= 1 if $45 \leqslant$ age $\leqslant 64$	0.377	0.485	206.039	19.988	0.000
Age2	= 1 if age <45 or age >64	Benchmark variable: dropped in fitting				
Intercept		-	-	3308.762	46.481	0.000

these departures. We carried out statistical tests and analysed diagnostic plots. The left panel of Figure 4 displays a diagnostic plot of Lange & Ryan (1989) using standardised empirical Bayes estimates of the random effects in a weighted normal QQ plot; it supports the adequacy of the normality assumption. Moreover, the p-values of Kolmogorov–Smirnov and Shapiro–Wilk tests, which are 0.997 and 0.944, confirm this conclusion. Regarding the normality of errors, the other two panels of Figure 4 present Cholesky residuals (Jacqmin-Gadda *et al.*, 2007). They are constructed by multiplying $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ by the Cholesky square root of the variance matrix. A right tail is visible in both panels. The p-values of Kolmogorov–Smirnov and Shapiro–Wilk tests are <0.0001. In the centre of Figure 4, we can see that the kernel density of the skewed errors has a long, but not a thick tail. Our simulation results in Table 3 indicate that such departure is not problematic for our bootstrap-based SPI. In fact, our SPI is quite robust to these departures, and a good coverage probability is still provided in comparison with other methods. We do not assess the robustness of bootstrap CPIs to the departure from normality of errors because they are not the topic of this article; we present them only for illustrative reasons.

Figure 5 displays bootstrap CPI as developed by Chatterjee *et al.* (2008), together with BS SPI for the county-level averages of monthly household income in Galicia. We can see a lot of variability over the estimates. Evaluating the results of CPI (dark blue) versus SPI (light blue), it is apparent that the cluster-wise prediction intervals are not adequate to address either a joint consideration or a comparison of the counties. If we consider, for example, the counties of A Fisterra and Noia (7th and 8th regions of the second panel in the black rectangle), the CPIs indicate significantly different incomes, whereas the SPIs do not support this claim. Moreover,

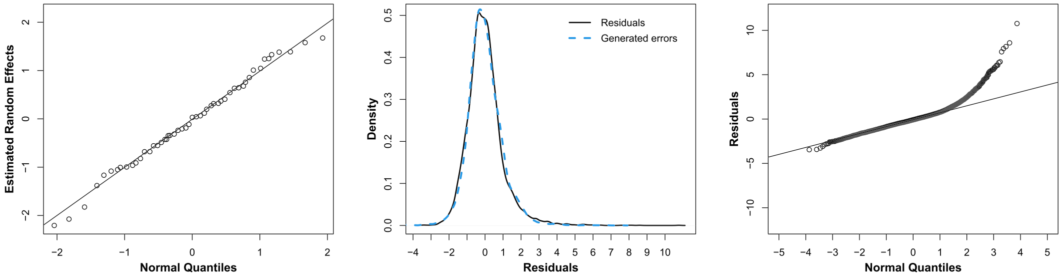

FIGURE 4. *REML empirical Bayes estimates of random effects: (left) QQ plot; Cholesky REML residuals: kernel density (centre) and QQ plot (right)*

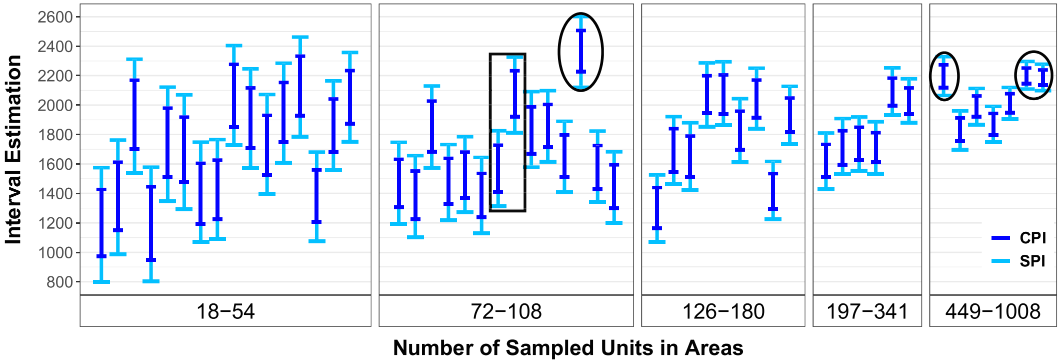

FIGURE 5. *95% bootstrap CPI and SPI for the county-level averages of the household income in Galicia*

there are other counties (practically in each panel) for which CPIs would insinuate significant differences whereas statistically valid SPIs do not confirm this conclusion. Nevertheless, SPIs are not unnecessarily wide for practical use. We detect significant and valid differences between several interval estimates.

Figure 6 presents maps with lower and upper limits of bootstrap SPI. The boundaries are classified into one of five categories which were built using 0.2, 0.4, 0.6 and 0.8 quantiles of the point estimates. We observe a substantial variation of average household income over the counties. Lower and upper boundaries of the interval estimates for the counties of A Coruña, Lugo, Vigo (with a large number of units) and Lower Miño are classified into the richest category; they are indicated with ellipsoid in the second and the last panel of Figure 5. In contrast, there is a group of eight counties (three in the centre, one in the west and four in the south of Galicia) which are classified to the poorest category in the left panel and the second poorest category in the middle panel. Right panel of Figure 6 presents CPIs and two different types of SPIs for eight poorest counties. In particular, SPI refers to the interval estimate constructed for 52 counties (estimates for all 52 counties are plotted in Figure 5). In contrast, SPI-S refers to the interval estimate constructed for a subset of eight poorest counties, but using all data to estimate fixed parameters and predict random effects (cf. Corollary 2 and simulations in Table 2). As expected, SPI-S is still wider than CPI, but much narrower than SPI. Moreover, SPI-S allows for a valid comparative inference for a subset of these poorest areas. In fact, we can conclude that the differences in the average household income are not statistically significant. Finally, CPI should not be used to make maps like in Figure 6, as this would suggest that we were allowed to compare them.

Finally, it would be interesting to investigate whether the monthly income of the households which reported difficulties in coming to the end of the month is significantly different from the monthly income of the households which did not struggle with this issue. For example, in touristic areas, households spend more such that they might face some difficulties without being poorer. The outcome of such test might then be used to develop a more targeted policy. Our MT procedure can be readily applied to support or disprove the hypothesis of no difference in monthly income between the two mentioned types of households. To test this hypothesis we take clusters created from a cross-section of counties and difficulty status. Therefore we apply our developed methodology to $2 \times 52 = 104$ counties by difficulty status. More specifically, we consider $\boldsymbol{\mu} \in \mathbb{R}^{104}$ and test $H_0 : \boldsymbol{A\mu} = \boldsymbol{0}_{52}$ versus $H_1 : \boldsymbol{A\mu} \neq \boldsymbol{0}_{52}$, where $\boldsymbol{A} \in \mathbb{R}^{52 \times 104}$ with rows that are composed of 104-dimensional vectors $\boldsymbol{a}$ with a 1 on the $2d - 1$ place, -1 on

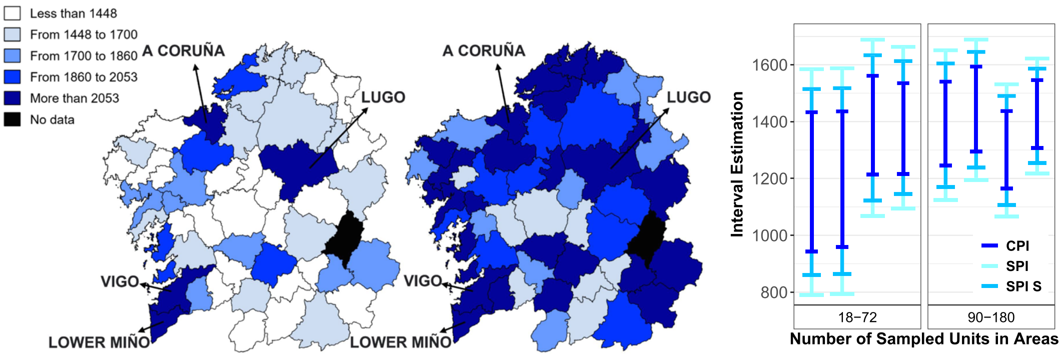

FIGURE 6. *95% bootstrap SCI for the county-level averages of the household income in Galicia: (left) lower boundary, (centre) upper boundary; (right) SPI, SPI-S and CPI for a subset of the poorest areas*

the $2d$ place but 0 otherwise, where d stands for a particular county. The test statistic is $t_H = \max_{d=1,\dots,D}|t_{H_d}| = 7.569$ whereas the critical value is $c_{BH_0}(1 - \alpha) = 4.220$. That is, we clearly reject H_0 of no difference in the household income between the two groups of household in each county. The rejection of the null hypothesis is consistent with the variable selection procedure which has suggested to find this test outcome.

8 Conclusions

We introduce a practical bootstrap-based method to construct SPI and MT procedures for mixed parameters under LMM. We illustrate its use and relevance in simulation studies and a data application within the framework of SAE. We theoretically derive two techniques based on bootstrap approximation of the distribution of the max-type statistic and the volume-of-tube formula. However, we proved that the latter is not directly operational. We further discussed various alternatives and assessed their empirical performance in the simulation study. Though slightly conservative for very small samples, the bootstrap-based SPI yields the most satisfactory results in our simulations. Moreover, it is quite robust to certain deviations from normality assumptions. In addition, our bootstrap based max-type statistic is readily applicable for testing multiple statistical hypotheses which was illustrated in our simulation studies.

Accounting for the joint coverage probability (or the Type I error) for several or all cluster-level parameters makes SPI wider than CPI. However, only SPI are statistically valid for joint statements or comparisons between cluster-level parameters. Moreover, if one conducted studies with several surveys, SPI would contain all true parameters in $100(1 - \alpha)\%$ of all studies, whereas CPI would not cover about $D\alpha$ of them in each survey. Our tools are equally applicable to any subset of the clusters while using all data for estimation and prediction. The possibility to apply our procedure to any subset of D' areas while using all areas for estimation makes it particularly appealing in practice.

Our method can be extended to account for more complex data structures such as LMM with spatial and/or temporal dependencies (Pratesi & Salvati, 2008; Morales & Santamaría, 2019) or highly skewed response variables (Moura *et al.*, 2017). Furthermore, our general idea could be also applied for a comparative analysis of nonlinear indicators of inequality or poverty obtained after the transformation of the dependent variable (Rojas-Perilla *et al.*, 2020) or benchmarked estimators under restrictions (Ugarte *et al.*, 2009). Yet, for these cases, the estimation of the variance components or MSE would have to be adjusted. Furthermore, within more complex modelling frameworks one would need to extend the asymptotic theory. In addition, different bootstrap schemes might be necessary to mimic the data-generation process in a suitable way (Field *et al.*, 2008). The above-mentioned extensions are beyond the scope of this paper, but are an open field for future research.

ACKNOWLEDGEMENTS

Open access funding provided by Universite de Geneve.

REFERENCES

Basu, R., Ghosh, J.K. & Mukerjee, R. (2003). Empirical Bayes prediction intervals in a normal regression model: higher order asymptotics. *Stat. Probab. Lett.*, **63**(2), 197–203.

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. *J. Am. Stat. Assoc.*, **83**(401), 28–36.

Beran, R. (1988). Balanced simultaneous confidence sets. *J. Am. Stat. Assoc.*, **83**(403), 679–686.

Chandra, H., Sud, U.C. & Gharde, Y. (2015). Small area estimation using estimated population level auxiliary data. *Commun. Stat. – Simul. Comput.*, **44**, 1197–1209.

Chatterjee, S., Lahiri, P. & Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Ann. Statist.*, **36**(3), 1221–1245.

Chernozhukov, V., Chetverikov, D. & Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, **41**(6), 2786–2819.

Cox, D.R. (1975). Prediction intervals and empirical Bayes confidence intervals. In *Perspectives in probability and statistics (papers in honour of m. s. bartlett on the occasion of his 65th birthday)*, pp. 47–55. Applied Probability Trust, Univ. Sheffield, Sheffie.

Datta, G.S., Rao, J.N.K. & Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, **92**(1), 183–196.

Fay, R.E. & Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Am. Stat. Assoc.*, **74**(366), 269–277.

Field, C.A., Pang, Z. & Welsh, A.H. (2008). Bootstrapping data with multiple levels of variation. *Can. J. Stat.*, **36**, 521–539.

Francq, B.G., Lin, D. & Hoyer, W. (2019). Confidence, prediction, and tolerance in linear mixed models. *Stat. Med.*, **38**(30), 5603–5622.

Ganesh, N. (2009). Simultaneous credible intervals for small area estimation problems. *J. Multivariate Anal.*, **100**(8), 1610–1621.

González-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D. & Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *J. Stat. Comput. Simul.*, **78**(5), 443–462.

Gosh, M. (1992). Constrained Bayes estimation with applications. *J. Am. Stat. Assoc.*, **87**, 533–540.

Hall, P. & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *J. R. Statist. Soc. B*, **68**(2), 221–238.

Henderson, C.R. (1950). Estimation of genetic parameters. *Ann. Math. Statist.*, **21**(2), 226–252.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**(2), 423–447.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.-M. & Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Comput. Statist. Data Anal.*, **51**(10), 5142–5154.

Jiang, J. (1998). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Stat. Sin.*, **8**(1), 861–885.

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

Kackar, R.N. & Harville, D.A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communs Statist. Theor. Meth.*, **10**(13), 1249–1261.

Kramlinger, P., Krivobokova, T. & Sperlich, S. 2018. Marginal and conditional multiple inference in linear mixed models. arXiv:1812.09250.

Krivobokova, T., Kneib, T. & Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. *J. Am. Stat. Assoc.*, **105**(490), 852–863.

Kubokawa, T. (2010). Corrected empirical Bayes confidence intervals in nested error regression models. *J. Korean Stat. Soc.*, **39**(2), 221–236.

Lange, N. & Ryan, L. (1989). Assessing normality in random effects models. *Ann. Statist.*, **17**(2), 624–642.

Lombardía, M.J., López-Vizcaíno, E. & Rueda, C. (2017). Mixed generalized Akaike information criterion for small area models. *J. R. Statist. Soc. A*, **180**(4), 1229–1252.

Lombardía, M.J., López-Vizcaíno, E. & Rueda, C. (2018). Selection of small area estimators. *Stat. Appl.*, **16**(1), 269–288.

Maringwa, J.T., Geys, H., Shkedy, Z., Faes, C., Molenberghs, G., Aerts, M., Ammel, K.V., Teisman, A. & Bijnens, L. (2008). Application of semiparametric mixed models and simultaneous confidence bands in a cardiovascular safety experiment with longitudinal data. *J. Biopharm. Stat.*, **18**(6), 1043–1062.

McCulloch, C.E. & Neuhaus, J.M. (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Stat. Sci.*, **26**(3), 388–402.

Morales, D. & Santamaría, L. (2019). Small area estimation under unit-level temporal linear mixed models. *J. Stat. Comput. Simul.*, **89**(9), 1592–1620.

Morris, C.N. (1983). Parametric empirical Bayes inference: theory and applications. *J. Am. Stat. Assoc.*, **78**(381), 47–55.

Moura, F.A.S., Neves, A.F. & Britz do N. Silva, D. (2017). Small area models for skewed brazilian business survey data. *J. R. Statist. Soc. A*, **180**(4), 1039–1055.

Prasad, N.G.N. & Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *J. Am. Stat. Assoc.*, **85**(409), 163–171.

Pratesi, M. & Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Stat. Methods Appl.*, **17**(1), 113–141.

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

Reluga, K. (2020). Simultaneous and post-selection inference for mixed parameters. Ph.D. Thesis, University of Geneva, Switzerland.

Reluga, K., Lombarda, M.-J. & Sperlich, S. (2021). Simultaneous inference for empirical best predictors with a poverty study in small areas. *J. Am. Statist. Ass.*, **2021**, 1–33. https://doi.org/10.1080/01621459.2021.1942014

Rojas-Perilla, N., Pannier, S., Schmid, T. & Tzavidis, N. (2020). Data-driven transformations in small area estimation. *J. R. Statist. Soc. A*, **183**(1), 121–148.

Romano, J.P. & Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Am. Stat. Assoc.*, **100**(469), 94–108.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric regression*. Cambridge University Press.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**(1-2), 87–110.

Shen, W. & Louis, T.A. (1998). Triple-goal estimates in two-stage hierarchical models. *J. R. Statist. Soc. B*, **60**, 455–471.

Stoline, M.R. & Ury, H.K. (1979). Tables of the studentized maximum modulus distribution and an application to multiple comparisons among means. *Technometrics*, **21**(1), 87–93.

Sun, J. & Loader, C.R. (1994). Simultaneous confidence bands for linear regression and smoothing. *Ann. Statist.*, **22**(3), 1328–1345.

Sun, J., Raz, J. & Faraway, J.J. (1999). Confidence bands for growth and response curves. *Stat. Sin.*, **61**(2), 679–698.

Tukey, J.W. 1953. The problem of multiple comparisons. Unpublished manuscript.

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. Statist. Soc. A*, **181**, 927–979.

Ugarte, M.D., Militino, A.F. & Goicoa, T. (2009). Benchmarked estimates in small areas using linear mixed models with restrictions. *Test*, **18**, 342–364.

Verbeke, G. & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer.

Weyl, H. (1939). On the volume of tubes. *Am. J. Math.*, **61**(2), 461–472.

Working, H. & Hotelling, H. (1929). Applications of the theory of error to the interpretation of trends. *J. Am. Stat. Assoc.*, **24**(165A), 73–85.

Yoshimori, M. (2015). Numerical comparison between different empirical prediction intervals under the Fay-Herriot model. *Comm. Statist. Simul. Comput.*, **44**(5), 1158–1170.

Yoshimori, M. & Lahiri, P. (2014). A second-order efficient empirical Bayes confidence interval. *Ann. Statist.*, **42**(4), 1233–1261.

APPENDIX A

A.1 Regularity Conditions

R.1 $\boldsymbol{X}_d$ and $\boldsymbol{Z}_d$ are uniformly bounded such that $\sum_{d=1}^{D} \boldsymbol{X}_d^t \boldsymbol{V}_d^{-1} \boldsymbol{X}_d = \{O(D)\}_{(p+1)\times(p+1)}$.

R.2 Covariance matrices $\boldsymbol{G}_d$ and $\boldsymbol{R}_d$ have a linear structure in $\boldsymbol{\theta}$.

R.3 Convergence: $D \rightarrow \infty$, $\sup_{d \geqslant 1} n_d < < \infty$ and $\sup_{d \geqslant 1} q_d < < \infty$.

R.4 To ensure the nonsingularity of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, $0 < \inf_{d \leqslant 1} \sigma_{e_d}^2 \leqslant \sup_{d \leqslant 1} \sigma_{e_d}^2 < \infty$ and $\sigma_u^2 \in (0, \infty)$.

R.5 $\boldsymbol{b}_d^t = \boldsymbol{k}_d^t - \boldsymbol{o}_d^t \boldsymbol{X}_d$ with $b_{di} = O(1)$ for $i = 1, \ldots, p+1$.

R.6 $\left\{\frac{\partial}{\partial \theta_j} \boldsymbol{o}_d^t \boldsymbol{X}_d\right\}_i = O(1)$ for $j = 1, \ldots, h$ and $i = 1, \ldots, p+1$.

R.7 $\widehat{\boldsymbol{\theta}}$ satisfies: (i) $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(D^{-1/2})$, (ii) $\widehat{\boldsymbol{\theta}}(\boldsymbol{y}) = \widehat{\boldsymbol{\theta}}(-\boldsymbol{y})$ and (iii) $\widehat{\boldsymbol{\theta}}(\boldsymbol{y} + \boldsymbol{X}\boldsymbol{r}) = \widehat{\boldsymbol{\theta}}(\boldsymbol{y})$ for any $\boldsymbol{r} \in \mathbb{R}^{p+1}$.

Furthermore, we will evoke Assumptions 1–4 from Chatterjee *et al.* (2008), which are quite technical but largely irrelevant in practice. For the sake of completeness, they are provided in our supporting information.

A.2 Proof of Proposition 1

We concentrate on the consistency of the bootstrap SPI. The consistency of the MT procedure follows straightforwardly with some changes of notation due to the correspondence between tests and interval estimates (for more details, see Corollary 2 in Reluga *et al.*, 2021). To demonstrate Proposition 1, we make use of the result in Theorem CLL of Chatterjee *et al.* (2008). In this section, we use some notation from their paper if it is not in conflict with ours. Consider S_{0d} in (8) and S^*_{Bd} in (11) and let $\mathcal{L}_d(q) = P(S_{0d} \leqslant q)$ and $\mathcal{L}^*_d(q) = P^*(S^*_{Bd} \leqslant q)$, where $P^*(\cdot)$ stands for a probability measure induced by a parametric bootstrap. Under suitable regularity conditions, Chatterjee *et al.* (2008) proved that $\mathcal{L}_d(q) = \Phi(q) + \gamma(q, \boldsymbol{\beta}, \boldsymbol{\theta}, n) + O(n^{-1})$ where $\gamma(\cdot)$ is some smooth function. Furthermore, $\mathcal{L}^*_d(q)$ admits almost identical, equally short expansion with $\boldsymbol{\beta}, \boldsymbol{\theta}$ replaced by $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}$. Because $\mathcal{L}_d(q) \approx \Phi(q)$, we can follow the same steps as Reluga *et al.* (2021) to prove the consistency of SPI. In particular, observe that

$$P(S_0 \leqslant q) = P\{\max_{d=1,\dots,2D}(-S_{01}, \dots, -S_{0D}, S_{01}, \dots, S_{0D}) \leqslant q\} = \prod_{d=1}^{2D} \mathcal{L}_d(q) \approx \prod_{d=1}^{2D} \Phi(q). \quad \text{(A1)}$$

The same arguments follow for S^*_B with P replaced by P^*. Moreover, the cdf of the standardised maxima of the normal distribution is in the domain of attraction of the Gumbel law. If we notice that the last term in (A1) can be approximated by the Gumbel distribution if suitably standardised, the consistency of SPI follows by applying Poyla's theorem which relates the convergence in law with sup-norm convergence (see Reluga *et al.*, 2021, for more details).

Remark *Chatterjee* et al. *(2008) estimated fixed parameters using an ordinary least squares method, whereas in our paper we use generalized least squares. As pointed out by the authors, an asymptotic expansion still holds as soon as the weighting matrices are smooth functions of* $\boldsymbol{\theta}$*, which we assume in R.2 in Appendix A.1.*

Remark *In their original proof, Chatterjee* et al. *(2008) considered a modified version of* S_{0d}*, that is* $S_{0d} = \widehat{\sigma}_T^{-1}(T - \widehat{\mu}_T)$ *where* $T = \boldsymbol{f}^t(\boldsymbol{X\beta} + \boldsymbol{Zu})$ *for a given fixed vector* $\boldsymbol{f}$*, and* $\widehat{\mu}_T = \boldsymbol{f}^t(\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{u}})$*. This modification leads to some smoothing effects and results in a faster convergence rate. Nevertheless, as pointed out in Remark 6 in their paper, the analysis of the area-specific mixed effects leads to a sightly slower convergence, but is equally valid, and can be carried out along the same lines.*

[*Received September 2021; accepted August 2022*]