A Random Effect Block Bootstrap for Clustered Data

Raymond CHAMBERS and Hukum CHANDRA

Random effects models for hierarchically dependent data, for example, clustered data, are widely used. A popular bootstrap method for such data is the parametric bootstrap based on the same random effects model as that used in inference. However, it is hard to justify this type of bootstrap when this model is known to be an approximation. In this article, we describe a random effect block bootstrap approach for clustered data that is simple to implement, free of both the distribution and the dependence assumptions of the parametric bootstrap, and is consistent when the mixed model assumptions are valid. Results based on Monte Carlo simulation show that the proposed method seems robust to failure of the dependence assumptions of the assumed mixed model. An application to a realistic environmental dataset indicates that the method produces sensible results. Supplementary materials for the article, including the data used for the application, are available online.

Key Words: Confidence interval; Consistency; Correlated clusters; Hierarchical data; Nonparametric bootstrap; Variance components.

1. INTRODUCTION

The bootstrap technique (Efron 1979; Efron and Tisbshirani 1993) was originally developed for parametric inference, given independent and identically distributed (iid) data. However, random effects models for hierarchically dependent data, for example, clustered or multilevel data, are now in wide use. With such data, it is important to use bootstrap techniques that replicate the hierarchical dependence structure of the data. A popular way of achieving this is to use a parametric bootstrap based on the assumed hierarchical random effects model. This is usually very effective, provided this model is correctly specified. On the other hand, if the stochastic assumptions of the model, for example, the assumption that the random effects are iid Gaussian random variables, are violated, then it is hard to justify use of the parametric bootstrap; see, for example, Rasbash et al. (2000). This is of particular concern since the bootstrap is often recommended as an alternative approach that is likely to lead to confidence intervals with better coverage in situations where the

Raymond Chambers, Professor, Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW 2522, Australia (E-mail: *ray@uow.edu.au*). Hukum Chandra, Senior Scientist, Indian Agricultural Statistics Research Institute, New Delhi 110012, India (E-mail: *hchandra@iasri.res.in*).

Journal of Computational and Graphical Statistics, Volume 22, Number 2, Pages 452–470
DOI: 10.1080/10618600.2012.681216

distribution assumptions that underlie analytical methods are questionable (Davison and Hinkley 1997).

Much of the early research on bootstrapping clustered data was within the design-based framework for sample survey inference, where the main focus is on replicating the impact of various forms of cluster sampling on repeated sampling inference for finite population parameters; see Rao and Wu (1988) and Canty and Davison (1999). However, our approach in this article is model based, in the sense that we treat the clusters as part of the data-generation mechanism rather the sampling scheme, and so is similar to the approach set out by Field and Welsh (2007). In particular, we consider inference with respect to the population model rather than the sampling mechanism, and so, our focus is on bootstrap inference for model parameters that accommodates the hierarchical dependence structure in the data. In this context, Carpenter, Goldstein, and Rasbash (2003, hereafter referred to as CGR) described a two-level bootstrap for a random effects model, while Field and Welsh (2007) reviewed various approaches to bootstrapping clustered data.

In what follows, we propose a random effects block (REB) bootstrap method for clustered, hierarchical, or multilevel data. Our approach is semiparametric, in the sense that the marginal model is generated parametrically within the bootstrap while the dependence structure of the model residuals is generated nonparametrically. Furthermore, the proposed bootstrap is simple to implement and seems free of both the distribution and the dependence assumptions of the parametric bootstrap, with its main assumption being that the marginal model is correctly specified. Note that the block bootstrap itself is not new, since this method has been used extensively with spatial and time series data; see Clark and Allingham (2011), Hutson (2004), and Hall, Horowitz, and Jing (1995). However, to the best of our knowledge, there have been no previous applications of the block bootstrap idea to multilevel data.

The remainder of this article is structured as follows. In Section 2, we describe how the parametric bootstrap and the CGR bootstrap can be used to construct bootstrap confidence intervals for multilevel data. We then describe a REB bootstrap method for such data and prove the consistency of the bootstrap confidence intervals obtained under this approach. Empirical results from model-based simulations of these different bootstrap methods are described in Section 3. In Section 4, we present results from the application of these bootstrap methods to a realistic environmental dataset where the hierarchical model is at best an approximation. Finally, Section 5 concludes the article with a summary of our major findings and a discussion of avenues for future research, and supplementary materials for this article are available online.

2. BOOTSTRAP METHODS FOR MULTILEVEL DATA

In this section, we outline bootstrap methods for constructing confidence intervals for parameters of multilevel models, focusing on the two-level case. To this end, consider the situation where we have data on a variable of interest y and a set of covariates $\mathbf{x}$ for n individuals clustered within D groups. Following standard practice, we refer to individuals as level 1 units and clusters as level 2 units. There are n_i ($i = 1, \ldots, D$) level 1 units making up level 2 unit i in the sample, with overall sample size $n = \sum_{i=1}^{D} n_i$. Such hierarchically

structured data are commonly modeled using random effects. In this article, we focus on a linear random intercepts model of the form

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij}, \quad j = 1, \ldots, n_i;\ i = 1, \ldots, D, \tag{1}$$

where y_{ij} denote the value of y for unit j in group i, $\mathbf{x}_{ij}$ is a $p \times 1$ column vector of auxiliary variables for unit j in group i, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, u_i denotes a cluster-specific (level 2) random effect, and e_{ij} is an individual (level 1) random effect. We assume that $\mathbf{x}_{ij}$ contains an intercept term as its first component. It is standard practice to model the random effects as Gaussian, and so, we further assume that these effects are mutually independent between individuals and between clusters, with $u_i \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. It follows that $E(y_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta}$ and $\text{cov}(y_{ij}, y_{ik}) = \sigma_u^2 + \sigma_e^2 I(j = k)$, where $I(A)$ is the indicator function for the event A. Let $\mathbf{y}$ denote the $n \times 1$ vector of values y_{ij}, with $\mathbf{x}$ denoting the corresponding $n \times p$ matrix defined by the $\mathbf{x}_{ij}$. Then, $E(\mathbf{y}) = \mathbf{x}\boldsymbol{\beta}$ and $\mathbf{V} = \text{var}(\mathbf{y}) = \text{diag}\{V_i = \sigma_u^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T + \sigma_e^2 \mathbf{I}_{n_i}; i = 1, \ldots, D\}$, where $\mathbf{I}_t$ and $\mathbf{1}_t$ denote the identity matrix of order t and a $t \times 1$ vector of ones, respectively. The parameters $\delta = (\sigma_u^2, \sigma_e^2)$ are typically referred to as the variance components of (1). Standard methods such as maximum likelihood (ML) or restricted maximum likelihood (REML) are used for estimating the unknown parameters of (1); see Harville (1977). In what follows, we use a "hat" to denote these estimates, that is, we let $\hat{\theta} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$ denote the ML or REML estimates of $\theta = (\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2)$. These allow us to define empirical best linear unbiased predictors (EBLUPs) $\hat{u}_i$ for the area effects u_i.

Given a bootstrap distribution for a component of $\hat{\theta}$, there are a number of methods that can be used to construct corresponding bootstrap confidence intervals for the parameter in θ corresponding to that component. For reviews of these methods, see Efron and Tibshirani (1993), DiCiccio and Efron (1996), Davison and Hinkley (1997), and DiCiccio and Romano (1988). Here, we use the percentile method, where a $100(1-\alpha)\%$ bootstrap confidence interval for any component of θ is constructed as the interval between upper and lower $\alpha/2$ percentile values of the bootstrap distribution of that component. Taking some liberties with notation, let $\hat{\theta}_{L,\alpha/2}$ denote the bootstrap estimate for a parameter in θ such that a fraction $\alpha/2$ of all its bootstrap estimates are smaller than $\hat{\theta}_{L,\alpha/2}$, with $\hat{\theta}_{U,\alpha/2}$ denoting the bootstrap estimate such that a fraction $\alpha/2$ of all its bootstrap estimates are larger than $\hat{\theta}_{U,\alpha/2}$. Then, an approximate $100(1-\alpha)\%$ confidence interval for this parameter is $[\hat{\theta}_{L,\alpha/2}, \hat{\theta}_{U,\alpha/2}]$.

2.1 A TWO-LEVEL PARAMETRIC BOOTSTRAP

The parametric bootstrap method for the ML/REML estimates $\hat{\theta} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$ obtained by fitting the model (1) to data with two-level structure is defined as follows:

1. Generate independent level 2 errors for the D groups as $u_i^* \sim N(0, \hat{\sigma}_u^2)$, $i = 1, \ldots, D$ and generate independent level 1 errors for all n sampled units as $e_{ij}^* \sim N(0, \hat{\sigma}_e^2)$, $j = 1, \ldots, n_i;\ i = 1, \ldots, D$.
2. Simulate bootstrap sample data $(y_{ij}^*, \mathbf{x}_{ij})$ using the model $y_{ij}^* = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_i^* + e_{ij}^*$.

3. Fit the two-level random effects model (1) to the bootstrap sample data generated in Step (2) to obtain bootstrap parameter estimates $\hat{\theta}^* = (\hat{\boldsymbol{\beta}}^*, \hat{\sigma}_u^{2*}, \hat{\sigma}_e^{2*})$.

4. Repeat Steps 1–3 B times to obtain B sets of bootstrap parameter estimates.

As noted in Section 1, this method works very well, provided the model (1) holds. However, it is hard to justify this type of bootstrap if the stochastic assumptions of this model, for example, that the random effects are iid Gaussian random variables, are violated.

2.2 A TWO-LEVEL SEMIPARAMETRIC BOOTSTRAP

Carpenter, Goldstein, and Rasbash (2003) described a semiparametric bootstrap method, hereafter referred to as the CGR bootstrap, for multilevel data that is less sensitive to model assumptions than the parametric bootstrap. Suppose that we have estimates $\hat{\theta} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$ of $\theta = (\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2)$ as well as the corresponding EBLUPs $\hat{u}_i$ of $u_i (i = 1, \ldots, D)$. In what follows, we use the notation srswr(A, m) to indicate the outcome of taking a simple random sample of size m with replacement from the set A. The CGR bootstrap is then implemented as follows:

1. The D EBLUPs $\hat{u}_i$ of the random effects u_i and the corresponding n level 1 residuals $\hat{e}_{ij} = y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} - \hat{u}_i$ are first scaled to ensure that they have empirical variances equal to $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$, respectively. The scaled level 2 residuals are $\hat{u}_i^c = \hat{\sigma}_u \hat{u}_i \{D^{-1} \sum_{k=1}^{D} \hat{u}_k^2\}^{-1/2}$ and the scaled level 1 residuals are $\hat{e}_{ij}^c = \hat{\sigma}_e \hat{e}_{ij} \{n^{-1} \sum_{k=1}^{D} \sum_{l=1}^{n_k} \hat{e}_{kl}^2\}^{-1/2}$. Both sets of scaled residuals are then centered at zero.

2. Sample independently with replacement from $\hat{\mathbf{u}}^c = (\hat{u}_i^c)$ and $\hat{\mathbf{e}}^c = (\hat{e}_{ij}^c)$ to get bootstrap samples $\mathbf{u}^*$ and $\mathbf{e}^*$ of D level 2 residuals and n level 1 residuals, respectively. That is, $\mathbf{u}^* = (u_i^*) = \text{srswr}\{\hat{\mathbf{u}}^c, D\}$ and $\mathbf{e}^* = (e_{ij}^*) = \text{srswr}\{\hat{\mathbf{e}}^c, n\}$.

3. Simulate bootstrap sample data (y_{ij}^*, x_{ij}) using the model $y_{ij}^* = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_i^* + e_{ij}^*$.

4. Fit the two-level random effects model (1) to the bootstrap sample data generated in Step (3) to obtain bootstrap parameter estimates $\hat{\theta}^* = (\hat{\boldsymbol{\beta}}^*, \hat{\sigma}_u^{2*}, \hat{\sigma}_e^{2*})$.

5. Repeat Steps 2–4 B times to obtain B sets of bootstrap parameter estimates.

2.3 A TWO-LEVEL REB BOOTSTRAP

Although the bootstrap errors used in the CGR bootstrap are less sensitive to the stochastic assumptions of (1), they still rely on the model-based EBLUPs $\hat{u}_i$ of the level 2 random effects. In addition, both the parametric and the CGR bootstraps assume independent and identically distributed level 1 errors. In practice, this may not be the case. For example, these errors can be correlated in environmental and agricultural applications, reflecting unmeasured spatial variation. Provided the within-cluster variability is similar from cluster to cluster, we can use within-block bootstrapping to recreate this variability in our bootstrap. We therefore now describe a REB bootstrap that allows for such residual heterogeneity. This approach is semiparametric in the sense that although the marginal bootstrap model is based on the parametric fit to the sample data, the dependence structure in the model residuals is generated nonparametrically.

2.3.1 REB Bootstrap. We first describe a simple REB bootstrap for two-level data and then develop refinements to this method. The steps in the REB bootstrap are as follows:

1. Using the marginal residuals: $r_{ij} = y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}},\ j = 1, \ldots, n_i; i = 1, \ldots, D$, calculate the level 2 average residuals for each of the D groups: $\bar{r}_h = n_h^{-1}\sum_{j=1}^{n_h} r_{hj},\ h = 1, \ldots, D$, and the level 1 residuals within each group h as $r_{hj}^{(1)} = r_{hj} - \bar{r}_h,\ j = 1, \ldots, n_h;\ h = 1, \ldots, D$. Let $\bar{\mathbf{r}}^{(2)}$ and $\mathbf{r}_h^{(1)}$ denote the vector of D level 2 average residuals $\bar{r}_h$ and the vector of n_h level 1 residuals $r_{hj}^{(1)}$ for group h, respectively.
2. Sample independently and with replacement from these two sets of residuals to define bootstrap errors for levels 1 and 2. In particular, level 2 bootstrap errors are given by $\mathbf{r}^{*(2)} = (r_i^{*(2)}) = \text{srswr}(\bar{\boldsymbol{r}}^{(2)}, D)$, while level 1 bootstrap errors in cluster i are given by $\mathbf{r}_i^{*(1)} = (r_{ij}^{*(1)}) = \text{srswr}(\boldsymbol{r}_{h(i)}^{(1)}, n_i)$, where $h(i) = \text{srswr}(\{1, \ldots, D\}, 1)$.
3. Simulate bootstrap sample data $(y_{ij}^*, \mathbf{x}_{ij})$ using the model $y_{ij}^* = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + r_i^{*(2)} + r_{ij}^{*(1)}$.
4. Fit the two-level random effects model (1) to the bootstrap sample data generated in Step 3 to obtain bootstrap parameter estimates $\hat{\theta}^* = (\hat{\boldsymbol{\beta}}^*, \hat{\sigma}_u^{2*}, \hat{\sigma}_e^{2*})$.
5. Repeat Steps 2–4 B times to obtain B sets of bootstrap parameter estimates.

2.3.2 Variation 1: Use of Centered and Scaled Residuals Before REB Bootstrapping. In the REB bootstrap described above, $E^*(r_i^{*(2)}) \neq 0$ and $\text{var}^*(r_i^{*(2)}) \neq \hat{\sigma}_u^2$, where E^*, var^* denote expectation and variance, respectively, with respect to the bootstrap distribution. Consequently, $E^*(y_{ij}^*) \neq \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}$, implying that the bootstrap confidence intervals generated by the REB bootstrap are not consistent. A variation, which also satisfies the conditions for consistency (Shao and Tu 1995, chap. 4), is to zero center and scale residuals prior to their use in the bootstrap. That is, following the same procedure as used in the CGR bootstrap, the residuals $\bar{r}_h$ and $r_{hj}^{(1)}$ computed in Step 1 of the REB bootstrap are transformed to scaled values $\bar{r}_h^c = \hat{\sigma}_u \bar{r}_h \{D^{-1}\sum_{k=1}^{D} \bar{r}_k^2\}^{-1/2}$ and $r_{ij}^{(1)c} = \hat{\sigma}_e r_{ij}^{(1)} \{n^{-1}\sum_{k=1}^{D}\sum_{l=1}^{n_k} (r_{kl}^{(1)})^2\}^{-1/2}$, respectively, before they are zero centered. Steps 2–5 of the REB bootstrap are then carried out, in which case we have $E^*(r_i^{*(2)}) = E^*(r_{ij}^{*(1)}) = 0$, $\text{var}^*(r_i^{*(2)}) = \hat{\sigma}_u^2$, and $\text{var}^*(r_{ij}^{*(1)}) = \hat{\sigma}_e^2$. Zero centering and scaling residuals prior to initiating the bootstrap ensures that the confidence intervals generated by the REB bootstrap are consistent. We refer to this as a "prescaled" REB bootstrap in what follows.

Note that the prescaled REB bootstrap is the same as the CGR bootstrap, except that it uses empirical residuals instead of EBLUPs. It also bootstraps within blocks. This approach has two major advantages. First, the empirical residuals depend only on the mean structure of the model being correctly specified, while the EBLUPs depend also on correct specification of the covariance structure of the model. Second, the use of bootstrapping within blocks ensures the preservation of within-block residual variability. Consequently, we expect that the prescaled REB bootstrap will be more robust to model misspecification than the CGR bootstrap, for example.

2.3.3 Variation 2: Scaling and Bias Adjustments After REB Bootstrapping. Under Equation (1), one would intuitively expect estimators of the variance components to be uncorrelated, since they should essentially depend on the uncorrelated level 1 and level 2

errors. In particular, the asymptotic covariance between the ML/REML estimators of σ_u^2 and σ_e^2 is negligible compared with their corresponding variances under Equation (1). However, there is no guarantee that their joint bootstrap distribution under REB bootstrapping has the same property. Furthermore, although the REB bootstrap preserves within-cluster heterogeneity, there is no guarantee that the bootstrap distributions that it generates are located at the estimated parameter values. Both of these properties can be guaranteed by appropriately modifying the bootstrap distributions generated by the REB bootstrap. We therefore now describe two further steps in this procedure that ensure these properties:

- We first modify the REB bootstrap distributions of the logarithms of the variance components estimates so that they are empirically uncorrelated. The steps in this process are as follows:
  (a) Let $(\log \hat{\sigma}_u^{2*})$ and $(\log \hat{\sigma}_e^{2*})$ denote the vectors of B bootstrap values of $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$, respectively. Define the $B \times 2$ matrices

$$\begin{aligned}\mathbf{S}^* &= \left[\left(\log \hat{\sigma}_u^{2*}\right),\ \left(\log \hat{\sigma}_e^{2*}\right)\right],\\ \mathbf{M}^* &= \left[\mathrm{av}\left(\log \hat{\sigma}_u^{*2}\right) \times \mathbf{1}_B,\ \mathrm{av}\left(\log \hat{\sigma}_e^{*2}\right) \times \mathbf{1}_B\right],\\ &\text{and}\\ \mathbf{D}^* &= \left[\mathrm{sd}\left(\log \hat{\sigma}_u^{*2}\right) \times \mathbf{1}_B,\ \mathrm{sd}\left(\log \hat{\sigma}_e^{*2}\right) \times \mathbf{1}_B\right].\end{aligned}$$

  Here, av$S$ and sd$S$ denote the average and the standard deviation of the values in the vector S, $\mathbf{1}_B$ denotes a vector of ones of size B, and $\times$ denotes componentwise multiplication.
  (b) Calculate the 2×2 covariance matrix $\mathbf{C}^* = \mathrm{cov}(\mathbf{S}^*)$ and put

$$\mathbf{L}^* = \mathbf{M}^* + \{(\mathbf{S}^* - \mathbf{M}^*)\mathbf{C}^{*-1/2}\} \times \mathbf{D}^*$$

  (c) The modified bootstrap values of $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ (denoted $\hat{\sigma}_u^{*\,\mathrm{mod}\,2}$ and $\hat{\sigma}_e^{*\,\mathrm{mod}\,2}$ below) are then obtained by exponentiating the elements of $\mathbf{L}^*$.
- All bootstrap distributions of model parameter estimates (including the modified bootstrap distributions of the estimated variance components) are then centered at the original estimate values, using a mean correction for regression coefficients, that is,

$$(\hat{\beta}_k^{**}) = [\hat{\beta}_k \mathbf{1}_B + (\hat{\beta}_k^*) - \mathrm{av}(\hat{\beta}_k^*)],$$

  and a ratio correction for variance components, that is,

$$\begin{aligned}\left(\hat{\sigma}_u^{2**}\right) &= \left(\hat{\sigma}_u^{*\,\mathrm{mod}\,2}\right) \times \hat{\sigma}_u^2 \left\{\mathrm{av}\left(\hat{\sigma}_u^{*\,\mathrm{mod}\,2}\right)\right\}^{-1},\\ \left(\hat{\sigma}_e^{2**}\right) &= \left(\hat{\sigma}_e^{*\,\mathrm{mod}\,2}\right) \times \hat{\sigma}_e^2 \left\{\mathrm{av}\left(\hat{\sigma}_e^{*\,\mathrm{mod}\,2}\right)\right\}^{-1}.\end{aligned}$$

Note that we use a "**" superscript here to distinguish the values defining these adjusted bootstrap distributions from the original values generated by the REB bootstrap, which are denoted by a "*" superscript.

We refer to this as a "postscaled" REB bootstrap in what follows, and note that it represents an alternative way of modifying the REB bootstrap to ensure its consistency under a linear mixed model. Bootstrap confidence intervals as well as bootstrap distributions

for functions of model parameters (e.g., EBLUPs) are then based on these postscaled REB bootstrap distributions.

2.3.4 Fisher Consistency of Prescaled and Postscaled REB Bootstraps. The REB bootstrap and its two variations described above are covered by the random effects bootstrap framework described by Field and Welsh (2007). These authors showed that the random effects bootstrap gives asymptotically consistent results for the corresponding random effects model under joint asymptotics, that is, when the number of clusters and the number of observations in each cluster increase. Assuming certain regularity conditions, Shao, Kübler, and Pigeot (2000) showed that bootstrap percentile confidence intervals are asymptotically Fisher consistent. Carpenter, Goldstein, and Rasbash (2003) used the same arguments as in Shao, Kübler, and Pigeot (2000) to prove the corresponding consistency of CGR bootstrap percentile confidence intervals under Equation (1). This follows from showing that the bootstrap expectations of the ML estimating equations are zero. We now show that this result also holds for the prescaled and the postscaled REB bootstraps.

Consider the case of ML estimation under (1), where, up to an additive constant, the log-likelihood function is $l = (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) - \log|\mathbf{V}|$ and $\mathbf{V}$ is the variance-covariance matrix of $\mathbf{y}$. Differentiating this log-likelihood with respect to $\boldsymbol{\beta}$ leads to the ML estimating function for $\boldsymbol{\beta}$, that is, $\text{sc}(\boldsymbol{\beta}) = \mathbf{x}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})$. Since $E^*(y^* - \mathbf{x}\hat{\boldsymbol{\beta}}) = 0$ for the prescaled REB bootstrap, it follows that the expectation of this estimating function with respect to the bootstrap distribution is zero at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. This shows the consistency of $\hat{\boldsymbol{\beta}}^*$ for $\hat{\boldsymbol{\beta}}$ in this case. To demonstrate consistency of the bootstrap estimates of the variance components under the prescaled REB bootstrap, we note that $(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) = \text{tr}\{\mathbf{V}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^T\}$; see McCulloch and Searle (2001, p. 301). Put $\mathbf{R} = (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^T$. The log-likelihood function can then be expressed as $l = -\text{tr}(\mathbf{V}^{-1}\mathbf{R}) - \log|\mathbf{V}|$. The first derivative of this log-likelihood with respect to the variance components parameter $\delta = (\sigma_u^2, \sigma_e^2)$ defines their estimating function, $\text{sc}(\delta) = -\text{tr}(\frac{\partial\mathbf{V}^{-1}}{\partial\delta}\mathbf{R}) + \text{tr}(\mathbf{V}\frac{\partial\mathbf{V}^{-1}}{\partial\delta})$, so $E\{\text{sc}(\delta)\} = \mathbf{0}$. Put $\mathbf{R}^* = (\mathbf{y}^* - \mathbf{x}\hat{\boldsymbol{\beta}})(\mathbf{y}^* - \mathbf{x}\hat{\boldsymbol{\beta}})^T$ and note that $\hat{\mathbf{V}} = E^*(\mathbf{R}^*)$, where $\hat{\mathbf{V}}$ is the ML estimate of V. We then need to show that $E^*\{\text{sc}^*(\delta)\} = \mathbf{0}$, where $\text{sc}^*(\delta) = -\text{tr}(\frac{\partial\mathbf{V}^{-1}}{\partial\delta}\mathbf{R}^*) + \text{tr}(\mathbf{V}\frac{\partial\mathbf{V}^{-1}}{\partial\delta})$. This follows because

$$
\begin{aligned}
-E^*\left[\text{tr}\left(\frac{\partial\mathbf{V}^{-1}}{\partial\delta}\mathbf{R}^*\right)\right] + \text{tr}\left(\mathbf{V}\frac{\partial\mathbf{V}^{-1}}{\partial\delta}\right) &= -E^*\left[\text{tr}\left(\mathbf{R}^{*T}\frac{\partial\mathbf{V}^{-1}}{\partial\delta}\right)\right] + \text{tr}\left(\mathbf{V}\frac{\partial\mathbf{V}^{-1}}{\partial\delta}\right) \\
&= -\text{tr}\left(E^*(\mathbf{R}^{*T})\frac{\partial\mathbf{V}^{-1}}{\partial\delta}\right) + \text{tr}\left(\mathbf{V}\frac{\partial\mathbf{V}^{-1}}{\partial\delta}\right) \\
&= -\text{tr}\left(\hat{\mathbf{V}}\frac{\partial\mathbf{V}^{-1}}{\partial\delta}\right) + \text{tr}\left(\mathbf{V}\frac{\partial\mathbf{V}^{-1}}{\partial\delta}\right) \\
&\approx 0,
\end{aligned}
$$

where the last approximate equality is a consequence of the fact that $\hat{\mathbf{V}}$ and $\mathbf{V}$ are symmetric and $\hat{\mathbf{V}} \sim \mathbf{V}$, where $\sim$ denotes "asymptotically equal." That is, $\hat{\delta}^* = (\hat{\sigma}_u^{2*}, \hat{\sigma}_e^{2*})$ is consistent for $\hat{\delta} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ under prescaled REB bootstrapping. Since ML and REML estimates are asymptotically identical, these consistency properties also hold for REML estimation.

The bias adjustment step in postscaled REB bootstrapping ensures that similar consistency results hold in this case. To show this, we use a superscript of "**" to denote

postscaled bootstrap realizations, with E^{**} denoting the corresponding expectation. Then,

$$\begin{aligned} E^{**}(\mathbf{y}^{**} - \mathbf{x}\hat{\boldsymbol{\beta}}) &= E^{**}\{\mathbf{y}^{**} - \mathbf{x}\hat{\boldsymbol{\beta}}^{**} + \mathbf{x}(\hat{\boldsymbol{\beta}}^{**} - \hat{\boldsymbol{\beta}})\} \\ &= E^{*}(\mathbf{y}^{*} - \mathbf{x}\hat{\boldsymbol{\beta}}^{*}) + \mathbf{x}E^{**}(\hat{\boldsymbol{\beta}}^{**} - \hat{\boldsymbol{\beta}}) \\ &= \mathbf{0}, \end{aligned}$$

since under bias adjustment, $E^{**}(\hat{\boldsymbol{\beta}}^{**}) = \hat{\boldsymbol{\beta}}$, while the adjusted residuals $\mathbf{y}^{**} - \mathbf{x}\hat{\boldsymbol{\beta}}^{**}$ and the unadjusted residuals $\mathbf{y}^{*} - \mathbf{x}\hat{\boldsymbol{\beta}}^{*}$ are identical and $E^{*}(\mathbf{y}^{*} - \mathbf{x}\hat{\boldsymbol{\beta}}^{*}) = 0$. It immediately follows that the postscaled RBB bootstrap is consistent for $\hat{\boldsymbol{\beta}}$. To prove the corresponding consistency of this bootstrap for the estimated variance components, we show that $\hat{\mathbf{V}} \sim E^{**}(\mathbf{R}^{**})$, where $\mathbf{R}^{**} = (\mathbf{y}^{**} - \mathbf{x}\hat{\boldsymbol{\beta}})(\mathbf{y}^{**} - \mathbf{x}\hat{\boldsymbol{\beta}})^{T}$. This follows because we can write

$$\begin{aligned} E^{**}(\mathbf{R}^{**}) &= E^{**}\{\mathbf{y}^{**} - \mathbf{x}\hat{\boldsymbol{\beta}}^{**} + \mathbf{x}(\hat{\boldsymbol{\beta}}^{**} - \hat{\boldsymbol{\beta}})\}\{\mathbf{y}^{**} - \mathbf{x}\hat{\boldsymbol{\beta}}^{**} + \mathbf{x}(\hat{\boldsymbol{\beta}}^{**} - \hat{\boldsymbol{\beta}})\}^{T} \\ &= \hat{\mathbf{V}} + \mathbf{x}\mathrm{var}^{**}(\hat{\boldsymbol{\beta}}^{**})\mathbf{x}^{T}, \end{aligned}$$

where the last equality is a consequence of the fact that under bias adjustment,

$$E^{**}\{\mathbf{y}^{**} - \mathbf{x}\hat{\boldsymbol{\beta}}^{**}\}\{\mathbf{y}^{**} - \mathbf{x}\hat{\boldsymbol{\beta}}^{**}\}^{T} = \hat{\mathbf{V}},$$

and because of the independence of the bootstrap distributions of $\hat{\boldsymbol{\beta}}^{**}$ and $\mathbf{y}^{**} - \mathbf{x}\hat{\boldsymbol{\beta}}^{**}$. It only remains to note that $\mathrm{var}^{**}(\hat{\boldsymbol{\beta}}^{**}) = O(n^{-1})$.

2.3.5 Variation 3: Calibration to the Estimated Covariance Matrix of the Variance Components. By construction, the rescaling of residuals underpinning the prescaled REB bootstrap leads to bootstrap residuals with variances that are close to the corresponding variance component estimates. However, this does not mean that the covariance matrix of the bootstrap distribution of these variance components is close to the estimated asymptotic covariance matrix of the variance components estimators. This suggests that we may be able to improve on the prescaled REB bootstrap by calibrating the empirical covariance matrix of the bootstrap estimates of the variance components generated under this procedure to the ML/REML estimate of the asymptotic covariance matrix of the variance components estimators. This can be achieved by a suitable Cholesky decomposition. However, it is important to note that the performance of this second-order calibrated block bootstrap then depends on the accuracy of the estimated asymptotic covariance matrix of the variance components estimators. In the simulations reported in the next section, we observed that this extra level of calibration led to undesirable sensitivity to model assumptions. This was not unexpected since this second-order calibration depends on the model (1) being true. Results for this method are therefore not reported, but can be obtained from the authors.

3. EMPIRICAL EVALUATIONS

3.1 DESCRIPTION OF THE SIMULATION EXPERIMENTS

We now describe a series of simulation experiments that were used to evaluate the performance of the different bootstrap methods described in the previous section and that are set out in Table 1. The R code that we used to implement these different methods is available online as part of the supplementary materials for this article, and from now on, we use the

Table 1. Description of bootstrap methods used in the simulation studies

Name	Description of bootstrap method
PAR	Two-level parametric bootstrap—see Section 2.1
CGR	Carpenter, Goldstein, and Rasbash (2003) bootstrap—see Section 2.2
REB/0	REB bootstrap—see Section 2.3.1
REB/1	Prescaled REB bootstrap—see Section 2.3.2
REB/2	Postscaled REB bootstrap—see Section 2.3.3

acronyms defined in Table 1 without further comment. In the first two of these experiments, referred to as simulation sets A and B below, we used the standard random effects model (1) to generate clustered data. In particular, in both, we generated data using a two-level model of the form $y_{ij} = \alpha + \beta x_{ij} + u_i + e_{ij}, i = 1, \ldots, D; j = 1, \ldots, n_i;$ with $\alpha = 1$ and $\beta = 2$. We fixed the total number of clusters at $D = 100$ and within each experiment, simulated data for two sets of equal cluster sizes, $n_i = 5$ and $n_i = 20$. Values of x_{ij} were generated independently as $x_{ij} \sim \text{Uniform}(0, 1)$. The cluster-specific (level 2) random errors u_i and the individual-specific (level 1) random errors e_{ij} were generated as mutually independent and identically distributed random variables with zero means and with variances σ_u^2 and σ_e^2, respectively. In simulation set A, $u_i \sim N(0, \sigma_u^2 = 0.04)$ and $e_{ij} \sim N(0, \sigma_e^2 = 0.16)$. In simulation set B, we generated u_i from a χ^2 distribution with mean zero and variance $\sigma_u^2 = 0.04$ as $u_i \sim 0.2[(\chi_1^2 - 1)/\sqrt{2}]$. Similarly, we generated the individual-level errors e_{ij} independently of the cluster-level errors u_i from a χ^2 distribution with mean zero and variance $\sigma_e^2 = 0.16$ as $e_{ij} \sim 0.4[(\chi_1^2 - 1)/\sqrt{2}]$.

Note that in both set A and set B, units within a cluster are equicorrelated. Since our interest is in clustered data situations where this does not hold, we investigated an alternative to set A where the individual-level errors e_{ij} were generated so that within-cluster units are not equicorrelated. In this case, individual-level errors within a cluster were simulated so that they corresponded to a first-order autocorrelated series of the form $e_{ij} = \lambda e_{i(j-1)} + \varepsilon_{ij}, \ j = 1, \ldots, n_i,$ with $\lambda = 0.5$ and $\varepsilon_{ij} \sim N(0, 1)$. This is referred to as simulation set C below. Finally, we investigated the impact of correlation between units in different clusters in a fourth set of simulations, denoted simulation set D below, where we replicated simulation set C, except that all individual-level errors were now generated from the same first-order autocorrelated series of size $n = \sum_{i=1}^{D} n_i$ as $e_{ij} = \lambda e_{i(j-1)} + \varepsilon_{ij}, \ j = 1, \ldots, n$. This simulation therefore approximates the type of time series problem that motivated the development of the block bootstrap.

A total of $R = 1000$ Monte Carlo simulations were carried out for each simulation set, and within each simulation, we implemented each of the bootstrap methods set out in Table 1 using $B = 1000$ bootstrap replicates. This number of replications is suitable for evaluating percentile-type 95% confidence intervals; see Caers, Beirlant, and Vynckier (1998).

3.2 DISCUSSION OF SIMULATION RESULTS

Average coverage rates of nominal 95% bootstrap confidence intervals for the various model parameters were obtained for the different simulation sets. These coverage rates are

Table 2. Average coverage rates of nominal 95% bootstrap confidence intervals for model parameters, simulation sets *A–D*

Method	$n_i = 5$				$n_i = 20$			
	α	β	σ_u^2	σ_e^2	α	β	σ_u^2	σ_e^2
				Set *A*				
PAR	0.95	0.94	0.95	0.95	0.96	0.95	0.93	0.95
CGR	0.91	0.95	0.95	0.94	0.81	0.95	0.94	0.94
REB/0	0.95	0.93	0.23	0.26	0.97	0.95	0.89	0.86
REB/1	0.95	0.95	0.96	0.98	0.96	0.95	0.94	0.99
REB/2	0.95	0.93	0.86	0.99	0.97	0.95	0.94	0.99
				Set *B*				
PAR	0.95	0.95	0.77	0.59	0.94	0.95	0.61	0.54
CGR	0.93	0.95	0.84	0.87	0.81	0.95	0.82	0.86
REB/0	0.96	0.93	0.64	0.72	0.95	0.95	0.94	0.96
REB/1	0.96	0.95	0.94	0.96	0.94	0.95	0.82	0.99
REB/2	0.96	0.93	0.82	0.97	0.95	0.95	0.83	0.99
				Set *C*				
PAR	0.91	0.96	0.95	0.77	0.94	0.93	0.95	0.82
CGR	0.88	0.94	0.95	0.76	0.77	0.95	0.94	0.83
REB/0	0.93	0.92	0.28	0.08	0.95	0.94	0.90	0.64
REB/1	0.92	0.94	0.96	0.89	0.93	0.95	0.94	0.95
REB/2	0.93	0.92	0.88	0.93	0.95	0.94	0.93	0.97
				Set *D*				
PAR	0.90	0.96	0.94	0.88	0.94	0.95	0.94	0.88
CGR	0.87	0.94	0.94	0.86	0.75	0.95	0.95	0.87
REB/0	0.91	0.92	0.24	0.27	0.94	0.93	0.90	0.78
REB/1	0.91	0.94	0.96	0.96	0.93	0.94	0.94	0.95
REB/2	0.91	0.92	0.85	0.96	0.94	0.93	0.94	0.95

reported in Table 2. It is clear that there is not much difference in the coverage rates for the regression coefficients (i.e., α and β) between the different bootstrap methods and between the different simulation sets, with the notable exception that CGR recorded low coverage for α in our large cluster size ($n_i = 20$) simulations, indicating a potential bias problem with our implementation of this method.

It is well known that classical inference for the variance component parameters of (1) is sensitive to deviations from this model. As a consequence, we now focus on bootstrap coverage performance for the variance component parameters σ_u^2 and σ_e^2. In simulation set *A*, the assumed model is true, and so, the coverage rates of PAR and CGR are around 95%. In contrast, REB/0 records low coverage, especially for small ($n_i = 5$) cluster sizes. This is effectively corrected by REB/1 and REB/2, although there is evidence that for small cluster sizes, REB/1 is more effective than REB/2.

Turning to results from simulation set *B*, we observe that PAR fails. The performance of CGR is better, but is still unsatisfactory. In contrast, although REB/0 remains unsatisfactory for small cluster sizes, its performance for large cluster sizes is good. This performance is reversed for REB/1, which performs better for small cluster sizes than for large cluster sizes. Cluster size does not seem to impact as much on REB/2, which performs adequately and seems better than CGR in this simulation.

The performances of the different bootstrap methods in sets C and D were qualitatively similar to those recorded for sets A and B. REB/0 fails when cluster sizes are small and recovers somewhat as the cluster size increases. The performances of both PAR and CGR are similar, as are those of REB/1 and REB/2, with REB/1 the better performer for small cluster sizes. Overall, REB/1 appears to be the best performing of the five bootstrap methods that we investigated, with REB/2 a little behind. Both these bootstrap methods seem robust to the departures from model assumptions that we considered in our simulations.

Although we do not present these results here, we also carried out a number of simulation studies that examined the performance of the bootstrap methods set out in Table 1 in other situations, all of which have some relevance to real-life data:

- We replicated simulation sets A–C with a smaller number of clusters, that is, $D = 50$, and noted that the relative performances of the different bootstrap methods were almost identical to those observed when $D = 100$.
- We examined the impact of misspecification of the cluster structure in the block bootstrap by replicating simulation set A with data generation and model fitting based on $D = 100$ clusters, but with bootstrap data generated using a smaller number $D = 50$ clusters. This did not change the behavior of the block bootstrap methods.
- We also examined the impact of varying cluster sizes by replicating simulation set A with cluster sizes ranging from 1 to 100, with about half the clusters having 10 or fewer observations. Again, REB/1 and REB/2 performed satisfactorily.

4. APPLICATION TO ENVIRONMENTAL DATA MODELING

In this section, we apply the different bootstrap methods set out in Table 1 to the environmental data analyzed by Beare et al. (2011). These data are available online as part of the supplementary materials for this article. They consist of $n = 3177$ values of positive daily rainfall measured at a group of rain gauges over a period of approximately 4 months, together with the values of 37 covariates measuring daily meteorological conditions as well as the spatial characteristics of the different gauges. The data were collected as part of a trial of the effect of two ground-based cloud ionizing devices on downwind rainfall, so the covariates include measurements relating to the daily operational status of the two devices as well as the distance and downwind orientation of a gauge relative to each device on a day. Since the hypothesized impact of these devices is to enhance downwind rainfall, it is necessary to include terms in the model for observed rainfall that allow for the natural variation in rainfall due to the spatial and temporal nonhomogeneity of rain cloud movement over the target downwind area. In the analysis described by Beare et al. (2011), this was done by including random effects for 397 spatiotemporal clusters in the rainfall model, where these clusters were defined by groups of gauges that had similar relative orientations to the two devices on a day. The distribution of these spatiotemporal cluster sizes can be observed in Figure 1, and we note that they vary from minimum of 1 gauge to maximum of 57 gauges, with an average size of 8 gauges.

A more conservative approach to defining cluster random effects for these rain data is based on the fact that the random sequence used to control the operation of the two

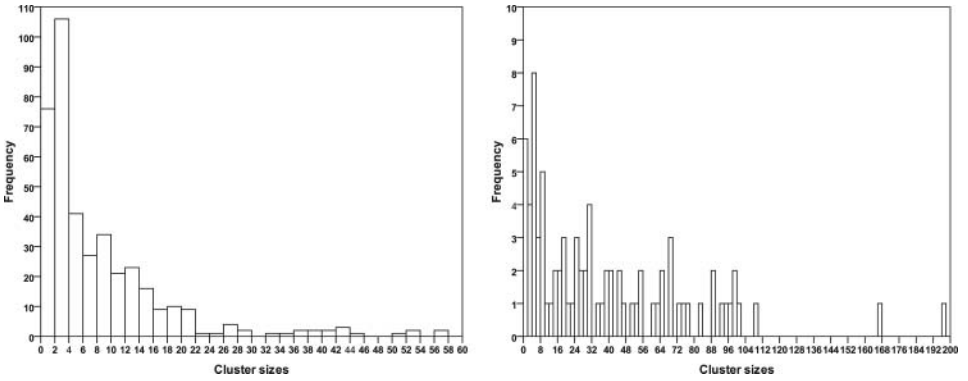

Figure 1. Distributions of cluster sizes for the environmental dataset, with spatiotemporal clusters on the left and 4-day downwind clusters on the right.

devices was essentially made up of nonoverlapping experimental "blocks," each consisting of a sequence of four trial days. Assuming that there could be significant unexplained between-block and between-device heterogeneity in rainfall then leads one to consider random effects based on clustering gauge-day rainfall measurements by both 4-day block and whether the rainfall measurement is for a gauge that is downwind of only one of the devices or downwind of both. We refer to these clusters as 4-day clusters in what follows. There are 83 such clusters in the data, and the distribution of their corresponding sizes is shown in Figure 1. Note that these sizes range from 1 to 197, with an average of 38.

The next issue that needs to be addressed is the scale at which the daily rainfall data is modeled using Equation (1). Clearly, we can fit this model to the actual rainfall values. However, given that rainfall measurements are strictly positive and heavily skewed, an obvious alternative is use Equation (1) as a model for the logarithm of rainfall. The marginal distributions of daily rainfall on the raw scale and on the log scale are shown in Figure 2. The apparently discrete nature of the distribution of log rainfall for small values of this variable evident in Figure 2 is because rainfall in gauges is measured in increments of

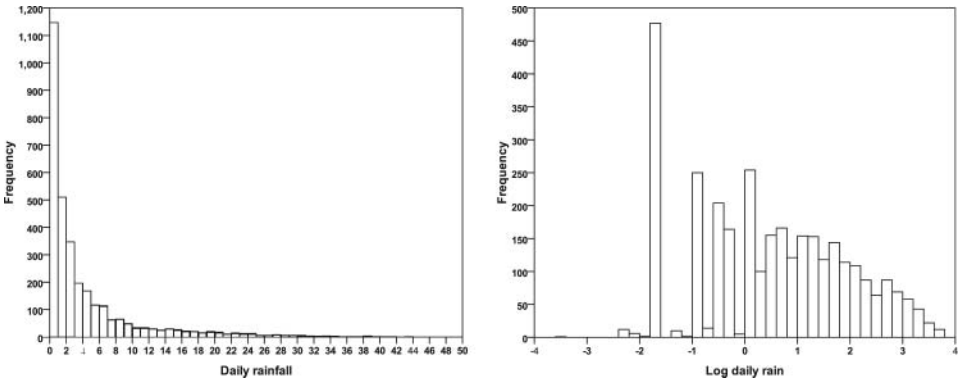

Figure 2. Distribution of daily rainfall for the environmental dataset—raw scale (left) and log scale (right).

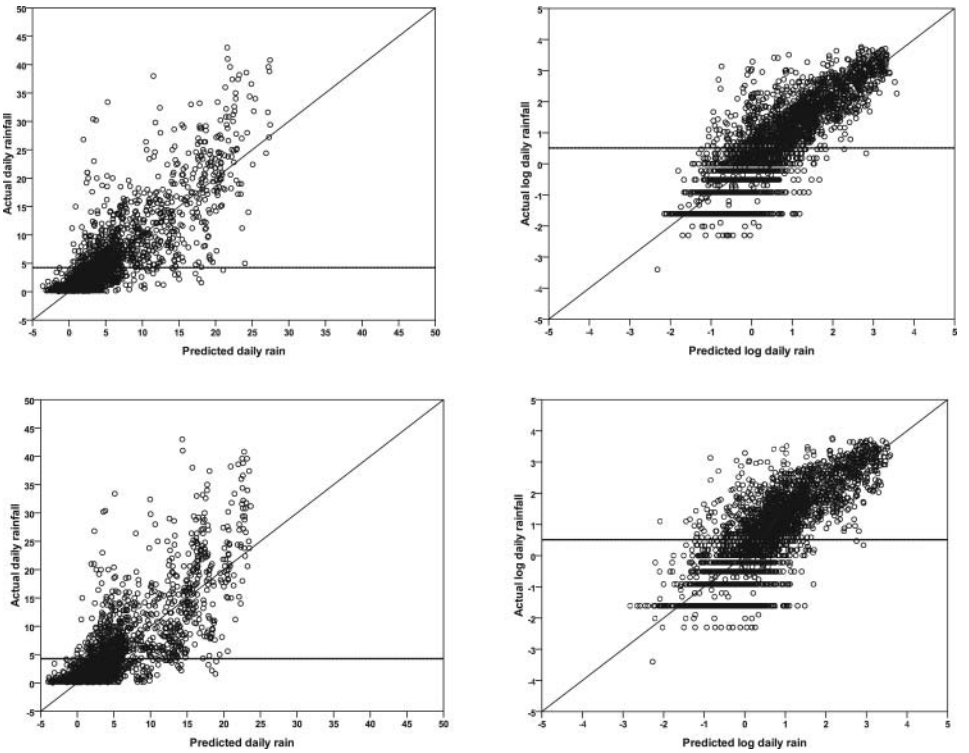

Figure 3. Plots of actual versus predicted values for linear mixed model fitted to the environmental dataset using daily rainfall values (left column) and log daily rainfall values (right column). Solid line is $y = x$ line and dotted line is average value of y. Top row corresponds to model with spatiotemporal clusters, while bottom row corresponds to model with 4-day downwind clusters.

0.2 mm. Figure 3 allows one to compare the predicted values (i.e., fitted values for fixed effects plus predicted random effects) generated by fitting (1) to both raw rainfall and log rainfall using both spatiotemporal clusters and 4-day clusters, with a REML fit used in all cases. The inadequacy of the raw scale fits is obvious. Also, the spatiotemporal log-scale model seems preferable since $-2 \times$ the log-likelihood values for the log-scale model based on the 4-day clusters and the spatiotemporal clusters were 8570 and 8348, respectively, with the corresponding EBLUP fitted values generating R^2 values of 64% and 71%.

In Section 3, we noted that the block bootstrap methods REB/1 and REB/2 should be robust to the assumption that level 1 and level 2 errors in Equation (1) are iid Gaussian variables. Although this assumption may be reasonable when (1) is fitted to log rainfall, it is clear from Figure 3 that it is hard to justify when (1) is fitted using raw rainfall values. We therefore examine the application of bootstrap methods to the rainfall data under both types of clustering as well as when (1) is fitted to raw rainfall and to log rainfall. This leads to four sets of analyses. These are reported in Table 3 and in Figures 4 and 5.

Our initial analysis focused on comparing the bootstrap tests of significance for the fixed effects in the model, where we decided that an effect is significant if its 95% confidence interval does not include zero. In no case did we observe a situation where the standard

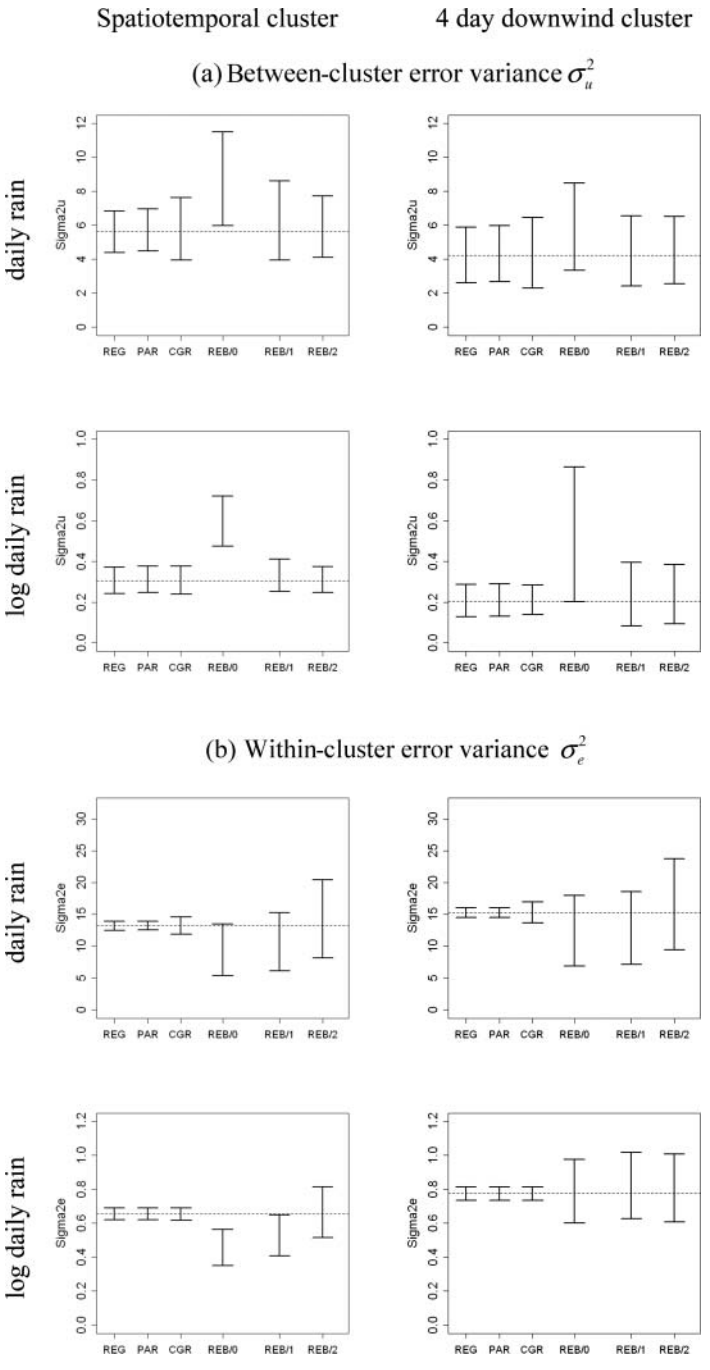

Figure 4. Nominal 95% confidence intervals for variance components for the environmental dataset. Horizontal line in each plot is the estimated value of parameter: (a) between-cluster error variance σ_u^2, and (b) within-cluster error variance σ_e^2.

parametric test (i.e., one based on the asymptotic REML-based confidence interval) led to a different conclusion about significance compared with any of the bootstrap tests. This is consistent with the results that we obtained in our simulations, and so we do not show them here. They can be obtained from the authors on request.

However, we did observe substantial differences between the different bootstrap methods as far as inferences about the variance components in the model are concerned. Table 3 shows the estimated standard errors and Figure 4 shows the associated 95% confidence intervals for these components generated by the different bootstrap methods under the

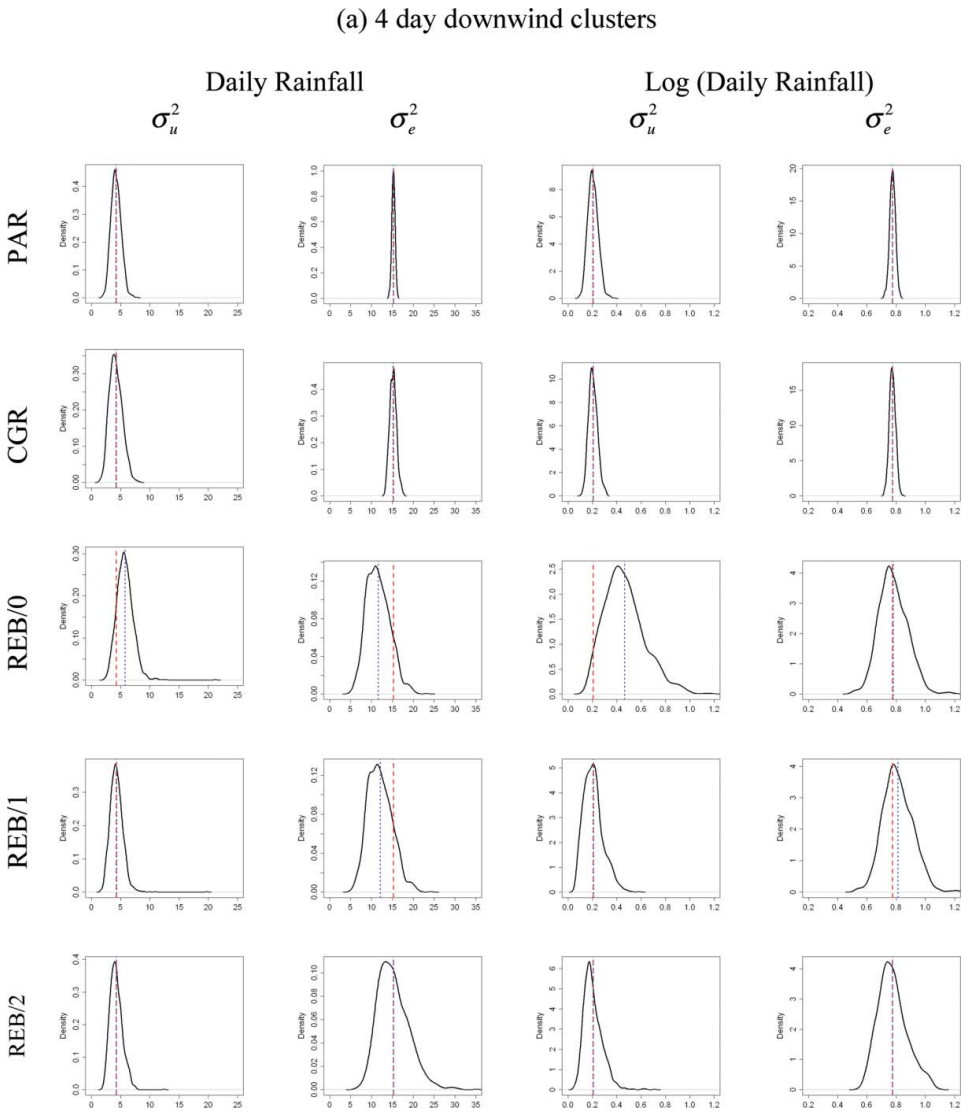

Figure 5. Bootstrap distributions of estimates of variance components for the environmental dataset. Dashed line shows the value of the estimate and the dotted line shows the mean of the bootstrap distribution: (a) 4-day downwind clusters, and (b) spatiotemporal clusters. The online version of this figure is in color. *(Continued)*

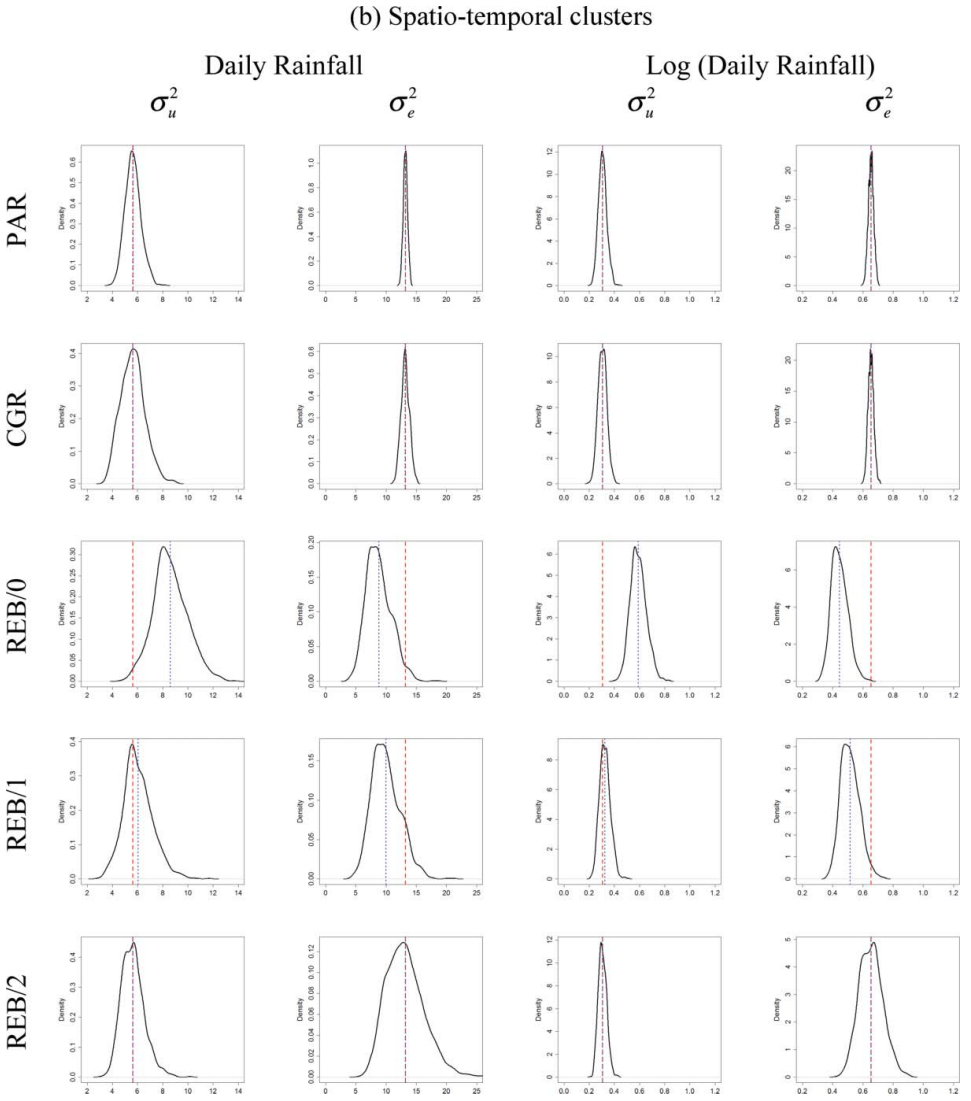

Figure 5. The online version of this figure is in color. *(Continued)*

four different model specifications. The corresponding bootstrap sampling distributions for these variance components under these model specifications are shown in Figure 5. We observe that the estimated standard errors generated by the REML fit of Equation (1) (denoted REG) and those generated by PAR are very close (see Table 3). The estimated standard errors generated by CGR are also very close to those generated by PAR and by REG when the model is fitted on the log scale. When the model is fitted on the raw scale, these estimated standard errors are larger. However, in all cases, the estimated standard errors generated by the REB bootstrap methods are much larger (often more than twice as large) than those generated by CGR, PAR, and REG. Since there is considerable doubt about (1) as a model for actual rainfall values, plus concern about the validity of the

Table 3. Bootstrap estimates of standard errors for estimates of variance components for the environmental dataset. Note that REG is the estimated asymptotic standard error produced under REML

Cluster	Model	Parameter	Estimate	Estimated standard error					
				REG	PAR	CGR	REB/0	REB/1	REB/2
Spatiotemporal	Rain	σ_u^2	5.622	0.610	0.625	0.957	1.366	1.176	0.915
		σ_e^2	13.207	0.352	0.352	0.701	2.081	2.357	3.172
	Log	σ_u^2	0.306	0.032	0.033	0.035	0.063	0.043	0.033
	Rain	σ_e^2	0.654	0.017	0.017	0.018	0.054	0.062	0.078
4-day downwind	Rain	σ_u^2	4.246	0.815	0.862	1.104	1.460	1.202	1.064
		σ_e^2	15.269	0.388	0.389	0.823	2.800	2.902	3.801
	Log	σ_u^2	0.206	0.040	0.042	0.036	0.166	0.081	0.078
	Rain	σ_e^2	0.775	0.020	0.020	0.021	0.099	0.103	0.099

homogeneous random effects assumptions when (1) is fitted on the log scale, these results imply that the more conservative estimated standard errors generated by the prescaled and postscaled REB bootstrap methods may be preferable. This conclusion is reinforced by the confidence intervals displayed in Figure 4. These show that the intervals defined by REG, PAR, and CGR are qualitatively very similar, and typically narrower than those generated by REB/1 and REB/2. These also show that the intervals generated by the unmodified REB bootstrap (REB/0) tend to be biased upward in the case of σ_u^2 and biased downward in the case of σ_e^2. Of more concern, however, is the extreme narrowness of the intervals for σ_e^2 generated by REG, PAR, and CGR. This concern is reinforced when we examine the bootstrap distributions for these methods shown in Figure 5, which indicate unwarranted precision as far as estimation of the variance components in the model is concerned. In contrast, the bootstrap distributions generated by REB/1 and REB/2 appear more realistic. These conclusions are consistent with the conclusions drawn from the simulations described in Section 3, where we noted that in the case of nonnormal data, both PAR and CGR lead to undercoverage, while both REB/1 and REB/2 lead to intervals with coverage that is much closer to nominal levels.

5. CONCLUDING REMARKS

Our aim in this article has been to describe and to evaluate an alternative REB bootstrap method for clustered data. The method itself is described in Section 2.3 and, when used with modified semiparametric level 1 and level 2 residuals (REB/1) or with additional postbootstrap processing (REB/2), appears to be a simple and robust alternative to the model-dependent bootstrap methods for clustered data that are presently available in the literature. Our simulations provide some evidence that, given that the first-order structure of the underlying linear mixed model is adequately specified, both REB/1 and REB/2 account for within-cluster heterogeneity as well as between-cluster dependence. This good performance is demonstrated in the application to an environmental dataset in Section 3, where we observe that it was only these REB bootstrap methods that provided realistic results across all four modeling scenarios that we investigated.

Extension of REB/1 and REB/2 to versions of the linear mixed model (1) that include random slope parameters is straightforward. Let $\mathbf{z}_{ij}$ be a $q \times 1$ vector of group-level covariates for unit j in cluster i and replace model (1) by

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T u_i + e_{ij}. \tag{2}$$

The only change to the REB bootstrap that is required in this case is the definition of the level 2 average residual $\bar{r}_h$ for cluster h. This can be replaced by the $q \times 1$ vector of level 2 average residuals for group h: $\mathbf{r}_h^{(2)} = (\mathbf{z}_h^T \mathbf{z}_h)^{-1} \mathbf{z}_h^T \mathbf{r}_h$, $h = 1, \ldots, D$, where $\mathbf{z}_h$ is the $n_h \times q$ matrix of $\mathbf{z}_{ij}$ for group h and $\mathbf{r}_h$ is the $n_h \times 1$ vector of marginal residuals. Investigation of the empirical performance of this extension is currently under way, as is research into extending REB/1 and REB/2 to generalized linear mixed models, and to M-quantile-based alternatives (Chambers and Tzavidis 2006) to (2) above.

SUPPLEMENTARY MATERIALS

R code: The supplementary materials for this article include the R code that can be used to implement the various bootstrap methods in Table 1 (PAR, CGR, REB/0, REB/1, and REB/2). See the .txt file JCGS 11–050 R Code.

Application data: The environmental data used for the application described in Section 4 are included in the supplementary materials for this article. See the .txt file JCGS 11–050 Example Data. Note that the data are in csv format, one observation per line, preceded by a description of the different variables making up this dataset.

ACKNOWLEDGMENTS

We thank two anonymous referees for their constructive comments. They have led to a substantially improved final version of this article.

[Received April 2011. Revised January 2012.]

REFERENCES

Beare, S., Chambers, R., Peak, S., and Ring, J. M. (2011), "Accounting for Spatiotemporal Variation of Rainfall Measurements When Evaluating Ground-Based Methods of Weather Modification," *Journal of Weather Modification*, 43, 44–63. [462]

Caers, J., Beirlant, J., and Vynckier, P. (1998), "Bootstrap Confidence Intervals for Tail Indices," *Computational Statistics and Data Analysis*, 26, 259–277. [460]

Canty, A. J., and Davison, A. C. (1999), "Resampling-Based Variance Estimation for Labour Force Surveys," *Statistician*, 48, 379–391. [453]

Carpenter, J. R., Goldstein, H., and Rasbash, J. (2003), "A Novel Bootstrap Procedure for Assessing the Relationship Between Class Size and Achievement," *Journal of the Royal Statistical Society,* Series C, 52, 431–443. [453,455,458,460]

Chambers, R., and Tzavidis, N. (2006), "M-Quantile Models for Small Area Estimation," *Biometrika*, 93, 255–268. [469]

Clark, R. G., and Allingham, S. (2011), "Robust Resampling Confidence Intervals for Empirical Variograms," *Mathematical Geosciences*, 43, 243–259. [453]

Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge: Cambridge University Press. [453,454]

DiCiccio, T. J., and Efron, B. (1996), "Bootstrap Confidence Intervals," *Statistical Science*, 11, 189–212. [454]

DiCiccio, T. J., and Romano, J. P. (1988), "A Review of Bootstrap Confidence Intervals," *Journal of the Royal Statistical Society*, Series B, 50, 338–354. [454]

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26. [452]

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall. [454]

Field, C. A., and Welsh, A. H. (2007), "Bootstrapping Clustered Data," *Journal of the Royal Statistical Society,* Series B, 69, 369–390. [453,458]

Hall, P., Horowitz, J. L., and Jing, B. (1995), "On Blocking Rule for the Bootstrap With Dependent Data," *Biometrika*, 82, 561–574. [452,453]

Harville, D. (1977), "The Use of Linear-Model Methodology to Rate High School or College Football Teams," *Journal of the American Statistical Association*, 72, 278–289. [454]

Hutson, A. D. (2004), "A Semiparametric Bootstrap Approach to Correlated Data Analysis Problems," *Computer Methods and Programs in Biomedicine*, 73, 129–134. [453]

McCulloch, C. E., and Searle, R. S. (2001), *Generalized, Linear, and Mixed Models*, New York: Wiley. [458]

Rao, J. N. K., and Wu, C. F. J. (1988), "Resampling Inference With Complex Survey Data," *Journal of the American Statistical Association*, 83, 231–241. [453]

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., and Lewis, T. (2000), *A User's Guide to MLwiN,* London: Institute of Education. [452]

Shao, J., Kübler, J., and Pigeot, I. (2000), "Consistency of the Bootstrap Procedure in Individual Bioequivalence," *Biometrika*, 87, 573–585. [458]

Shao, J., and Tu, D. (1995), *The Jackknife and Bootstrap*, New York: Springer. [456]