*J. R. Statist. Soc.* B (2006)
**68**, *Part* 2, *pp.* 221–238

# On parametric bootstrap methods for small area prediction

Peter Hall

*Australian National University, Canberra, Australia*

and Tapabrata Maiti

*Australian National University, Canberra, Australia, and Iowa State University, Ames, USA*

[Received February 2005. Revised October 2005]

**Summary.** The particularly wide range of applications of small area prediction, e.g. in policy making decisions, has meant that this topic has received substantial attention in recent years. The problems of estimating mean-squared predictive error, of correcting that estimator for bias and of constructing prediction intervals have been addressed by various workers, although existing methodology is still restricted to a narrow range of models. To overcome this difficulty we develop new, bootstrap-based methods, which are applicable in very general settings, for constructing bias-corrected estimators of mean-squared error and for computing prediction regions. Unlike existing techniques, which are based largely on Taylor expansions, our bias-corrected mean-squared error estimators do not require analytical calculation. They also have the property that they are non-negative. Our prediction intervals have a high degree of coverage accuracy, $O(n^{-3})$, where $n$ is the number of areas, if double-bootstrap methods are employed. The techniques do not depend on the form of the small area estimator and are applicable to general two-level, small area models, where the variables at either level can be discrete or continuous and, in particular, can be non-normal. Most importantly, the new methods are simple and easy to apply.

*Keywords*: Best linear unbiased predictor; Best predictor; Bootstrap; Double bootstrap; Mean-squared predictive error; Mixed effects; Prediction interval; Two-stage estimation

## 1. Introduction

Growing demand from public and private agencies has meant that small area estimation has become a particularly important field. However, the stochastic variability of small area estimators, which are based solely on data from the given area, can be unduly large. Partly as a result, the estimation of mean-squared error (MSE), and the correction of bias of those estimators, has become a centre-piece of small area inference.

In particular, there is strong demand for reliable, simple-to-apply methods for MSE estimation. See Rao (2003), and references therein, for detailed discussion. In this paper we propose bootstrap methods for MSE estimation. The new techniques are at least as accurate as existing ones, although valid in substantially more general settings, and do not require derivation of analytical expansions. Additionally the methods give non-negative, bias-corrected estimators of MSE. Moreover, the new approach is readily adapted to the problem of constructing prediction intervals, where third-order accuracy is achieved and the technique appears to have no real competitors.

*Address for correspondence*: Peter Hall, Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia.
E-mail: halpstat@maths.anu.edu.au

 1369–7412/06/68221

Resampling methods arguably have their genesis in survey sampling, e.g. through work of Hubback and Mahalanobis in India and Gurney and McCarthy in the USA, up to 80 years ago. More recent resampling contributions have been surveyed by Lahiri (2003a), who gave an authoritative account of the bootstrap for small area inference, and Lahiri (2003b), who reviewed jackknife methods in the small area estimation problem. Much of the existing small area bootstrap work, e.g. that of Butar Butar and Lahiri (2003) and Meza (2003), is founded on empirical Bayes arguments, whereas the approach that is suggested in this paper is distinctly frequentist. Of course, Bayesian methods have very important contributions to make; see Arora and Lahiri (1997), for example. Jiang (2003) gave a more extensive review of Bayesian approaches and developed frequentist, although not bootstrap, methodology which parallels existing hierarchical Bayes methods.

Standard small area models usually have two levels. The first describes sampling and, the second, the population. Methods for inference also have two steps. In the first, the predictor involves unknown parameters. In the second, these are replaced by suitable estimators. The final form is commonly known as an empirical predictor. Two approaches, giving respectively the empirical values of the best linear unbiased predictor (BLUP) and of the best predictor (BP), are in common usage. We shall consider both approaches, and discuss related work below.

Kackar and Harville (1984) studied approximations to the MSE of the empirical BLUP, assuming normality at both stages. Datta and Lahiri (2000) gave second-order approximations to mean-squared prediction errors and their estimators, also in the important case of mixed linear normal models. Prasad and Rao (1990) pointed out that if unknown model parameters are replaced by their estimators then significant underestimation of the true MSE can result. This difficulty can have significant impact on policy making.

To alleviate the problem, Prasad and Rao (1990) constructed second-order correct MSE estimators of the empirical BLUP under normal models. Various extensions of Prasad–Rao-type MSE estimators are available; see Rao (2003). Jiang *et al.* (2002) proposed a jackknife-based bias correction of the MSE estimator, based on an empirical BP. Here, normality is not required. Lahiri and Rao (1995) showed that the Prasad–Rao estimator of predictive MSE, in the Fay–Herriot model, can be robust with respect to non-normality of the distribution of small area effects.

To elucidate the differences between bias correction methods, such as those of Prasad and Rao (1990), Jiang *et al.* (2002) and ourselves, we briefly outline their construction. The MSE of a predictor can generally be written, up to terms of second order, as

$$M(\delta) = M_1(\delta) + n^{-1} M_2(\delta),$$

where $M_1$ and $M_2$ are smooth functions, $\delta$ is a vector of unknown parameters and $n$ denotes the number of areas or number of clusters. Replacing $\delta$ by an estimator $\hat{\delta}$, we obtain a naïve estimator of MSE, $M(\hat{\delta})$, which has expected value $M(\delta) + n^{-1} M_3(\delta) + O(n^{-2})$, where $M_3(\delta)$ denotes the coefficient of the term of order $n^{-1}$ in an expansion of $E\{M_1(\hat{\delta})\}$.

If we have an explicit formula for $M_3$ then we may correct the estimator $M(\hat{\delta})$ to $M(\hat{\delta}) - n^{-1} M_3(\hat{\delta})$, the expected value of which generally equals $M(\delta) + O(n^{-2})$. This is the approach that was taken by Prasad and Rao (1990). Booth and Hobert (1998) applied similar techniques for estimating conditional MSEs. However, the functions $M_2$ and $M_3$ must be rederived each time that a new model is used. There is also a potential for $M(\hat{\delta}) - n^{-1} M_3(\hat{\delta})$ to be negative, although not in the normal error case that was studied by Prasad and Rao (1990).

The method of Jiang *et al.* (2002) applies jackknife bias correction to $M_1(\hat{\delta})$ to adjust for the term $n^{-1} M_3(\delta)$. The jackknife method is readily accessible only for the BP, not for the BLUP, and often substantially inflates the variance. By way of contrast, our approach avoids the traditional decomposition $M = M_1 + n^{-1} M_2$. Instead we use bootstrap techniques to estimate MSE, to estimate the bias of that estimator and to correct for bias. We show how to do this in such a way that the bias-corrected estimator suffers relatively little from increased variance.

There is a significant literature on bootstrap methods for prediction, although not in the two-stage, small area, parametric context of the present paper. The majority of contributions to bootstrap prediction have been to nonparametric settings. Recent work includes that of Alonso *et al.* (2003), Sjöstedt-de Luna and Young (2003) and Kim (2004a, b).

## 2. Methodology

### *2.1. Model for two-stage sampling*

Let $Q(\mu,\xi)$ and $R(\mu,\eta)$ represent univariate distributions that are determined by a scalar parameter $\mu$, denoting expected value in each case, and by other, possibly vector, parameters $\xi$ and $\eta$. For an integer $n_i \geqslant 1$, denote by $f_i(\beta)$ a known smooth function of explanatory variables $X_{i1},\ldots,X_{in_i}$ and the vector $\beta$. For example, we might have

$$f_i(\beta) = c_i(X_{i1},\ldots,X_{in_i})^{\mathrm{T}}\beta,$$

where $c_i$ is a smooth, $q_1$-variate function of $n_i$ $q_2$-vectors $X_{ij}$ and $\beta$ is a $q_1$-vector. An example of $c_i$ that arises commonly in practice is that for which $q_1 = q_2$ and

$$c_i(x_1,\ldots,x_{n_i}) = n_i^{-1}\sum_j x_j.$$

Data pairs $(X_{ij}, Y_{ij})$, for $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant n_i$, with $Y_{ij}$ a scalar, are observed and are generated as follows. Random variables $\Theta_i$ are drawn from the distribution $Q\{f_i(\beta),\xi\}$, where $\beta$ is a $q_1$-vector. Given the values of $X_i = (X_{i1},\ldots,X_{in_i})$ and $\Theta_i$, $Y_{ij}$ for $1 \leqslant j \leqslant n_i$ are independent and drawn from $R\{\psi(\Theta_i),\eta_i\}$, where $\psi$ is a known link function. The parameters $\beta$ and $\xi$ are unknown. In many instances $\eta_i$ is known, typically as a known function of $X_i$, although in general $\eta_i$ would consist of both known and unknown components, the latter being the same for each $i$. Inference is conducted conditionally on the set $\mathcal{X} = \{X_{ij} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n_i\}$.

We may write

$$Y_{ij} = \psi\{f_i(\beta) + U_i\} + V_{ij}, \tag{2.1}$$

where $U_i = \Theta_i - f_i(\beta)$ and $V_{ij} = Y_{ij} - \psi(\Theta_i)$. The variable $U_i$ has zero mean, and, conditional on $U_i$, $V_{ij}$ has zero mean.

### *2.2. Specific examples of models*

We give five examples of popular models that are commonly used in applications.

#### *2.2.1. Fay–Herriot model*

In the Fay–Herriot model (model 1) $Q(\mu,\xi)$ and $R(\mu,\eta_i)$ are normal, with means $\mu$ and respective variances $\xi$ and $\eta_i$. The value of $\eta_i$ is known. Usually, each $n_i = 1$, $f_i(\beta) = X_i^{\mathrm{T}}\beta$ and $\psi$ is the identity function. Fay and Herriot (1979) employed this model for estimating *per capita* income for regions in the USA with populations less than 1000. Later the model became very popular for other applications.

*2.2.2. Battese–Harter–Fuller model*
In the Battese–Harter–Fuller setting (model 2), $\psi$, $Q(\mu,\xi)$ and $R(\mu,\eta_i)$ are as in model 1, although $n_i$ is allowed to exceed 1, and $\eta_i$ is assumed not to depend on $i$. The mean of $Q(\mu,\xi)$ is taken to be $\mu=f_i(\beta)=\bar{X}_i^{\mathrm{T}}\beta$. Battese *et al.* (1988) used this model to estimate crop areas in north-central Iowa counties.

*2.2.3. You–Rao model*
In the You–Rao model (model 3) $Q$ is log-linear and, in particular, $\log(\Theta_i)$ has the $N(X_i^{\mathrm{T}}\beta,\xi)$ distribution. As in model 1, the distribution $R(\mu,\eta_i)$ is normal $N(\mu,\eta_i)$ and the $\eta_i$s are known. However, $\psi$ is now a logarithmic link function. You and Rao (2002) called this the 'mismatched model' and used it to estimate undercoverage in the 1991 Canadian census.

*2.2.4. Logistic regression*
In logistic regression (model 4) $Q(\mu,\xi)$ is normal, $f_i(\beta)=X_i^{\mathrm{T}}\beta$, the distribution $R$ is binomial $\mathrm{Bi}(n_i,p_i)$ and $\psi$ is the logit link function. In particular, $\Theta_i=\mathrm{logit}(p_i)$ is drawn from the $N(X_i^{\mathrm{T}}\beta,\xi)$ distribution.

*2.2.5. Poisson regression*
In the Poisson regression case (model 5) the $\log(\Theta_i)$s are drawn from the $N(X_i^{\mathrm{T}}\beta,\xi)$ distribution. Given $\Theta_i$, $Y_i$ comes from the Poisson distribution with mean $\zeta_i\Theta_i$, where the $\zeta_i$s are offset parameters and the $\Theta_i$s represent relative risks. See Clayton and Kaldor (1987). The function $\psi$ is a logarithmic link. When the $\Theta_i$s are gamma distributed, the term 'conjugate' regression is sometimes used. See, for example, Lahiri and Maiti (2002).

## 2.3. Parameter estimation
Write $\eta'$ for the common unknown part of each vector $\eta_i$. If $\psi(x)\equiv x$ then $\beta$ can be estimated root $n$ consistently by least squares, noting formula (2.1), and $\xi$ and $\eta'$ can be estimated by moment methods. Maximum likelihood can also be employed, as follows. Let $q(\cdot|\mu,\xi)$ and $r(\cdot|\mu,\eta)$ denote the densities corresponding to the distributions $Q(\mu,\xi)$ and $R(\mu,\eta)$ respectively. Assuming that $\psi(x)\equiv x$, and conditioning on $\mathcal{X}$, the joint density of $W_{ij}=U_i+V_{ij}$, for $1\leqslant j\leqslant n_i$, is

$$\begin{aligned}
\phi_i(w_{i1},\ldots,w_{in_i}|\beta,\xi,\eta')
&=\prod_{j=1}^{n_i}\int q\{w_{ij}-t+f_i(\beta)|f_i(\beta),\xi\}\,r[t+\{w_{ij}-t+f_i(\beta)\}|w_{ij}-t+f_i(\beta),\eta_i]\,\mathrm{d}t\\
&=\prod_{j=1}^{n_i}\int q\{w_{ij}-t+f_i(\beta)|f_i(\beta),\xi\}\,r\{w_{ij}+f_i(\beta)|w_{ij}-t+f_i(\beta),\eta_i\}\,\mathrm{d}t.
\end{aligned}$$

In this notation the likelihood of $Y_1,\ldots,Y_n$, given $\mathcal{X}$, is

$$\begin{aligned}
L(\beta,\xi,\eta')&=\prod_{i=1}^{n}\phi_i\{Y_{i1}-f_i(\beta),\ldots,Y_{in_i}-f_{n_i}(\beta)|\beta,\xi\}\\
&=\prod_{i=1}^{n}\prod_{j=1}^{n_i}\int q\{Y_{ij}-t|f_i(\beta),\xi\}\,r(Y_{ij}|t,\eta_i)\,\mathrm{d}t.
\end{aligned}$$

The maximum likelihood approach, but not the least squares argument, is readily extended to the case where $\psi$ is not the identity. Denote by $\hat{\beta}$, $\hat{\xi}$ and $\hat{\eta}'$ the estimators of $\beta$, $\xi$ or $\eta'$ respectively, and write $\hat{\eta}_i$ for the value of $\eta_i$ when $\eta'$ is replaced by $\hat{\eta}'$.

### 2.4. *Predictors $\Theta_i^{\mathrm{BLUP}}$ and $\Theta_i^{\mathrm{BP}}$ of $\Theta_i$*

Following Ghosh and Maiti (2004), we take the BLUP of $\Theta_i$ to be

$$\Theta_i^{\mathrm{BLUP}} = f_i(\beta) + \frac{\mathrm{cov}(\bar{Y}_i, \Theta_i|\mathcal{X})}{\mathrm{var}(\bar{Y}_i|\mathcal{X})}\{\bar{Y}_i - f_i(\beta)\}, \tag{2.2}$$

where

$$\bar{Y}_i = n_i^{-1}\sum_j Y_{ij}.$$

Standard calculations show that the covariance–variance ratio in equation (2.2) is given by

$$\rho_i(\beta, \xi, \eta_i) = \frac{\gamma\{f_i(\beta), \xi\}}{\sigma_W\{f_i(\beta), \xi\}^2 + n_i^{-1}\,\sigma_V\{f_i(\beta), \xi, \eta_i\}^2},$$

where

$$\gamma(\mu, \xi) = \int u\,\psi(u)\,q(u|\mu, \xi)\,\mathrm{d}u - \left\{\int u\,q(u|\mu, \xi)\,\mathrm{d}u\right\}\int \psi(u)\,q(u|\mu, \xi)\,\mathrm{d}u,$$

$$\sigma_V(\mu, \xi, \eta)^2 = \int \sigma_R(u, \eta)^2\,q(u|\mu, \xi)\,\mathrm{d}u, \tag{2.3}$$

$\sigma_R(\mu, \eta)^2$ is the variance of the distribution $R(\mu, u)$ and $\sigma_W(\mu, \xi)^2$ denotes the variance of $W = \psi(\Theta)$ when $\Theta$ has the distribution $Q(\mu, \xi)$.

In the particular case where the link function $\psi$ is the identity there is a relatively simple formula for $\rho_i$. In fact, defining

$$\sigma_i(\mu, \xi, \eta)^2 = \sigma_Q(\mu, \xi)^2 + n_i^{-1}\,\sigma_V(\mu, \xi, \eta)^2,$$

where $\sigma_Q(\mu, \xi)^2$ is the variance of the distribution $Q(\mu, \xi)$, we have

$$\rho_i(\beta, \xi, \eta_i) = \frac{\sigma_Q\{f_i(\beta), \xi\}^2}{\sigma_i\{f_i(\beta), \xi, \eta_i\}^2}. \tag{2.4}$$

The empirical version, $\hat{\Theta}_i^{\mathrm{BLUP}}$, of $\Theta_i^{\mathrm{BLUP}}$ is obtained by replacing $\beta$, $\xi$ and $\eta_i$ by their estimators, $\hat{\beta}$, $\hat{\xi}$ and $\hat{\eta}_i$, in equation (2.2):

$$\hat{\Theta}_i^{\mathrm{BLUP}} = f_i(\hat{\beta}) + \rho_i(\hat{\beta}, \hat{\xi}, \hat{\eta}_i)\{\bar{Y}_i - f_i(\hat{\beta})\}. \tag{2.5}$$

For a general link function we usually do not have a closed form expression for $\gamma(\mu, \xi)$, and hence we do not have one for $\rho_i(\beta, \xi, \eta_i)$. In such cases, $\rho_i(\hat{\beta}, \hat{\xi}, \hat{\eta}_i)$ in equation (2.5) generally must be derived by using methods such as numerical integration. However, if $\psi(x) \equiv x$, and if in addition

> varying the parameter $\mu$ alters only the location of the distributions $Q(\mu, \xi)$ and $R(\mu, \eta)$, and in particular does not affect scale or shape, (2.6)

then simple explicit formulae are possible. Details will be given in Appendix A.

The best predictor, $\Theta_i^{\mathrm{BP}}$, is given by

$$\Theta_i^{\mathrm{BP}} = f_i(\beta) + E(U_i|\bar{Y}_i),$$

which we may write in the form

$$\Theta_i^{\mathrm{BP}} = f_i(\beta) + J_i(\bar{Y}_i|\beta, \xi, \eta_i),$$

where

$$J_i(u|\beta,\xi,\eta_i)=\frac{I_{i1}(u|\beta,\xi,\eta)}{I_{i0}(u|\beta,\xi,\eta)},$$

$$I_{ij}(u|\beta,\xi,\eta)=\int \theta^j\,\bar{r}_i\{u|\psi(\theta),\eta_i\}\,q\{\theta|f_i(\beta),\xi\}\,\mathrm{d}\theta,$$

and $\bar{r}_i(u|\mu,\eta)$ denotes the density of $n_i^{-1}\Sigma_{1\leqslant j\leqslant n_i} Z_j$, with the $Z_j$s being independent random variables having common density $r(u|\mu,\eta)$.

The empirical form of $\Theta_i^{\mathrm{BP}}$ is of course

$$\hat{\Theta}_i^{\mathrm{BP}}=f_i(\hat{\beta})+J_i(\bar{Y}_i|\hat{\beta},\hat{\xi},\hat{\eta}_i). \tag{2.7}$$

Again, calculation in the general case typically involves techniques such as numerical integration, unless $\psi$ is the identity and result (2.6) holds. We may express both $\Theta_i^{\mathrm{BLUP}}$ and $\Theta_i^{\mathrm{BP}}$ as

$$\Theta_i^{\mathrm{pred}}=K_i(\bar{Y}_i|\beta,\xi,\eta_i)$$

and hence, noting equations (2.4), (2.5) and (2.7), may write both $\hat{\Theta}_i^{\mathrm{BLUP}}$ and $\hat{\Theta}_i^{\mathrm{BP}}$ in the form

$$\hat{\Theta}_i^{\mathrm{pred}}=K_i(\bar{Y}_i|\hat{\beta},\hat{\xi},\hat{\eta}_i), \tag{2.8}$$

where $K_i$ denotes a known function.

### 2.5. Mean-square prediction error

The mean-squared predictive error of the 'ideal' predictor, $\Theta_i^{\mathrm{pred}}=\Theta_i^{\mathrm{BLUP}}$ or $\Theta_i^{\mathrm{pred}}=\Theta_i^{\mathrm{BP}}$, is given by

$$\mathrm{MSEpred}_i=E\{(\Theta_i^{\mathrm{pred}}-\Theta_i)^2|\mathcal{X}\}. \tag{2.9}$$

Of course, it differs somewhat from its counterpart for the practical predictor $\hat{\Theta}_i^{\mathrm{pred}}=\hat{\Theta}_i^{\mathrm{BLUP}}$ or $\hat{\Theta}_i^{\mathrm{pred}}=\hat{\Theta}_i^{\mathrm{BP}}$; we denote this MSE by

$$\mathrm{MSE}_i=E\{(\hat{\Theta}_i^{\mathrm{pred}}-\Theta_i)^2|\mathcal{X}\}.$$

A bootstrap estimator of $\mathrm{MSE}_i$ may be constructed as follows. Conditionally on the data $\mathcal{Z}=\{(X_{ij},Y_{ij}):1\leqslant i\leqslant n, 1\leqslant j\leqslant n_i\}$, draw $\Theta_i^*$ from the distribution $Q\{f_i(\hat{\beta}),\hat{\xi}\}$; given $\Theta_i^*$, draw $Y_{ij}^*$ from $R\{\psi(\Theta_i^*),\hat{\eta}_i^*\}$; and compute $\hat{\beta}^*$, $\hat{\xi}^*$ and $\hat{\eta}_i^*$ from the data

$$\mathcal{Z}^*=\{(X_{ij},Y_{ij}^*):1\leqslant i\leqslant n, 1\leqslant j\leqslant n_i\}.$$

Put

$$\bar{Y}_i^*=n_i^{-1}\sum_j Y_{ij}^*,$$

and let

$$\hat{\Theta}_i^{*\mathrm{pred}}=K_i(\bar{Y}_i^*|\hat{\beta}^*,\hat{\xi}^*,\hat{\eta}_i^*) \tag{2.10}$$

denote the bootstrap version of $\hat{\Theta}_i^{\mathrm{pred}}$. Then our bootstrap estimator of $\mathrm{MSE}_i$ is

$$\widehat{\mathrm{MSE}}_i=E\{(\hat{\Theta}_i^{*\mathrm{pred}}-\Theta_i^*)^2|\mathcal{Z}\}. \tag{2.11}$$

### 2.6. Bias of $\widehat{\mathrm{MSE}}_i$, and analytical bias correction

Let $(\Theta_i^{\mathrm{pred}}, \hat{\Theta}_i^{\mathrm{pred}})$ denote either $(\Theta_i^{\mathrm{BLUP}}, \hat{\Theta}_i^{\mathrm{BLUP}})$ or $(\Theta_i^{\mathrm{BP}}, \hat{\Theta}_i^{\mathrm{BP}})$, and recall the definition of $\widehat{\mathrm{MSE}}_i$ at equation (2.11). Now, $\mathrm{MSEpred}_i$ is a known function of the unknown parameters $\beta$, $\xi$ and $\eta_i$; let $\widehat{\mathrm{MSEpred}}_i$ denote the version of $\mathrm{MSEpred}_i$ that arises if, in this formula, $(\hat{\beta}, \hat{\xi}, \hat{\eta}_i)$ is replaced by $(\beta, \xi, \eta_i)$. We shall show in Section 4.2 that, in regular cases,

$$E(\widehat{\mathrm{MSE}}_i|\mathcal{X}) = \mathrm{MSE}_i + \frac{b_i(\beta, \xi, \eta_i)}{n} + O(n^{-2}), \tag{2.12}$$

where $b_i$ is a smooth function of its arguments.

In fact, the bias term $n^{-1} b_i$ appearing on the right-hand side of equation (2.12) is the same as that which arises when $\widehat{\mathrm{MSEpred}}_i$ is considered as an approximation to $\mathrm{MSEpred}_i$:

$$E(\widehat{\mathrm{MSEpred}}_i|\mathcal{X}) = \mathrm{MSEpred}_i + \frac{b_i(\beta, \xi, \eta_i)}{n} + O(n^{-2}). \tag{2.13}$$

See Section 4.2. Therefore, $\widehat{\mathrm{MSE}}_i$ implicitly corrects for the dominant bias term in $\mathrm{MSE}_i$ as an approximation to $\mathrm{MSEpred}_i$; the uncorrected bias comes principally from the fact that $\widehat{\mathrm{MSE}}_i$ is close to $\widehat{\mathrm{MSEpred}}_i$, and the latter is biased for $\mathrm{MSEpred}_i$.

It is of significant practical interest to correct for the principal bias contribution in equation (2.12). When $\psi$ is the identity and result (2.6) holds, the quantity $b_i(\beta, \xi, \eta_i)$ in equation (2.12) admits a reasonably simple formula. Replacing $\beta$, $\xi$ and $\eta_i$ in that formula by their estimators, we may construct an analytically bias-corrected estimator of $\mathrm{MSE}_i$:

$$\widehat{\mathrm{MSE}}_i^{\mathrm{bc}} \equiv \widehat{\mathrm{MSE}}_i - \frac{b_i(\hat{\beta}, \hat{\xi}, \hat{\eta}_i)}{n}. \tag{2.14}$$

This is the approach that was taken by, for example, Prasad and Rao (1990), Jiang *et al.* (2002) and Rao (2003), under the assumption that $\psi$ is the identity and the distributions $Q$ and $R$ are both normal, which implies property (2.6). Rao's calculations can be repeated for a variety of non-normal models that satisfy property (2.6).

Since

$$E\{b_i(\hat{\beta}, \hat{\xi}, \hat{\eta}_i)\} = b_i(\beta, \xi, \eta_i) + O(n^{-1})$$

then expressions (2.12) and (2.14) together imply that

$$E(\widehat{\mathrm{MSE}}_i^{\mathrm{bc}}|\mathcal{X}) = \mathrm{MSE}_i + O(n^{-2}), \tag{2.15}$$

i.e. bias is reduced from the standard level, $O(n^{-1})$, which is given in equation (2.12), to only $O(n^{-2})$.

### 2.7. Bias correction using bootstrap methods

In most cases an analytical approach to bias correction is precluded by either the sheer complexity of a formula for $b_i(\beta, \xi, \eta_i)$ or the unattractiveness, to practitioners, of working out a formula for $b_i$ in the context of a new model that they wish to use. These considerations motivate an alternative, double-bootstrap approach to bias correction, which may be developed as follows.

Recall that in Section 2 we described a method for creating the data $Y_{ij}^*$, from which we assembled the data set $\mathcal{Z}^*$ and then calculated the bootstrap version, $(\hat{\beta}^*, \hat{\xi}^*, \hat{\eta}_i^*)$, of $(\hat{\beta}, \hat{\xi}, \hat{\eta}_i)$. Conditionally on $\mathcal{Z}^*$, draw $\Theta_i^{**}$ by sampling randomly from the distribution $Q\{f_i(\hat{\beta}^*), \hat{\xi}^*\}$; given $\Theta_i^{**}$, draw $Y_{ij}^{**}$ from the distribution $R(\Theta_i^{**}, \hat{\eta}_i^*)$; and compute $(\hat{\beta}^{**}, \hat{\xi}^{**}, \hat{\eta}_i^{**})$ from the data in

$$\mathcal{Z}^{**} = \{(X_{ij}, Y_{ij}^{**}) : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n_i\}.$$

Put

$$\bar{Y}_i^{**} = n_i^{-1} \sum_j Y_{ij}^{**}.$$

Analogously to equations (2.10) and (2.11), define

$$\hat{\Theta}_i^{**\mathrm{pred}} = K_i(\bar{Y}_i^{**} | \hat{\beta}^{**}, \hat{\xi}^{**}, \hat{\eta}_i^{**}),$$
$$\widehat{\mathrm{MSE}}_i^* = E\{(\hat{\Theta}_i^{**\mathrm{pred}} - \Theta_i^{**})^2 | \mathcal{Z}^*\},$$

where the function $K_i$ is as at equations (2.8) and (2.10). We view $\widehat{\mathrm{MSE}}_i^*$ as an estimator, in the bootstrap world, of $\hat{u} \equiv \widehat{\mathrm{MSE}}_i$. Its conditional mean, $\hat{v} \equiv E(\widehat{\mathrm{MSE}}_i^* | \mathcal{Z})$, is an estimator of $E(\hat{u}) = E(\widehat{\mathrm{MSE}}_i)$.

Conventional additively and multiplicatively bias-corrected estimators are

$$\begin{aligned} \widehat{\mathrm{MSE}}_i^{\mathrm{add-bc}} &= 2\hat{u} - \hat{v}, \\ \widehat{\mathrm{MSE}}_i^{\mathrm{mult-bc}} &= \hat{u}^2/\hat{v} \end{aligned} \tag{2.16}$$

respectively. Both have property (2.15); see Section 4.3. The estimator $\widehat{\mathrm{MSE}}_i^{\mathrm{add-bc}}$ is attractive when there is a suggestion that $\widehat{\mathrm{MSE}}_i$ is positively biased, i.e. when $\hat{u} \geqslant \hat{v}$. However, when $\hat{u} < \hat{v}$ the additive approach can give too great a degree of correction, and even a negative value of $2\hat{u} - \hat{v}$. Moreover, when there is an indication that the estimator $\hat{u}$ is negatively biased, using the multiplicative approach can give unreliable results because of the effect of dividing by the stochastically variable quantity $\hat{v}$. This difficulty is reduced if we employ, instead of $\hat{u}^2/\hat{v}$, $\exp\{-(\hat{v} - \hat{u})/\hat{v}\}\hat{u}$, which is strictly greater than $\hat{u}^2/\hat{v}$ but preserves the same degree of first-order bias correction.

Together, these considerations suggest the bias-corrected estimator

$$\widehat{\mathrm{MSE}}_i^{\mathrm{bc1}} = \begin{cases} 2\hat{u} - \hat{v} & \text{if } \hat{u} \geqslant \hat{v}, \\ \exp\{-(\hat{v} - \hat{u})/\hat{v}\}\hat{u} & \text{if } \hat{u} < \hat{v}. \end{cases} \tag{2.17}$$

Of all the many variants of the estimators at expression (2.16) that we considered, and also $\widehat{\mathrm{MSE}}_i^{\mathrm{bc}}$, at expression (2.14), in the particular case where $Q$ and $R$ are normal distributions, $\widehat{\mathrm{MSE}}_i^{\mathrm{bc1}}$ gave, in almost all cases, the greatest degree of bias correction. However, bias reduction almost invariably increases the variance and can lead to a consequent increase in MSE. The estimator $\widehat{\mathrm{MSE}}_i^{\mathrm{bc1}}$ sometimes suffers from this difficulty.

To overcome this problem we sought an approach which gave less bias reduction but suffered less from inflation of variance or MSE. To see how this might be done, note that when correcting bias additively the total amount of correction is $\hat{u} - \hat{v}$, and variance inflation can be diminished if this quantity is reduced when $|\hat{u} - \hat{v}|$ is relatively large. That can be effected by, for example, taking the bias-corrected estimator to be $\hat{u} + \chi_n(\hat{u} - \hat{v})$, instead of $2\hat{u} - \hat{v}$, where $\chi_n$ denotes a bounded symmetric function which admits the approximation $\chi_n(t) \approx t$ when $t$ is close to 0.

Since the amount of bias that we are correcting is expected to be of size $n^{-1}$, then it is appropriate to take $\chi_n(t) = n^{-1}\chi(nt)$, where $\chi$ is a fixed function and $\chi(t) \approx t$ when $t$ is close to 0. We shall report results when $\chi$ is the inverse tangent function, which leads to the bias-corrected estimator

$$\widehat{\mathrm{MSE}}_i^{\mathrm{bc2}} = \begin{cases} \hat{u} + n^{-1} \tan^{-1}\{n(\hat{u} - \hat{v})\} & \text{if } \hat{u} \geqslant \hat{v}, \\ \hat{u}^2/[\hat{u} + n^{-1} \tan^{-1}\{n(\hat{v} - \hat{u})\}] & \text{if } \hat{u} < \hat{v}. \end{cases} \tag{2.18}$$

An alternative approach would be to use the estimator on the first line of expression (2.18) regardless of the value of $\hat{u}-\hat{v}$. However, this has the potential to give a negative value when $\hat{u}<\hat{v}$.

### 2.8. Prediction intervals

Let $(n_0,x_0)$ denote an existing, or future, value of $(n_i,X_i)$, where $x_0=(x_{01},\ldots,x_{0n_0})$ is a candidate for $X_i=(X_{i1},\ldots,X_{in_i})$. Assuming that the model in Section 1 that generated the random variables $\Theta_i$ still prevails, the value $\Theta_0$ of $\Theta$ that is associated with $(n_0,x_0)$ has the distribution $Q\{f_0(\beta),\xi\}$, where $f_0$ is a known function. Then, an $\alpha$-level, two-sided, equal-tailed prediction interval for the value of $\Theta_0$, given $(n_0,x_0)$, is

$$\begin{aligned}I_\alpha &= I_\alpha(n_0,x_0|\beta,\xi)\\ &\equiv [z_{(1-\alpha)/2}(n_0,x_0|\beta,\xi),\ z_{(1+\alpha)/2}(n_0,x_0|\beta,\xi)],\end{aligned}$$

where $z_\alpha(n_0,x_0|\beta,\xi)$ denotes the $\alpha$-level quantile of the distribution $Q\{f_0(\beta),\xi\}$:

$$\int_{z_\alpha(n_0,x_0|\beta,\xi)}^{\infty} q\{u|f_0(\beta),\xi\}\,\mathrm{d}u=\alpha.$$

Of course, $I_\alpha$ is impractical, since $\beta$ and $\xi$ are not known. Its empirical form, $\hat{I}_\alpha=I_\alpha(n_0, x_0|\hat{\beta},\hat{\xi})$, has coverage that converges to $\alpha$ as $n$ increases. In theory, the coverage error of $\hat{I}_a$ equals $O(n^{-1})$:

$$P(\Theta_0\in\hat{I}_\alpha)=\alpha+O(n^{-1}) \tag{2.19}$$

as $n\to\infty$. See Section 4.4 for a derivation of this and the other coverage properties that are noted below. Nevertheless, for finite $n$ the level of error might not be inconsiderable. To remedy this problem we suggest calibrating on $\alpha$, using the bootstrap.

Specifically, construct the bootstrap estimators $\hat{\beta}^*$ and $\hat{\xi}^*$ that were discussed in Section 2, put $\hat{I}_a^*=I_a(n_0,x_0|\hat{\beta}^*,\hat{\xi}^*)$ and let $a=\hat{a}$ denote the solution of the equation

$$P(\Theta_0^*\in\hat{I}_a^*|\mathcal{Z})=\alpha,$$

where $\Theta_0^*$ has the distribution $Q\{f_0(\hat{\beta}),\hat{\xi}\}$. The bootstrap-calibrated prediction interval is $\hat{I}_{\hat{a}}$.

We may calibrate at a second level, by first choosing $a=\hat{a}^*$ to solve

$$P(\Theta_0^{**}\in\hat{I}_a^{**}|\mathcal{Z}^*)=\alpha,$$

where $\hat{I}_a^{**}=I_a(n_0,x_0|\hat{\beta}^{**},\hat{\xi}^{**})$ and $\Theta_0^{**}$ has the distribution $Q\{f_0(\hat{\beta}^*),\hat{\xi}^*\}$; then selecting $a=\hat{a}_1$ to solve

$$P(\Theta_0^*\in\hat{I}_{\hat{a}^*+a}^*|\mathcal{Z})=\alpha;$$

and, finally, taking our interval to be $\hat{I}_{\hat{a}+\hat{a}_1}$. The intervals $\hat{I}_{\hat{a}}$ and $\hat{I}_{\hat{a}+\hat{a}_1}$ have coverages equal to $\alpha+O(n^{-2})$ and $\alpha+O(n^{-3})$ respectively, when viewed as prediction intervals for $\Theta_0$.

## 3. Numerical results

### 3.1. Simulation study: mean-squared error estimation

We conducted a simulation study to assess the small sample performance of our methods, and to compare them with their competitors where available. We took $n=15$ throughout. Below, the bias correction method of Prasad and Rao (1990) is denoted by PR, and the jackknife approach of Jiang *et al.* (2002) is indicated by JK.

First we considered the case where both populations are normal. In this setting, all three classes of method (PR, JK and our own) can be used to correct MSE estimators for bias. Specifically, we assumed that $Q(\mu,\xi)$ and $R(\theta_i,\eta_i)$ were normal with $\xi=1$ and $\mu=0$, and took each $n_i=1$. The $n=15$ areas were divided into three groups of 5, with equal numbers of areas and equal values of $\eta_i$ within each group. Two different choices of the $\eta_i$s were made. This led to two models, denoted below by $\mathrm{M}_1, (0.7, 0.5, 0.3)$, and $\mathrm{M}_2, (4, 0.5, 0.1)$. Each is a special case of the Fay–Herriot model that was discussed in Section 2.2. The set-up is similar to that considered by Datta *et al.* (2005). Model $\mathrm{M}_1$ is nearly balanced, but model $\mathrm{M}_2$ has high variability.

Our third model, $\mathrm{M}_3$, was similar except that we allowed the $R$-distribution to have relatively heavy tails. Specifically, we took $Q(\mu,\xi)$ to be normal with mean $\mu=0$ and $R(\Theta_i,\eta_i)$ to have Student's $t$-distribution (translated so that its mean was $\Theta_i$) with 4, 6 or 8 degrees of freedom, each repeated five times to produce the $n=15$ data values. For this model we used the BLUP to construct small area predictors, and we did not consider the JK method since it is not readily accessible in this setting. Note that, though the BLUP has a closed form expression here, BP does not have this advantage.

Our fourth model, $\mathrm{M}_4$, was the logistic regression that was discussed in Section 2.2. We took $R$ to be a binomial distribution and $Q$ to be normal, with $\psi$ the logistic function. The binomial population had 15 areas, of respective sizes $n_i=48$, 55, 60, 61, 68, 98, 135, 136, 181, 187, 210, 228, 232, 275, 287, based on occupational category sizes that were used by Dartigues *et al.* (1992). To compute the binomial probabilities we first generated design variables $X_i$ from the uniform distribution on $[0, \frac{1}{2}]$, then generated standard normal variables $U_i$ (playing the same role as $U_i$ at equation (2.1)) and defined

$$p_i=\frac{\exp(\beta_0+\beta_1 X_i+U_i)}{1+\exp(\beta_0+\beta_1 X_i+U_i)}$$

for $i=1,\ldots,15$. We took $\beta=(\beta_0,\beta_1)^{\mathrm{T}}=(0,1)^{\mathrm{T}}$. In this case, neither the BLUP nor the BP has a closed form expression. We took our small area predictor to be the BP and did not consider the PR method since it does not have a readily available form in this setting. However, we applied the JK method and calculated the jackknife MSE estimates as outlined by Rao (2003).

We used empirical measures of relative bias and coefficient of variation to quantify the performances of various methods. The relative absolute bias of the MSE estimator was defined by

$$\mathrm{RB}_i=\left|\frac{E\{\widehat{\mathrm{MSE}}(\hat\Theta_i)\}-\mathrm{SMSE}(\hat\Theta_i)}{\mathrm{SMSE}(\hat\Theta_i)}\right|,$$

for $i=1,\ldots,n$, where $E\{\widehat{\mathrm{MSE}}(\hat\Theta_i)\}$ was estimated empirically as the average of values of $\widehat{\mathrm{MSE}}(\hat\Theta_i)$ over replications. Likewise, $\mathrm{SMSE}(\hat\Theta_i)$ was defined as the average value of $(\hat\Theta_i-\Theta_i)^2$. The coefficient of variation of the MSE estimator was taken to be

$$\mathrm{CV}_i=\frac{[E\{\widehat{\mathrm{MSE}}(\hat\Theta_i)-\mathrm{SMSE}(\hat\Theta_i)\}^2]^{1/2}}{\mathrm{SMSE}(\hat\Theta_i)}$$

for $i=1,\ldots,n$, where $E\{\widehat{\mathrm{MSE}}(\hat\Theta_i)-\mathrm{SMSE}(\hat\Theta_i)\}^2$ was computed by averaging $\{\widehat{\mathrm{MSE}}(\hat\Theta_i)-\mathrm{SMSE}(\hat\Theta_i)\}^2$ over replicates.

Except in the case of model $\mathrm{M}_4$, where we used maximum likelihood (see for example Slud (2000)), the method of moments (see for example Prasad and Rao (1990)) was employed to

**Table 1.** Relative biases (upper figure in each row) and coefficient of variations (lower figure), for bias-corrected estimators of the predictive MSE

| *Method* | *Results for the following models:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $M_1$ | | $M_2$ | | $M_3$ | | $M_4$ | |
| | *Median* | *Mean* | *Median* | *Mean* | *Median* | *Mean* | *Median* | *Mean* |
| PR | 0.057 | 0.077 | 1.270 | 2.289 | 0.197 | 0.270 | — | — |
| | 0.154 | 0.161 | 1.972 | 4.519 | 0.429 | 0.392 | — | — |
| JK | 0.043 | 0.074 | 0.191 | 0.253 | — | — | 0.079 | 0.145 |
| | 0.442 | 0.471 | 1.047 | 1.693 | — | — | 1.089 | 1.306 |
| $\widehat{\mathrm{MSE}}_i^{\mathrm{mult-bc}}$ | 0.046 | 0.070 | 0.195 | 0.169 | 0.113 | 0.162 | 0.128 | 0.129 |
| | 0.304 | 0.311 | 0.537 | 0.594 | 0.647 | 0.767 | 0.489 | 0.512 |
| $\widehat{\mathrm{MSE}}_i^{\mathrm{bc1}}$ | 0.062 | 0.074 | 0.216 | 0.188 | 0.128 | 0.294 | 0.075 | 0.128 |
| | 0.290 | 0.296 | 0.510 | 0.562 | 0.690 | 0.899 | 0.415 | 0.434 |
| $\widehat{\mathrm{MSE}}_i^{\mathrm{bc2}}$ | 0.078 | 0.079 | 0.240 | 0.206 | 0.185 | 0.241 | 0.078 | 0.128 |
| | 0.277 | 0.283 | 0.482 | 0.531 | 0.586 | 0.589 | 0.407 | 0.431 |

estimate parameters. Other approaches would have been feasible, but by using the method of moments we made our methodology directly comparable with that of Prasad and Rao (1990) and Lahiri and Rao (1995), for example. In particular, the PR method is based on the method of moments. Of course, our bootstrap algorithm can be used in conjunction with any technique for parameter estimation.

Each reported result was based on 1000 synthetic samples, to each of which we applied each method. For the bootstrap we used $B=100$ and $C=50$ simulations in the first and second stages respectively. We ran some of the analyses for $(B,C)=(200,100)$ but obtained virtually identical results, so for the full study we confined attention to the smaller number of simulations.

Summary results are presented in Table 1. There, PR and JK denote the respective methods that were discussed earlier in this section, and $\widehat{\mathrm{MSE}}_i^{\mathrm{mult-bc}}$, $\widehat{\mathrm{MSE}}_i^{\mathrm{bc1}}$ and $\widehat{\mathrm{MSE}}_i^{\mathrm{bc2}}$ are the estimators that are defined at expressions (2.16), (2.17) and (2.18) respectively. The body of Table 1 gives averages, over the $n=15$ different small areas, of the values of $\mathrm{RB}_i$ or $\mathrm{CV}_i$, where 'average' is measured in terms of the median (given in the first column for each model $\mathrm{M}_j$) or the mean (in the second column).

As can be seen from Table 1, the estimator $\widehat{\mathrm{MSE}}_i^{\mathrm{bc2}}$ performs particularly well in terms of minimizing relative bias, as indicated by the tabulated average values of $\mathrm{RB}_i$. This estimator also has a low coefficient of variation in non-normal cases, and in the high variability normal model $\mathrm{M}_2$. In the simpler model $\mathrm{M}_1$ the method PR, which was designed for this type of setting, gives the best results. Nevertheless, in terms of relative bias, each of the bootstrap methods outperforms each of its non-bootstrap rivals for the high variability normal model $\mathrm{M}_2$ and for each of the non-normal models $\mathrm{M}_3$ and $\mathrm{M}_4$.

The estimator $\widehat{\mathrm{MSE}}_i^{\mathrm{bc1}}$ also reduces bias significantly, often outperforming $\widehat{\mathrm{MSE}}_i^{\mathrm{bc2}}$ in this respect. However, it has greater tendency than $\widehat{\mathrm{MSE}}_i^{\mathrm{bc2}}$ to increase the variance; the latter estimator arguably achieves a better balance, relative to either of its bootstrap competitors, between the two opposing goals of reducing bias and not exacerbating variance.

For each model where the JK method is applicable, the estimator that it produces has inferior performance in terms of coefficient of variation. The bias-corrected estimator that is given by the PR method has a lower coefficient of variation than $\widehat{\mathrm{MSE}}_i^{\mathrm{bc2}}$ in the cases of models $\mathrm{M}_1$ and $\mathrm{M}_3$, but not for $\mathrm{M}_2$, and is not readily accessible for model $\mathrm{M}_4$.

In summary, the PR method gives low relative bias and variance for normal models with low variance ratios, but not in other cases. The JK method can produce low relative bias but suffers from high variability. The new bootstrap estimators, especially $\widehat{\mathrm{MSE}}_i^{\mathrm{bc2}}$, provide very good compromises between low bias and high variability, and have the advantage of being usable particularly widely.

We also addressed the case where the distributions $Q$ and $R$ were normal and exponential respectively, and prediction was done by using the BLUP. Here the median relative bias and median coefficient of variation under the PR method were 39% and 62% respectively. However, for the bootstrap methods $\widehat{\mathrm{MSE}}_i^{\mathrm{bc1}}$ and $\widehat{\mathrm{MSE}}_i^{\mathrm{bc2}}$ they dropped to 0% and 42% and to 8% and 40% respectively. The JK method is not readily accessible in this setting.

### 3.2. *Simulation study: prediction intervals*

For the same models and parameter choices as in Section 3.1, and again for $n=15$, Table 2 provides results for prediction intervals. BootN, Boot1 and Boot2 denote prediction interval methods that are based on the naïve bootstrap and on one and two applications respectively of the bootstrap for bias correction. The BootN, Boot1 and Boot2 prediction intervals are $\hat{I}_\alpha$, $\hat{I}_{\hat{a}}$ and $\hat{I}_{\hat{a}+\hat{a}_1}$ respectively. (Their respective coverage errors are of orders $n^{-1}$, $n^{-2}$ and $n^{-3}$.) We used three different nominal coverages: $\alpha=0.80$, 0.90, 0.95. Standard normal percentile points were used to construct the prediction intervals under methods PR and JK.

**Table 2.** Coverage probabilities for various methods

| *Method* | *Results for the following models:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *M*$_1$ | | *M*$_2$ | | *M*$_3$ | | *M*$_4$ | |
| | *Median* | *Mean* | *Median* | *Mean* | *Median* | *Mean* | *Median* | *Mean* |
| *(a) $\alpha=0.80$* | | | | | | | | |
| PR | 0.796 | 0.796 | 0.889 | 0.860 | 0.835 | 0.822 | — | — |
| JK | 0.796 | 0.796 | 0.728 | 0.731 | — | — | 0.780 | 0.793 |
| BootN | 0.651 | 0.651 | 0.592 | 0.586 | 0.940 | 0.931 | 0.784 | 0.789 |
| Boot1 | 0.796 | 0.796 | 0.808 | 0.810 | 0.795 | 0.807 | 0.791 | 0.800 |
| Boot2 | 0.799 | 0.796 | 0.804 | 0.822 | 0.800 | 0.805 | 0.810 | 0.800 |
| *(b) $\alpha=0.90$* | | | | | | | | |
| PR | 0.899 | 0.899 | 0.955 | 0.937 | 0.920 | 0.913 | — | — |
| JK | 0.893 | 0.894 | 0.840 | 0.834 | — | — | 0.889 | 0.886 |
| BootN | 0.797 | 0.797 | 0.735 | 0.731 | 0.998 | 0.994 | 0.820 | 0.820 |
| Boot1 | 0.899 | 0.899 | 0.878 | 0.886 | 0.912 | 0.902 | 0.900 | 0.901 |
| Boot2 | 0.902 | 0.910 | 0.916 | 0.916 | 0.912 | 0.910 | 0.905 | 0.910 |
| *(c) $\alpha=0.95$* | | | | | | | | |
| PR | 0.952 | 0.952 | 0.981 | 0.981 | 0.949 | 0.946 | — | — |
| JK | 0.943 | 0.943 | 0.862 | 0.860 | — | — | 0.940 | 0.941 |
| BootN | 0.865 | 0.865 | 0.819 | 0.818 | 0.999 | 0.999 | 0.899 | 0.920 |
| Boot1 | 0.945 | 0.945 | 0.925 | 0.933 | 0.963 | 0.945 | 0.940 | 0.940 |
| Boot2 | 0.945 | 0.950 | 0.957 | 0.953 | 0.961 | 0.951 | 0.952 | 0.957 |

Panels (a), (b) and (c) of Table 2 treat the cases where the nominal coverage was $\alpha = 0.80, 0.90, 0.95$ respectively. Since the conservatism, or otherwise, of the different methods was consistent across the values of different small areas, then we summarize results by giving the average values (defined in terms of the median or mean, given respectively in the first and second columns of blocks in Table 2, and taken over the 15 small areas) of coverage, rather than by taking the averages of absolute values of departures from nominal levels.

Each prediction interval method, except the method that was based on the naïve bootstrap, performed well under model $M_1$. Under model $M_2$, the PR and JK intervals tended to overcover and undercover respectively. They continued to have this difficulty for models $M_3$ and $M_4$, unless $\alpha = 0.95$, where both did reasonably well. The bootstrap prediction method Boot1 performed uniformly well throughout. However, despite performing better than its competitors in each case, it had a little difficulty for model $M_2$ when $\alpha = 0.90$ or $\alpha = 0.95$, where it tended to undercover. Taking the bootstrap to an extra level, giving the results that are listed as Boot2 in Table 2, this difficulty was overcome. In other cases, where Boot1 already gave good performance, Boot2 had little or no effect. In the few instances where Boot2 produced worse coverage accuracy than Boot1, both methods enjoyed good performance.

### *3.3. Simulation study: robustness*

To asses the robustness of different approaches to estimation of the mean-squared prediction error, we simulated from models $M_5$, $M_6$ and $M_7$ where $Q(\mu, \xi)$ was normal, exponential and double exponential respectively. However, in each case we conducted inference as though the model was normal. Reflecting the approach of Lahiri and Rao (1995), two different patterns of parametric combinations were used. For each pattern we took $n = 20$ and $n = 30$. In the case of pattern (a) the random-effects variance was kept fixed at 0.2. For that pattern, and when $n = 20$, the 20 areas were divided into four groups of five areas. The four values of $\eta_i$ (which was kept fixed within each group) were $(0.5, 0.33, 0.25, 0.2)$. Still in the case of pattern (a), when $n = 30$ the number of groups was increased to 5, the new $\eta_i$ was 0.17 and group sizes were increased to 6. For pattern (b), the random-effects variance was kept fixed at 1; the respective number and size of groups remained at 4 and 5 when $n = 30$, and 5 and 6 when $n = 30$, and the sampling variances $\eta_i$ were $(1, 0.67, 0.5, 0.4)$ and $(1, 0.67, 0.5, 0.4, 0.33)$ for $n = 20$ and $n = 30$ respectively.

Results for these respective cases are given in panels (a) and (b) respectively of Table 3. For brevity, only the median and mean areas are reported. In Table 3, HM denotes the bootstrap method, JK and PR are as before and Md and Mn denote the median and mean respectively. Formula (2.18) was used to implement the bootstrap. In each part of Table 3, the first set of three blocks of columns headed PR, JK and HM correspond to pattern (a), and the second set to pattern (b). The relative biases in Table 3 are signed, rather than absolute, in keeping with the biases that were tabulated by Lahiri and Rao (1995).

It can be seen from Table 3 that, in all cases, the coefficient of variation for method HM is lower than that for JK but larger than that for PR. In several instances, the coefficient of variation for method HM is less than half that for JK. The bias for method HM is less than that for PR, but more than that for JK, for three of the 12 models and parameter settings that are represented in Table 3. In another three instances, the bias for method HM is less than that for JK but greater than that for PR. In five cases the bias of method HM is less than that for either of the other two methods, and in the remaining case the bias for method HM is equal to that for PR, and both are less than that for JK.

Taken together, these results suggest that the robustness of the bootstrap method to model misspecification is similar to that of the jackknife, and also to that of the Prasad–Rao approach. It is generally argued that bias is the litmus test of robustness, and from this viewpoint it can be

**Table 3.** Relative biases (upper figure in each row) and coefficients of variation (lower figure), for bias-corrected estimators of the predictive MSE

| *Model* | *Results for the following methods:* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *PR* | | *JK* | | *HM* | | *PR* | | *JK* | | *HM* | |
| | *Md* | *Mn* | *Md* | *Mn* | *Md* | *Mn* | *Md* | *Mn* | *Md* | *Mn* | *Md* | *Mn* |
| *(a) n = 20* | | | | | | | | | | | | |
| $M_5$ | 0.146 | 0.135 | −0.022 | −0.022 | −0.038 | −0.037 | 0.021 | 0.019 | 0.057 | 0.057 | 0.017 | 0.016 |
| | 0.278 | 0.292 | 0.928 | 0.938 | 0.470 | 0.475 | 0.145 | 0.154 | 0.372 | 0.398 | 0.293 | 0.299 |
| $M_6$ | 0.190 | 0.201 | 0.072 | 0.073 | −0.030 | −0.028 | 0.032 | 0.031 | 0.140 | 0.163 | 0.036 | 0.030 |
| | 0.356 | 0.358 | 1.129 | 1.229 | 0.543 | 0.549 | 0.219 | 0.216 | 1.057 | 1.132 | 0.373 | 0.372 |
| $M_7$ | 0.142 | 0.163 | 0.028 | 0.038 | −0.058 | −0.054 | 0.029 | 0.036 | 0.130 | 0.142 | 0.017 | 0.011 |
| | 0.297 | 0.323 | 1.036 | 1.097 | 0.516 | 0.525 | 0.187 | 0.191 | 0.676 | 0.734 | 0.337 | 0.338 |
| *(b) n = 30* | | | | | | | | | | | | |
| $M_5$ | 0.043 | 0.042 | −0.006 | −0.003 | 0.013 | 0.019 | 0.003 | 0.006 | 0.023 | 0.023 | 0.012 | 0.010 |
| | 0.269 | 0.261 | 0.608 | 0.616 | 0.455 | 0.449 | 0.127 | 0.141 | 0.200 | 0.209 | 0.254 | 0.262 |
| $M_6$ | 0.082 | 0.084 | 0.117 | 0.110 | 0.037 | 0.034 | 0.003 | 0.009 | 0.109 | 0.110 | 0.019 | 0.018 |
| | 0.283 | 0.304 | 0.935 | 1.103 | 0.483 | 0.489 | 0.180 | 0.197 | 0.607 | 0.658 | 0.301 | 0.314 |
| $M_7$ | 0.055 | 0.065 | 0.039 | 0.040 | 0.030 | 0.042 | −0.004 | −0.004 | 0.053 | 0.058 | 0.028 | 0.025 |
| | 0.288 | 0.285 | 0.730 | 0.732 | 0.474 | 0.502 | 0.159 | 0.167 | 0.346 | 0.429 | 0.280 | 0.285 |

concluded that neither method PR nor method JK is more robust than the bootstrap; the latter improves on PR on eight out of 12 occasions, and on JK in the same number.

### 3.4. Real data application

To illustrate our method, and to compare it with others, we used an example involving data from the US National Resources Inventory surveys. See Nusser and Goebel (1997) and Wang and Fuller (2003) for background and discussion of the data. We used road area in Missouri as the response variable, taking the small areas to correspond to the 114 counties in the state. Following Wang and Fuller (2003) we treated the within-counties data as a simple random sample, and we took the variable of interest to be the mean road area, in acres. In each instance we used the BLUP for small area prediction, and the method of moments for estimating model parameters. The fitted model was that of a simple random intercept model without covariate, and the estimated parameter values were 3.156 and 0.345 for the grand mean and the variance component associated with small area effects respectively.

The direct survey estimates $\bar{Y}_i$ had mean 3.23, sampling variances between 0.022 and 1.153, and very skew distributions. Because of the skew distributions a translated exponential model was attractive for the sampling error distribution, although we also considered a normal model. For both choices, a normal model was used for the $\Theta_i$s. Results for our bootstrap method in the exponential–normal and normal–normal cases are indicated by BootEN and BootNN respectively in Table 4, which reports estimates of mean-squared prediction errors. The 15 counties that were selected for illustration were chosen by using the procedure that was suggested by Wang and Fuller (2003). From each of their three groups of 38 counties we selected those with indices 1, 10, 19, 28 and 37.

As can be seen from Table 4, the normal–normal bootstrap and Prasad–Rao methods generally give similar estimates of the mean-squared prediction error, the bootstrap estimates being

**Table 4.** Analysis of US National Resources Inventory road area in Missouri†

| *County* | $n_i$ | $\bar{Y}_i$ | $\hat{\Theta}_i^{\text{pred}}$ | $\hat{\eta}_i$ | *PR* | *BootNN* | *BootEN* |
|---|---|---|---|---|---|---|---|
| 1 | 56 | 1.663 | 1.949 | 0.082 | 0.067 | 0.067 | 0.026 |
| 2 | 99 | 1.998 | 2.188 | 0.068 | 0.057 | 0.055 | 0.026 |
| 3 | 124 | 2.585 | 2.646 | 0.042 | 0.038 | 0.035 | 0.034 |
| 4 | 70 | 2.804 | 2.903 | 0.135 | 0.099 | 0.087 | 0.082 |
| 5 | 60 | 2.922 | 3.021 | 0.253 | 0.150 | 0.169 | 0.077 |
| 6 | 48 | 2.929 | 3.003 | 0.166 | 0.115 | 0.110 | 0.133 |
| 7 | 148 | 3.153 | 3.153 | 0.022 | 0.021 | 0.016 | 0.021 |
| 8 | 51 | 3.218 | 3.187 | 0.348 | 0.178 | 0.147 | 0.145 |
| 9 | 61 | 3.236 | 3.222 | 0.072 | 0.060 | 0.042 | 0.020 |
| 10 | 72 | 3.379 | 3.327 | 0.105 | 0.082 | 0.081 | 0.045 |
| 11 | 92 | 3.478 | 3.394 | 0.122 | 0.092 | 0.087 | 0.036 |
| 12 | 81 | 3.607 | 3.544 | 0.056 | 0.049 | 0.047 | 0.018 |
| 13 | 46 | 3.772 | 3.468 | 0.335 | 0.175 | 0.165 | 0.124 |
| 14 | 45 | 4.342 | 3.913 | 0.196 | 0.128 | 0.115 | 0.117 |
| 15 | 57 | 4.542 | 3.615 | 0.697 | 0.237 | 0.254 | 0.187 |

†The last three columns show mean-squared prediction error estimators based on the Prasad–Rao (PR), normal–normal bootstrap (BootNN) and exponential–normal bootstrap (BootEN) methods. Earlier columns show the quantities indicated.

lower in all except one of the 15 cases. The exponential–normal bootstrap often gives much lower error estimates, reflecting the greater suitability of the exponential model for sampling-error distributions.

## 4. Theoretical details

We assume throughout that the cluster sizes $n_i$ are uniformly bounded, which implies that the number of different possible values of $n_i$ is bounded. In this case there is no loss of generality in supposing that each $n_i = 1$. This we do below, writing $X_{ij}$ and $Y_{ij}$ as simply $X_i$ and $Y_i$ respectively. We require the densities of the distributions $Q$ and $R$, in either the continuous or the discrete case, to be sufficiently smooth functions of the unknown parameters. It is this smoothness that dictates the smoothness that we need of the various functions that arise in our proofs. Details of the proofs are given in Hall and Maiti (2005).

The main assumption that is needed is that the estimators $\hat{\beta}$, $\hat{\xi}$ and $\hat{\eta}_i$ should be closely approximable by smooth functions of sums of independent and identically distributed random variables. In particular,

$$\hat{\beta} = \beta^0 + b_n + \frac{1}{n}\sum_{i=1}^{n} a_{in}(Y_i) + \sum_{j=1}^{k}\left\{\frac{1}{n}\sum_{i=1}^{n} a_{jin}(Y_i)\right\}^2 + \Delta_n, \tag{4.1}$$

where $\beta^0$, $b_n$ and $\Delta_n$ will be discussed in the next paragraph, the uniformly bounded functions $a_{in}$ and $a_{jin}$ depend on the set $\mathcal{X}$ of explanatory variables and satisfy $E\{a_{in}(Y_i)\} = E\{a_{jin}(Y_i)\} = 0$, and

$$\sup_i [E\{|a_{in}(Y_i)|^k\} + E\{|a_{jin}(Y_i)|^k\}] < \infty \tag{4.2}$$

for each $k \geqslant 1$. Here and below, all expectations and probabilities are interpreted as conditional on $\mathcal{X}$. Expansion (4.1) holds when the distributions $Q(\mu, \xi)$ and $R(\mu, \eta)$ come from exponential families, but it also applies in many other cases, including cases where the distribution $Q(\mu, \xi)$ is discrete. A case in point is that where $\Theta_i$ equals the proportion of 0s in a sequence of Bernoulli trials; see Section 3 for numerical results in this setting.

In expansion (4.1), $\beta^0$ denotes the true value of $\beta$, the vector $b_n$, having the same length as $\beta$, satisfies $b_n = n^{-1}C + O(n^{-2})$, for a constant $C>0$, the remainder $\Delta_n$ satisfies

$$P(|\Delta_n| > n^{\varepsilon-3/2}) = O(n^{-C}) \tag{4.3}$$

for all $\varepsilon, C>0$, and analogues of expansion (4.1) apply to the estimators $\hat{\xi}$ and $\hat{\eta}_i$, with, in the latter case, moment and probability bounds applying uniformly in $i$. Property (4.1) is a conventional Taylor series expansion representation of an estimator, in which $\hat{\beta}$ is expressed as a series in linear and quadratic terms, with cubic and higher order contributions going into the remainder $\Delta_n$. It is also conventional that an expansion of $E(\hat{\beta})$ consists of terms in $n^{-j}$ for $j \geqslant 0$. The properties that were mentioned in the two previous sentences imply the property $b_n = n^{-1}C + O(n^{-2})$. The requirements that inequality (4.2) holds for all $k$ and condition (4.3) be true for all $\varepsilon$ and $C$ are more stringent than necessary for the Taylor series expansion arguments that we shall use below. However, we shall consider those expansions in a variety of settings, including the case where expansion (4.1) is extended to include cubic and quartic terms in means of sums of independent random variables. The relatively strong form of conditions (4.2) and (4.3) avoids our having to account for the numbers of finite moments that are needed in each case.

The predictor $\hat{\Theta}_i^{\text{pred}}$, denoting either $\hat{\Theta}_i^{\text{BLUP}}$ or $\hat{\Theta}_i^{\text{BP}}$, is expressible as a smooth function of $(\hat{\beta}, \hat{\xi}, \hat{\eta}_i)$ and $Y_i$; see equations (2.5) and (2.7). Taylor expanding this function for $(\hat{\beta}, \hat{\xi}, \hat{\eta}_i)$ in a neighbourhood of the true value, $(\beta^0, \xi^0, \eta_i^0)$, of $(\beta, \xi, \eta_i)$, using expansion (4.1), and its analogues for $\hat{\xi}$ and $\hat{\eta}_i$, to represent the difference between $(\hat{\beta}, \hat{\xi}, \hat{\eta}_i)$ and $(\beta^0, \xi^0, \eta_i^0)$, and noting that, if $b_i$ and $c_i$ are functions, such as $a_{in}$ and $a_{jin}$, satisfying $E\{b_i(Y_i)\} = E\{c_i(Y_i)\} = 0$, then

$$n\,E\left[\left\{\frac{1}{n}\sum_{i=1}^{n} b_i(Y_i)\right\}\frac{1}{n}\sum_{i=1}^{n} c_i(Y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} E\{b_i(Y_i)\,c_i(Y_i)\};$$

we deduce from expansion (4.1) that

$$\begin{aligned}\text{MSE}_i &= E\{(\hat{\Theta}_i^{\text{pred}} - \Theta_i)^2 | \mathcal{X}\} \\ &= T + \frac{a_i(\beta, \xi, \eta_i)}{n} + o(n^{-1}),\end{aligned} \tag{4.4}$$

where $a_i$ is a bounded function and $T$ denotes the version of $E\{(\hat{\Theta}_i^{\text{pred}} - \Theta_i)^2 | \mathcal{X}\}$ that would arise if, in the definition of $\hat{\Theta}_i^{\text{pred}}$, we were to replace $(\hat{\beta}, \hat{\xi}, \hat{\eta}_i)$ by $(\beta^0, \xi^0, \eta_i^0)$. If we were to carry the representation (4.1) to greater length, so that it included cubic and quartic terms in means of sums of independent random variables, then we could replace the $o(n^{-1})$ term on the right-hand side of equation (4.4) by $O(n^{-2})$.

It follows directly from the definition of $T$ that $T = \text{MSEpred}_i$, the latter being defined at equation (2.9). Therefore, by equation (4.4),

$$\text{MSE}_i = \text{MSEpred}_i + \frac{a_i(\beta, \xi, \eta_i)}{n} + o(n^{-1}).$$

## Acknowledgements

We are grateful to Dr Wayne Fuller for providing the real data set, and to three reviewers for helpful comments. The second author's research was partially funded by a co-operative agreement between the US Department of Agriculture National Resources Conservation Service and Iowa State University.

## Appendix A

Here we discuss simplifications that are possible when $\mu$ affects only the location of the distributions $Q(\mu,\xi)$ and $R(\mu,\eta)$. If property (2.6) holds then both $\sigma_Q(\mu,\xi)$ and $\sigma_V(\mu,\xi,\eta)$ depend only degenerately on $\mu$. In the case of $\sigma_Q(\mu,\xi)$ this is obvious. For $\sigma_V(\mu,\xi)$ the property follows from the fact that, if property (2.6) holds, $q(u|\mu,\xi)=q(u-\mu|0,\xi)$, and so, by equation (2.3),

$$\begin{aligned}\sigma_V(\mu,\xi,\eta)^2 &= \int \sigma_R(u+\mu,\eta)^2\, q(u|0,\xi)\,\mathrm{d}u\\ &= \sigma_R(0,\eta)^2 \int q(u|0,\xi)\,\mathrm{d}u\\ &= \sigma_R(0,\eta)^2.\end{aligned}$$

Therefore, property (2.6) implies that $\sigma_V(\mu,\xi,\eta)$ depends only on $\eta$; the roles of $\mu$ and $\xi$ are degenerate.

In such cases we may drop the quantity $f_i(\beta)$ from equation (2.4) or, more formally, replace it by any known value (such as 0), without influencing the value of $\rho_i$. Similarly, when property (2.6) is valid we may drop $f_i(\hat\beta)$ and $f_i(\hat\beta^*)$ from the analogues of equation (2.4) that are appropriate in the empirical and bootstrap cases respectively. Of course, these remarks are relevant only when the link function $\psi$ is the identity, since only then is equation (2.4) correct.

Property (2.6) is common. It holds when the distributions $Q(\mu,\xi)$ and $R(\mu,\eta)$ are normal with means equal to $\mu$ and variances which do not depend on $\mu$. Nevertheless, property (2.6) is not universal. For example, it fails if either $Q(\mu,\xi)$ or $R(\mu,\eta)$ denotes simply the exponential distribution with mean $\mu$.

## References

Alonso, A. M., Peña, D. and Romo, J. (2003) On sieve bootstrap prediction intervals. *Statist. Probab. Lett.*, **65**, 13–20.

Arora, V. and Lahiri, P. (1997) On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statist. Sin.*, **7**, 1053–1063.

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988) An error component model for prediction of county crop areas using survey and satellite data. *J. Am. Statist. Ass.*, **83**, 28–36.

Booth, J. G. and Hobert, J. P. (1998) Standard errors of prediction in generalized linear mixed models. *J. Am. Statist. Ass.*, **93**, 262–272.

Butar Butar, F. and Lahiri, P. (2003) On measures of uncertainty of empirical Bayes small-area estimators: model selection, model diagnostics, empirical Bayes and hierarchical Bayes. *J. Statist. Planng Inf.*, **112**, 63–76.

Clayton, D. and Kaldor, J. (1987) Empirical Bayes estimates of age standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

Dartigues, J. F., Gagnon, M., Letenneur, L., Barberger-Gateau, P., Commenges, D., Evaldre, M. and Salamon, R. (1992) Principle lifetime occupation and cognitive impairment in a French Elderly Cohort. *Am. J. Epidem.*, **135**, 981–988.

Datta, G. S. and Lahiri, P. (2000) A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statist. Sin.*, **10**, 613–627.

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

Fay, R. E. and Herriot, R. A. (1979) Estimation of income from small places: an application of James–Stein procedures to census data. *J. Am. Statist. Ass.*, **74**, 269–277.

Ghosh, M. and Maiti, T. (2004) Small area estimation based on natural exponential family-quadratic variance function models and survey weights. *Biometrika*, **91**, 95–112.

Hall, P. and Maiti, T. (2005) A general approach to small-area prediction. *Manuscript*. Australian National University, Canberra.

Jiang, J. (2003) Empirical best prediction for small-area inference based on generalized linear mixed models: model selection, model diagnostics, empirical Bayes and hierarchical Bayes. *J. Statist. Planng Inf.*, **111**, 117–127.

Jiang, J., Lahiri, P. and Wan, S.-M. (2002) A unified jackknife theory for empirical best prediction with M-estimation. *Ann. Statist.*, **30**, 1782–1810.

Kackar, R. and Harville, D. A. (1984) Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *J. Am. Statist. Ass.*, **79**, 853–862.

Kim, J. H. (2004a) Bias-corrected bootstrap prediction regions for vector autoregression. *J. Forecast.*, **23**, 141–154.
Kim, J. H. (2004b) Bootstrap prediction intervals for autoregression using asymptotically mean-unbiased estimators. *Int. J. Forecast.*, **20**, 85–97.
Lahiri, P. (2003a) On the impact of bootstrap in survey sampling and small-area estimation. *Statist. Sci.*, **18**, 199–210.
Lahiri, P. (2003b) A review of empirical best linear unbiased prediction for the Fay-Herriot small-area model. *Philip. Statistn*, **52**, 1–15.
Lahiri, P. and Maiti, T. (2002) Empirical Bayes estimation of relative risks in disease mapping. *Bull. Calc. Statist. Ass.*, **53**, 213–223.
Lahiri, P. and Rao, J. N. K. (1995) Robust estimation of mean squared error of small area estimators. *J. Am. Statist. Ass.*, **82**, 758–766.
Meza, J. L. (2003) Empirical Bayes estimation smoothing of relative risks in disease mapping. *J. Statist. Planng Inf.*, **112**, 43–62.
Nusser, S. M. and Goebel, J. J. (1997) The National Resources Inventory: a multi-resource monitoring program. *Ecol. Environ. Statist.*, **4**, 181–204.
Prasad, N. G. N. and Rao, J. N. K. (1990) The estimation of mean squared error of small area estimators. *J. Am. Statist. Ass.*, **85**, 163–171.
Rao, J. N. K. (2003) *Small Area Estimation*. Hoboken: Wiley.
Sjöstedt-de Luna, S. and Young, A. (2003) The bootstrap and kriging prediction intervals. *Scand. J. Statist.*, **30**, 175–192.
Slud, E. V. (2000) Comparison of aggregate versus unit-level models for small-area estimation. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 563–568.
You, Y. and Rao, J. N. K. (2002) Small area estimation using unmatched sampling and linking models. *Can. J. Statist.*, **30**, 3–15.
Wang, J. and Fuller, W. A. (2003) The mean squared error of small area predictors constructed with estimated area variances. *J. Am. Statist. Ass.*, **98**, 716–723.