Journal of Official Statistics, Vol. 25, No. 3, 2009, pp. 379–395

Multipurpose Weighting for Small Area Estimation

Hukum Chandra[1] *and Ray Chambers*[2]

Sample surveys are generally multivariate, in the sense that they collect data on more than one response variable. In theory, each variable can then be assigned an optimal weight for estimation purposes. However, it is a distinct practical advantage to have a single weight for all variables collected in the survey. This article describes how such multipurpose sample weights can be constructed when small area estimates of the survey variables are required. The approach is based on the model-based direct (MBD) method of small area estimation described in Chandra and Chambers (2005). Empirical results reported in this article show that MBD estimators for small areas based on multipurpose weights perform well across a range of variables that are often of interest in business surveys. Furthermore, these results show that the proposed approach is robust to model misspecification when applied to variables (e.g., those that contain a significant proportion of zeros) that are not suited to linear model-based small area estimation methods.

Key words: Multivariate surveys; multipurpose sample weights; MBD approach; mixed model; EBLUP.

1. Introduction

The weights that define the best linear unbiased predictor (BLUP) for the population total of a variable of interest (see Royall 1976) depend on the population level conditional covariance matrix for that variable, where the conditioning is with respect to the values of auxiliary variables. Unless this matrix is always proportional to a known matrix, this optimality is variable-specific. However, most surveys are multivariate, and it is often an advantage to have a common weight for all response variables. This is especially true where linear estimates are produced using the survey data. In what follows we refer to such weights as "multipurpose."

When a sufficiently rich set of auxiliary variables exist, and response variables can be assumed to be conditionally uncorrelated given these variables, multipurpose weights can be constructed by fitting a linear model for each response variable in terms of the complete set of auxiliary variables (see Chambers 1996). An essentially equivalent idea is to use a calibrated set of sample weights, where the calibration is with respect to these auxiliary variables (see Deville and Särndal 1992).

[1] Indian Agricultural Statistics Research Institute, Library Avenue, PUSA Campus, New Dehli-110012, India. Email: hchandra@iasri.res.in
[2] University of Wollongong, Centre for Statistical and Survey Methodology, Wollongong, NSW 2522, Australia. Email: ray@uow.edu.au
Acknowledgments: The first author gratefully acknowledges the financial support provided by a PhD scholarship from the U.K. Commonwealth Scholarship Commission. Constructive comments from an Associate Editor and three referees are also gratefully acknowledged. They resulted in the revised version of the article representing a considerable improvement on the original.

Small area estimation is now widely used in sample surveys. Many of the methods currently in use are variable-specific and based on the application of mixed models (Rao 2003). Weighted direct estimation for small areas based on these models is described in Chandra and Chambers (2005), who refer to this approach as the C-EBLUP method. Here we more accurately refer to it as the model-based direct (MBD) method of small area estimation. Since the weights used in MBD estimation are based on the second order properties of linear mixed models fitted to the survey variables, they are variable-specific. However, as noted above, there are obvious practical advantages to having a single multipurpose weight that can be used for small area estimation for all the survey variables. Consequently, in Section 2 of this article we replace the variable-specific BLUP optimality criterion that underlies the mixed model weights used in the MBD approach by a multivariate criterion that leads to a single set of optimal multipurpose weights for use in MBD estimation for small areas. Section 3 then presents empirical results on the performance of this approach. Finally, in Section 4 we summarize our results and make suggestions for further research.

2. Optimal Multipurpose Sample Weighting

2.1. Basic Concepts and Notation

Consider a population U consisting of N units, each of which has a value of a characteristic of interest y associated with it. The population vector $y_U = (y_1, \ldots, y_N)'$ is treated as a realization of a random vector $Y_U = (Y_1, \ldots, Y_N)'$, and our aim is estimation of the total $T_y = \sum_{j \in U} y_j$ (or mean $\bar{Y} = N^{-1} \sum_{j \in U} y_j$) of the values making up y_U. A sample s of n units is selected from U, and the y-values of the sample units are observed. We denote the set of $N - n$ nonsampled population units by r. We assume the availability of X_U, an $N \times p$ matrix of values of p auxiliary variables that are related, in some sense, to the values in y_U. In particular, y_U and X_U are related by the general linear model

$$E(y_U) = X_U \beta \text{ and } Var(y_U) = V_U \tag{1}$$

where β is a $p \times 1$ vector of unknown parameters and V_U is a positive definite covariance matrix. Without loss of generality, we arrange the vector y_U so that the first n elements correspond to the sample units, writing $y_U' = (y_s' y_r')$. We similarly partition X_U and V_U according to sample and nonsample units as

$$X_U = \begin{bmatrix} X_s \\ X_r \end{bmatrix} \text{ and } V_U \begin{bmatrix} V_{ss} & V_{sr} \\ V_{rs} & V_{rr} \end{bmatrix}$$

Here X_s is the $n \times p$ matrix of sample values of the auxiliary variable, V_{ss} is the $n \times n$ covariance matrix associated with the n sample units that make up the $n \times 1$ sample vector y_s. Corresponding nonsample quantities are denoted by a subscript of r, while V_{rs} denotes the $(N - n) \times n$ matrix defined by $Cov(y_r, y_s)$. It is known (see Royall 1976) that among linear prediction unbiased estimators $\hat{T}_y = w_s' y_s$ of T_y the variance of the prediction error, $Var(\hat{T}_y - T_y)$, is minimized by weights of the form

$$w_s = 1_n + H'\left(X_U' 1_N - X_s' 1_n\right) + \left(I_n - H' X_s'\right) V_{ss}^{-1} V_{sr} 1_{N-n} \tag{2}$$

Here $H = \left(X_s' V_{ss}^{-1} X_s\right)^{-1} X_s' V_{ss}^{-1}$, 1_m is a vector of ones of order m and I_n is the identity matrix of order n. The weights (2) define the best linear unbiased predictor (BLUP) for T_y given y_s, assuming (1) holds. In what follows, we will refer to them simply as the BLUP weights. By definition, these weights are calibrated on the variables in X_U and so exactly reproduce the known population totals defined by the columns of this matrix, i.e., $w_s' X_s = 1_N' X_U = T_x$. Furthermore, under the assumption that a mixed linear model can be used to specify the covariance matrix components V_{ss} and V_{sr} in (2), the MBD approach to small area estimation then uses these weights, with V_{ss} and V_{sr} replaced by suitable estimates, to define direct estimates of small area quantities.

2.2. *Optimal Multipurpose Weighting for Uncorrelated Variables*

Suppose we have K response variables and a common set of auxiliary variables with values defined by the population matrix X_U, and that Model (1) holds for each of them (although with different parameter values). Suppose initially that these variables are mutually uncorrelated. We use an extra subscript $k\,(k = 1, \ldots, K)$ to denote quantities associated with the kth response variable, for example V_{kss} and w_{ks} denote respectively the $n \times n$ covariance matrix and $n \times 1$ vector of sample weights that are associated with the $n \times 1$ vector y_{ks} of sample values of the kth response variable. With this notation, our aim is to derive an optimal set of multipurpose weights $w_s = \{w_j; j \in s\}$ for the K response variables measured in the survey. Let $T_k = 1_N' y_k$ denote the population total of y_k, with estimator $\hat{T}_k = w_s' y_{ks}$ based on these multipurpose weights. The weights w_s are then said to be ϕ -optimal if (a) $E(\hat{T}_k - T_k) = 0$ for each value of k, and (b) the ϕ-weighted total prediction variance $\sum_k \phi_k Var(\hat{T}_k - T_k)$ is minimized at w_s. Here ϕ_k is a user-specified nonnegative scalar quantity that reflects the relative importance attached to the kth response variable, with $\sum_k \phi_k = 1$.

Put $a_s = w_s - 1_s$. In order to derive an explicit expression for the ϕ-optimal multipurpose weights we first note that under (a)

$$E(\hat{T}_k - T_k) = E(a_s' y_{ks} - 1_{N-n}' y_{kr}) = E(a_s' X_s - 1_{N-n}' X_r)\beta_k = 0 \Rightarrow a_s' X_s = 1_{N-n}' X_r \tag{3}$$

Furthermore, the prediction variance for the estimator $\hat{T}_k = w_s' y_{ks}$ is then

$$Var(\hat{T}_k - T_k) = E(a_s' y_{ks} - 1_{N-n}' y_{kr})^2 = Var(a_s' y_{ks} - 1_{N-n}' y_{kr}) + [E(a_s' y_{ks} - 1_{N-n}' y_{kr})]^2$$

The second term on the right hand side above vanishes under (3), so that

$$\begin{aligned} Var(\hat{T}_k - T_k) &= a_s' Var(y_{ks}) a_s - 2a_s' Cov(y_{ks}, y_{kr}) 1_{N-n} + 1_{N-n}' Var(y_{kr}) 1_{N-n} \\ &= a_s' V_{kss} a_s - 2a_s' V_{ksr} 1_{N-n} + 1_{N-n}' V_{krr} 1_{N-n} \end{aligned} \tag{4}$$

We use the method of Lagrange multipliers to minimize (4) subject to (3). The corresponding Lagrangian loss function is

$$\Phi^{(1)} = \sum_{k=1}^{K} \phi_k \left\{ a_s' V_{kss} a_s - 2a_s' V_{ksr} 1_{N-n} \right\} + 2(a_s' X_s - 1_{N-n}' X_r)\lambda \tag{5}$$

where λ is a vector of Lagrange multipliers. Differentiating (5) with respect to a_s and setting the result equal to zero leads to

$$\begin{aligned}\frac{\partial \Phi^{(1)}}{\partial a_s} &= \sum_{k=1}^{K} \phi_k \{2V_{kss}a_s - 2V_{ksr}1_{N-n}\} + 2X_s\lambda = 0 \\ &\Rightarrow X_s\lambda = \sum_{k=1}^{K} \phi_k V_{ksr}1_{N-n} - \sum_{k=1}^{K} \phi_k V_{kss}a_s \\ &\Rightarrow a_s = \left(\sum_{k=1}^{K} \phi_k V_{kss}\right)^{-1} \left\{\sum_{k=1}^{K} \phi_k V_{ksr}1_{N-n} - X_s\lambda\right\} \end{aligned} \tag{6}$$

Multiplying both sides of (6) on the left by X_s' and using (3), we see that

$$\begin{aligned} X_s'a_s &= X_s'\left(\sum_{k=1}^{K} \phi_k V_{kss}\right)^{-1} \left(\sum_{k=1}^{K} \phi_k V_{ksr}1_{N-n}\right) - X_s'\left(\sum_{k=1}^{K} \phi_k V_{kss}\right)^{-1} X_s\lambda \\ &\Rightarrow X_r'1_{N-n} = X_s'U_1^{-1}W_1 1_{N-n} - X_s'U_1^{-1}X_s\lambda \\ &\Rightarrow \lambda = \left(X_s'U_1^{-1}X_s\right)^{-1}\left\{X_s'U_1^{-1}W_1 - X_r'\right\}1_{N-n} \end{aligned} \tag{7}$$

where $U_1 = \sum_{k=1}^{K} \phi_k V_{kss}$ and $W_1 = \sum_{k=1}^{K} \phi_k V_{ksr}$. Substituting (7) in (6) then yields the optimal value of a_s:

$$\begin{aligned} a_s^{(1)} &= U_1^{-1}W_1 1_{N-n} - U_1^{-1}X_s\lambda \\ &= \left[U_1^{-1}W_1 - U_1^{-1}X_s\left(X_s'U_1^{-1}X_s\right)^{-1}\left\{X_s'U_1^{-1}W_1 - X_r'\right\}\right]1_{N-n} \\ &= U_1^{-1}X_s\left(X_s'U_1^{-1}X_s\right)^{-1}\left(X'1_N - X_s'1_n\right) + \left[I_n - U_1^{-1}X_s\left(X_s'U_1^{-1}X_s\right)^{-1}X_s'\right]U_1^{-1}W_1 1_{N-n} \end{aligned}$$

That is, the optimal multipurpose sample weights for uncorrelated response variables are given by

$$w_s^{(1)} = 1_n + H_1'\left(X_U'1_N - X_s'1_n\right) + \left[I_n - H_1'X_s'\right]U_1^{-1}W_1 1_{N-n} \tag{8}$$

where

$$H_1 = \left(X_s'U_1^{-1}X_s\right)^{-1}X_s'U_1^{-1} = \left\{X_s'\left(\sum_{k=1}^{K} \phi_k V_{kss}\right)^{-1}X_s\right\}^{-1} X_s'\left(\sum_{k=1}^{K} \phi_k V_{kss}\right)^{-1}$$

Observe that the analytical form of the optimal multipurpose weights (8) is similar to the variable-specific BLUP weights (2), except that V_{kss} and V_{ksr} are replaced by the weighted sums $U_1 = \sum_k \phi_k V_{kss}$ and $W_1 = \sum_k \phi_k V_{ksr}$ respectively. Clearly (8) reduces to (2) for $K = 1$.

2.3. *Optimal Multipurpose Weighting for Correlated Variables*

Survey variables are correlated in general. Let $C_{kl} = Cov(y_k, y_l)$. The obvious generalization of the ϕ-weighted total prediction variance to this case leads to the loss function

$$\left(\sqrt{\phi_1}, \sqrt{\phi_2}, \ldots \sqrt{\phi_K}\right)' \Delta \left(\sqrt{\phi_1}, \sqrt{\phi_2}, \ldots \sqrt{\phi_K}\right) \tag{9}$$

where elements of the matrix $\Delta = \{\Delta_{kl}\}$ are given by

$$\Delta_{kl} = \begin{cases} Var(\hat{T}_k - T_k) & if\ k = l \\ Cov(\hat{T}_k - T_k, \hat{T}_l - T_l) & if\ k \neq l \end{cases}$$

and we now have

$$Cov(\hat{T}_k - T_k, \hat{T}_l - T_l) = a_s' C_{klss} a_s - 2a_s' C_{klsr} 1_{N-n} + 1_{N-n}' C_{klrr} 1_{N-n}$$

The Lagrange function to be minimized in this case is

$$\begin{aligned} \Phi^{(2)} &= \left(\sqrt{\phi_1}, \sqrt{\phi_2}, \ldots \sqrt{\phi_K}\right)' \Delta \left(\sqrt{\phi_1}, \sqrt{\phi_2}, \ldots \sqrt{\phi_K}\right) \\ &\quad + 2\left(a_s' X_s - 1_{N-n}' X_r\right)\lambda \\ &= \sum_k \phi_k Var(\hat{T}_{y_k} - T_{y_k}) + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} Cov(\hat{T}_{y_k} - T_{y_k}, \hat{T}_{y_l} - T_{y_l}) \\ &\quad + 2\left(a_s' X_s - 1_{N-n}' X_r\right)\lambda \\ &= \sum_k \phi_k \left\{a_s' V_{kss} a_s - 2a_s' V_{ksr} 1_{N-n} + 1_{N-n}' V_{krr} 1_{N-n}\right\} \\ &\quad + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} \left\{a_s' C_{klss} a_s - 2a_s' C_{klsr} 1_{N-n} + 1_{N-n}' C_{klrr} 1_{N-n}\right\} \\ &\quad + 2(a_s' X_s - 1_{N-n}' X_r)\lambda \end{aligned} \tag{10}$$

Differentiating (10) with respect to a_s and setting the result equal to zero yields

$$\begin{aligned} &\left\{\sum_k \phi_k V_{kss} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klss}\right\} a_s \\ &\quad - \left\{\sum_k \phi_k V_{ksr} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klsr}\right\} 1_{N-n} + X_s \lambda = 0 \\ &\Rightarrow U_2 a_s - W_2 1_{N-n} + X_s \lambda = 0 \Rightarrow a_s = U_2^{-1}(W_2 1_{N-n} - X_s \lambda) \end{aligned} \tag{11}$$

where $U_2 = \sum_k \phi_k V_{kss} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klss}$ and $W_2 = \sum_k \phi_k V_{ksr} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klsr}$. The same argument as in the uncorrelated case then leads to the optimal multipurpose weights for correlated survey variables

$$w_s^{(2)} = 1_n + H_2'\left(X_U' 1_N - X_s' 1_n\right) + \left[I_n - H_2' X_s'\right] U_2^{-1} W_2 1_{N-n} \tag{12}$$

where $H_2 = \left(X_s' U_2^{-1} X_s\right)^{-1} X_s' U_2^{-1}$. As in the uncorrelated case, we note that the weights defined by (12) have the same analytic form as the BLUP weights (2), except that in this

case V_{kss} and V_{ksr} are replaced by

$$U_2 = \sum_k \phi_k V_{kss} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klss}$$

and

$$W_2 = \sum_k \phi_k V_{ksr} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klsr}$$

respectively.

2.4. Application to Small Area Estimation

Following Chandra and Chambers (2005), we use the multipurpose weights (8) and (12) to construct model-based direct (MBD) estimates for small area means. In this case we assume that the population can be partitioned into m nonoverlapping small areas or domains, indexed by i in what follows. Thus, for example, the population size of area i is denoted by N_i and so on. The variable-specific MBD estimate of the mean of the kth response variable with values y_{kj} in area i is then

$$\hat{\bar{Y}}_{k,i}^{MBD} = \frac{\sum_{j \in s_i} w_{kj} y_{kj}}{\sum_{j \in s_i} w_{kj}} \tag{13}$$

where s_i denotes the sample (of size n_i) in area i and the weights w_{kj} are calculated using (2), substituting estimated values $\hat{V}_{kss}$ and $\hat{V}_{ksr}$ for the corresponding components of the covariance matrix of the population values of this variable. In order to define these estimates, we assume that these population values follow the linear mixed model

$$Y_{kU} = X_U \beta_k + Z_U u_k + e_{kU} \tag{14}$$

where $Y_{kU} = (Y'_{k,1}, \ldots, Y'_{k,m})'$, $X_U = (X'_1, \ldots, X'_m)'$, $Z_U = diag(Z_i; 1 \leq i \leq m)$, $u_k = (u'_{k,1}, \ldots, u'_{k,m})'$ and $e_{kU} = (e'_{k,1}, \ldots, e'_{k,m})'$ denote partitioning into area "components." Here $u_{k,i}$ is a vector-valued random effect associated with area i, with $Var(u_{k,i}) = \Sigma_{u,k} I_{N_i}$, and $e_{k,i}$ is the vector of individual random effects for area i, with $Var(e_{k,i}) = \Sigma_{e,k} I_{N_i}$. It follows that $Var(Y_{k,i}) = V_{k,i} = \Sigma_{e,k} I_{N_i} + Z_i \Sigma_{u,k} Z'_i$. The variance components $\Sigma_{e,k}$ and $\Sigma_{u,k}$ can be estimated from the sample data using standard methods (maximum likelihood, restricted maximum likelihood, i.e., REML, or method of moments). Substituting these estimated variance components back into the definition of $V_{k,i}$ and noting that $V_k = diag(V_{k,i}; 1 \leq i \leq m)$ then leads to a corresponding estimate of this population level covariance matrix. This can be appropriately partitioned into sample and nonsample components to give the estimated values $\hat{V}_{kss}$ and $\hat{V}_{ksr}$. We refer to the weights (2) with these estimated values substituted as the (variable-specific) EBLUP weights.

In order to use the multipurpose weights (8) and (12) in MBD estimation, we assume that the survey variables all follow the linear mixed model (14), with different parameter values and with normal random effects. Furthermore, for any two variables of interest, say

the kth and lth, area and individual random effects remain uncorrelated but now

$$\begin{pmatrix} u_{ki} \\ u_{li} \end{pmatrix} \sim MVN(0, \Sigma_u) \text{ with } \Sigma_u = \begin{bmatrix} Var(u_{ki}) & Cov(u_{ki}, u_{li}) \\ Cov(u_{li}, u_{ki}) & Var(u_{li}) \end{bmatrix} = \begin{bmatrix} \Sigma_{u,kk} & \Sigma_{u,kl} \\ \Sigma_{u,kl} & \Sigma_{u,ll} \end{bmatrix} \tag{15}$$

where u_{ki} and u_{li} are the $q \times 1$ vectors of random area effects, so that all variances and covariances (i.e., $\Sigma_{u,kk}$, $\Sigma_{u,kl}$, $\Sigma_{u,kl}$ and $\Sigma_{u,ll}$) have dimension $q \times q$, and

$$\begin{pmatrix} e_{kij} \\ e_{lij} \end{pmatrix} \sim MVN(0, \Sigma_e) \text{with} \Sigma_e = \begin{bmatrix} Var(e_{kij}) & Cov(e_{kij}, e_{lij}) \\ Cov(e_{lij}, e_{kij}) & Var(e_{lij}) \end{bmatrix} = \begin{bmatrix} \Sigma_{e,kk} & \Sigma_{e,kl} \\ \Sigma_{e,kl} & \Sigma_{e,ll} \end{bmatrix} \tag{16}$$

Here Σ_e is a 2×2 matrix. Hence

$$V_{k,i} = Var(Y_{k,i}) = \Sigma_{e,kk} I_{N_i} + Z_i \Sigma_{u,kk} Z_i'$$

$$V_{l,i} = Var(Y_{l,i}) = \Sigma_{e,ll} I_{N_i} + Z_i \Sigma_{u,ll} Z_i'$$

and

$$C_{kl,i} = Cov(Y_{k,i}, Y_{l,i}) = \Sigma_{e,kl} I_{N_i} + Z_i \Sigma_{u,kl} Z_i'$$

Given these definitions, we put $U_1 = diag(U_{1i}; 1 \le i \le m)$ and $W_1 = diag(W_{1i}; 1 \le i \le m)$ in (8) and $U_2 = diag(U_{2i}; 1 \le i \le m)$ and $W_2 = diag(W_{2i}; 1 \le i \le m)$ in (12). Here

$$U_{1i} = \sum_k \phi_k V_{kss,i} = \sum_k \phi_k \left(\Sigma_{e,kk} I_{n_i} + Z_{s,i} \Sigma_{u,kk} Z_{s,i}' \right)$$

$$W_{1i} = \sum_k \phi_k V_{ksr,i} = \sum_k \phi_k \left(Z_{s,i} \Sigma_{u,kk} Z_{r,i}' \right)$$

and

$$\begin{aligned} U_{2i} &= \sum_k \phi_k V_{kss,i} + \sum_k \sum_{l \ne k} \sqrt{\phi_k} \sqrt{\phi_l} C_{klss,i} \\ &= \sum_k \phi_k \left(\Sigma_{e,kk} I_{n_i} + Z_{s,i} \Sigma_{u,kk} Z_{s,i}' \right) + \sum_k \sum_{l \ne k} \sqrt{\phi_k} \sqrt{\phi_l} \left(\Sigma_{e,kl} I_{n_i} + Z_{s,i} \Sigma_{u,kl} Z_{s,i}' \right) \\ W_{2i} &= \sum_k \phi_k V_{ksr,i} + \sum_k \sum_{l \ne k} \sqrt{\phi_k} \sqrt{\phi_l} C_{klsr,i} \\ &= \sum_k \phi_k \left(Z_{s,i} \Sigma_{u,kk} Z_{r,i}' \right) + \sum_k \sum_{l \ne k} \sqrt{\phi_k} \sqrt{\phi_l} \left(Z_{s,i} \Sigma_{u,kl} Z_{r,i}' \right) \end{aligned}$$

In practice, the bivariate variance components $\Sigma_{u,kk}$, $\Sigma_{u,kl}$, $\Sigma_{e,kk}$ and $\Sigma_{e,kl}$ (see (15) and (16)) are unknown and must be estimated from the survey data. For example, in the empirical study described in the next section these components were estimated using the method of moments (Henderson's Method 3). In any case, substituting estimates for these components in the formulae above then enables us to compute U_1, W_1, U_2, and W_2, and hence the multipurpose weights (8) and (12). Computation of MBD estimates for the small

area means of the different survey variables is then straightforward using (13), with these multipurpose weights replacing the variable-specific EBLUP weights there.

As noted earlier, the multipurpose weights (8) and (12) are essentially EBLUP type weights based on "importance averaging" of the variance and covariance components associated with the different survey variables. This motivates us to consider a second approach to deriving multipurpose weights based on corresponding "importance averaging" of the variable-specific EBLUP sample weights (2) for these variables. That is, we simply define our multipurpose weights as the importance-weighted average of the variable-specific weights (2) across all K survey variables. This leads to weights

$$w_s^{(3)} = \sum_k \phi_k w_{sk} \tag{17}$$

where w_{sk} denotes the value of (2) for the kth survey variable and ϕ_k denotes the relative importance of this variable, with $\sum_k \phi_k = 1$.

3. An Empirical Study

In this section we report on a design-based simulation study that illustrates the performance of small area MBD estimation combined with multipurpose weights. Our basic data come from the same sample of 1,652 Australian broadacre farms that participated in the annual Australian Agricultural and Grazing Industries Survey (AAGIS) that was carried out by the Australian Bureau of Agricultural and Resource Economics in the late 1980s and were used in the simulation study reported in Chambers (1996). Here we used these sample farms to generate a target population of 81,982 farms by sampling with replacement from them with probabilities proportional to their sample weights. We drew 1,000 independent stratified random samples from this (fixed) population, with total sample size in each simulation equal to the original sample size (1,652) and with strata defined by the 29 different Australian broadacre agricultural regions. Sample sizes within these regions were fixed to be the same as in the original sample. Note that these varied from a low of 6 to a high of 117, allowing an evaluation of the performance of different small area estimation methods across a range of realistic small area sample sizes. (See Chandra and Chambers (2005) for more details).

We consider $K = 8$ variables of interest. These are (i) TCC = total cash costs (in Australian dollars, A$) of the farm business over the surveyed year, (ii) TCR = total cash receipts (A$) of the farm business over the surveyed year, (iii) FCI = farm cash income (A$), defined as TCR − TCC, (iv) Crops = area under crops (in hectares), (v) Cattle = number of cattle on the farm, (vi) Sheep = number of sheep on the farm, (vii) Equity = total farm equity (A$), and (viii) Debt = total farm debt (A$). Our aim is to estimate the averages of these variables in each of the 29 regions. In doing so, we use the fact that these regions can be grouped into three zones (Pastoral, Mixed Farming, and Coastal), with farm area (hectares) known for each farm in the population. This auxiliary variable is referred to as Size in what follows.

Although the linear relationship between the eight target variables and Size is rather weak in the population, this improves when separate linear models are fitted within six post-strata, defined by splitting each zone into small farms (farm area less than zone

median) and large farms (farm area greater than or equal to zone median). The mixed model (14) was therefore specified so that the matrix X_U of auxiliary variable values included an effect for Size, effects for the post-strata and effects for interactions between Size and the post-strata. Two different specifications for Z_U (corresponding to whether a random slope on Size was included or not) were considered. This leads to two specifications for (14), a random intercept specification (where Z_i is vector of 1's) and a random slope specification (where both the model intercept and the slope parameter for Size are considered as random, so Z_i has an additional column defined by values of this auxiliary variable). We refer to these as Model I and as Model II, respectively, below. We use REML estimates of random effects parameters, obtained via the *lme* function in R (Bates and Pinheiro 1998) when fitting (14) to individual survey variables. When fitting the multivariate mixed models defined by (15) and (16) we use the method of moments (Rao 2003).

The simulation study investigated the performance of five different estimators of the 29 regional means, along with corresponding estimators of their mean squared error. These are:

- EBLUP – EBLUP weights (Equation 14);
- MBD-E – MBD estimator (13) with EBLUP weights (2);
- MBD-I – MBD estimator (13) with multipurpose weights (8) for independent variables;
- MBD-C – MBD estimator (13) with multipurpose weights (12) for correlated variables;
- MBD-W – MBD estimator (13) with multipurpose weights (17) defined by weight averaging.

Mean squared errors for the EBLUP were estimated using the approach of Prasad and Rao (1990), while mean squared errors for the various MBD estimators were estimated using the robust method described in Chandra and Chambers (2005), which itself is an application of the heteroskedasticity robust method of prediction variance estimation described in Royall and Cumberland (1978). Note that this robust MSE estimator is a "plug in" estimator, in the sense that unknown parameters in the actual MSE are replaced by sample estimates. As a consequence, it does not include an adjustment for the extra variability introduced by this substitution. Given that this method of MSE estimation has been empirically demonstrated to have good model-based as well as repeated sampling properties (see Chambers and Tzavidis 2006; Chandra, Salvati, and Chambers 2007; Tzavidis, Salvati, Pratesi, and Chambers 2007), we do not anticipate that such an adjustment will make a substantial difference in realistic applications.

The simulation study was carried out in five stages. In the first stage, Model I was assumed and the performance of the three estimators MBD-E, MBD-I, and MBD-C for two variables (TCC and TCR) was investigated to see if there were gains to be had from exploiting correlations among the survey variables. Table 1 sets out the average and median values of various summary performance measures for MBD-E, MBD-I and MBD-C under Model I when used with TCC and TCR. We observed that the weights generated under all three methods were very similar in our simulations, and, as a consequence, the results set out in this table are also very similar for these variables (regional specific results generated by these methods were also virtually identical).

Table 1. Average relative bias (ARB), median relative bias (MRB), average relative root mean squared error (ARRMSE), median relative root mean squared error (MRRMSE) and average coverage rate (ACR) generated by MBD-E, MBD-I and MBD-C for TCC and TCR under Model I. All averages and medians are expressed as percentages and are over the 29 regions of interest

Variable	Criterion	MBD-E	MBD-I	MBD-C
TCC	ARB	−2.99	−2.67	−2.71
	ARRMSE	20.32	20.39	20.39
	ACR	92	92	92
	MRB	−0.92	−0.85	−0.86
	MRRMSE	14.29	14.36	14.35
TCR	ARB	−2.38	−2.62	−2.67
	ARRMSE	21.21	21.13	21.12
	ACR	92	92	92
	MRB	−0.52	−0.56	−0.57
	MRRMSE	13.28	13.27	13.27

Furthermore, since Spearman's rho for TCC and TCR was 0.92 in the AAGIS sample that underpinned the study population, the results also indicate that the MBD method based on the multipurpose weights (8) is not sensitive to high correlations between the target variables. Although not presented here, results from model-based simulations of target variables with different levels of correlation (varying between 0.0 to 0.75) support this conclusion. The simulation results discussed below therefore focus on MBD-I.

In the second stage of the study we compared the performance of the four estimation methods EBLUP, MBD-E, MBD-I, and MBD-W under Models I and II for the five response variables (TCC, TCR, FCI, Cattle and Sheep) where both models can be fitted. Tables 2 and 3 show the summary performances generated by these four methods for the five variables TCC, TCR, FCI, Cattle and Sheep under both these models. Under the better-fitting Model II (Table 3), all three MBD weighting methods perform similarly with respect to both measures (ARB and MRB) of bias performance, with EBLUP performing somewhat worse in this regard. As far as mean squared error performance is concerned, we see that the multipurpose method MBD-I slightly outperforms the variable-specific method MBD-E on both AARMSE and MRRMSE, with the multipurpose method MBD-W on a par with these two methods as far as MRRMSE is concerned, but notably less efficient with respect to AARMSE. All three MBD methods clearly dominate the EBLUP in terms of mean squared error performance under Model II. Under Model I (Table 2) the picture is a little less clear, with the two multipurpose methods MBD-I and MBD-W recording substantially better bias performances than the variable-specific MBD-E and EBLUP, and comparable performances to MBD-E with respect to mean squared error. In this case the EBLUP performs somewhat better as far as mean squared error is concerned, but still exhibits large instability on occasion (e.g., AARMSE for the Cattle and Sheep variables).

The fact that both the "covariance smoothed" multipurpose MBD-I and the "weight smoothed" multipurpose MBD-W perform somewhat better in terms of mean squared error than the variable-specific MBD-E in Table 3 deserves comment, since one would expect that an estimation method (MBD-E) that is tuned to a particular variable should

Table 2. Average relative bias (ARB), median relative bias (MRB), average relative root mean squared error (ARRMSE), median relative root mean squared error (MRRMSE) and average coverage rate (ACR) for the five variables best suited to linear mixed modelling. All averages and medians are expressed as percentages and are over the 29 regions of interest. Model I is assumed

Criterion	Method	TCC	TCR	FCI	Cattle	Sheep
ARB	EBLUP	4.24	5.48	6.93	138.48	304.24
	MBD-E	−2.49	−9.25	−13.80	−15.05	−7.33
	MBD-I	−1.54	−1.30	−0.50	−1.78	0.69
	MBD-W	−1.29	−1.02	−0.04	−1.35	0.98
MRB	EBLUP	1.55	0.55	−2.08	0.95	−0.23
	MBD-E	−0.82	−3.87	−2.83	−4.79	−4.48
	MBD-I	−0.61	−0.42	−0.56	−0.97	−0.35
	MBD-W	−0.52	−0.39	−0.54	−0.75	−0.30
ARRMSE	EBLUP	19.92	21.76	63.93	304.74	906.18
	MBD-E	20.56	23.34	54.42	37.45	24.88
	MBD-I	20.86	21.77	59.72	33.29	30.24
	MBD-W	20.85	21.77	60.07	33.36	30.64
MRRMSE	EBLUP	15.74	14.83	40.41	25.97	13.00
	MBD-E	14.45	16.20	35.85	30.34	15.50
	MBD-I	14.69	13.41	42.09	30.55	14.67
	MBD-W	14.74	13.46	42.45	30.56	14.67
ACR	EBLUP	90	88	87	86	91
	MBD-E	92	91	94	93	94
	MBD-I	92	92	94	95	96
	MBD-W	92	92	94	95	96

have the edge when used to estimate that variable. However, this expectation is based on the assumption that the model used in the variable-specific estimator is "correct." For the population used in our simulations this is certainly not the case, with the assumption of linearity representing a convenient approximation that holds better for some variables than others. In this context a multipurpose approach to weighting is qualitatively more robust than variable-specific weighting since it does not put too much importance on any particular variable and its underlying model assumptions.

In Figure 1 we show the regional level performances of EBLUP, MBD-E, and MBD-I when estimating average TCC under Model II. We do not show results for MBD-W since these were very similar to those shown for MBD-I. Also, we do not show results for Model I since these were only marginally different from those for Model II. A considerable reduction in relative biases under multipurpose weighting can be seen in most regions. A similar pattern of results was observed for TCR, FCI, Cattle, and Sheep.

From Figure 1 we see that the weighting methods (MBD-E and MBD-I) do not perform well in Region 21, recording very large values of relative RMSE. Inspection of the data indicates that this is because of a small number of outlying estimates that were generated during the simulations, due to the selection into sample of a large outlier (TCC $>$ A\$30,000,000) in this region. When we discard these outlying estimates, all three weighting methods, and particularly MBD-I and MBD-W, perform well for TCC across all 29 regions. Similar results were observed for the other four variables TCR, FCI,

Table 3. Average relative bias (ARB), median relative bias (MRB), average relative root mean squared error (ARRMSE), median relative root mean squared error (MRRMSE) and average coverage rate (ACR) for the five variables best suited to linear mixed modelling. All averages and medians are expressed as percentages and are over the 29 regions of interest. Model II is assumed

Criterion	Method	TCC	TCR	FCI	Cattle	Sheep
ARB	EBLUP	2.98	2.85	16.70	131.66	2.63
	MBD-E	−2.13	−1.25	0.50	−0.29	3.66
	MBD-I	−1.67	−1.29	0.74	−1.95	1.10
	MBD-W	−1.30	−0.72	3.17	−1.29	0.93
MRB	EBLUP	0.61	1.37	3.98	0.62	0.00
	MBD-E	−0.47	−0.51	0.35	−0.31	0.00
	MBD-I	−0.65	−0.50	0.24	−0.30	−0.15
	MBD-W	−0.52	0.01	0.53	−0.22	−0.09
ARRMSE	EBLUP	19.87	20.28	68.85	231.08	630.01
	MBD-E	20.15	21.46	65.43	30.80	37.82
	MBD-I	19.06	21.03	64.03	30.09	32.04
	MBD-W	27.13	34.84	129.29	45.16	34.99
MRRMSE	EBLUP	16.40	15.61	33.89	22.64	11.73
	MBD-E	13.16	12.39	37.64	28.79	14.68
	MBD-I	12.84	12.18	37.92	24.84	14.77
	MBD-W	12.84	12.71	37.62	24.93	14.72
ACR	EBLUP	85	86	84	86	89
	MBD-E	93	93	90	95	96
	MBD-I	93	93	94	95	96
	MBD-W	93	93	94	95	96

Cattle, and Sheep. The outlier in Region 21 also impacts on the EBLUP MSE estimator, leading to low coverage for the EBLUP in this region. This low coverage is evident in Regions 3 and 7 as well, mainly because the highly skewed distribution of Area in these regions leads to negative values for this estimator. The same phenomenon underpins the high values of relative bias and relative root mean squared error of the EBLUP for the Cattle and Sheep variables that are evident in Tables 2 and 3. Here zero values in the data for these variables tended to generate negative values for the EBLUP.

In the third stage of the simulation study, we used the multipurpose weights derived in the second stage (i.e., weights based on the $K = 5$ variables TCC, TCR, FCI, and Cattle and Sheep) in MBD-I to evaluate the performance of this estimator for the three variables Crops, Equity and Debt that were impossible to fit using model II because they contained many zero values. In particular, in Table 4 we contrast the performances of the variable-specific estimators EBLUP and MBD-E with that of MBD-I for these three variables. Note that these results are based on Model I, since Model II cannot be used here. We see that MBD-I is again clearly the method of choice, with EBLUP performing particularly badly – as one might expect given the large number of zero values in the data for Crops, Equity, and Debt. This is evident when we look at Figure 2, which shows the regional specific performances of the three methods for Crops. Here we see that the EBLUP method fails in Regions 2, 6, 9, and 18. These are regions where there are a large number of zero values for this variable.

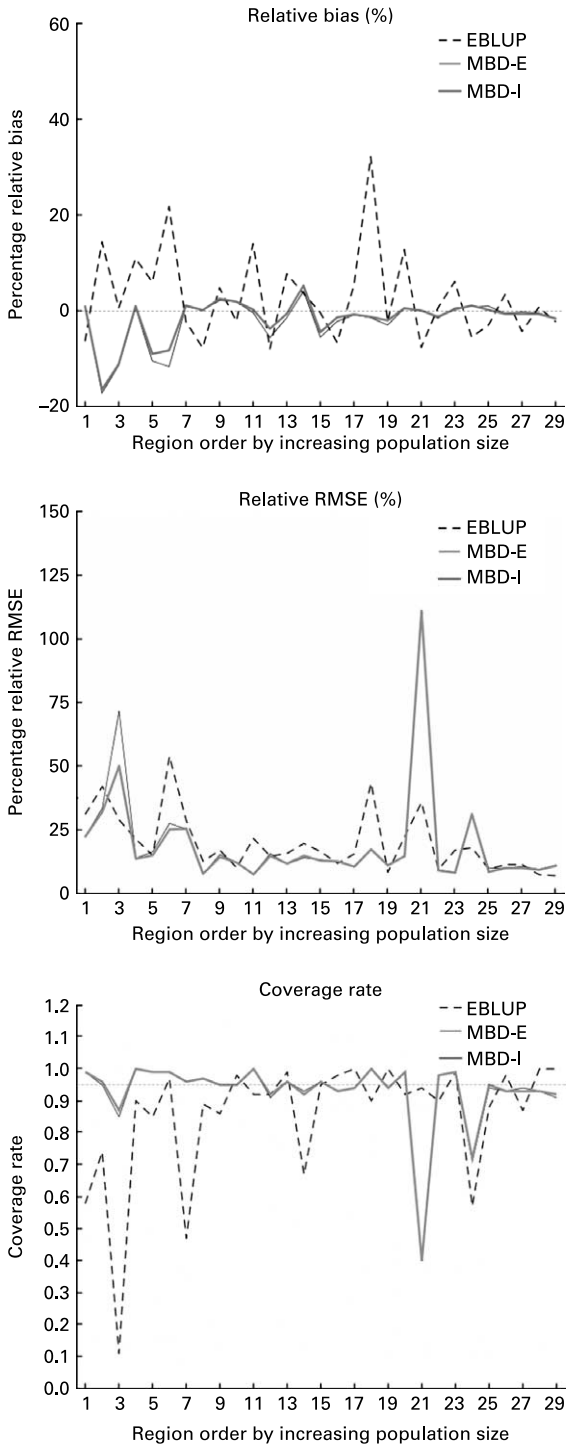

Fig. 1. Regional performance of EBLUP (dashed line), MBD-E (thin line) and MBD-I (thick line) for TCC under Model II

Table 4. Average relative bias (ARB), median relative bias (MRB), average relative root mean squared error (ARRMSE), median relative root mean squared error (MRRMSE) and average coverage rate (ACR) for EBLUP, MBD-E and MBD-I for Crops, Equity and Debt under Model I. All averages are expressed as percentages and are over the 29 regions of interest

Criterion	Methods	Crops	Equity	Debt
ARB	EBLUP	90.31	4.36	8.39
	MBD-E	0.00	−9.32	−4.94
	MBD-I	−0.21	−1.20	−0.96
MRB	EBLUP	0.00	−0.28	1.16
	MBD-E	−0.84	−3.51	−2.36
	MBD-I	0.00	−0.32	−0.61
ARRMSE	EBLUP	123.96	18.51	29.02
	MBD-E	23.53	19.14	27.71
	MBD-I	22.92	17.05	28.57
MRRMSE	EBLUP	15.10	12.32	21.49
	MBD-E	15.76	16.18	23.70
	MBD-I	15.80	13.52	24.88
ACR	EBLUP	95	88	91
	MBD-E	96	92	93
	MBD-I	96	94	93

Since Model I can be fitted to all eight variables, and so can be used to define multipurpose weights that depend on all of them, in stage four of our simulations we computed MBD-I using weights defined by both the limited ($K = 5$) and full ($K = 8$) set of target variables in (8). However, we again noted that these weights were not substantially different, and led to very similar estimates. Consequently we do not present these results here.

Note that in all four of the simulation stages so far, we assign equal importance to all variables included in derivation of the multipurpose weights. In the final stage of our simulations we therefore replicated the stage two simulations for MBD-I assuming Model I, but this time also considered three other ways of assigning importance factors to the different variables. The first assigned an importance factor to variable k inversely proportional to its estimated individual level variance component $\hat{\sigma}^2_{e,k}$. The second assigned an importance factor inversely proportional to $\hat{\sigma}^2_{u,k} + \hat{\sigma}^2_{e,k}$, i.e., it also took account of estimated between region variability. Finally, the third (denoted *EI* in Table 5) allocated values for these importance factors proportional to a subjective assessment that financial variables TCC, TCR, and FCI are five times more important, in terms of prediction variance, than the production variables Cattle and Sheep in the AAGIS. Table 5 provides summary details of the performance of the MBD-I weighting method when the multipurpose weights (based on TCC, TCR, FCI, and Cattle and Sheep) are computed using these alternative importance factors. The four different weighting methods showed remarkably small differences in the weights that they generated. That is, for the population considered in the simulation study, there is little to choose between these different

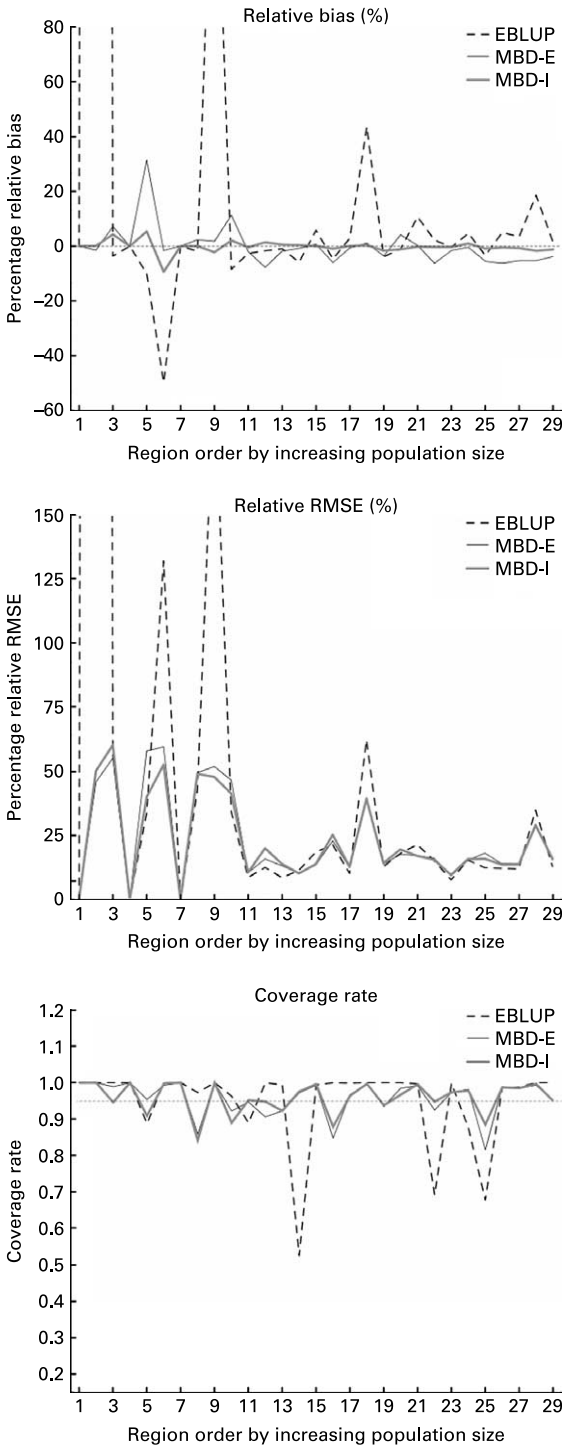

Fig. 2. Regional performances of EBLUP (dashed line), MBD-E (thin line) and MBD-I (thick line) for Crops under Model I

Table 5. Average relative bias (ARB), average relative root mean squared error (ARRMSE) and average coverage rate (ACR) for multi-purpose weighting (MBD-I) based on different choices of ϕ_k for $K = 5$ target variables (TCC, TCR, FCI, Cattle, and Sheep) under Model I. Note that the EI (economic importance) option corresponds to $\phi_k = 100/340$ for the economic performance variables TCC, TCR and FCI (the main target of the AAGIS), and $\phi_k = 20/340$ for the production variables Cattle and Sheep

Criterion	ϕ_k^{-1}	TCC	TCR	FCI	Cattle	Sheep
ARB	K	−1.54	−1.30	−0.50	−1.78	0.69
	$\hat{\sigma}_{e,k}^2$	−1.69	−1.48	−0.82	−2.03	0.52
	$\hat{\sigma}_{u,k}^2 + \hat{\sigma}_{e,k}^2$	−1.64	−1.42	−0.70	−1.95	0.57
	EI	−1.54	−1.30	−0.50	−1.79	0.68
ARMSE	K	20.86	21.77	59.72	33.29	30.24
	$\hat{\sigma}_{e,k}^2$	20.83	21.71	58.00	33.19	29.99
	$\hat{\sigma}_{u,k}^2 + \hat{\sigma}_{e,k}^2$	20.85	21.75	58.15	33.25	30.11
	EI	20.86	21.77	58.30	33.29	30.24
ACR	K	92	92	94	95	96
	$\hat{\sigma}_{e,k}^2$	92	92	94	95	96
	$\hat{\sigma}_{u,k}^2 + \hat{\sigma}_{e,k}^2$	92	92	94	95	96
	EI	92	92	94	95	96

approaches to assessing the relative importance of the variables that define the multipurpose weights, and our decision to treat all five as being of equal importance seems a reasonable overall choice.

4. Concluding Comments

In this article we develop two loss functions that can be used to compute optimal multipurpose weights suitable for use in small area estimation using MBD estimators. The first (8) ignores the correlations between the survey variables, while the second (12) takes these into account. For the population considered in our simulation studies the performances of the MBD estimators MBD-I and MBD-C based on "covariance smoothed" multipurpose weighting are virtually identical, i.e., there are no gains from taking account of the correlations between the survey variables when constructing the multipurpose weights. We also investigated an alternative "weight smoothing" approach to constructing multipurpose weights. However, our empirical results indicate that this method is marginally less efficient than the loss function based MBD-I method. Our simulations also indicate that these multipurpose weights remain efficient across a wide range of variables, even variables that have not been used in the definition of the multipurpose weights. This can be important in some situations (e.g., where variables have many zero values) where standard mixed models cannot be fitted and the usual EBLUP-based methods are inappropriate.

A final comment concerns the performance of the EBLUP in our simulations. By definition, the EBLUP is the most efficient linear estimator provided its model assumptions hold. However, this efficiency was not evident in our simulations, in all probability because this estimator's underlying model assumptions fail when applied to

the AAGIS data. This does raise the issue of when is it appropriate to use the EBLUP in small area estimation. To argue that this should only be when its assumptions hold seems to be somewhat impractical, since the reality is that it is never possible to know if this is the case. What happens in practice is that the EBLUP is computed because a mixed linear model provides a reasonable fit to the data, with significant small area effects. This was certainly the case for the AAGIS sample data that underpinned our simulations. Undoubtedly a better fit to the AAGIS data could have been obtained by adoption of a more complex model. For example, a mixed linear model on the logarithmic scale is a much better fit for the strictly positive values of TCC, TCR, and Cattle and Sheep in these data. However, taking advantage of this nonlinearity leads to an extra level of technical complexity in estimation, which we hope to report on elsewhere. Our concern here is to demonstrate that combining a simple linear method (MBD) of model-based small area estimation with a multipurpose approach to sample weighting leads to a robust estimation method, i.e., one that works well in a variety of situations, even those where underlying model assumptions are approximate at best.

5. References

Bates, D.M. and Pinheiro, J.C. (1998). Computational Methods for Multilevel Models. Available from http://franz.stat.wisc.edu/pub/NLME/

Chambers, R.L. (1996). Robust Case-weighting for Multipurpose Establishment Surveys. Journal of Official Statistics, 12, 3–32.

Chambers, R.L. and Tzavidis, N. (2006). M-quantile Models for Small Area Estimation. Biometrika, 93, 255–268.

Chandra, H. and Chambers, R.L. (2005). Comparing EBLUP and C-EBLUP for Small Area Estimation. Statistics in Transition, 7, 637–648.

Chandra, H., Salvati, N., and Chambers, R.L. (2007). Small Area Estimation for Spatially Correlated Populations. A Comparison of Direct and Indirect Model-based Methods. Statistics in Transition, 8, 887–906.

Deville, J.C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Prasad, N.G.N. and Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. Journal of the American Statistical Association, 85, 163–171.

Rao, J.N.K. (2003). Small Area Estimation. New York: Wiley.

Royall, R.M. (1976). The Linear Least-squares Prediction Approach to Two-stage Sampling. Journal of the American Statistical Association, 71, 657–664.

Royall, R.M. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. Journal of the American Statistical Association, 73, 351–358.

Tzavidis, N., Salvati, N., Pratesi, M., and Chambers, R.L. (2008). M-quantile Models with Application to Small Area Estimation and Poverty Mapping. Statistical Methods and Applications, 17, 393–411.

Received August 2006
Revised January 2009