Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

06-08

Multipurpose Small Area Estimation

Hukum Chandra and Ray Chambers

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Multipurpose Small Area Estimation

Hukum Chandra[1] and Ray Chambers[2]

1. Southampton Statistical Sciences Research Institute
University of Southampton
Highfield, Southampton, SO17 1BJ, UK

2. Centre for Statistical and Survey Methodology
University of Wollongong
Wollongong, NSW, 2522, Australia

August 15 2006

Abstract

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 paper 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 Chambers and Chandra (2006). Empirical results reported in this paper 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 and also efficient when used with variables that are not suited to standard methods of small area estimation (e.g. variables that contain a significant proportion of zeros).

Keywords: 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 variance/covariance matrix for that variable. 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).

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 Chambers and Chandra (2006), who refer to this approach 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 from having a single multipurpose weight that can be used for small area estimation for all the survey variables. Consequently, in section 2 of this paper we replace the variable specific BLUP optimality criterion that underlies the mixed model weights used in the MBD approach by a modified 'total variability' 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 summarise 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 $y_U = (y_1,...,y_N)'$ is treated as the realisation of a random vector $Y_U = (Y_1,...,Y_N)'$, 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 defining 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 $N - n$ non-sampled 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

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

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 non-sample 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 non-sample 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 minimised by weights of the form

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

Here $H = \left(X_s'V_{ss}^{-1}X_s\right)^{-1}X_s'V_{ss}^{-1}$, 1_m is a vectors of ones of order m and I_n is the identity matrix of order n. We refer to the weights (2) as the best linear unbiased prediction (BLUP) weights for y. 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 (see Chambers and Chandra, 2006) 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 further that these variables are mutually uncorrelated. We use an extra subscript k $(k = 1,....,K)$ to denote quantities associated with the k^{th} 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 k^{th} 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 minimised at w_s. Here ϕ_k is a user-specified non-negative scalar quantity that reflects the relative importance attached to the k^{th} 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)

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

Furthermore, the prediction variance for 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

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

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

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

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

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

$$\Rightarrow \qquad X_s \lambda = \sum_{k=1}^{K} \phi_k V_{ksr} 1_{N-n} - \sum_{k=1}^{K} \phi_k V_{kss} a_s$$

$$\Rightarrow \qquad a_s = \left(\sum_{k=1}^{K} \phi_k V_{kss} \right)^{-1} \left\{ \sum_{k=1}^{K} \phi_k V_{ksr} 1_{N-n} - X_s \lambda \right\} \qquad (6)$$

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

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

$$\Rightarrow \qquad X_r' 1_{N-n} = X_s' U_1^{-1} W_1 1_{N-n} - X_s' U_1^{-1} X_s \lambda$$

$$\Rightarrow \qquad \lambda = \left(X_s' U_1^{-1} X_s \right)^{-1} \left\{ X_s' U_1^{-1} W_1 - X_r' \right\} 1_{N-n} \qquad (7)$$

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:

$$a_s^{(1)} = U_1^{-1} W_1 1_{N-n} - U_1^{-1} X_s \lambda = \left[U_1^{-1} W_1 - U_1^{-1} X_s \left(X_s' U_1^{-1} X_s \right)^{-1} \left\{ X_s' U_1^{-1} W_1 - X_r' \right\} \right] 1_{N-n}$$

$$= U_1^{-1} X_s \left(X_s' U_1^{-1} X_s \right)^{-1} \left(X' 1_N - X_s' 1_n \right) + \left[I_n - U_1^{-1} X_s \left(X_s' U_1^{-1} X_s \right)^{-1} X_s' \right] U_1^{-1} W_1 1_{N-n}.$$

That is, the optimal multipurpose sample weights are given by

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

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

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

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

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

where $U_2 = \sum_k \phi_k V_{kss} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klss}$ and $W_2 = \sum_k \phi_k V_{ksr} + \sum_k \sum_{l \neq k} \sqrt{\phi_k}\sqrt{\phi_l} C_{klsr}$.

Proceeding 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 variables 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 Chambers and Chandra (2006), 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* non-overlapping 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 k^{th} response variable with values y_{kj} in area i is then

$$\hat{\bar{Y}}_{k,i}^{MBD} = \sum\nolimits_{j \in s_i} w_{kj} y_{kj} \Big/ \sum\nolimits_{j \in s_i} w_{kj} \tag{13}$$

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 $Y_{kU} = (Y'_{k,1},.....,Y'_{k,m})'$, $X_U = (X'_1,.....,X'_m)'$, $Z_U = diag(Z_i; 1 \le i \le m)$, $u_k = (u_{k,1},...,u_{k,m})'$ and $e_{kU} = (e_{k,1},...,e_{k,m})'$ denote partitioning into area 'components'. Here $u_{k,i}$ is a 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 $V_k = diag(V_{k,i}; 1 \le i \le m)$ then leads to a corresponding estimate of this population level covariance matrix. This can be appropriately partitioned into sample and non-sample 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 normal random effects. Furthermore, for any two variables of interest, say the k^{th} and l^{th}, area and individual random effects remain uncorrelated but now

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

and

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

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. 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 k^{th} 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. The basis of this study is the same target population of N = 81982 farms, the same 1000 independent replications of a stratified random sampling design with overall sample size $n = 1652$ and the same m = 29 small areas of interest (defined by agricultural regions) that underpin the simulation results reported in Chandra and Chambers (2005). Note that regional sample sizes in this design are fixed from simulation to simulation but vary between regions, ranging from a low of 6 to a high of 117, and hence allowing an evaluation of the performance of the different methods considered across a range of realistic small area sample sizes. See Chandra and Chambers (2005) for more details.

Here we consider K = 8 variables of interest. These are (i) TCC = total cash costs (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

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 average of these variables in each of the 29 different 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. These post-strata are 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. 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 the variable specific EBLUP under (14), referred to as EBLUP below; the MBD estimator (13) based on variable specific EBLUP weights (2), referred to as MBD0 below; the MBD estimator (13) based on multipurpose weights (8), referred to as MBD1-A below; the MBD estimator (13) based on multipurpose weights (12), referred to as MBD1-B below; and the MBD estimator (13) based on multipurpose weights (17), referred to as MBD2 below. 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 Chambers and Chandra (2006), which itself is an application of the heteroskedasticity robust method of prediction variance estimation described in Royall and Cumberland (1978).

The simulation study was carried out in five stages. In the first stage, model I was assumed and the performance of the three estimators MBD0, MBD1-A and MBD1-B for two variables (TCC and TCR) was investigated to see if there were gains to be had from exploiting correlations among the survey variables. As noted earlier, we used the method of moments (Henderson's Method 3) to estimate model parameters in this case. Results from this stage are set out in Table 1. In the second stage of the study we compared the performance of the four estimation methods EBLUP, MBD0, MBD1-A and MBD2 under models I and II for the 5 response variables (TCC, TCR, FCI, Cattle and Sheep) where both models can be fitted. Results from this stage are presented in Tables 2 and 3 and in Figure 1. Note that the remaining three target variables in the study (Crops, Equity and Debt) are not suited to linear modeling via (14) under model II because of the presence of large numbers of zeros. Consequently, in the third stage of the study, we used the multipurpose weights derived in the second phase (i.e. weights based on the K = 5 variables TCC, TCR, FCI, Cattle and Sheep) in MBD1-A to evaluate the performance of this estimator for the three variables Crops, Equity and Debt that were impossible to model using model II. Results from this stage are shown in Table 4 and in Figure 2. In the fourth stage we used the fact that model I can be fitted to all eight variables to define multipurpose weights that we then use in MBD1-A. Results from this stage are presented in Table 5 and in Figure 2. Note that in all four of these simulation stages, we assign equal importance to all variables included in derivation of the multipurpose weights. Consequently, in the final simulation (stage five) we replicated the stage two simulation for MBD1-A, but this time assigned weights to each variable proportional to its population variability.

Table 1 about here

For the two variables TCC and TCR, Table 1 sets out the average and median values of various summary measures of estimation performance for the three methods MBD0, MBD1-A and MBD1-B under model I. These results clearly show that all three methods perform equivalently for this data set (regional specific results generated by these methods are virtually identical as well). This is evidence that the MBD method based on the multipurpose weights (8) is not sensitive to correlations between the target variables. Although not presented here, results from model-based simulations of target variables

with different levels of correlation support this conclusion. Consequently the simulation results presented below focus on MBD1-A.

Tables 2 and 3 about here

In the second stage of the simulation study, we compared the two variable specific methods EBLUP and MBD0 with the two multipurpose methods MBD1-A and MBD2. Tables 2 and 3 show the summary performances generated by these four methods for the five variables TCC, TCR, FCI, Cattle and Sheep under Models I and II respectively. Under the better fitting Model II (Table 3), multipurpose method MBD1-A performs marginally better than multipurpose method MBD2, which in turn is slightly better than the variable specific MBD0. All three are often substantially better than EBLUP for these data. Under Model I (Table 2), the two multipurpose methods MBD1-A and MBD2 record substantially better bias performances than the variable specific MBD0 and EBLUP, and better to comparable performances with respect to mean squared error. Overall, the multipurpose method MBD1-A seems the weighting method of choice for these five variables and these data.

In Figure 1 we show the regional level performances of EBLUP, MBD0, MBD1-A and MBD2 when estimating average TCC under model I and model II. Note the relatively better performance of all methods under model II. A considerable reduction in relative biases under multipurpose weighting can also be seen in most regions. A similar pattern of results was observed for TCR, FCI, Cattle and Sheep.

Figure 1 about here

From Figure 1 we see that in two regions (3 and 21) the weighting methods (MBD0, MBD1-A and MBD2) fail. Inspection of the data indicates that this is because of a small number of outlying estimates that were generated during the simulations. In region 21 for example these outlying estimates are due to the presence of a single massive outlier (TCC>A$30,000,000) in the sample data. When we discard these outlying estimates then the weighting methods, particularly MBD1-A and MBD2, perform well for TCC across all regions. Similar results were observed for the other four variables TCR, FCI, Cattle and Sheep.

The unstable performance of EBLUP for the Cattle and Sheep variables in Tables 2 and 3 is also noteworthy. Upon investigation we found that these anomalous results were due to the presence of large numbers of negative estimates in some of the regions, which in turn were caused by zero values in the data.

Table 4 about here

As noted earlier, our results suggest that multipurpose estimation based on MBD1-A is preferable to that based on MBD2. Consequently, in Table 4 we contrast the performances of the variable specific estimators EBLUP and MBD0 with that of the multipurpose estimator MBD1-A for the three variables (Crops, Equity and Debt) that contain a large number of zeros, and so were not included in calculation of the multipurpose weights used in MBD1-A. Note that these results are based on model I, since model II cannot be used for these variables. We see that MBD1-A 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.

Figure 2 about here

In the results presented so far, the multipurpose weights used in the MBD1-A method have been based on the $K = 5$ target variables that were 'suited' to linear mixed modeling with the model II specification. However, if a model I specification is used, we can use all $K = 8$ target variables to define these weights via (8). In Table 5 therefore we compare the performance of the MBD1-A method under this model with weights obtained by using both the limited ($K = 5$) and full ($K = 8$) set of target variables in (8). This shows that these weights are quite insensitive to this choice. The almost imperceptible regional differences between the Crops estimates defined by these two sets of weights (see Figure 2) reinforces this observation. Similar region-specific performances were observed for Equity and Debt as well.

Table 5 about here

So far, when computing the multipurpose weights, we have assigned equal importance to all K target variables that are used to define them. However, a reasonable alternative approach would be to assign importance factors based on the intrinsic variability of these variables. Two natural options in this regard are $\phi_k = 1/\sigma^2_{e,k}$ and $\phi_k = 1/V_k$, where $\sigma^2_{e,k}$ and V_k are the individual and total variability of the k^{th} target variable. Table 6 provides summary details of the performance of the MBD1-A method when the multipurpose weights (based on TCC, TCR, FCI, Cattle and Sheep) are computed using these alternative importance weighting factors. These results show that, for the population considered in the simulation study, there is little to choose between these different importance weighting factors.

Table 6 about here

4. Summary and Further Research

In this paper 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 performance of the corresponding multipurpose weighting based MBD1-A and MBD1-B estimators are almost identical, i.e. there are no real gains from taking account of the correlations between the survey variables when constructing the multipurpose weights. We also investigated an alternative approach to constructing multipurpose weights for use in MBD small area estimation by suitably averaging the variable specific EBLUP weights. Here again, our empirical results demonstrate that this method is somewhat less efficient than the loss function based MBD1-A method. We also show 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 methods do not work. An alternative in such cases is extend the EBLUP approach to mixtures of linear mixed models. The authors are currently working on this issue, and results obtained so far are encouraging.

Acknowledgements

The first author gratefully acknowledges the financial support provided by a PhD scholarship from the U.K. Commonwealth Scholarship Commission.

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. and Chandra, H. (2006). Improved direct estimators for small areas. Submitted for publication.

Chandra, H. and Chambers, R. (2005). Comparing EBLUP and C-EBLUP for small area estimation. *Statistics in Transition*, **7**, 637-648.

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.

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 MBD0, MBD1-A and MBD1-B 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	MBD0	MBD1-A	MBD1-B
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

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
	MBD0	-2.49	-9.25	-13.80	-15.05	-7.33
	MBD1-A	-1.54	-1.30	-0.50	-1.78	0.69
	MBD2	-1.29	-1.02	-0.04	-1.35	0.98
MRB	EBLUP	1.55	0.55	-2.08	0.95	-0.23
	MBD0	-0.82	-3.87	-2.83	-4.79	-4.48
	MBD1-A	-0.61	-0.42	-0.56	-0.97	-0.35
	MBD2	-0.52	-0.39	-0.54	-0.75	-0.30
ARRMSE	EBLUP	19.92	21.76	63.93	304.74	906.18
	MBD0	20.56	23.34	54.42	37.45	24.88
	MBD1-A	20.86	21.77	59.72	33.29	30.24
	MBD2	20.85	21.77	60.07	33.36	30.64
MRRMSE	EBLUP	15.74	14.83	40.41	25.97	13.00
	MBD0	14.45	16.20	35.85	30.34	15.50
	MBD1-A	14.69	13.41	42.09	30.55	14.67
	MBD2	14.74	13.46	42.45	30.56	14.67
ACR	EBLUP	90	88	87	86	91
	MBD0	92	91	94	93	94
	MBD1-A	92	92	94	95	96
	MBD2	92	92	94	95	96

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
	MBD0	-2.13	-1.25	0.50	-0.29	3.66
	MBD1-A	-1.67	-1.29	0.74	-1.95	1.10
	MBD2	-1.30	-0.72	3.17	-1.29	0.93
MRB	EBLUP	0.61	1.37	3.98	0.62	0.00
	MBD0	-0.47	-0.51	0.35	-0.31	0.00
	MBD1-A	-0.65	-0.50	0.24	-0.30	-0.15
	MBD2	-0.52	0.01	0.53	-0.22	-0.09
ARRMSE	EBLUP	19.87	20.28	68.85	231.08	630.01
	MBD0	20.15	21.46	65.43	30.80	37.82
	MBD1-A	19.06	21.03	64.03	30.09	32.04
	MBD2	27.13	34.84	129.29	45.16	34.99
MRRMSE	EBLUP	16.40	15.61	33.89	22.64	11.73
	MBD0	13.16	12.39	37.64	28.79	14.68
	MBD1-A	12.84	12.18	37.92	24.84	14.77
	MBD2	12.84	12.71	37.62	24.93	14.72
ACR	EBLUP	85	86	84	86	89
	MBD0	93	93	90	95	96
	MBD1-A	93	93	94	95	96
	MBD2	93	93	94	95	96

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, MBD0 and MBD1-A 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
	MBD0	0.00	-9.32	-4.94
	MBD1-A	-0.21	-1.20	-0.96
MRB	EBLUP	0.00	-0.28	1.16
	MBD0	-0.84	-3.51	-2.36
	MBD1-A	0.00	-0.32	-0.61
ARRMSE	EBLUP	123.96	18.51	29.02
	MBD0	23.53	19.14	27.71
	MBD1-A	22.92	17.05	28.57
MRRMSE	EBLUP	15.10	12.32	21.49
	MBD0	15.76	16.18	23.70
	MBD1-A	15.80	13.52	24.88
ACR	EBLUP	95	88	91
	MBD0	96	92	93
	MBD1-A	96	94	93

Table 5 Average relative bias (ARB), average relative root mean squared error (ARRMSE) and average coverage rate (ACR) for multi-purpose weighting (MBD1-A) based on original $K = 5$ and extended $K = 8$ variable sets under model I.

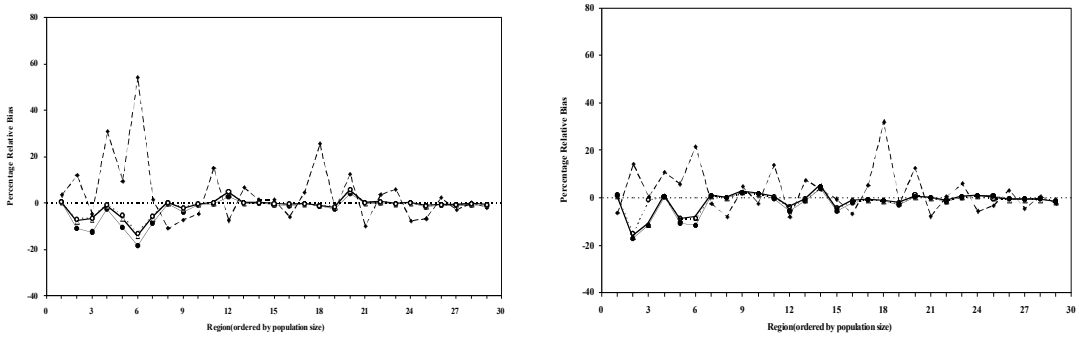
Variable	$K = 5$			$K = 8$		
	ARB	ARRMSE	ACR	ARB	ARRMSE	ACR
TCC	-1.54	20.86	92	-1.08	20.91	92
TCR	-1.30	21.77	92	-0.80	21.83	92
FCI	-0.50	59.72	94	0.21	60.22	94
Cattle	-1.78	33.29	95	-1.05	33.49	95
Sheep	0.69	30.24	96	1.24	31.06	96
Crops	-0.21	22.92	96	-0.20	22.97	96
Equity	-1.20	17.05	94	-0.72	17.14	94
Debt	-0.96	28.57	93	-0.68	28.74	93

Table 6 Average relative bias (ARB), average relative root mean squared error (ARRMSE) and average coverage rate (ACR) for multi-purpose weighting (MBD1-A) under $\phi_k = 1/K$, $\phi_k = 1/\sigma^2_{e,k}$ and $\phi_k = 1/V_k$ for $K = 5$ target variables (TCC, TCR, FCI, Cattle, Sheep) under model I.

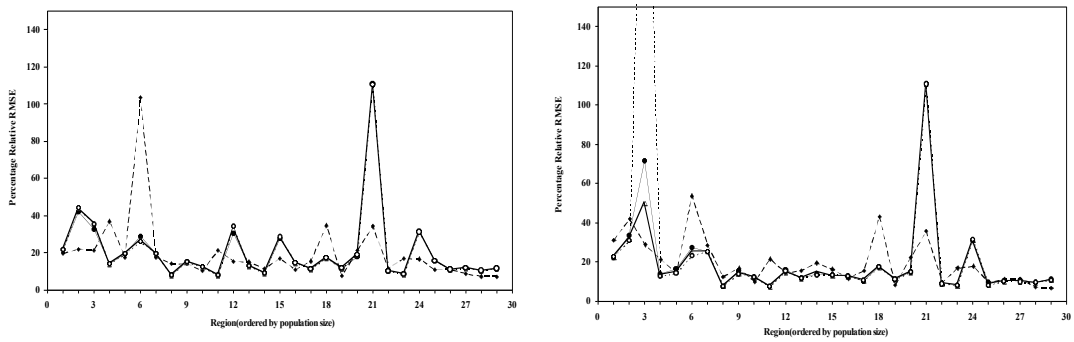
Criterion	ϕ_k^{-1}	TCC	TCR	FCI	Cattle	Sheep
ARB	K	-1.54	-1.30	-0.50	-1.78	0.69
	$\sigma^2_{e,k}$	-1.69	-1.48	-0.82	-2.03	0.52
	V_k	-1.64	-1.42	-0.70	-1.95	0.57
ARMSE	K	20.86	21.77	59.72	33.29	30.24
	$\sigma^2_{e,k}$	20.83	21.71	58.00	33.19	29.99
	V_k	20.85	21.75	58.15	33.25	30.11
ACR	K	92	92	94	95	96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

Figure 1 Regional performance of EBLUP (dashed line), MBD0 (thin line), MBD1-A (thick line) and MBD2 (dotted line) for TCC under model I (left) and model II (right).

Relative Bias (%)



Relative RMSE (%)



Coverage Rate

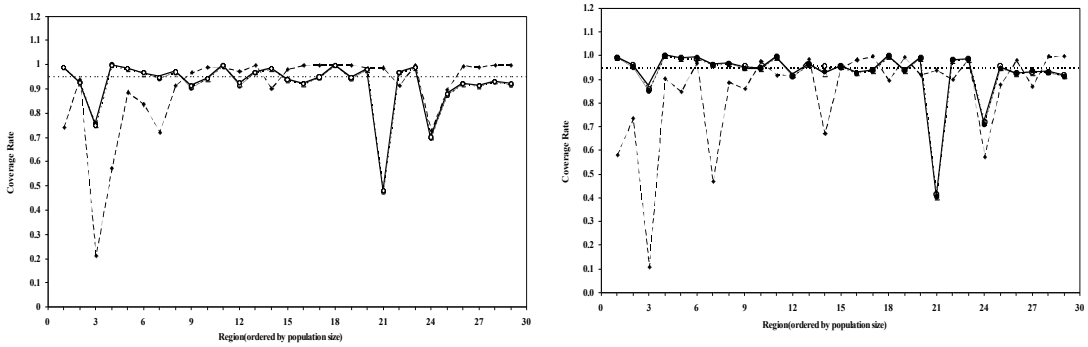
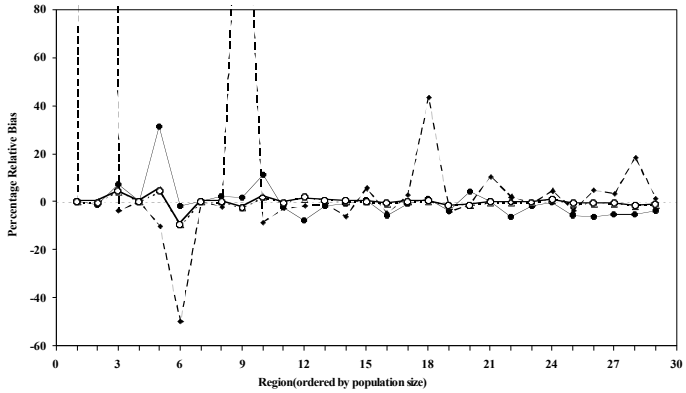
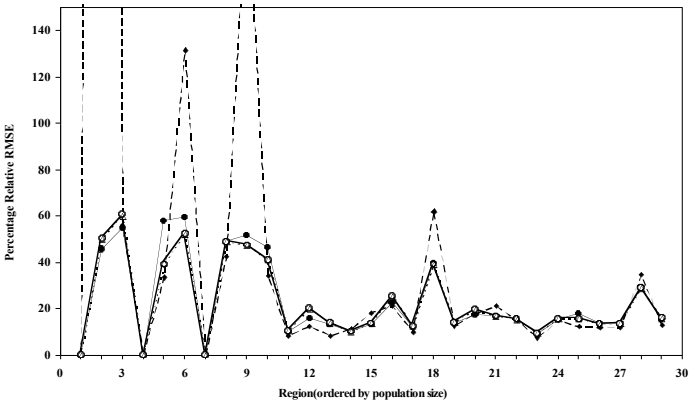


Figure 2 Regional performances of EBLUP (dashed line), MBD0 (thin line), MBD1-A under $K = 5$ (thick line) and MBD1-A under $K = 8$ (dotted line) for Crops under model I.

Relative Bias (%)



Relative RMSE (%)



Coverage Rate