# Multipurpose Small Area Estimation

**Hukum Chandra**
University of Southampton, U.K.

**Ray Chambers**
University of Wollongong, Australia

## Weighting and Small Area Estimation

Sample surveys are generally **multivariate**, in the sense that they collect data on more than one response variable

- In theory, each variable can be assigned an optimal weight
- Advantageous to have a **common weight** for all variables
- **Multipurpose** sample weights when small area estimates of the survey variables are required

**The model-based direct** (MBD) approach of SAE **(Chambers and Chandra, 2006)**: weighted direct estimator for small areas, the EBLUP weights used are **variable specific**, derived under linear mixed model and borrows strength via this model

**Multipurpose SAE**: replace the **variable specific** BLUP optimality criterion by modified 'total variability' criterion that leads to a single set of optimal multipurpose weights

# Population Level Estimation: The General Linear Model

$$y_U = X_U \beta + \varepsilon_U \text{ with } E(\varepsilon_U) = 0_N \text{ , } Var(\varepsilon_U) = \sigma^2 V_U$$

**BLUP weights** for Population Total of $Y$ **(Royall, 1976)**

$$w_{BLUP} = 1_n + H'\left(X'_U 1_N - X'_s 1_n\right) + \left(I_n - H'X'_s\right)V_{ss}^{-1} V_{sr} 1_{N-n}$$
$$H = \left(X'_s V_{ss}^{-1} X_s\right)^{-1} X'_s V_{ss}^{-1}$$

- Depends on the population level conditional variance/covariance matrix for that variable

- This BLUP optimality is variable specific

**MBD approach** to SAE: a mixed linear model is used to specify the covariance matrix to derive the EBLUP weights

# Multipurpose Sample Weighting

- K- response variables and a common set of auxiliary variables $X_U$, subscript $k = 1,..,K$ denote quantities associated with the $k^{\text{th}}$ response variable

- 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 the **multipurpose weights** $w_s = \{w_j ; j \in s\}$

- The weights $w_s$ are said to be **$\phi$-optimal** if

  **(a)** $E(\hat{T}_k - T_k) = 0, \forall k$, and

  **(b)** the $\phi$-weighted total prediction variance is minimised at $w_s$

where $\sum_k \phi_k = 1$ is a user-specified non-negative scalar quantity, that reflects the **relative importance** attached to the $k^{\text{th}}$ response variable

# Multipurpose Sample Weighting

The **optimal multipurpose** sample weights are

## 1. Uncorrelated Variables

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

$$H_1 = \left( X_s' U_1^{-1} X_s \right)^{-1} X_s' U_1^{-1} \text{ with } U_1 = \sum_{k=1}^{K} \phi_k V_{kss} \text{ and } W_1 = \sum_{k=1}^{K} \phi_k V_{ksr}$$

## 2. Correlated Variables: $C_{kl} = Cov(y_k, y_l)$

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

$$H_2 = \left( X_s' U_2^{-1} X_s \right)^{-1} X_s' U_2^{-1} \text{ with } U_2 = \sum_k \phi_k V_{kss} + \sum_k \sum_{l \neq k} \sqrt{\phi_k} \sqrt{\phi_l} C_{klss}$$

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

# Application to Small Area Estimation

- The **multipurpose weights** $w_s^{(1)}$ and $w_s^{(2)}$ are essentially EBLUP type weights based on 'importance averaging' of the variance and covariance components
- **A second approach** to deriving multipurpose weights based on corresponding 'importance averaging' of the variable specific EBLUP sample weights: $w_s^{(3)} = \sum_k \phi_k w_{sk}$
- In order to use the multipurpose weights $w_s^{(1)}$, $w_s^{(2)}$ and $w_s^{(3)}$ in MBD methods, we assume that the variables follow the **linear mixed model**
- The variable-specific MBD estimate of the mean of the $k^{th}$ response variable in area *i*

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

- **Multipurpose SAE:** replace variable-specific EBLUP sample weights by multipurpose sample weights ($w_s^{(1)}$, $w_s^{(2)}$ or $w_s^{(3)}$)

# An Empirical Study

- Sample of 1652 Australian broadacre farms

- Target population of **81982** farms obtained by sampling with replacement from this sample with probabilities proportional to their sample weights

- 1000 independent stratified random samples from this (fixed) population, with total sample size in each simulation equal to the original sample size (1652) and with strata defined by the **29** different Australian broadacre agricultural regions. **Sample sizes varied from 6 to 117** within these strata and were fixed to be the same as in the original sample

## Response Variables (*K* = 8)

| Variable | Description |
| --- | --- |
| **TCC** | Total cash costs (A$) |
| **TCR** | Total cash receipts (A$) |
| **FCI** | Farm cash income (A$), defined as TCR – TCC |
| **Crops** | Area under crops (in hectares) |
| **Cattle** | Number of Cattle on the farm |
| **Sheep** | Number of sheep on the farm |
| **Equity** | Total farm equity (A$), and |
| **Debt** | Total farm debt (A$) |

**Auxiliary variable:** Farm size (referred as **Size**)

**Target:** Estimate the average of these variables in each of the 29 regions

## Exploring the Data

- Regions can be grouped into 3 **zones** (Pastoral, Mixed Farming, and Coastal), with farm size(ha) known for each farm in the population
- The linear relationship between the 8 target variables and **Farm Size** is rather **weak**, however this improves when separate linear models are fitted within six post strata
- **Post-strata** are defined by splitting each zone into small farms (farm size < than zone median) and large farms (farm size>= zone median)
- **Fixed Effects Specification:** include an effect for **farm size**, effects for the **post-strata** and effects for **interactions** between farm size and the post strata
- **Random Effects Specification**
  - Random intercepts **(Model I)**
  - Random intercepts + random slopes on Size term **(Model II)**

# Estimators Investigated in Empirical Studies

| Estimator | Description |
|---|---|
| **MBD1-A** | MBD estimator based on multipurpose weights $w_s^{(1)}$ |
| **MBD1-B** | MBD estimator based on multipurpose weights $w_s^{(2)}$ |
| **MBD2** | MBD estimator based on multipurpose weights $w_s^{(3)}$ |
| **MBD0** | MBD estimator based on variable specific EBLUP weights |
| **EBLUP** | variable specific EBLUP under linear mixed model |

- **MSE for the EBLUP:** follow the approach of **Prasad and Rao (1990)**
- **MSE for the various MBD estimators:** Adapt standard methods for estimating the MSE of a weighted linear estimator

**(Chambers and Chandra, 2006; Chandra and chambers, 2005; and Royall and Cumberland, 1978)**

## Performances (%) for 2 Variables under Model-I, Exploiting Correlation

| 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 |

# Performances (%) for 5 'Well Behaved' Variables under Model I

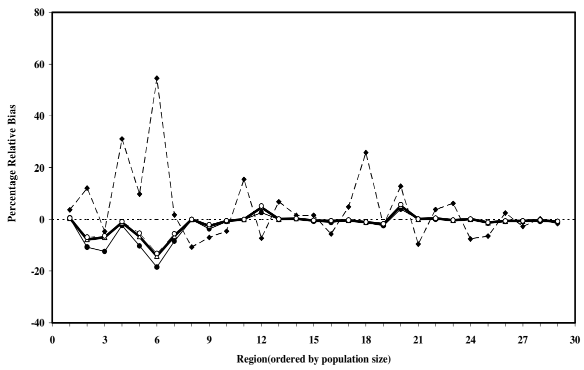
| Criterion | Method | TCC | TCR | FCI | Cattle | Sheep |
|---|---|---|---|---|---|---|
| ARB | EBLUP | 4.24 | 5.48 | 6.93 | 138.48 | 304.24 |
| | MBD0 | -2.49 | -9.25 | -13.80 | -15.05 | -7.33 |
| | MBD1-A | -1.54 | -1.30 | -0.50 | -1.78 | 0.69 |
| | MBD2 | -1.29 | -1.02 | -0.04 | -1.35 | 0.98 |
| MRB | 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 |

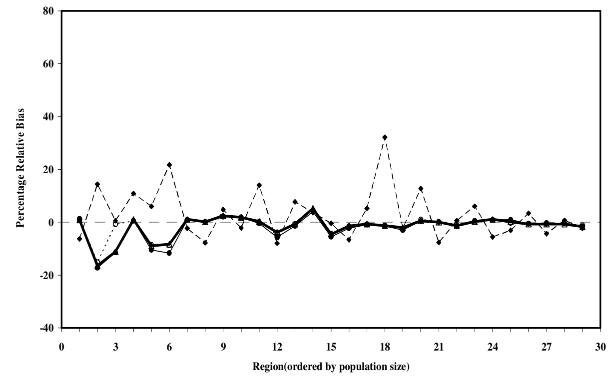
# Performances (%) for 5 'Well Behaved' Variables under Model II

| 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 |

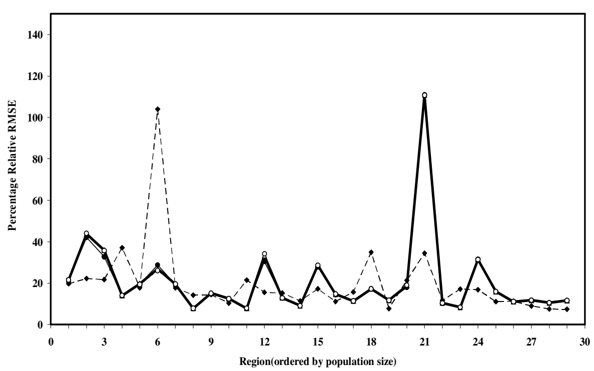
# Regional performance of EBLUP (dashed line), MBD0 (thin line), MBD1-A (thick line) and MBD2 (dotted line) for TCC under model I (left) and model II (right)

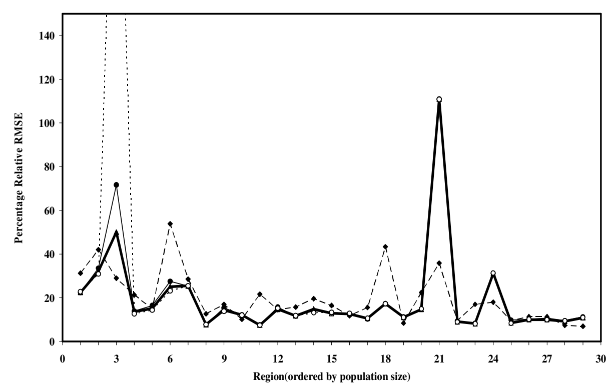
## Relative Bias (%)




## Relative RMSE (%)



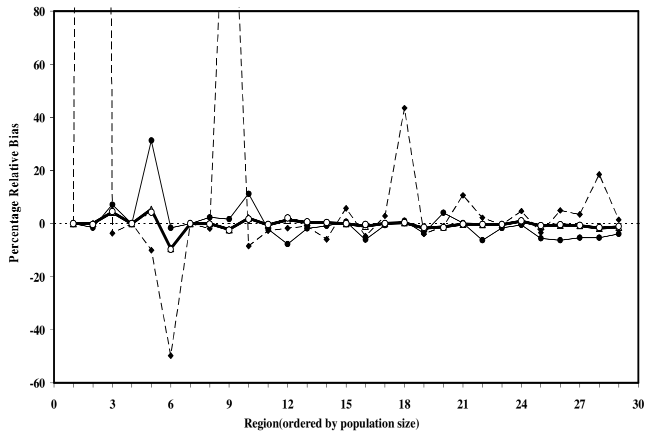



## Average performance measures (%) for ‘Zero Contaminated’ Variables (Model I is assumed)

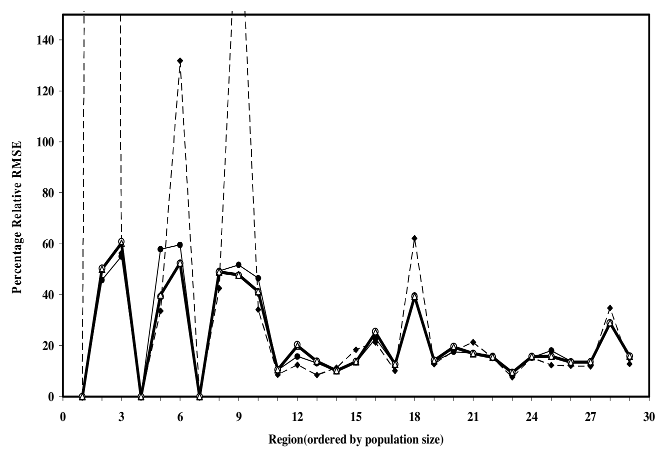
| Criterion | Method | Crops | Equity | Debt |
|---|---|---|---|---|
| ARB | EBLUP | **90.31** | **4.36** | **8.39** |
| | MBD0 | **0.00** | **-9.32** | **-4.94** |
| | MBD1-A | **-0.21** | **-1.20** | **-0.96** |
| ARRMSE | EBLUP | **123.96** | **18.51** | **29.02** |
| | MBD0 | **23.53** | **19.14** | **27.71** |
| | MBD1-A | **22.92** | **17.05** | **28.57** |
| ACR | EBLUP | **95** | **88** | **91** |
| | MBD0 | **96** | **92** | **93** |
| | MBD1-A | **96** | **94** | **93** |

# 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

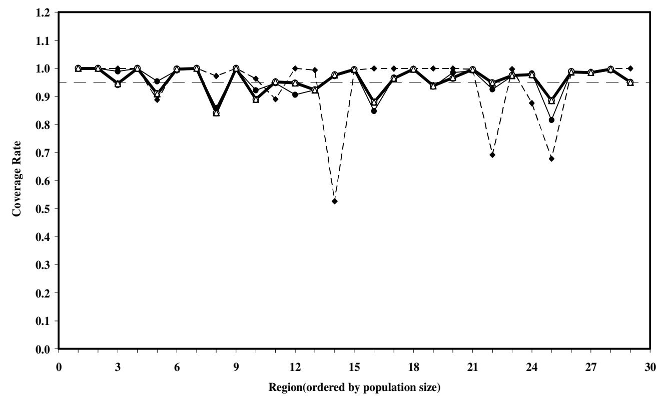
## Relative Bias (%)




## Relative RMSE (%)




## Coverage Rate



## Average performance (%) for multipurpose 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 |

# Conclusions

- For the population considered in our simulation studies there are no real gains from taking account of the **correlations** between the variables

- **An alternative approach** to constructing multipurpose weights for use in MBD SAE by suitably averaging the variable specific EBLUP weights
  - Empirical results demonstrate that this method is somewhat less efficient than the loss function based MBD1-A method

- The 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:** extend the EBLUP approach to mixtures of linear mixed models

# References


1. Chambers, R. and Chandra, H. (2006). Improved direct estimators for small areas. Submitted.

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

3. 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.

4. Royall, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association,* **71**, 657-664.

5. Royall, R.M. and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association,* **73**, 351-358.