SMALL AREA ESTIMATION WITH SKEWED DATA

HUKUM CHANDRA, RAY CHAMBERS

ABSTRACT

In business surveys, data typically are skewed and the standard approach for small area estimation based on linear mixed models lead to inefficient estimates. In this paper, we discuss small area estimation techniques for skewed data that are linear following a suitable transformation. In this context, implementation of the empirical best linear unbiased prediction (EBLUP) approach under transformation to a linear mixed model is complicated. However, this is not the case with the model-based direct (MBD) approach (Chambers and Chandra, 2006), which is based on weighted linear estimators. We extend the MBD approach to skewed data using sample weights derived via model calibration based on a log transform model with random area effects. Our results show this estimator is both efficient and robust with respect to the distribution of these random effects. An application to real data demonstrates the satisfactory performance of the method.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M06/05

University of Southampton

Small Area Estimation with Skewed Data

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

1. Southampton Statistical Sciences Research Institute
University of Southampton, Southampton, SO17 1BJ, UK
Email: hchandra@soton.ac.uk

2. Centre for Statistical and Survey Methodology
University of Wollongong, Wollongong, NSW, 2522, Australia
Email: ray@uow.edu.au

Abstract

In business surveys, data typically are skewed and the standard approach for small area estimation based on linear mixed models lead to inefficient estimates. In this paper, we discuss small area estimation techniques for skewed data that are linear following a suitable transformation. In this context, implementation of the empirical best linear unbiased prediction (EBLUP) approach under transformation to a linear mixed model is complicated. However, this is not the case with the model-based direct (MBD) approach (Chambers and Chandra, 2006), which is based on weighted linear estimators. We extend the MBD approach to skewed data using sample weights derived via model calibration based on a log transform model with random area effects. Our results show this estimator is both efficient and robust with respect to the distribution of these random effects. An application to real data demonstrates the satisfactory performance of the method.

Key Words: Small areas, Skewed data, Model Calibration, Expected value model, Calibrated sample weights, MBD approach, EBLUP.

1. Introduction

Small area estimation (SAE) is typically an increasingly important secondary objective of many sample surveys, and several methods exist in the literature (Rao, 2003). However, research is continuing on several important practical problems related to small area estimation. Standard methods for SAE such as the empirical best linear unbiased prediction (EBLUP) approach (Prasad and Rao, 1990) and the model-based direct (MBD) approach (Chambers and Chandra, 2006) assume a linear mixed model can be used to characterize the small areas of interest. However, it happens (typically for skewed data) that the variable of interest Y is linear on some transformed scale (e.g. in business surveys, often variables are linear on logarithmic scale). In this context, estimation based on linear model for Y leads to inefficient estimates. In such situation, an appropriate technique for SAE should essentially be based on a linear mixed model for a transformed variable. The use of transform variables for survey estimation with skewed data has been investigated by Carroll and Ruppert (1988), Chen and Chen (1996), Karlberg (2000) and Chambers and Dorfman (2003). In this paper we explore transform variable based estimation in context of SAE for skewed data, focussing on the widely used logarithmic (log) transformation function. Implementation of the EBLUP approach under transformation to a linear mixed model is quite complicated. However, this is not the case with the MBD approach, which is based on weighted linear estimators. In this paper we extend the MBD approach of Chambers and Chandra (2006) to small area estimation for skewed data. In particular, we consider the use of sample weights derived via model calibration (Wu and Sitter, 2001) based on a log transform model with random area effects. A simple MSE estimator for weighted small area estimation is also developed. We also relax the usual normality assumption for random errors in order to examine robustness with respect to this assumption.

In the following section we summarize the model calibration approach for estimation of population quantities. In section 3 we then discuss the expected value model derived from a transform linear mixed model for small area estimation of skewed data. Section 4 introduces the survey weights based on expected value model derived from a transform linear mixed model and describes the MBD estimator for SAE in this case. In section 5 we provide illustrative empirical results that contrast the proposed MBD estimator for skewed data with the MBD and EBLUP method under a linear mixed model. Finally, in section 6 some concluding remarks are made and some related issues that needs further attention are discussed.

2. Model Calibration for Population Estimation

In this section we briefly review model calibration for estimation of population level quantities. To start, we fix our notation. Let *Y* denote an *N*-vector of population values of a characteristic of interest, and suppose that our primary aim is estimation of the total T_y of the values in *Y* (or their mean $\overline{Y}$). In order to assist us in this objective, we shall assume that we have 'access' to *X*, an $N \times p$ matrix of values of *p* auxiliary variables that are related, in some sense, to the values in *Y*. In particular, we assume that the individual sample values in *X* are known. The non-sample values in *X* may not be individually known, but are assumed known at some aggregate level. At a minimum, we know the population totals T_x of the columns of *X.* Given this set up, Deville and Särndal (1992) introduce the notation of a calibration estimator of population total of *Y* as $\hat{T}_{y,c} = \sum_{j \in s} w_j y_j$, where the calibration weights w_j's are chosen to minimise their average distance (Φ_s, say) from the basic design weights, $d_j = \pi_j^{-1}$ with $\pi_j = \Pr(j \in s)$, that are used in Horvitz-Thompson (HT) estimator $\hat{T}_{y,HT} = \sum_{j \in s} d_j y_j$, subject to the calibration constraint

$$\sum_{j \in s} w_j x_j = \sum_{j=1}^{N} x_j = T_x \qquad (1)$$

Deville and Särndal (1992) argue "weights that perform well for the auxiliary variables also should perform well for the study variable". However, there is an implicit underlying assumption that *Y* and *X* are linearly related that makes this a valid argument, i.e. the conventional calibration approach (Deville and Särndal, 1992, Chambers, 1997) implicitly relies on the assumption that the survey variable and the auxiliary variables are linearly related. Thus, if the underlying model is non-linear then the calibrated estimator derived under a linearity assumption cannot be very efficient. Let us assume the relationship between *Y* and *X* can be described by a super population model

$$E_\xi(y_j \mid x_j) = h(x_j;\eta),\ V_\xi(y_j \mid x_j) = \sigma^2 \omega_j;\ j = 1,...,N, \qquad (2)$$

where η, typically vector-valued, and σ^2 are model parameters, and the mean function $h(x_j;\eta)$ is a known function of x_j and η, the variance function ω_j is a known function of x_j and $h(x_j;\eta)$. Here E_ξ and V_ξ denotes the expectation and variance with respect to super population model. In matrix notation we write (2) as

$$E_\xi(Y \mid X) = h(X;\eta) \text{ and } V_\xi(Y \mid X) = \Omega \qquad (3)$$

The model (3) is quite general and includes linear, non-linear, and generalized linear models as special cases. In this context, Wu and Sitter, (2001) proposed the use of sample weights derived via model calibration. They defined the calibration estimator for population mean of Y as $\hat{\bar{Y}}_c = N^{-1}\sum_{j\in s} w_j y_j$ with weights sought to minimize the distance measure Φ_s under the constraints:

$$\sum_{j\in s} w_j = N \text{ and } \sum_{j\in s} w_j h(x_j;\hat{\eta}) = \sum_{j=1}^{N} h(x_j;\hat{\eta}) \tag{4}$$

where $\hat{\eta}$ is a design consistent estimator for η. That is calibration is performed with respect to the population mean of the 'fitted values' $\hat{h}_j = h(x_j;\hat{\eta})$ of $h(x_j;\hat{\eta})$. Provided the model (3) is a reasonable one, y_j is then (at least approximately) a linear function of its 'fitted values' $h(x_j;\hat{\eta})$ under this model. The basic idea of this approach is then we can carry out linear estimation using these 'fitted or expected values' as auxiliary variables.

The above discussion represents what might be referred to the design-based interpretation of model calibration. A model-based perspective on model calibration can be described as follows. We assume that Y and $h(X;\eta)$ are related by the linear model of the form

$$Y = \alpha_0 1_N + \alpha_1 h(X;\eta) + \varepsilon = \alpha J + \varepsilon \tag{5}$$

where J denotes the 'design matrix' for the linear model (5) linking Y and $h(X;\eta)$, $\alpha = (\alpha_0, \alpha_1)'$ is a vector of unknown parameters, ε denotes a N-vector of random variables with $E_\xi(\varepsilon) = 0$ and $V_\xi(\varepsilon) = \Omega = [\omega_{jk}]$. We called model (5) the 'expected value' or 'fitted value' model defined by (3). For $\alpha_0 = 0$ in model (5) we refer as ratio specification of this model, otherwise regression specification. The model (5) can have either ratio or regression specification. Without loss of generality, we arrange the vector Y so that the first n elements correspond to the sample units, and partition Y, J and Ω according to sample and non-sample units:

$$Y = \begin{bmatrix} Y_s \\ Y_r \end{bmatrix},\ J = \begin{bmatrix} J_s \\ J_r \end{bmatrix} \text{ and } \Omega = \begin{bmatrix} \Omega_{ss}\ \Omega_{sr} \\ \Omega_{rs}\ \Omega_{rr} \end{bmatrix}.$$

Here J_s is the $n \times 1$ vector of 'fitted values' of the auxiliary variables and Ω_{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. A subscript of r is used to denote corresponding quantities defined by the $N-n$ non-sample units, with Ω_{rs} denoting the $(N-n)\times n$ matrix defined by $Cov(Y_r, Y_s)$. In what follows we denote 1_N, 1_n and 1_r as vectors of 1's and I_N, I_n and I_r as identity matrices of order N, n and $N-n$ respectively. In practice the variance components that define covariance matrix Ω are unknown and so need to be estimated from the sample data. We use a “hat” to denote such an estimate. Further, throughout this paper we assume that sampling is uninformative, so the sample data also follow the population model.

Given this notation, the sample weights that define the BLUP for population total of Y under a general linear ‘fitted value’ model (5) are

$$w_{BLUP}^{h} = 1_n + H_h'(J'1_N - J_s'1_n) + (I_n - H_h'J_s')\Omega_{ss}^{-1}\Omega_{sr}1_r \tag{6}$$

where $H_h = (J_s'\Omega_{ss}^{-1}J_s)^{-1}J_s'\Omega_{ss}^{-1}$. See Royall (1976). The sample weights (6) derived via model calibration are calibrated on J, i.e. $J_s'w_{BLUP}^{h} = J'1_N$. The weights (6) are based on a model appropriate for estimation of population as a whole (i.e. population weighting) and using these weights for small area estimation will be inefficient. The most commonly used class of models for small area estimation model is essentially a mixed model, i.e. model implied by the covariance structure that includes the random area effect components. The next section describes the model that includes the random area effects and suitable for small area estimation.

3. Small Area Models under Transformation

3.1 Linear Mixed Model

Let Y_i be the $N_i \times 1$ vector of values of variable of interest in small area i $(i = 1,....,m)$ and let X_i be the $N_i \times p$ matrix of values of the auxiliary variables associated with Y_i. We assume that Y_i and X_i are not related by a linear model on themselves, but they are linearly related on logarithm (natural) transform model. We consider the following linear mixed model specification for the distribution of $l_i = \log(Y_i)$ given Z_i:

$$l_i = Z_i\beta + G_iu_i + e_i \tag{7}$$

where $Z_i = (1_{N_i}, \log(X_i))$ is the $N_i \times (p+1)$ matrix of values of the auxiliary variables in area *i*, β is a $(p+1)\times 1$ vector of fixed effects, G_i is a $N_i \times q$ matrix of known covariates characterising differences between small areas, N_i is the number of population units in the small area *i*, 1_{N_i} is a vector of 1’s of order N_i, u_i is a random area effect associated with the i^{th} small area and e_i is a $N_i \times 1$ vector of individual level random errors. The two random variables u_i and e_i are assumed to be independently normally distributed, with zero means and with variances $V(u_i) = \Sigma$ and $V(e_i) = \sigma_e^2 I_{N_i}$ respectively. The covariance matrix of l_i is

$V_i = Var(l_i) = G_i\Sigma(\theta)G_i' + \sigma_e^2 I_{N_i}$, with $v_{ijj} = Var(l_{ij}) = G_{ij}\Sigma(\theta)G_{ij}' + V(e_{ij})$ and $v_{ijk} = Cov(l_{ij}, l_{ik}) = G_{ij}\Sigma(\theta)G_{ik}'$; $j,k = 1,.....,N_i$. The covariance of l_i depends on a vector of fixed parameters θ, usually called the variance components of the model.

By grouping the area-specific models (7) over the population, we are led to the population level model:

$$l = Z\beta + Gu + e \qquad (8)$$

where $l = (l_1',......,l_m')'$, $Z = (Z_1',.......,Z_m')'$, $G = diag(G_i; 1 \le i \le m)$, $u = (u_1',.....,u_m')'$ and $e = (e_1',\ldots,e_m')'$. The variance-covariance matrix of *l* is $V = diag(V_i; 1 \le i \le m)$. We assume that *Z* has full column rank. In practice the variance components of the model that define the covariance matrix *V* are unknown and we estimate them from the sample data under the model (8) with suitable estimation methods such as maximum likelihood (ML), restricted maximum likelihood (REML) or method of moments (Harville, 1977). The estimated variance-covariance matrix of *l* is $\hat{V} = diag(\hat{V}_i; 1 \le i \le m)$ with $\hat{V}_i = \hat{\sigma}_e^2 I_{N_i} + G_i\hat{\Sigma}G_i'$. Again, we consider the decomposition of *l*, *Z*, *G* and *V* into sample and non-sample components as mentioned before (6). We use similar notation at the small area level by introducing an extra subscript *i* to denote small area. For example, we denote by s_i the set of n_i sample units in area *i*, r_i the corresponding $N_i - n_i$ non-sampled units in the area and put $\hat{V}_{iss} = \hat{\sigma}_e^2 I_{n_i} + G_{is}\hat{\Sigma}G_{is}'$ and $\hat{V}_{isr} = G_{is}\hat{\Sigma}G_{ir}'$.

With this notation, and assuming (8) holds, the empirical best linear unbiased estimator of β is $\hat{\beta} = \left(\sum_{i=1}^m Z_{is}'\hat{V}_{iss}^{-1}Z_{is}\right)^{-1}\left(\sum_{i=1}^m Z_{is}'\hat{V}_{iss}^{-1}l_{is}\right)$ with $E_\xi(\hat{\beta}) = \beta$ and $V_\xi(\hat{\beta}) = \left(\sum_{i=1}^m Z_{is}'\hat{V}_{iss}^{-1}Z_{is}\right)^{-1}$, so that $E_\xi(l_i' - l_i) \approx 0$ for large *n*. We denote $\hat{\phi}_i = Z_i\hat{\beta}$ with $E_\xi(\hat{\phi}_i) = Z_i\beta$ and $V_\xi(\hat{\phi}_i) = Z_i\left(\sum_{i=1}^m Z_{is}'\hat{V}_{iss}^{-1}Z_{is}\right)^{-1} Z_i'$, where $a_{ijk} = Z_{ij}\left(\sum_{i=1}^m Z_{is}'\hat{V}_{iss}^{-1}Z_{is}\right)^{-1} Z_{ik}' \to 0$ as $n \to \infty$. We denote by $a_i = (a_{i11},....,a_{iN_iN_i})'$ and $v_i = (v_{i11},...,v_{iN_iN_i})'$, the vectors of diagonal elements of the covariance matrices $V_\xi(\hat{\phi}_i)$ and $V_\xi(l_i)$ respectively.

In order to use the Chambers and Chandra (2006) MBD method to get estimates for small areas we require sample weights. For skewed data that follows a linear mixed model on the log scale (8), the sample weights can be derived via model calibration, so first we need to evaluate ‘expected value’ model (Section 2). In other words, we need to evaluate the first and second moments, i.e. h, Ω under the model (8) to derive the sample weights (6). We can use parameter estimates derived under model (8) to obtain the predicted values of the transform variable and then back-transform to get predicted values of *Y*. These lead to the naïve-lognormal predictor. However, this predictor is biased (Chambers and Dorfman, 2003). Bias corrected first and second order moments that define the expected value model are expressed below.

3.2 An Expected Value Model for Small Area Estimation

Let us consider

$$E_\xi(Y_{ij}) = E_\xi\left[\exp(l_{ij})\right] = \exp\left(Z_{ij}\beta + v_{ijj}/2\right) = \varphi_i(\eta) \neq E_\xi\left[\exp(\hat{l}_{ij})\right] = E_\xi(\hat{Y}_{ij}) \qquad (9)$$

Thus, we need to adjust this bias. To this end we write $\varphi_i = \varphi_i(\eta) = \exp[Z_{ij}\beta + (v_{ijj}/2)]$ and then by a two-step Taylor series approximation:

$$\varphi_i(\hat{\eta}) \cong \varphi_i(\eta) + \varphi_i'(\hat{\eta}-\eta) + \frac{1}{2}(\hat{\eta}-\eta)'\varphi_i''(\hat{\eta}-\eta),$$

so that $E_\xi[\varphi_i(\hat{\eta})] \cong \varphi_i(\eta) + \varphi_i' E_\xi(\hat{\eta}-\eta) + \frac{1}{2}tr\left[E_\xi\left\{\varphi_i''(\hat{\eta}-\eta)(\hat{\eta}-\eta)'\right\}\right].$

Here, φ_i' and φ_i'' are the first and second derivatives of $\varphi_i(\eta)$ with respect to η at $\eta = \hat{\eta}$, $\hat{\eta} = (\hat{\beta}, \hat{v}_{ijj})'$ is the estimate of vector of unknown fixed parameters $\eta = (\beta, v_{ijj})'$ such that $E_\xi(\hat{\eta}-\eta) \approx 0$ for large n. Further, $\hat{\beta}$ and $\hat{v}_{ijj}$ are independent (McCulloch and Searle, 2001) and thus

$$tr\left\{E_\xi\left[\varphi_i''(\hat{\eta}-\eta)(\hat{\eta}-\eta)'\right]\right\} = tr\left\{\varphi_i'' E_\xi[(\hat{\eta}-\eta)(\hat{\eta}-\eta)']\right\}$$

$$= e^{(Z_{ij}\beta + \frac{v_{ijj}}{2})}\left[Z_{ij}\left(\sum\nolimits_{i=1}^{m} Z'_{is}\hat{V}_{iss}^{-1}Z_{is}\right)^{-1} Z'_{ij} + \frac{1}{4}Var(\hat{v}_{ijj})\right]$$

with $\varphi_i' = \begin{pmatrix} Z_{ij}e^{Z_{ij}\beta+\frac{v_{ijj}}{2}} \\ \frac{1}{2}e^{Z_{ij}\beta+\frac{v_{ijj}}{2}} \end{pmatrix}$ and $\varphi_i'' = \begin{pmatrix} Z_{ij}^2 e^{Z_{ij}\beta+\frac{v_{ijj}}{2}} & \frac{1}{2}Z_{ij}e^{Z_{ij}\beta+\frac{v_{ijj}}{2}} \\ \frac{1}{2}e^{Z_{ij}\beta+\frac{v_{ijj}}{2}} & \frac{1}{4}e^{Z_{ij}\beta+\frac{v_{ijj}}{2}} \end{pmatrix}$.

Substituting these expressions, we get

$$E_\xi\left[\varphi_i(\hat{\eta})\right] \cong \mathrm{e}^{(Z_{ij}\beta+\frac{v_{ijj}}{2})}\left\{1 + \frac{1}{2}\left[a_{ijj} + \frac{1}{4}V(\hat{v}_{ijj})\right]\right\} \neq E_\xi\left[\varphi_i(\eta)\right] = \mathrm{e}^{(Z_{ij}\beta+\frac{v_{ijj}}{2})}.$$

This indicates that transformation leads to biased estimator. A second order bias corrected estimate of $E_\xi(Y_{ij})$ is defined as

$$\hat{Y}_{ij} = h(Z_{ij};\hat{\eta}) = \hat{k}_{ij}^{-1} e^{(Z_{ij}\hat{\beta}+\frac{\hat{v}_{ijj}}{2})},\ i = 1,....,m;\ j = 1,....,N_i \qquad (10)$$

so that $E_{\xi}(\hat{Y}_{ij}) \approx \exp\left(Z_{ij}\beta + v_{ijj}/2\right) = E_{\xi}(Y_{ij}) = h(Z_{ij};\eta)$, i.e. $\hat{Y}_{ij}$ is an approximately ξ unbiased predictor of Y_{ij}. Here $\hat{k}_{ij} = \left[1+\frac{1}{2}\left(a_{ijj} + \frac{\hat{V}ar(\hat{v}_{ijj})}{4}\right)\right]$ is the bias correction and $Var(\hat{v}_{ijj})$ is the asymptotic covariance matrix of $\hat{v}_{ijj}$ given by inverse of the relevant information matrix. Note that the bias adjustment in (10) has same form as Karlberg (2000). However, Karlberg (2000) assumes uncorrelated variables. In contrast, predictor (10) is defined for general case allowing correlated variables.

Under normality of the random errors u_i and e_i, covariance between Y_{ij} and Y_{ik} in small area i is

$$Cov_{\xi}(Y_{ij}, Y_{ik}) = \omega_{ijk} = e^{\left(Z_{ij}+Z_{ik}\right)\beta}\left\{E_{\xi}(e^{G_{ij}u_i + e_{ij}}e^{G_{ik}u_i + e_{ik}}) - E_{\xi}(e^{G_{ij}u_i + e_{ij}})E_{\xi}(e^{G_{ik}u_i + e_{ik}})\right\}$$

$$= \begin{cases} e^{(Z_{ij}+Z_{ik})\beta}[e^{\frac{1}{2}(v_{ijj}+v_{ikk})}(e^{v_{ijk}}-1)] & if\ \ j \neq k \\ e^{2Z_{ij}\beta}[e^{v_{ijj}}(e^{v_{ijj}}-1)] & if\ \ j = k \end{cases} \qquad (11)$$

We group the bias corrected predictor (10) and the covariance (11) at the small area level as

$$\hat{Y}_i = h(Z_i;\hat{\eta}) = (\hat{Y}_{i1},....,\hat{Y}_{iN_i})' = \hat{k}_i^{-1}\exp(Z_i\hat{\beta} + \frac{\hat{v}_i}{2}) \qquad (12)$$

with $\hat{k}_i = 1 + \frac{1}{2}\left(Z_i\left(\sum_{i=1}^{m} Z'_{is}\hat{V}_{iss}^{-1}Z_{is}\right)^{\text{-1}} Z'_i + \frac{\hat{V}ar(\hat{v}_{ijj})}{4}\right)$ and

$$Var_{\xi}(Y_i) = \Omega_i = [\omega_{ijk}] = A_i \Delta_i A'_i \qquad (13)$$

where $A_i = \left\{diag(e^{Z_{ij}\beta});1 \le j \le N_i\right\}$ and Δ_i is $N_i \times N_i$ positive definite matrix with $(j,k)^{th}$ elements as $\delta_{ijk} = \left[\exp\left(v_{ijj} + v_{ikk}\big/2\right)\left\{\exp(v_{ijk}) - 1\right\}\right]$.

For example, under random intercept model (i.e. model specification-I described in Chandra and Chambers, 2005): $V(u_i) = \sigma_u^2$, $V(e_i) = \sigma_e^2$ and $V_i = \sigma_e^2 I_{N_i} + \sigma_u^2 1_{N_i}1'_{N_i}$ with $v_{ijj} = \sigma_e^2 + \sigma_u^2$, $v_{ijk} = \sigma_u^2$, and then $V_{\xi}\left(Y_i\right) = \Omega_i = A_i\left[e^{(\sigma_e^2+\sigma_u^2)}\left\{\exp(\sigma_e^2 I_{N_i} + \sigma_u^2 1_{N_i}1'_{N_i}) - 1_{N_i}1'_{N_i}\right\}\right]A'_i$.

The area-specific approximately bias corrected estimator (12) and variance-covariance matrix (13), grouped at population level define the population level version of ‘expected value’ model

$$E_{\xi}(Y \mid h) = \alpha_0 1_N + \alpha_1 h(Z;\eta) = \alpha J \text{ and } Var_{\xi}(Y \mid h) = \Omega \qquad (14)$$

where $Y = (Y_1',.....,Y_m')'$, $h = (h_1',....,h_m')'$ and $\Omega = diag(\Omega_i; 1 \le i \le m)$. Note that the ‘expected value’ models (5) and (14) have same form. However, model (5) is suitable for the population estimation, while model (14) includes the random area effects and is suitable for small area estimation.

4. Small Area Estimator under the Expected Value Model (14)

With appropriate sample and non-sample partition of Y, J and Ω, as in section 2, the EBLUP version of sample weights (6) under the model (14) are

$$w_{EBLUP}^{h} = 1_n + \hat{H}_h'(J'1_N - J_s'1_n) + (I_n - \hat{H}_h'J_s')\hat{\Omega}_{ss}^{-1}\hat{\Omega}_{sr}1_r \qquad (15)$$

where $\hat{H}_h = (J_s'\hat{\Omega}_{ss}^{-1}J_s)^{-1}J_s'\hat{\Omega}_{ss}^{-1}$. We note that the sample weights (15) depend on random area effects of the mixed model (7) via the covariance structure of model (14) and are thus suitable for small area estimation. We now use the MBD approach of Chambers and Chandra (2006) to define estimator for small areas. They only consider the Hájek form of the MBD estimator for small areas using sample weights derived under a linear mixed model. However, the weights (15) are derived via model calibration under the expected value model (14) where estimator is defined as the HT form (Section 2). Thus, we consider both forms of MBD estimators. The sample weights (15) associated with the sample units in the small area *i* can be used to define the following model-based direct (MBD) estimators for the i^{th} small area mean $\overline{Y}_i$:

- The Hájek form of the weighted sample for area *i*

$$\hat{\overline{Y}}_i^{Hájek} = \sum_{s_i} w_j y_j \Big/ \sum_{s_i} w_j \qquad (16)$$

- The Horvitz-Thompson form of the weighted sample for area *i*

$$\hat{\overline{Y}}_i^{HT} = \sum_{s_i} w_j y_j \Big/ N_i \qquad (17)$$

Both estimators (16) and (17) also depend on how the model calibration weights (15) are specified. In particular, we consider two different specifications for the expected value model (14), the ratio and the regression specification (see below equation (5)). This leads to four different MBD estimators that are set out below.

Estimator	Estimator type	Model specification
TrMBD1	Hájek type	Ratio specification
TrMBD2	Horvitz-Thompson type	Ratio specification
TrMBD3	Hájek type	Regression specification
TrMBD4	Horvitz-Thompson type	Regression specification

Estimation of mean squared error (MSE) of (16) and (17) follows the approach of Chambers and Chandra (2006), and treats these expressions as simple weighted domain mean estimates under the population level model (5). Under this approach the sample weights derived from (15) are treated as fixed and the prediction variance of (16) or (17) is estimated using a standard robust variance estimator. See Royall and Cumberland (1978). A "plug-in" estimate of the squared bias of (16) and (17) under this model is added to this estimated prediction variance to finally define a simple estimate of the MSE. Note that under this approach the EBLUP weights underlying (16) and (17) "borrow strength" via the assumed small area model (14), but this model is not used in inference. In particular, we treat the expected value model (14) as a vehicle for generating estimation weights, but base inference on the model (5), thus ensuring consistency with the way mean squared errors are estimated at population level. See Chambers and Chandra (2006) and Chandra and Chambers (2005).

5. Simulation Study

In this section we illustrate the performance of seven different small area estimators. These are the proposed MBD estimators (TrMBD1-TrMBD4) for skewed data (Section 4), the Hájek type (MBD1), and HT type (MBD2) MBD estimators based on sample weights derived under a linear mixed model (Chambers and Chandra, 2006) and the EBLUP under a linear mixed model (Prasad and Rao, 1990).

We consider two types of simulation studies. The first type of study uses model-based simulation to generate artificial population and sample data. These data are then used to contrast the performance of different estimators. We carried out two sets of model-based simulations, labelled A and B. In first set of simulations (simulation set-A), we investigate the performance of these estimators. However, in second set of simulations (simulation set-B), we examine the robustness of proposed method under wrong model choices. The second type of simulation study was carried out using real data and design-based simulations to test these estimators in the context of a real population and realistic sampling methods.

Four measures of estimation performance were computed using the estimates generated in the simulation study. These were the relative mean error (or relative bias) and the relative root mean squared error (RMSE), both expressed as percentages, of regional mean estimates and the coverage rate (CR) of nominal 95 per cent confidence intervals and the width of interval (Width) for regional means.

5.1 The Model Based Simulation Study

In model-based simulations, we consider a population size $N = 15{,}000$ and generated randomly the small area population sizes N_i, $i = 1,\dots,m = 30$, so that $\sum_i N_i = N$ and was kept fixed throughout the simulations. Further, we consider the sample size $n =$ 600 and generated the small area sample sizes as $n_i = N_i(n/N)$ so that $\sum_i n_i = n$ and kept fixed for all simulations (i.e. simulation set-A and set-B).

In simulation set-A, we generated the population values y_{ij} from a multiplicative model $y_{ij} = 5.0 x_{ij}^{\beta} u_i e_{ij}$. The generated population is skewed on the raw scale and linear on the log transform scale. The random errors e_{ij} were independently generated from a lognormal distribution with parameter $\mu_e = 0$ and σ_e, denoted by LN $(0, \sigma_e)$. The random area effects u_i were generated from LN $(0, \sigma_u)$. The covariate values x_{ij} were generated from LN $(6, \sigma_x)$. The values of parameter σ_e and σ_u were fixed up so that intra-area correlation varies between 0.20-0.25. We used six different sets of parameter to bring different level of variation in generated data as shown below:

Parameter	β	σ_u	σ_e	σ_x
ParA1	0.5	0.30	0.50	3.00
ParA2	0.8	0.35	0.60	2.50
ParA3	1.0	0.40	0.70	2.25
ParA4	1.3	0.45	0.80	1.75
ParA5	1.5	0.50	0.90	1.50
ParA6	2.0	0.60	1.00	1.20

From this multiplicative model, values of the response variable y_{ij} were generated for 25 small areas of sizes N_i and then random samples of sizes n_i were drawn from each area. Using this generated data we estimated the parameters using the *lme* function in R (Bates and Pinheiro, 1998), and then calculated the estimates for small areas (Section 4). The process of generating population and sample data and estimation of model parameters were independently replicated 1000 times. The results from this simulation study are reported in Table 1.

In simulation set B, population data were generated from the model $y_{ij} = 5.0x_{ij}^{1.0}\left[\exp\left(\log(x_{ij})\right)^2\right]^{\gamma} u_i e_{ij}$. The generated population is non-linear in the raw scale and quadratic on the log scale. Here, independent random errors e_{ij} and the random area effects u_i were generated from LN (0, 1.0) and LN (0, 0.5) respectively. The covariate values x_{ij} were generated from a LN (3, 0.2). We used five different values for parameter γ (-1.0, -0.5, 0.0, 0.5 and 1.0) to bring different degree of curvature in generated data, these parameter sets are denoted by ParB1-ParB5. Rest of the process was similar to simulation set-A. Table 2 presents the results from this simulation study.

5.2 The Design Based Simulation Study

In design-based simulations, our basic data come from the same sample of 1652 Australian broadacre farms (AAGIS) that were used in the simulation study reported in Chambers and Chandra (2006) and Chandra and Chambers (2005). In particular, we use the same target population of 81982 farms (obtained by sampling with replacement from the original sample of 1652 farms with probabilities proportional to their sample weights). The same 1000 independent stratified random samples as used in Chambers and Chandra (2005) were then drawn from this (fixed) population, with total sample size in each draw equal to the original sample size (1652) and with the small areas of interest defined by the 29 Australian agricultural regions represented in this population. 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 the different methods considered across a range of realistic small area sample sizes. Here, our aim is to estimate average annual farm costs (A$) in these regions with farm size (hectares) as auxiliary variable. We used random intercept model specification of the mixed model. Details of this simulated population are described in Chambers and Chandra (2006) and Chandra and Chambers (2005). Table 3 set out the results from this simulation study.

5.3 Results of the Simulation Studies

5.3.1 *Model Based Simulations*

These results show that the average relative mean errors and the average relative RMSEs for

Hajek type of estimators (TrMBD1 and TrMBD3) under expected value model (14) are significantly large for all parameter choices. Further, high coverage rates under these estimators (TrMBD1 and TrMBD3) are the consequence of large biases and wider intervals (Table 1). These estimators are severely biased since under model calibration an appropriate estimator is HT type (Section 2). However, the HT type estimators (TrMBD2 and TrMBD4) derived under ratio and regression specifications for the expected value model are almost identical. Among conventional calibration weighting based MBD estimators, both Hajek type (MBD1) and HT type (MBD2) estimators are identical. Therefore, in further discussion we drop the Hajek type of estimator under model calibration and HT type estimator under classical calibration.

Table 1 shows that the average relative mean errors and the average relative RMSEs for TrMBD2 estimator are consistently lower than both MBD1 and EBLUP estimator for all choices of parameters. However, with same order of average relative mean errors, the relative RMSE of EBLUP estimator is lower order than MBD1. The average coverage rates for TrMBD2 estimator are relatively higher with smaller width of 2-sigma confidence intervals as compare to MBD1 and EBLUP. However, with almost same coverage rates, the EBLUP has smaller average widths than MBD1.

Figure 1-2 shows the region-specific performance measures generated by three estimators (TrMBD2, MBD1 and EBLUP) for simulation set-A. These results show that both the relative mean error and the relative RMSEs of TrMBD2 are smaller than MBD1 and EBLUP method in all regions. The relative biases and the RRMSE of MBD1 and EBLUP increases proportionately with non-linearity (ParA1 to ParA6). Figure 2 indicates that the coverage rate increases and the interval width decreases, hence accuracy increases in transformation-based methods. Further, the relative interval width under TrMBD2 reduced more rapidly as non-linearity in data increases. The results indicate a significant gain due to transformation based method of small area estimation for skewed data. Further, this gain is proportionate to non-linearity in the data. Between MBD1 and EBLUP methods, the EBLUP appears to perform better.

The results from simulation set-B correspond to population data that is non-linear on raw as well as log transform scale. Here, with same justification as mentioned earlier, we consider the results generated by three estimators (TrMBD2, MBD1 and EBLUP) only. These results show when transform model is not linear then the average biases under TrMBD2 are larger than MBD1 and EBLUP and difference increases as values of δ moves away from zero. On the other hand, MBD1 and EBLUP have same order of mean errors. However, the relative

RMSEs of TrMBD2 method are lower than MBD1, but neither estimator dominates between TrMBD2 and EBLUP. The average coverage rates of EBLUP are higher than both MBD1 and TrMBD2. However, EBLUP has larger average widths than TrMBD2 (Table 2).

Figure 3-4 summarizes the region-specific performance measures generated by three estimators (TrMBD2, MBD1 and EBLUP) for simulation set-B. Figure 3 shows that for parameter set ParB1 and ParB5 (with quadratic rate γ = -1 and +1 respectively), the relative biases of TrMBD2 are larger than both MBD1 and EBLUP. However, for small values of γ (±0.5 i.e. near to zero), the relative biases are marginally same order for all methods. The relative RMSEs of TrMBD2 are lower than both MBD1 and EBLUP in most of the areas for all parameter sets except the parameter sets ParB2 and ParB3, where EBLUP is marginally better. Figure 4 demonstrates that although coverage rates of TrMBD2 are marginally lower for ParB2-ParB5 but interval widths are consistently smaller for all parameter choices (ParB1-ParB5). We noticed that in regional estimation loss in terms of coverage are marginal, however, gain in terms of reduced width is significant.

5.3.2 *Design Based Simulations*

The results from the design-based simulation using the real data (AAGIS) show that the average relative bias of TrMBD2 is smaller than EBLUP and but larger than MBD1. The relative RMSE of TrMBD2 is marginally larger and the average coverage rate higher overall (Table 3). However, Figure 5 indicates that the high relative bias and RRMSE of TrMBD2 estimator is due to an outlier in region 21. The estimator TrMBD2 is more affected by this outlying point. If we discard the outlier contaminated estimates in region 21 and examine the average based on 28 regions then the TrMBD2estimator seems to be performing better. Overall transform variable based small area estimation methods for AAGIS data appears to provide efficient set of estimates.

Note that the TrMBD2 estimator provides significant gain under linearity on transform model. However, gain may not be significant if linearity does not hold. At the same time, we noticed that if the transform model is approximately linear then it is in safer to use TrMBD2 method. For the AAGIS data, the fitted model on the transform scale (on log scale) is not exactly linear (but linear in many areas) overall. Thus, overall TrMBD2 estimator performs marginally better and provides a gain in those areas where linearity holds, not in all areas.

6. Conclusions and Further Research

Our results show that transformed variable based method for small area estimation of skewed data performs well. We note that the gain in efficiency by accounting non-linearity in data via log transform linear model is quite significant, and thus we propose to use this method for small area estimation of skewed data. Further, even though assumed model deviates slightly from linearity on transform scale, the proposed method still works well with marginal gain. These results are based on normality assumption of random errors. However, we also investigated the method assuming a gamma distribution for the random errors and noticed that the form of the estimators remain the same. This indicates that method is robust with respect to distribution of random errors. The application of proposed SAE techniques to real data from AAGIS provides a satisfactory performance. The proposed method is advisable for skewed data but identification of appropriate transform model is crucial in application of this method, otherwise results can be misleading.

In the proposed method for SAE under log transform model, the survey variables only can have strictly positive values. However, the survey variables can take zero or negative values as well and therefore it would be useful to generalise the estimation procedure for skewed data that includes these cases. We are currently working on this issue, and results obtained so far are very encouraging.

Acknowledgement

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. http://franz.stat.wisc.edu/pub/NLME/.

Carroll, R. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Chambers, R.L. (1997). Weighting and Calibration in Sample Survey Estimation. *Conference on Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth* (Editors C. Malaguerra, S. Morgenthaler and E. Ronchetti). Basel: Birkhäuser Verlag.

Chambers, R.L. and Chandra, H. (2006). Improved Direct Estimators for Small Areas. Submitted.

Chambers, R.L. and Dorfman, A.H. (2003). Transformed Variables in Survey Sampling. *Southampton Statistical Sciences Research Institute*, University of Southampton, U.K., WP/M03/21, http://www.s3ri.soton.ac.uk/publications.

Chandra, H. and Chambers, R.L. (2005). Comparing EBLUP and C-EBLUP for Small Area Estimation. *Statistics in Transition*, **7(3)**, 637-648.

Chen, G. and Chen, J. (1996). A Transformation Method for Finite Population Sampling Calibrated with Empirical Likelihood. *Survey Methodology*, **22**, 139-146.

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

Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, **72**, 320–338.

Karlberg, F. (2000). Population Total Prediction Under a Lognormal Superpopulation Model. *Metron*, 53-80.

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

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.

Wu, C. and Sitter, R.R. (2001). A Model Calibration Approach to Using Complete Auxiliary Information from Survey Data. *Journal of the American Statistical* Association, **96**, 185-193.

Table 1 Average relative mean error (ARME), average relative RMSE (ARRMSE), average coverage rate (ACR) and average 2-sigma confidence interval width (AW) for simulation set-A

Criterion	Estimator	ParA1	ParA2	ParA3	ParA4	ParA5	ParA6
ARME	TrMBD1	-86.02	-96.54	-98.43	-98.58	-98.45	-99.06
	TrMBD2	-0.01	-0.05	0.27	0.09	-0.43	0.76
	TrMBD3	-75.2	-95.97	-97.97	-98.55	-98.12	-98.66
	TrMBD4	0.02	-0.07	0.28	0.11	-0.39	0.75
	MBD1	10.98	4.11	-0.29	-6.28	-7.81	-9.59
	MBD2	12.63	5.47	0.48	-5.91	-7.58	-9.5
	EBLUP	12.65	5.44	0.49	-5.85	-7.68	-9.32
ARRMSE	TrMBD1	0.92	1.13	1.2	1.29	1.43	1.56
	TrMBD2	0.15	0.29	0.39	0.52	0.7	0.88
	TrMBD3	7.98	1.25	1.22	1.3	1.44	1.59
	TrMBD4	0.15	0.29	0.39	0.52	0.7	0.88
	MBD1	1.03	1.47	1.79	1.89	1.98	2.78
	MBD2	1.16	1.6	1.83	1.91	1.99	2.79
	EBLUP	0.76	0.69	0.61	0.75	0.98	1.29
ACR	TrMBD1	0.99	0.98	0.96	0.95	0.94	0.92
	TrMBD2	0.94	0.9	0.89	0.89	0.89	0.89
	TrMBD3	0.99	0.98	0.96	0.95	0.94	0.92
	TrMBD4	0.94	0.91	0.89	0.89	0.89	0.89
	MBD1	0.87	0.85	0.85	0.87	0.88	0.87
	MBD2	0.87	0.85	0.85	0.87	0.88	0.87
	EBLUP	0.85	0.85	0.85	0.87	0.87	0.87
AW	TrMBD1	1265	22389	140563	27x10^{4}	35 x10^{5}	44x10^{6}
	TrMBD2	208	4326	33228	7x10^{4}	11x10^{5}	15x10^{6}
	TrMBD3	1753	22487	141001	27x10^{4}	35 x10^{5}	43x10^{6}
	TrMBD4	220	4426	33722	8x10^{4}	11x10^{5}	16x10^{6}
	MBD1	1007	19318	139346	28x10^{4}	38x10^{5}	56x10^{6}
	MBD2	1033	19677	140626	28x10^{4}	38x10^{5}	56 x10^{6}
	EBLUP	380	7253	55498	13x10^{4}	20x10^{5}	31 x10^{6}

Table 2 Average relative mean error (ARME), average relative RMSE (ARRMSE), average coverage rate (ACR) and average 2-sigma confidence interval width (AW) for simulation set-B

Criterion	Estimator	ParB1	ParB2	ParB3	ParB4	ParB5
ARME	TrMBD2	3.46	0.37	0.14	-0.9	-7.54
	MBD1	-0.21	0.04	0.12	0.16	-0.85
	EBLUP	-0.19	0.04	0.13	0.17	-0.77
ARRMSE	TrMBD2	0.35	0.33	0.33	0.34	0.39
	MBD1	0.56	0.36	0.34	0.53	1.2
	EBLUP	0.38	0.3	0.29	0.36	0.56
ACR	TrMBD2	0.93	0.92	0.92	0.91	0.86
	MBD1	0.91	0.92	0.92	0.92	0.9
	EBLUP	0.93	0.94	0.94	0.93	0.92
AW	TrMBD2	0.04	2.4	207	26409	5077959
	MBD1	0.06	2.7	214	38660	12659988
	EBLUP	0.05	2.6	214	33442	9929767

Table 3 Average relative mean error (ARME), average relative RMSE (ARRMSE) and average coverage rate (ACR) for AAGIS data

	TrMBD2			MBD1			EBLUP		
	ARME	ARRMSE	ACR	ARME	ARRMSE	ACR	ARME	ARRMSE	ACR
Average of 29 areas	3.00	22.00	0.99	-2.49	20.55	0.92	4.24	19.92	0.90
*Average of 28 areas	2.54	17.15	0.99	-2.58	17.33	0.93	4.74	19.40	0.90

*excluding region number 21

Figure 1 Area-specific relative biases and RRMSE for simulation set-A

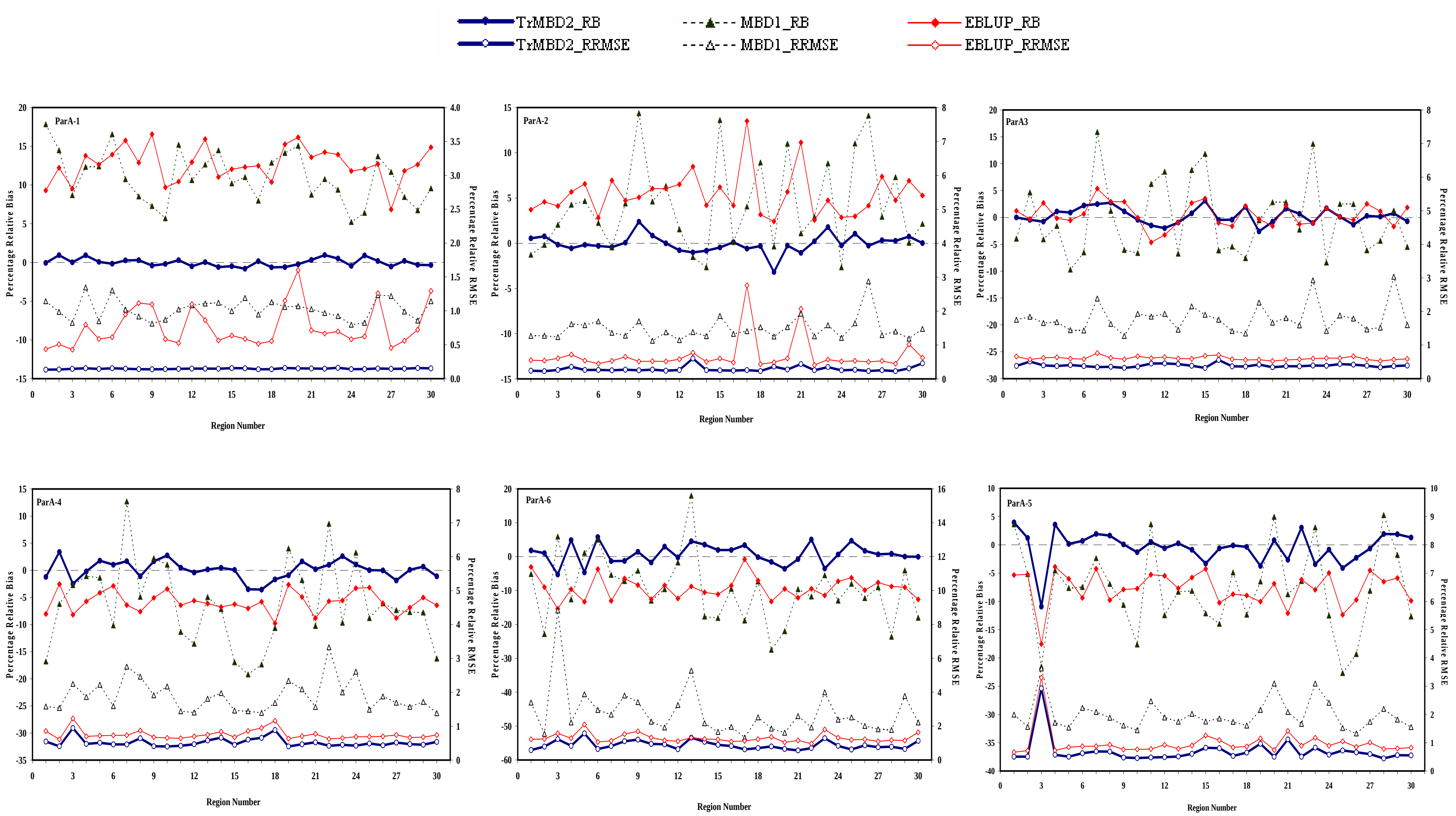


Figure 2 Area-specific coverage rates and widths of CI for simulation set-A

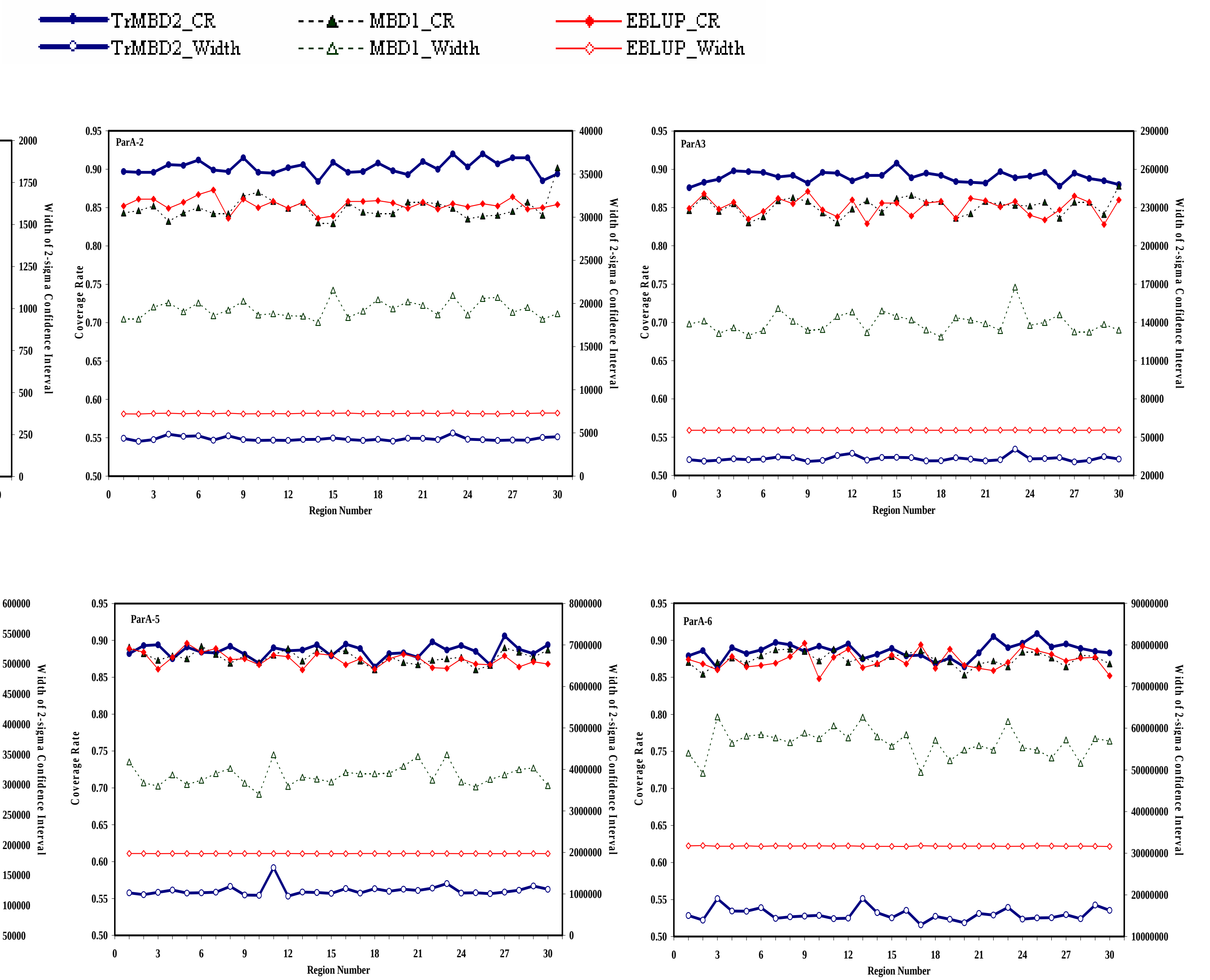


Figure 3 Area-specific relative biases and RRMSE for simulation set-B

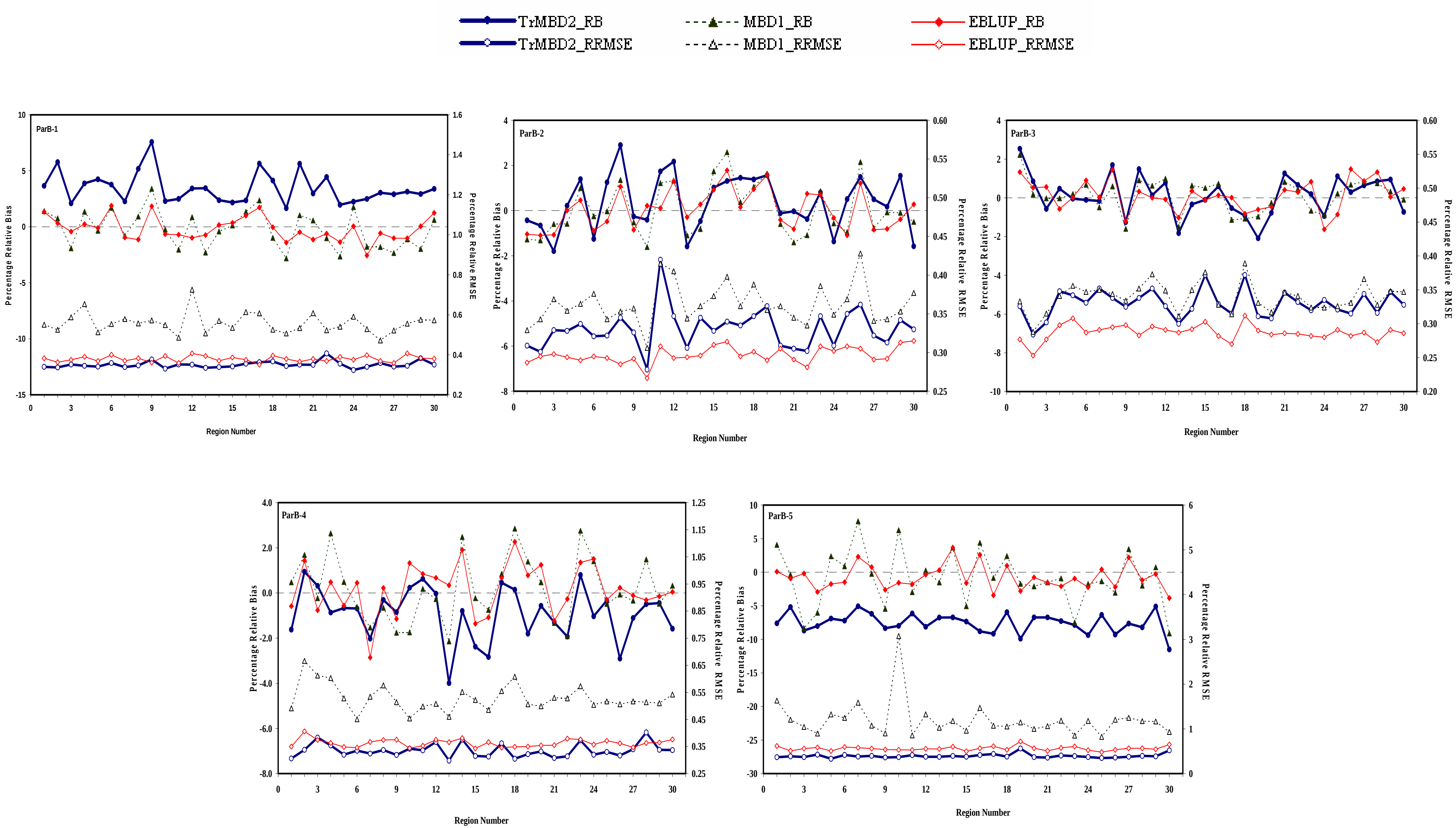


Figure 4 Area-specific coverage rates and widths of CI for simulation set-B

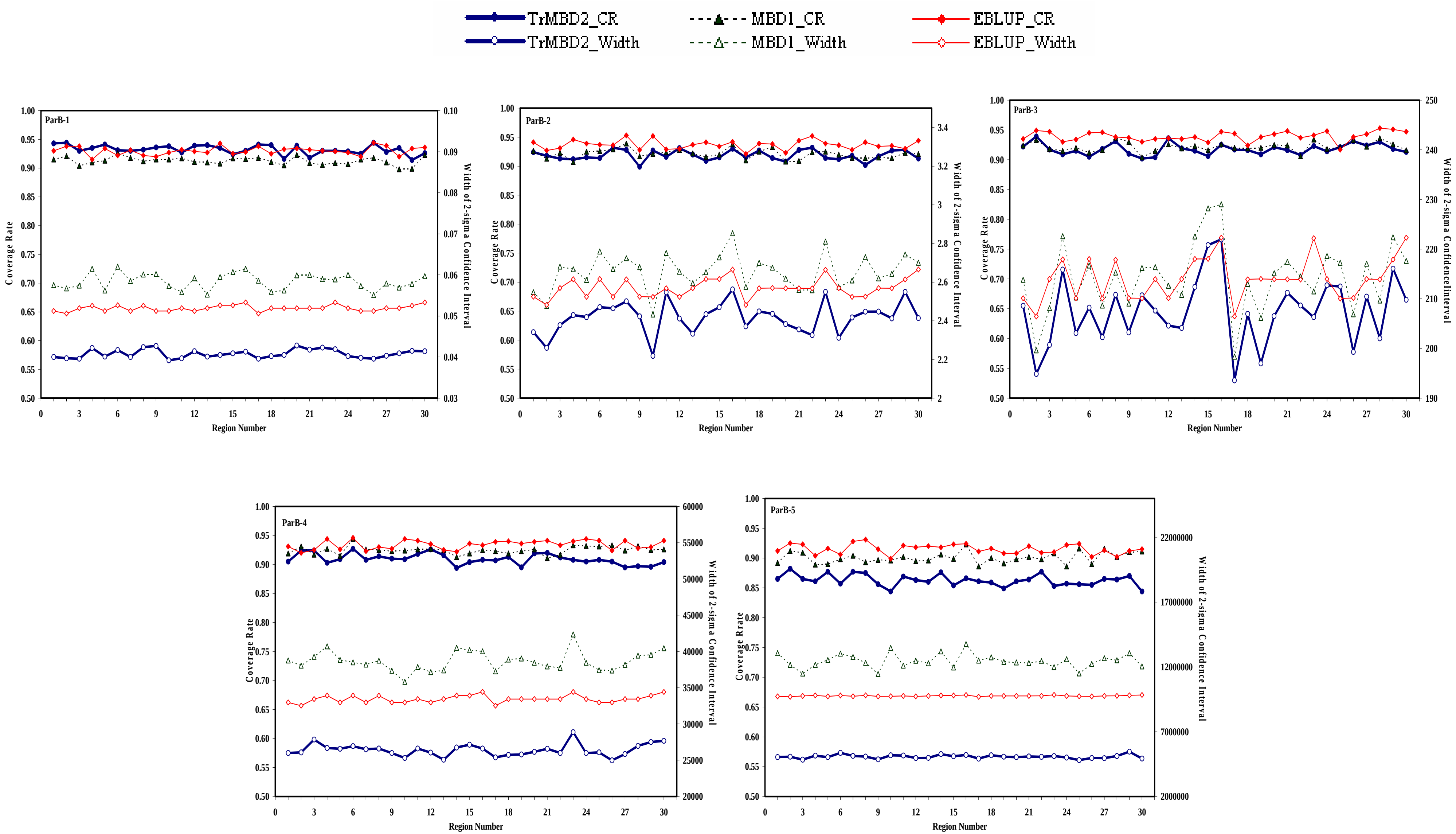
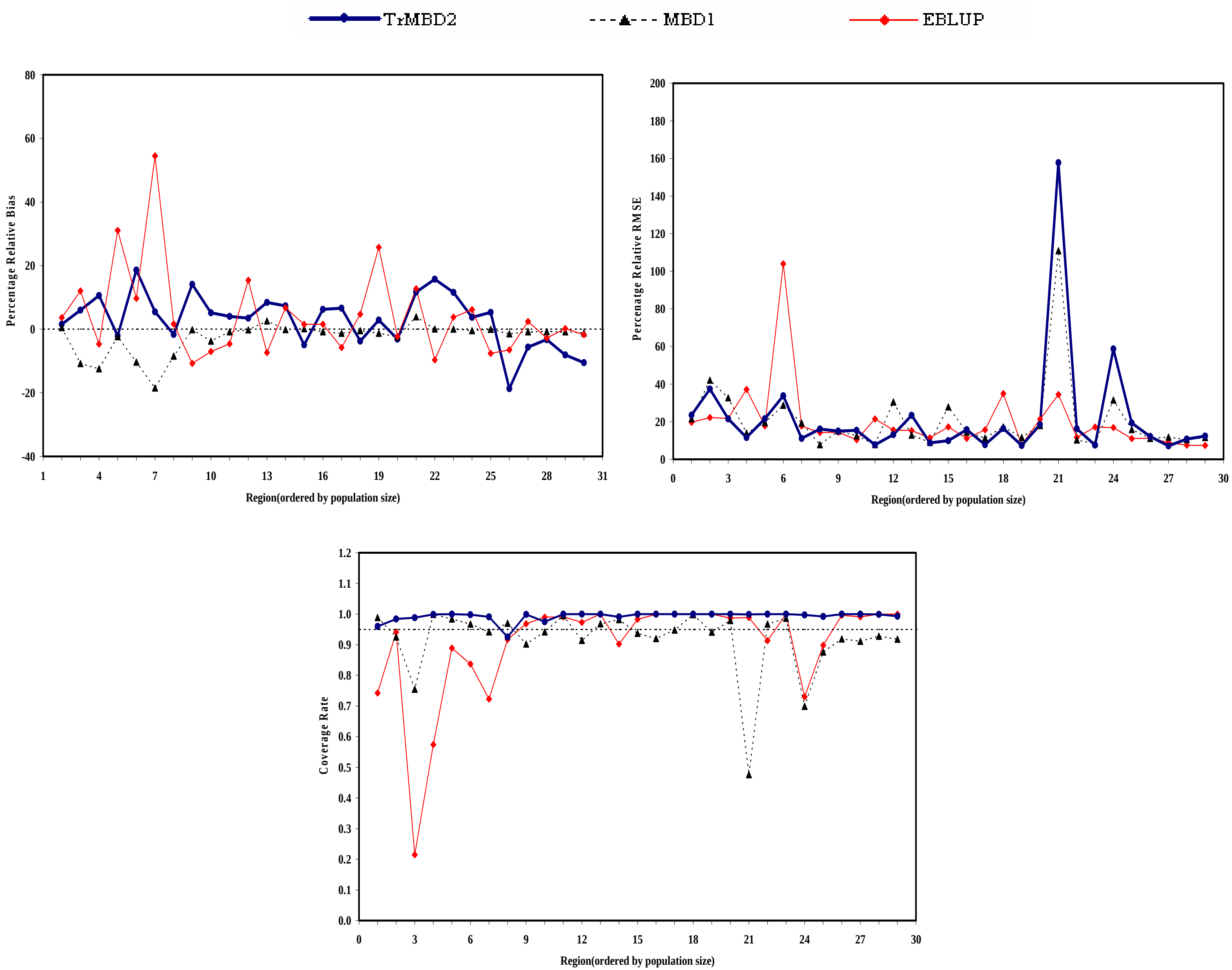


Figure 5 Area-specific relative biases and RRMSE for AAGIS data


TrMBD2
MBD1
EBLUP
Percentage Relative Bias
Region(ordered by population size)
Percenatge Relative RMSE
Region(ordered by population size)
Coverage Rate
Region(ordered by population size)