CALIBRATED WEIGHTING FOR SMALL AREA ESTIMATION

R. L. CHAMBERS

ABSTRACT

Calibrated weighting methods for estimation of survey population characteristics are widely used. At the same time, model-based prediction methods for estimation of small area or domain characteristics are becoming increasingly popular. This paper explores weighting methods based on the mixed models that underpin small area estimates to see whether they can deliver equivalent small area estimation performance when compared with standard prediction methods and superior population level estimation performance when compared with standard calibrated weighting methods. A simple MSE estimator for weighted small area estimation is also developed.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M05/04

Calibrated Weighting for Small Area Estimation

R. L. Chambers

Southampton Statistical Sciences Research Institute
University of Southampton
Highfield, Southampton SO17 1BJ
United Kingdom

Abstract

Calibrated weighting methods for estimation of survey population characteristics are widely used. At the same time, model-based prediction methods for estimation of small area or domain characteristics are becoming increasingly popular. This paper explores weighting methods based on the mixed models that underpin small area estimates to see whether they can deliver equivalent small area estimation performance when compared with standard prediction methods and superior population level estimation performance when compared with standard calibrated weighting methods. A simple MSE estimator for weighted small area estimation is also developed.

Key Words: Sample survey, sample weighting, survey estimation, model-based estimation, generalised regression estimation, MSE estimation, mixed model.

1. Overview

This paper explores the use of calibrated sample weights for small area estimation. Its motivation lies in the observation that the dominant paradigm in survey estimation for populations is weighted linear estimation, while the rapidly expanding field of small area estimation is currently dominated by a model-based predictive approach where the survey weights have little or no relevance. See Rao (2003). Many of the practical advantages of weighted linear estimation are lost when one adopts a predictive approach. Perhaps the most important of these are the simplicity of the estimation process (as well as estimation of mean square errors) and the fact that one can provide "general purpose" weights for straightforward secondary analysis of public use data sets derived from the survey data. This type of analysis has become very common with the increased availability of statistical analysis software that can accommodate weighted survey data. A further advantage is that calibration constraints are readily included in a weighted approach, allowing survey analysts who prefer a design-based approach to inference to obtain weighted estimates that have good design-based properties (Hidiroglou *et al*, 2000) at some level of aggregation.

In the following section we summarise the calibrated approach to survey weighting for population quantities. In Section 3 we then discuss issues that arise when weights that also reflect small area or local characteristics are required, and focus on the construction of calibrated survey weights under the popular linear mixed model that underpins much of small area estimation methodology. In section 4 we provide illustrative empirical results that contrast estimation using calibrated weights based on linear mixed models with estimates based on standard linear model-based calibrated weights as well as with estimates obtained via standard small area predictive estimation. Finally, in Section 5 we discuss the important issues that arise when a weighting approach is used in small area estimation and identify related topics that require further research.

2. Calibrated Sample Weighting for Population Estimation

In this section we briefly review calibrated sample weighting 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 $\bar{y}$). 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, and these values define a full rank matrix of order $n \times p$ that we denote by $\mathbf{X}_s$. 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 $\mathbf{t}_x$ of the columns of **X**.

Given this set up, the "industry standard" method of survey estimation is to estimate the population total and population mean of **y** by

$$\hat{t}_{wy} = \sum_s w_i y_i \tag{1}$$

and

$$\hat{\bar{y}}_w = \sum_s w_i y_i / \sum_s w_i \qquad (2)$$

respectively, where the weights $\{w_i\}$ reflect the relationship between the values in **y** and **X**, typically via some form of statistical model. In addition, many survey applications require weights that are calibrated on **X**, in the sense that they exactly reproduce the known population totals defined by the columns of **X**, i.e.

$$\sum_s w_i \mathbf{x}_i = \hat{\mathbf{t}}_{wx} = \mathbf{t}_x . \qquad (3)$$

Note that in some applications of calibration the population totals on the right hand side of (3) are not known, and are replaced by "reliable" estimates. This adds an extra degree of complexity to the procedure because of possible biases due to errors in the estimation of the totals. We do not consider this issue here, focusing instead on the case where these constraints are specified without error.

There are two basic approaches to constructing weights that satisfy (3). The first is by what we refer to as Linear Unbiased (LU) Weighting. That is, we assume that **y** and **X** are related by the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (4)$$

where $\boldsymbol{\varepsilon}$ denotes a *N*-vector of random variables with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$, with

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}$$

a known positive definite matrix. Here $\mathbf{V}_{ss}$, $\mathbf{V}_{sr}$, $\mathbf{V}_{rr}$ define the sample/non-sample decomposition of **V**. Royall (1976) shows that the Best Linear Unbiased Predictor (BLUP) of t_y under (4) is given by (1) with weights

$$\mathbf{w}_L = \mathbf{1}_s + \mathbf{H}_L'\left(\mathbf{X}'\mathbf{1} - \mathbf{X}_s'\mathbf{1}_s\right) + \left(\mathbf{I}_s - \mathbf{H}_L'\mathbf{X}_s'\right)\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\mathbf{1}_r \qquad (5)$$

where $\mathbf{I}_s$ is the identity matrix of order *n*, $\mathbf{1}$, $\mathbf{1}_s$, $\mathbf{1}_r$ are vectors of one's with dimensions *N*, *n* and *N* - *n* respectively, and

$$\mathbf{H}_L = \left(\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\mathbf{V}_{ss}^{-1} .$$

Let $\mathbf{I}_p$ denotes the identity matrix of order *p* and let **K** and **d** denote arbitrary *n*-vectors of constants. Following Chambers (1997), we refer to any matrix H satisfying $\mathbf{H}\mathbf{X}_s = \mathbf{I}_p$ as a LU (linear unbiased) matrix since it allows us to define a class of linear unbiased predictors (1) with weights defined by (5), but replacing $\mathbf{H}_L$ by **H**, $\mathbf{1}_s$ by **d** and $\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\mathbf{1}_r$ by **K**. These are weights of the form

$$\mathbf{w}_{He} = \mathbf{d} + \mathbf{H}'\left(\mathbf{X}'\mathbf{1} - \mathbf{X}_s'\mathbf{d}\right) + \left(\mathbf{I}_s - \mathbf{H}'\mathbf{X}_s'\right)\mathbf{K} . \qquad (6)$$

It is easy to see that the weights (6) are always calibrated on $\mathbf{X}$, i.e. $\mathbf{X}_s'\mathbf{w}_{He} = \mathbf{X}'\mathbf{1} = \mathbf{t}_x$. They also define unbiased linear predictors of t_y (provided $\mathbf{H}$ is a function of $\mathbf{X}$ and not $\mathbf{y}$) since

$$E\left(\hat{t}_{wy} - t_y\right) = E\left(\mathbf{w}_{He}'\mathbf{y}_s - \mathbf{1}'\mathbf{y}\right) = \left(\mathbf{w}_{He}'\mathbf{X}_s - \mathbf{1}'\mathbf{X}\right)\boldsymbol{\beta} = 0.$$

In fact, this unbiasedness result provides us with a useful perspective on calibration - unbiasedness with respect to the linear model (4) defined by $\mathbf{X}$ and calibration on $\mathbf{X}$ are equivalent. That is, any linear estimator with weights that are calibrated on $\mathbf{X}$ will be unbiased under (4), and conversely, any linear estimator that is unbiased under (4) will have weights that are calibrated on $\mathbf{X}$.

The above discussion represents what might be referred to the model-based interpretation of calibration. From a design-based perspective it is more usual to develop calibration weighting via the concept of "closest" calibrated weights (Deville and Särndal, 1992). Under this approach we assume the existence of an initial set of sample weights $\mathbf{d} = \{d_i\}$ (typically these are the inverses of the inclusion probabilities of the sample units, but they do not have to be – see Chambers (1996) where they are defined by a nonparametric regression model), and we construct final sample weights that are as "close" as possible to these initial weights but at the same time are calibrated on $\mathbf{X}$ (i.e. they correspond to a linear estimator that is unbiased under a linear model for the expected value of $\mathbf{y}$ given $\mathbf{X}$). There are many metrics that can be used to define "close" here, but, from a design-based perspective all are equivalent asymptotically when $\mathbf{d}$ equals the vector of inverses of inclusion probabilities of the sample units, since they then lead to the same weights as those obtained when we choose $\mathbf{w}$ to minimise $Q = (\mathbf{w} - \mathbf{d})'\boldsymbol{\Omega}(\mathbf{w} - \mathbf{d})$, where $\boldsymbol{\Omega}$ is a known positive definite matrix.

Minimising Q subject to calibration on $\mathbf{X}$, i.e. (3), leads to sample weights of the form

$$\mathbf{w}_\Omega(\mathbf{d}) = \mathbf{d} + \mathbf{H}_\Omega'(\mathbf{X}'\mathbf{1} - \mathbf{X}_s'\mathbf{d}) \tag{7}$$

where $\mathbf{H}_\Omega$ is the LU matrix $\left(\mathbf{X}_s'\boldsymbol{\Omega}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\boldsymbol{\Omega}_s^{-1}$. Again we note that $\mathbf{H}_\Omega$ can be replaced by an arbitrary LU matrix and the calibration property still holds, though now $\mathbf{w}$ and $\mathbf{d}$ are no longer necessarily "close" with respect to the metric defined by Q. Also, setting $\mathbf{d} = \mathbf{1}_s$ and $\boldsymbol{\Omega} = \mathbf{V}_s$ in (7) leads to the BLUP weights (5) under diagonal $\mathbf{V}$, but not for general $\mathbf{V}$. Similarly, setting $\mathbf{d}_s = (\pi_i^{-1})$ and $\boldsymbol{\Omega} = diag(\pi_i v_{ii})$, where π_i denotes the inclusion probability of unit i and $\mathbf{V}_s = diag(v_{ii})$, leads to the less efficient (from a model-based perspective) Generalised Regression (GREG) weights.

Irrespective of how the calibrated sample weights are derived, it is now well known that a key issue in calibration is the number of constraints. The reason for this is simple - the greater the number p of explanatory variables in (4), the greater the variability of a set of LU weights based on (4). Consequently, the more calibration constraints one imposes, the higher the variability of the resulting calibrated weights. This increased variability has unfortunate side effects, including larger standard errors and the creation of extreme weights, particularly negative weights, thus raising the possibility of negative estimates for strictly positive quantities, especially in domain analyses. Balanced against this however is the fact that increasing the number of calibration constraints also increases the explanatory power of the linear model implied by the constraints (i.e. (4) fits individual behaviour better), implying that

we can never increase (and often we decrease) the bias of a weighted estimator by increasing the number of constraints. This bias-variance trade-off in the choice of calibration constraints is an important practical problem (see Chambers *et al*, 1999).

3. Calibrated Sample Weighting for Small Area Estimation

The primary target of most surveys is estimation of population level quantities, and so calibrated sample weights are usually based on models appropriate for the population as a whole (population weighting). In particular, small area and individual level variation are assumed to "average out" over the population, in the sense that if in fact $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}$ where $\mathbf{X}\boldsymbol{\beta}$ denotes the contribution from population level effects, $\mathbf{Z}\boldsymbol{\gamma}$ denotes the contribution from small area effects and $\mathbf{e}$ denotes the contribution from individual effects, then $\mathbf{1}'\mathbf{X}\boldsymbol{\beta} >> \mathbf{1}'(\mathbf{Z}\boldsymbol{\gamma} + \mathbf{e})$ so that weights based on the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (i.e. population weighting) will still give almost unbiased estimates at population level.

However, estimation at small area level is typically an increasingly important secondary objective of many sample surveys, and in this context the above argument fails. This is because small area effects do not average out at small area level. Consequently using population weights for small area estimation (e.g. estimating the mean of $\mathbf{y}$ in small area d via the weighted mean $\hat{\bar{y}}_d = \sum_{s_d} w_i y_i / \sum_{s_d} w_i$) will be inefficient, maybe even biased. An immediate consequence is that some form of local weighting is required if survey weights are going to be used to construct small area estimates, where we define local weighting as the capacity to differentiate between the small areas that make up the population. This requirement is in addition to the calibration constraints necessary for population estimation, so sample weights become more variable, leading to the possibility of greater mean squared errors at the population level.

The simplest way to take account of differences in distribution of $\mathbf{y}$ between small areas is to assume that area effects are constant within a small area (d denotes one of D small areas). This suggests we extend (4) to

$$\mathbf{y}_d = \mathbf{X}_d\boldsymbol{\beta} + \gamma_d\mathbf{1}_d + \boldsymbol{\varepsilon}_d \qquad (8)$$

where a subscript of d denotes restriction to small area d. It is easy to see that unbiased weighting under this model requires us to calibrate both on $\mathbf{X}$ and on the small area population counts $\{N_d\}$. Assuming $\mathbf{X}$ contains an intercept term, this equates to $p+D-1$ calibration constraints, i.e. an additional $D - 1$ constraints.

There are two problems with (8). The first is that it implicitly contains the assumption that the relationship between $\mathbf{y}$ and $\mathbf{X}$ is essentially the "same" in each small area. The second is that D is not so large that fitting (8) becomes very difficult using the sample data. If we believe that the relationship between $\mathbf{y}$ and $\mathbf{X}$ varies between areas we could consider extending (8) (again assuming $\mathbf{X}$ contains an intercept term) to

$$\mathbf{y}_d = \mathbf{X}_d\boldsymbol{\beta}_d + \boldsymbol{\varepsilon}_d . \qquad (9)$$

This is the small area post-stratification model, and is equivalent to calibrating on **X** at small area, rather than population, level (i.e. *pD* constraints). It can only be used if we know the area level values of the calibration constraints and is clearly even more problematic than (8) when *D* is large.

There is a different way to build small area effects into weights, however, and this is by basing them on mixed models. To this end we observe that the preceding development implicitly assumed that small area effects are included in the fixed part of our model for **y**, with uncorrelated individual effects. However, we can also build small area effects into our weights by explicitly allowing for the possibility of correlations between individuals, both within small areas and between small areas. That is, we use the BLUP specification (5), with **V** defined by an appropriate model that allows this correlation.

The most commonly used class of models with these characteristics is the class of mixed linear models. These are models for the values $\mathbf{y}_d$ of the survey variable in small area *d* of the form

$$\mathbf{y}_d = \mathbf{X}_d\boldsymbol{\beta} + \mathbf{Z}_d\boldsymbol{\gamma}_d + \mathbf{e}_d \tag{10}$$

where $\mathbf{Z}_d$ denotes a matrix of order $N_d \times q$ and $\boldsymbol{\gamma}_d$, $\mathbf{e}_d$ are independent random vectors, of dimension q and N_d respectively, both with zero mean vectors and with $Var(\boldsymbol{\gamma}_d) = \boldsymbol{\Sigma}$, $Var(\mathbf{e}_d) = \sigma_e^2\mathbf{I}_d$. There is a huge literature on the use of (10) in small area estimation, see Rao (2003). Our interest, however, is in its use in sample weighting via, for example, (5). In this context we assume that the sample data also satisfy (10) and so we can estimate $\boldsymbol{\Sigma}$ and σ_e^2 from these data. Denote these estimates by $\hat{\boldsymbol{\Sigma}}$ and $\hat{\sigma}_e^2$. An estimate $\hat{\mathbf{V}}_d = \hat{\sigma}_e^2\mathbf{I}_d + \mathbf{Z}_d\hat{\boldsymbol{\Sigma}}\mathbf{Z}_d'$ of the variance-covariance matrix $\mathbf{V}_d$ of $\mathbf{y}_d$ follows immediately. For any sequence $\{\mathbf{A}_k\}$ of matrices, let $blk-diag(\mathbf{A}_k)$ denote the block diagonal matrix defined by $\{\mathbf{A}_k\}$. Then $\hat{\mathbf{V}} = blk-diag(\hat{\mathbf{V}}_d)$ is the resulting estimated variance-covariance matrix for the population vector **y**. It follows we can approximate the BLUP sample weights (5) under (10) by substituting this estimated variance-covariance matrix for **V**, leading to

$$\mathbf{w}_{EBLUP} = \mathbf{1}_s + \mathbf{H}'_{EBLUP}\left(\mathbf{X}'\mathbf{1} - \mathbf{X}'_s\mathbf{1}_s\right) + \left(\mathbf{I}_s - \mathbf{H}'_{EBLUP}\mathbf{X}'_s\right)\hat{\mathbf{V}}_{ss}^{-1}\hat{\mathbf{V}}_{sr}\mathbf{1}_r \tag{11}$$

where

$$\mathbf{H}_{EBLUP} = \left(\mathbf{X}'_s\hat{\mathbf{V}}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}'_s\hat{\mathbf{V}}_{ss}^{-1}.$$

It is easy to see that these "EBLUP" weights are a special case of (6) and so are calibrated on **X**. Furthermore, since they only depend on (10) via the covariance structure in the sample/population, extension to more complex covariance structures (e.g. spatial correlation between population units) only requires $\hat{\mathbf{V}}_{ss}$ and $\hat{\mathbf{V}}_{sr}$ to be computed under these more complex models. We do not pursue this extension in this paper however.

Put $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_s'\hat{\mathbf{V}}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\hat{\mathbf{V}}_{ss}^{-1}\mathbf{y}_s = \mathbf{H}_{EBLUP}\mathbf{y}_s$, $\hat{\boldsymbol{\beta}}_d = \left(\mathbf{X}_{sd}'\hat{\mathbf{V}}_{ssd}^{-1}\mathbf{X}_{sd}\right)^{-1}\mathbf{X}_{sd}'\hat{\mathbf{V}}_{ssd}^{-1}\mathbf{y}_{sd}$ and $\mathbf{K}_d = \hat{\mathbf{V}}_{ssd}^{-1}\hat{\mathbf{V}}_{srd}\mathbf{1}_{rd}$, where a subscript of d denotes restriction to area d. Then it is not difficult to see that the estimator of the area d total $t_{dy} = \mathbf{1}_d'\mathbf{y}_d$ that results when we use the weights (11) is

$$\begin{aligned}\hat{t}_{wdy} &= \mathbf{w}'_{EBLUP,d}\mathbf{y}_{sd} \\ &= \mathbf{1}_{sd}'\mathbf{y}_{sd} + \left(\mathbf{1}'\mathbf{X} - \mathbf{1}_s'\mathbf{X}_s\right)\left(\mathbf{X}_s'\hat{\mathbf{V}}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_{sd}'\hat{\mathbf{V}}_{ssd}^{-1}\mathbf{y}_{sd} + \mathbf{K}_d'\left(\mathbf{y}_{sd} - \mathbf{X}_{sd}\hat{\boldsymbol{\beta}}\right) \\ &= \mathbf{1}_{sd}'\mathbf{y}_{sd} + \left(\mathbf{1}'\mathbf{X} - \mathbf{1}_s'\mathbf{X}_s\right)\left(\mathbf{X}_s'\hat{\mathbf{V}}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\left(\mathbf{X}_{sd}'\hat{\mathbf{V}}_{ssd}^{-1}\mathbf{X}_{sd}\right)\hat{\boldsymbol{\beta}}_d + \mathbf{K}_d'\left(\mathbf{y}_{sd} - \mathbf{X}_{sd}\hat{\boldsymbol{\beta}}\right)\end{aligned} \tag{12}$$

where the first two terms on the right hand side of (12) can be interpreted as the contribution of the fixed effect part of (10) to the estimate and the final term is essentially an area specific residual, reflecting the contribution of the random effects in (10) to the estimate. Thus, for the special case where $\mathbf{Z}_d$ is a vector of one's and so γ_d in (10) is scalar (sometimes referred to as the Random Intercepts Model), this final term is $(N_d - n_d)\hat{\gamma}_d$ where $\hat{\gamma}_d$ is the estimated effect for area d, i.e. the "shrunken residual" $\hat{\gamma}_d = (1 + \hat{\phi}n_d)^{-1}\hat{\phi}n_d\bar{r}_{sd}$. Here $\hat{\phi} = \hat{\Sigma}/\hat{\sigma}_e^2$ and $\bar{r}_{sd}$ is the average of the area d residuals $\mathbf{y}_{sd} - \mathbf{X}_{sd}\hat{\boldsymbol{\beta}}$.

An interesting generalisation of (11) leads to a version of GREG weighting based on mixed models. These are weights of the form

$$\mathbf{w}_{EGREG} = \mathbf{d} + \mathbf{H}'_{EGREG}\left(\mathbf{X}'\mathbf{1} - \mathbf{X}_s'\mathbf{d}\right) + \left(\mathbf{I}_s - \mathbf{H}'_{EGREG}\mathbf{X}_s'\right)\tilde{\mathbf{K}} \tag{13}$$

where $\mathbf{H}_{EGREG} = \left(\mathbf{X}_s'\tilde{\mathbf{V}}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\tilde{\mathbf{V}}_{ss}^{-1}$ and $\tilde{\mathbf{K}} = \tilde{\mathbf{V}}_{ss}^{-1}\tilde{\mathbf{V}}_{sr}\mathbf{1}_{rd}$, with $\tilde{\mathbf{V}}_{ss}$, $\tilde{\mathbf{V}}_{sr}$ defined by the pseudo-likelihood ($\mathbf{d}$-weighted) estimates of the parameters of the mixed model.

An important point that needs be made at this stage is that "EBLUP weighting" is not the same as "EBLUP prediction" of the population total in area d. This is easily seen when we compare (12) with the corresponding predictor, which is

$$\hat{t}_{dy}^{EBLUP} = \mathbf{1}_{sd}'\mathbf{y}_{sd} + \left(\mathbf{1}_d'\mathbf{X}_d - \mathbf{1}_{sd}'\mathbf{X}_{sd}\right)\hat{\boldsymbol{\beta}} + \mathbf{K}_d'\left(\mathbf{y}_{sd} - \mathbf{X}_{sd}\hat{\boldsymbol{\beta}}\right). \tag{14}$$

Thus, although (12) and (14) lead to identical population level estimates, there is in fact no unique representation of (14) as a weighted linear combination of the values in $\mathbf{y}_{sd}$.

We finally turn to estimation of the MSE of the weighted estimate for small area d. To start, we observe that when small area effects are part of the mean structure of a linear model for $\mathbf{y}$, e.g. via fixed area effects, see (8) and (9), MSE estimation is relatively straightforward. Well known results indicate that robust model-based methods as well as appropriately conditioned design-based methods lead to estimators of the MSE that are essentially of the form $\hat{V}_d = \sum_{s_d} w_i^2(y_i - \hat{y}_i)^2 + \text{lower order terms}$, where $\hat{y}_i$ denotes the fitted value for y_i under the linear model implied by the calibration constraints. When the assumed model includes random effects, e.g. (10), then we need to decide whether we wish to estimate the conditional

or unconditional MSE, i.e. whether we wish to treat the random effect γ_d in (10) as fixed or not. Estimation of the conditional MSE of the "true" EBLUP can be extremely complicated (Prasad and Rao, 1990). However, this is not the case when one considers estimating the unconditional MSE of the "weighted EBLUP" (12) since this is still a simple weighted estimate (albeit with weights that are rather complex in structure). Furthermore, such an approach is consistent with the way the MSE is estimated at the population level. Our strategy, therefore, is to estimate the unconditional MSE treating these weights as fixed. Thus, we write down the prediction variance for the area *d* weighted mean (2) as

$$Var(\hat{\bar{y}}_{wd} - \bar{y}_d) = N_d^{-2}\left(\sum\nolimits_{s_d} u_i^2 Var(y_i) + \sum\nolimits_{r_d} Var(y_i)\right) \tag{15}$$

where

$$u_i = \left(\sum\nolimits_{s_d} w_j\right)^{-1}\left(N_d w_i - \sum\nolimits_{s_d} w_j\right).$$

A robust model-based estimate of (15) is obtained by substituting the squared residual $(y_i - \hat{y}_i)^2$ for $Var(y_i)$ in the first (leading) term on the right hand side of (15). If these squared sample residuals are also used to estimate the second term, the resulting estimator of (15) is

$$\hat{V}_d = \sum\nolimits_{s_d} \theta_i (y_i - \hat{y}_i)^2 \tag{16}$$

where $\theta_i = N_d^{-2}\left(u_i^2 + (N_d - n_d)/(n_d - 1)\right)$. Since we are interested in an unconditional MSE, (15) is an unconditional variance, and so we use the residuals $y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ in (16).

Using (16) to estimate the MSE of $\hat{\bar{y}}_{wd}$ implicitly assumes that this weighted mean is unconditionally unbiased for $\bar{y}_d$. However, this is not generally the case, since $E(\hat{\bar{y}}_{wd} - \bar{y}_d) = (\bar{\mathbf{x}}_{wd} - \bar{\mathbf{x}}_d)'\boldsymbol{\beta}$ under (10), where $\bar{\mathbf{x}}_{wd}$ denotes the weighted average of the sample $\mathbf{x}_i$ in area *d*. Calibration on $\mathbf{X}$ ensures that this term vanished at population level, but not necessarily at small area level. A simple estimate of this bias is $\hat{B}_d = (\bar{\mathbf{x}}_{wd} - \bar{\mathbf{x}}_d)'\hat{\boldsymbol{\beta}}$. Our suggested estimator of the unconditional MSE of $\hat{\bar{y}}_{wd}$ is therefore

$$\hat{M}_d = \hat{V}_d + \hat{B}_d^2 . \tag{17}$$

Note that one could alternatively directly "bias correct" the estimate $\hat{\bar{y}}_{wd}$ using $\hat{B}_d$. However, this is not recommended since this bias estimator increases the variability of our estimator much more than it reduces its bias. Using it in (17) is a more conservative, and safer, approach.

4. An Empirical Evaluation

In this section we illustrate the performance of small area estimation based on the weighting approach. We use design-based simulation since this represents a standard way of evaluating

the performance of a survey estimation procedure. The basic data for this study came from the same sample of 1652 Australian broadacre farms that were the basis of the exploration of weighting methods reported in Chambers (1996). Here however we use these farms to generate a target population of 81982 farms by sampling with replacement from them with probabilities proportional to their sample weights. We then drew 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 within these strata were fixed to be the same as in the original sample. Various characteristics of the target population are set out in Table 1.

In our analysis we treated the regions as the small areas of interest and focused on estimating the average value of Y = annual farm costs (A$) in these regions. Regions are grouped into zones (Pastoral, Mixed Farming, Coastal), with farm size (hectares) assumed known for each farm in the population. Analysis of the data in the original sample indicated that, although the linear relationship between Y and farm size is rather weak, this improves when separate linear models are fitted within six poststrata, 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). However, as Figure 1 clearly shows, the relationship between Y and farm size in each of these six poststrata is still extremely heteroskedastic. In many ways this population is very much like the populations that are the focus of business surveys, and our simulation should provide a perspective on how some widely used small area estimation methods might work in this environment.

To start, we consider three specifications for $\mathbf{X}$ (and hence sets of calibration constraints). These are Mean (the only constraint is that the weights sum to N = 81982), Area+Region (weights constrained to reproduce the population total for farm area and the population size in each region) and SizeZone*Area (weights constrained to reproduce population and farm area totals in each of the six poststrata). In addition to the fixed effects model (4), weights were constructed based on the random effects model (10) under two specifications for $\mathbf{Z}_d$. These are a Random Intercepts specification ($\mathbf{Z}_d$ equal to a vector on one's) and a Random Slopes specification ($\mathbf{Z}_d$ equal to the design matrix for a linear regression on farm area). Both BLUP/EBLUP and GREG/EGREG versions of weights were computed, as were EBLUP predictors under the Random Intercepts and Random Slopes specifications. REML estimates of random effects parameters, based on default values output by the lme function in R (Bates and Pinheiro, 1998, R Development Core Team, 2004), were used throughout.

Table 2 sets out the population level empirical biases and RMSEs that were generated by the different weighting methods in the simulations. No results are presented for the EBLUP prediction methods since these coincide with EBLUP weighting results at population level. Two things stand out in Table 2. The first is that under both the Mean and Area+Region specifications introduction of the mixed model (10) does not always lead to better population estimates. It is only under the (preferred) SizeZone*Area specification that introduction of (10) leads to improved estimation. Secondly, GREG/EGREG weighting is much less efficient than BLUP/EBLUP weighting, particularly under a random effects specification. In fact, on the basis of these results one would be extremely cautious about using (13) in estimation. Such an assessment, however, needs to take into account that GREG estimation via (13) might have been be destabilised in our simulations by the fact that the sample design strata and the small areas of interest were the same. It should also be remembered that REML estimation was used for the random effects parameters in (13), rather than pseudo-likelihood estimation. Further research is clearly needed to identify situations where (13) is appropriate.

In the following analysis we therefore no longer consider GREG/EGREG weighting (because of its poor performance in our study) and focus on BLUP/EBLUP weighting under the SizeZone*Area model.

The main reason for using EBLUP weighting defined by (11) is to allow regional estimation with performance comparable with that achieved when one applies standard prediction methods based on (10). Figure 2 shows boxplots of the distributions of both weighted and predictive estimates under the SizeZone*Area specification in each of the 29 regions in our application. These plots show clearly that the regional estimates that result from application of weighting methods are not the same as those that result from taking a predictive approach. They also show that neither approach dominates the other, though, with the exception of two regions, it seems that Random Slopes weighting performs marginally better overall. In the two regions (3 and 21) where this weighting approach fails, inspection of Figure 2 indicates that this is the consequence of a few outlying estimates. In fact, the outlying estimates for region 21 are all caused by presence of a single massive outlier (y_i > A$30,000,000) from the original sample that was included in the simulation population (twice) and then selected (in one case, twice) in 37 of the 1000 simulation samples. Figure 3 shows the boxplot of distributions of estimate values when these samples are excluded. This is very different from the results for region 21 shown in Figure 2 and is closer to reality, in the sense that standard data editing procedures would normally be used to detect and exclude (or at least modify) a sample outlier prior to publication of the survey estimates. The impact of such quality control can be substantial, reducing the overall relative RMSE for the Random Slopes weighting estimator from the 4.54% reported in Table 2 to 3.30%.

Figures 4 to 6 summarise the regional variation in bias and RMSE of the estimators contributing to Figure 2. In Figure 4 we compare the Fixed Effects, Random Intercepts and Random Slopes versions of the BLUP/EBLUP weighting estimators. This confirms the generally better performance of the Random Slopes weighting method. In Figures 5 and 6 we compare the weighting and predictive approaches under Random Intercepts and Random Slopes separately. As pointed out earlier, neither approach is generally dominant, though if one discounts the outlier driven results for region 21, then the Random Slopes version of EBLUP weighting seems the method of choice for regional estimation in our simulation population (among the methods considered in this study).

In Table 3 we show how the simple unconditional MSE estimator (17) performed in the simulation study with respect to the design-based coverage of "2 sigma" confidence intervals based upon it. These are intervals with nominal coverage of approximately 95%, defined by the estimate plus or minus twice the square root of the estimated MSE. Our results are very encouraging, particularly for the MSE estimator based on Random Slopes weighting. Its population coverage is almost perfect (94.3%) and in only 3 of the 29 regions does it lead to intervals with coverage less than 90%, with one of these already identified as problematic because of outliers. The corresponding coverage performances for this estimator under Fixed Effects and Random Intercept weighting are not as good, but still reasonable. Under Fixed Effects weighting the population coverage drops to 90.3%, with 7 of the 29 regions recording coverages less than 90%, while under Random Intercepts weighting the population coverage is 88.3%, with 4 of the 29 regions recording coverages less than 90%. Median coverage over the 29 regions is 91.7% for Fixed Effect weighting, 94.1% for Random Intercept weighting and 94.6% for Random Slopes weighting.

Before concluding this discussion of our simulation study there is one further important observation that needs to be made. The more diligent reader will no doubt have realised that the boxplots in Figure 2 display negative estimates on a number of occasions. That such estimates occur should not come as a surprise, since there is nothing in the construction of the BLUP/EBLUP (and GREG/EGREG) weights to prevent the occurrence of negative weights, and consequently the possibility of negative regional estimates. Similarly, there is nothing in the definition of the EBLUP (14) to stop this estimate becoming negative. In fact, over our 1000 simulations we observed that under the SizeZone*Area model, Random Intercepts prediction generated a single negative estimate in region 1, Random Slopes prediction generated 14 negative estimates in region 2 and Random Slopes weighting generated 3 negative estimates in region 3. These numbers are not excessive, but they do show that there is a problem, particularly in cases where the sample size in the small area of interest is very low. It was particularly acute for EGREG weighting, which generated many more negative regional estimates. For example, under Random Slopes EGREG weighting based on the SizeZone*Area model, 599 out of the 1000 estimates in region 13 were negative! Furthermore, negative weights don't just lead to negative estimates. Since under (2) the sum of sample weights in a region d is effectively an estimate of the population size N_d of the region, negative weights can skew this estimate towards zero and consequently lead to a gross overestimate of the region mean. This was the reason for the large positive outlier under Random Slopes EBLUP weighting in region 3. In this case, of the total of 35 (out of 1652) sample units with negative weights under this method in this simulation, 22 were concentrated in region 3. Given a total sample size of 30 in the region, this lead to an implied population estimate of approximately 9 for region 3, which is patently ridiculous (the true population is 189), and a grossly inflated estimate of average regional farm costs.

Conclusions and Further Research

In this paper we have explored the possibility of using mixed models to construct calibrated sample weights and have investigated the performance of such weights in both population level as well as small area level estimation. Our results show that EBLUP weighting based on a suitable mixed model (i.e. one that adequately reflects the between area variability in the population) can work well, leading to small area estimates that are comparable to the usual prediction-based EBLUP estimates. However, it also shows that issues such as the occurrence of negative weights, which are typically of less concern when constructing population level estimates, can become much more important when small area estimates are also required. To illustrate, in the situation discussed in the previous paragraph the 35 sample units with negative weights had virtually no impact on the population estimate of average farm costs, but seriously destabilised the corresponding estimate for region 3. Methods for dealing with negative weights under "standard" regression models have been discussed in the literature (Huang and Fuller, 1978; Bardsley and Chambers, 1984; Deville and Sarndal, 1992; Chambers, 1996) but their application in the context of mixed models remains to be explored. This is particularly important if one wishes to adopt a design-based approach to weighting (i.e. use GREG/EGREG weights) since such weights are particularly susceptible to taking negative values.

We have also suggested a simple estimator for the unconditional MSE of our weighted small area estimator. In our simulation, this MSE estimator provided good coverage performance. However, this may have been because the highly heteroskedastic nature of our simulation population meant that only the leading term in the MSE (the target of our MSE estimator) was

relevant. Further work is needed with less variable populations to see whether it continues to perform well when variability from estimation of population parameters, particularly the variance components) starts to become important.

A large number of further areas of research remain. These include further development of GREG/EGREG weighting to make this approach less affected by extreme weights. They also include the rather important problem of extending the weighting method to nominal and ordinal variables. Linear mixed models for such variables are typically in a transformed scale (e.g. generalised linear mixed models) and so linear estimators like (1) cannot be expected to perform well. We anticipate that the concept of model calibration (Wu and Sitter, 2001) will play a major role in resolving this problem.

Finally, we should point out that the EBLUP weights (11) are, strictly speaking, variable specific since they depend on estimated variance components for a particular *Y*. The issue of how to develop a "general purpose" mixed model specification that allows these weights to also become "general purpose" therefore remains.

References

Bardsley, P. and Chambers, R. L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics* **33**, 290 - 299.

Bates, D.M. and Pinheiro, J.C. (1998). Computational methods for multilevel models. Available in PostScript or PDF formats at http://franz.stat.wisc.edu/pub/NLME/

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics* **12**, 3 - 32.

Chambers, R.L. (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., Skinner, C.J. and Suojin Wang. (1999). Intelligent calibration? Invited Paper, *Proceedings of the International Association of Survey Statisticians*, *52nd Session of the International Statistical Institute*, Helsinki, August 10-18.

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

Hidiroglou, M. A., Estavao, V. M. and Arcaro, C. (2000). Generalised estimation system and future enhancements. In *ICES-II Proceedings of the Second International Conference on Establishment Surveys*, pp 687-696. Alexandra, Virginia: American Statistical Association.

Huang, E. T. & Fuller, W. A. (1978). Nonnegative regression estimation for survey data. *Proceedings of the American Statistical Association*, 300 – 305.

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.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.

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.

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 Regional characteristics of simulation population. Regions are numbered in order of increasing population size.

Region	Population	Sample	Average Farm Area	Average Farm Costs
1	79	6	297958	467964
2	115	10	55731	171414
3	189	30	359383	670926
4	330	25	178355	186984
5	388	36	108038	208142
6	465	19	16717	130316
7	604	36	131544	302583
8	729	40	21976	242836
9	737	30	23083	179112
10	964	30	23712	180467
11	1586	51	2213	116965
12	1778	62	891	114442
13	1984	55	1066	96162
14	2182	47	4398	233171
15	2607	79	1239	97839
16	2683	60	581	93202
17	2689	60	701	84790
18	2847	34	373	36979
19	3056	74	799	101101
20	3139	51	3200	87919
21	3910	73	563	78509
22	4486	117	4635	164889
23	4550	80	960	86218
24	4587	95	1862	184153
25	5368	83	1838	198156
26	5528	103	1013	105151
27	6489	108	1403	134169
28	6980	81	812	95617
29	10933	77	360	66285
Population	81982	1652	5475	118997

Table 2 Empirical biases and RMSEs (both expressed as percentages of the target value) for the population estimates generated by the different weighting methods.

Model/Calibration Constraints	Fixed Effects		Random Intercepts		Random Slopes	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
BLUP/EBLUP weighting						
Mean	19.9468	20.9360	2.8234	8.5190		
Area+Region	0.0348	5.4221	0.4436	5.8628	1.6685	5.4813
SizeZone*Area	-1.0523	4.9556	-0.1790	4.7726	0.0763	4.5406
GREG/EGREG weighting						
Mean	0.0091	5.4631	-34.1876	36.9257		
Area+Region	0.0684	5.4204	3.4529	14.1892	-85.2070	94.4300
SizeZone*Area	0.1830	5.1093	3.0184	5.8319	-3.4054	8.5548

Table 3 Empirical coverages of "2-sigma" confidence intervals for population mean and regional means generated using (17) and a SizeZone*Area specification for **X**. Regions are numbered in order of increasing population size.

Region	Fixed Effects	Random Intercepts	Random Slopes
1	0.990	0.989	0.993
2	0.917	0.928	0.955
3	0.622	0.760	0.854
4	0.998	0.999	1.000
5	0.915	0.984	0.990
6	0.926	0.968	0.992
7	0.917	0.942	0.962
8	0.965	0.970	0.965
9	0.903	0.902	0.945
10	0.931	0.941	0.946
11	0.996	0.996	0.995
12	0.861	0.914	0.911
13	0.963	0.968	0.962
14	0.978	0.982	0.934
15	0.909	0.937	0.959
16	0.903	0.920	0.928
17	0.943	0.948	0.939
18	0.998	0.998	0.998
19	0.902	0.941	0.937
20	0.970	0.979	0.990
21	0.510	0.476	0.404
22	0.962	0.967	0.981
23	0.990	0.986	0.985
24	0.681	0.699	0.708
25	0.864	0.876	0.944
26	0.915	0.919	0.925
27	0.902	0.911	0.938
28	0.898	0.928	0.930
29	0.875	0.918	0.912
Population	0.903	0.883	0.943

Figure 1 Farm costs vs. farm area for the simulation population. Rows correspond to zones (1 = Pastoral, 2 = Wheat-Sheep and 3 = High Rainfall) and columns are "smaller" farms (left) and "larger" farms (right). Line is least squares regression fit.

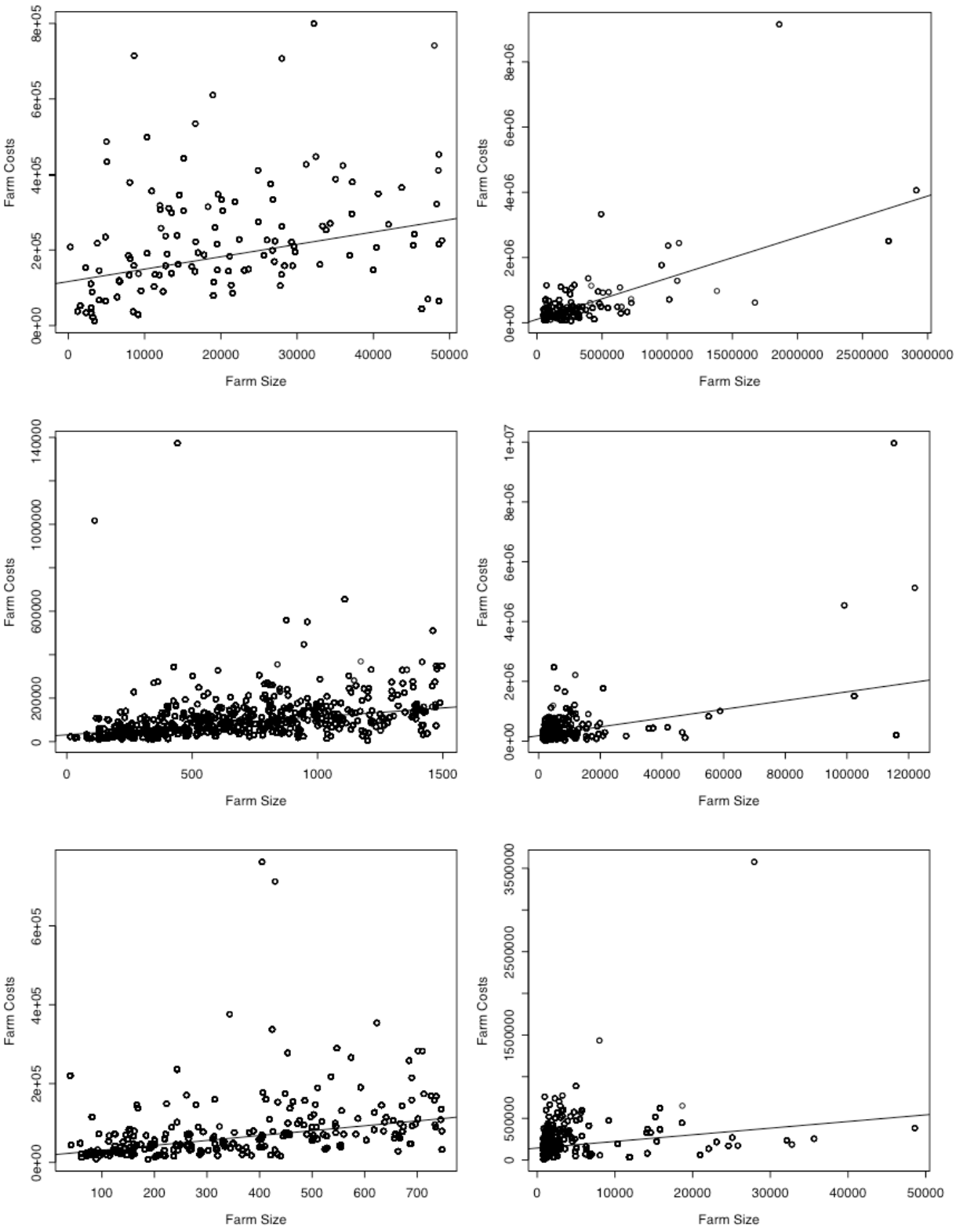


Figure 2 Distributions of estimates for regions 1 to 6 produced under the SizeZone*Area specification. FEW = Fixed Effect weighting, RIW = Random Intercept weighting, RIP = Random Intercept prediction, RSW = Random Slope weighting, RSP = Random Slope Prediction. Plots are ordered left to right and top down by increasing region population size. Dotted horizontal line is true region mean.

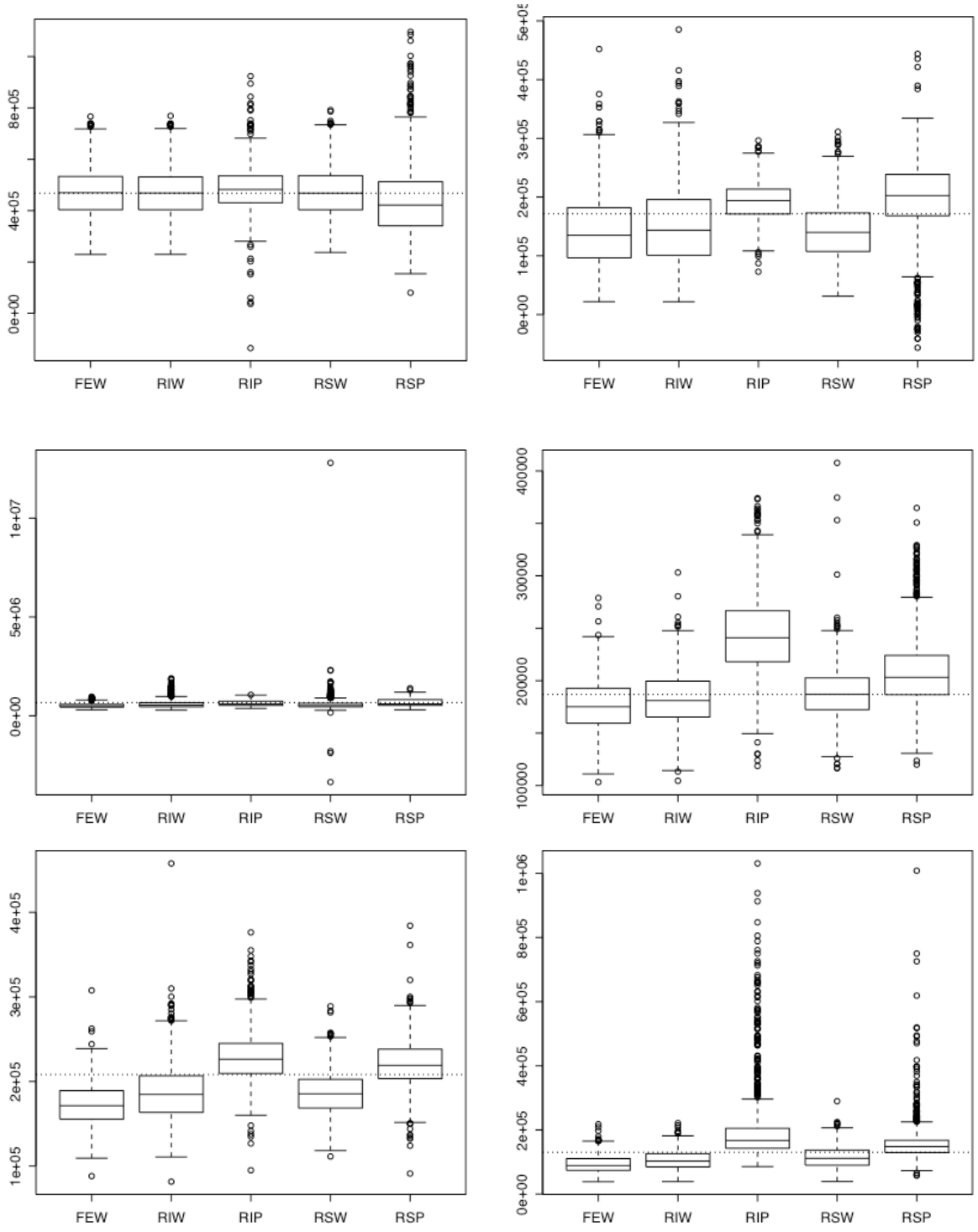


Figure 5 (continued) Distributions of estimates for regions 7 to 12 produced under the SizeZone*Area specification. FEW = Fixed Effect weighting, RIW = Random Intercept weighting, RIP = Random Intercept prediction, RSW = Random Slope weighting, RSP = Random Slope Prediction. Plots are ordered left to right and top down by increasing region population size. Dotted horizontal line is true region mean.

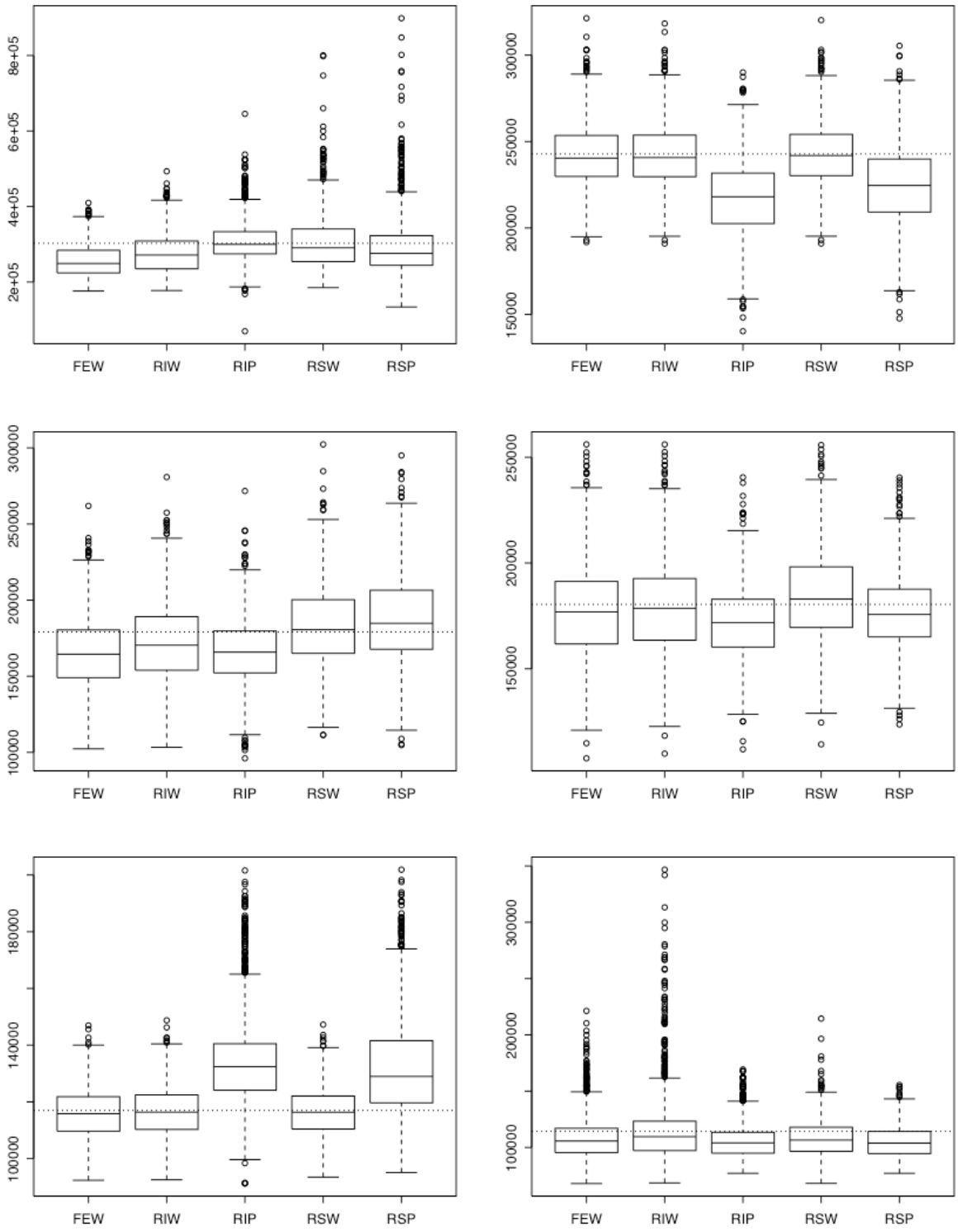


Figure 5 (continued) Distributions of estimates for regions 13 to 18 produced under the SizeZone*Area specification. FEW = Fixed Effect weighting, RIW = Random Intercept weighting, RIP = Random Intercept prediction, RSW = Random Slope weighting, RSP = Random Slope Prediction. Plots are ordered left to right and top down by increasing region population size. Dotted horizontal line is true region mean.

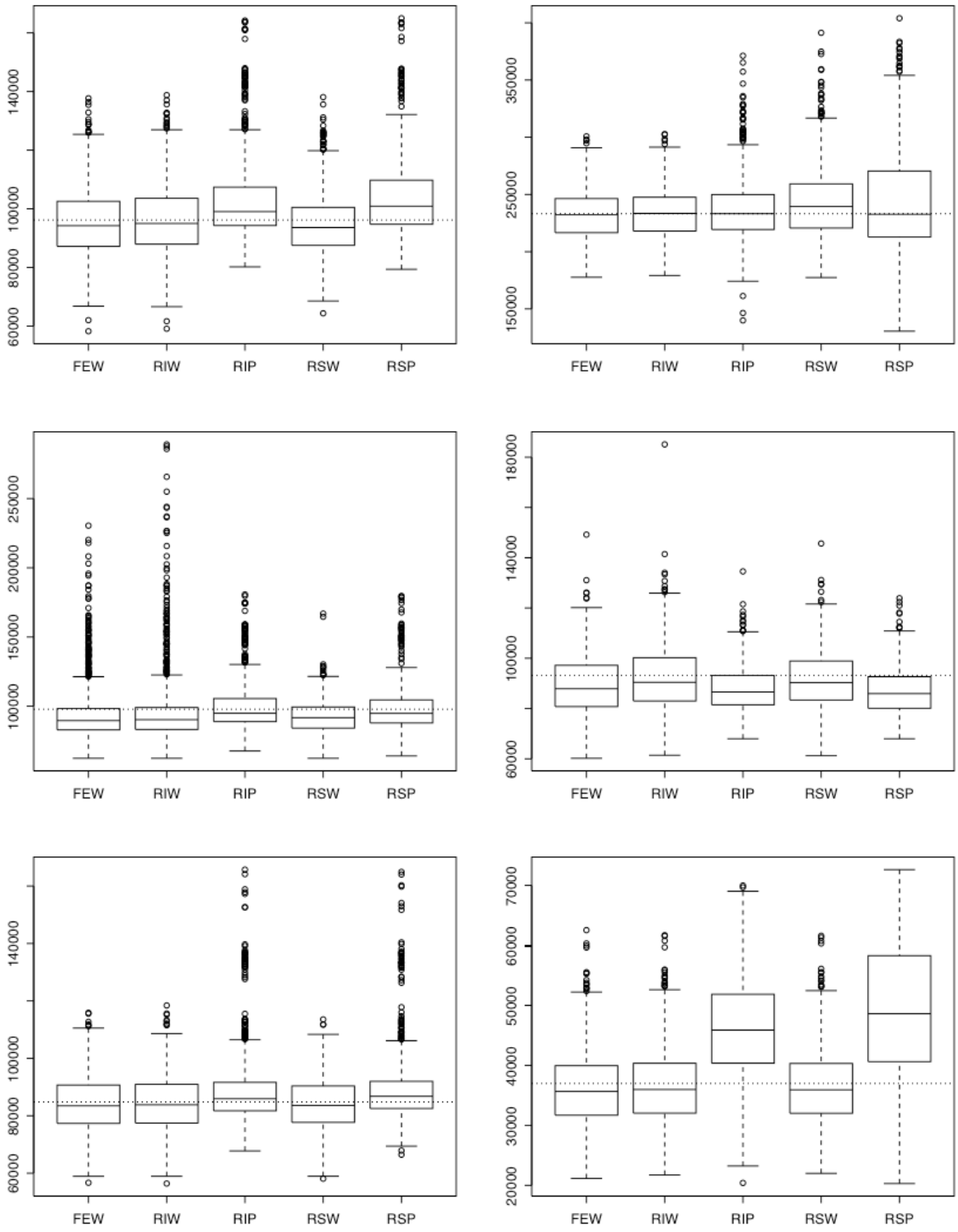


Figure 5 (continued) Distributions of estimates for regions 19 to 24 produced under the SizeZone*Area specification. FEW = Fixed Effect weighting, RIW = Random Intercept weighting, RIP = Random Intercept prediction, RSW = Random Slope weighting, RSP = Random Slope Prediction. Plots are ordered left to right and top down by increasing region population size. Dotted horizontal line is true region mean.

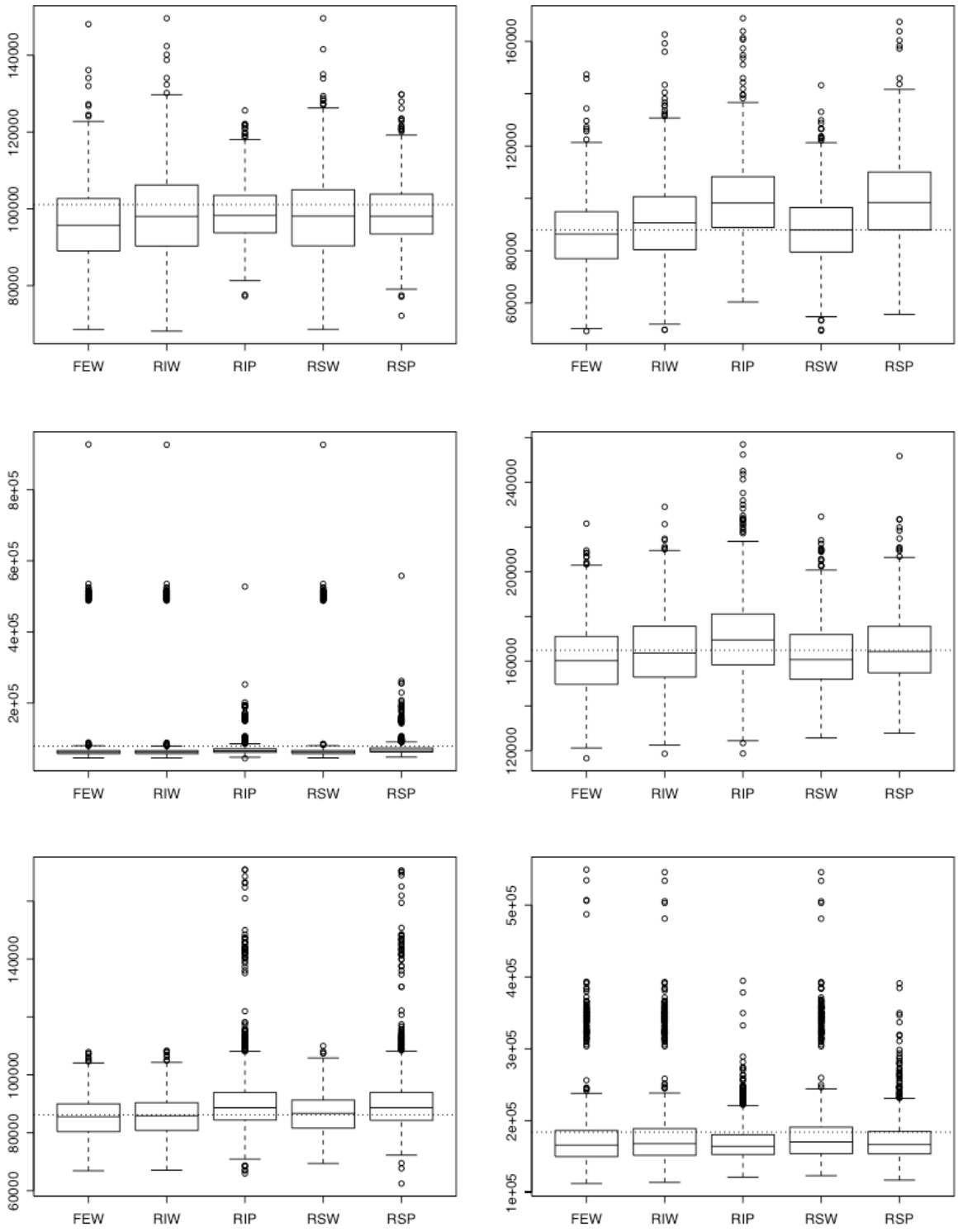


Figure 5 (continued) Distributions of estimates for regions 25 to 29 produced under the SizeZone*Area specification. FEW = Fixed Effect weighting, RIW = Random Intercept weighting, RIP = Random Intercept prediction, RSW = Random Slope weighting, RSP = Random Slope Prediction. Plots are ordered left to right and top down by increasing region population size. Dotted horizontal line is true region mean.

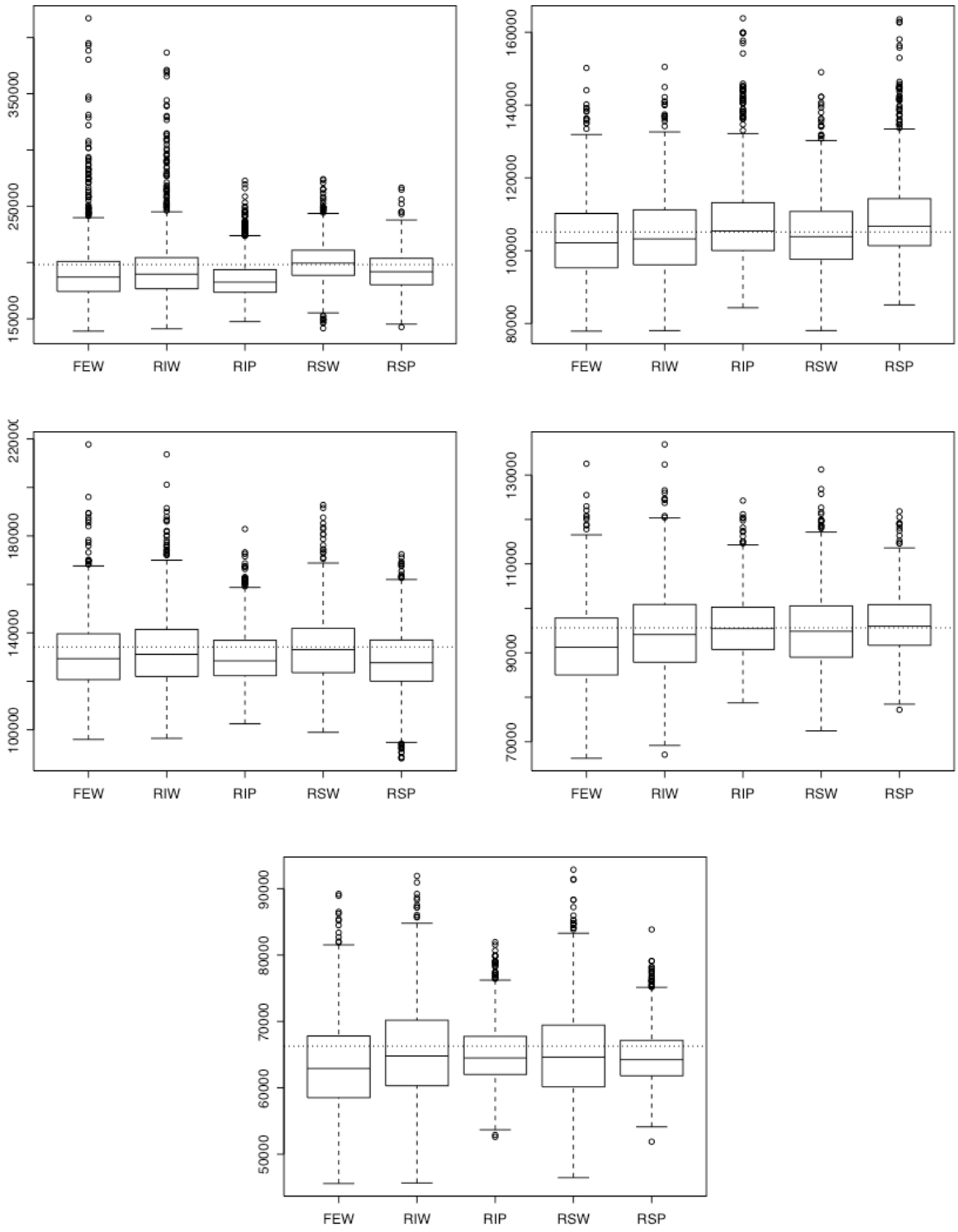


Figure 3 Distributions of estimates for region 21 produced under the SizeZone*Area specification when 37 samples that include the massive outlier in this region are excluded. FEW = Fixed Effect weighting, RIW = Random Intercept weighting, RIP = Random Intercept prediction, RSW = Random Slope weighting, RSP = Random Slope Prediction. Dotted horizontal line is region mean including outlier, dashed horizontal line is region mean excluding outlier.

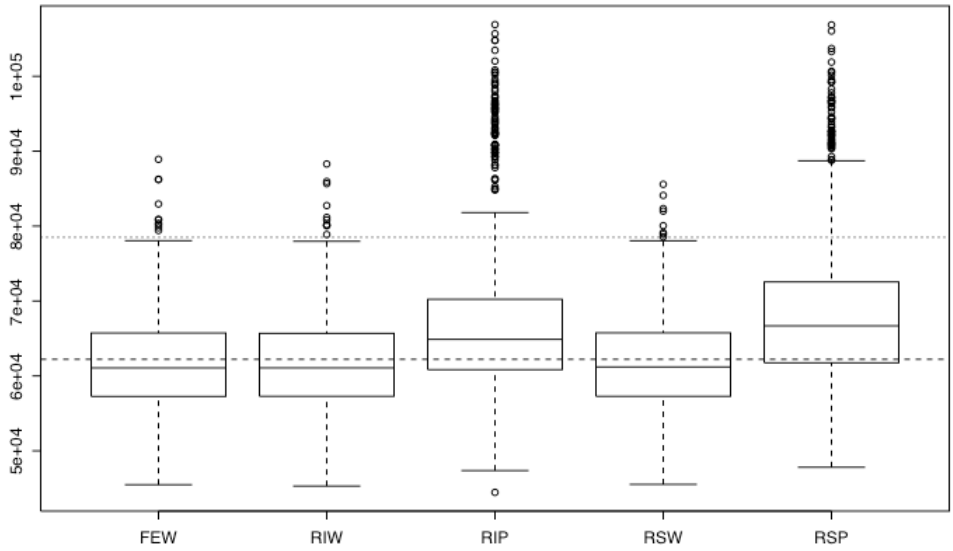
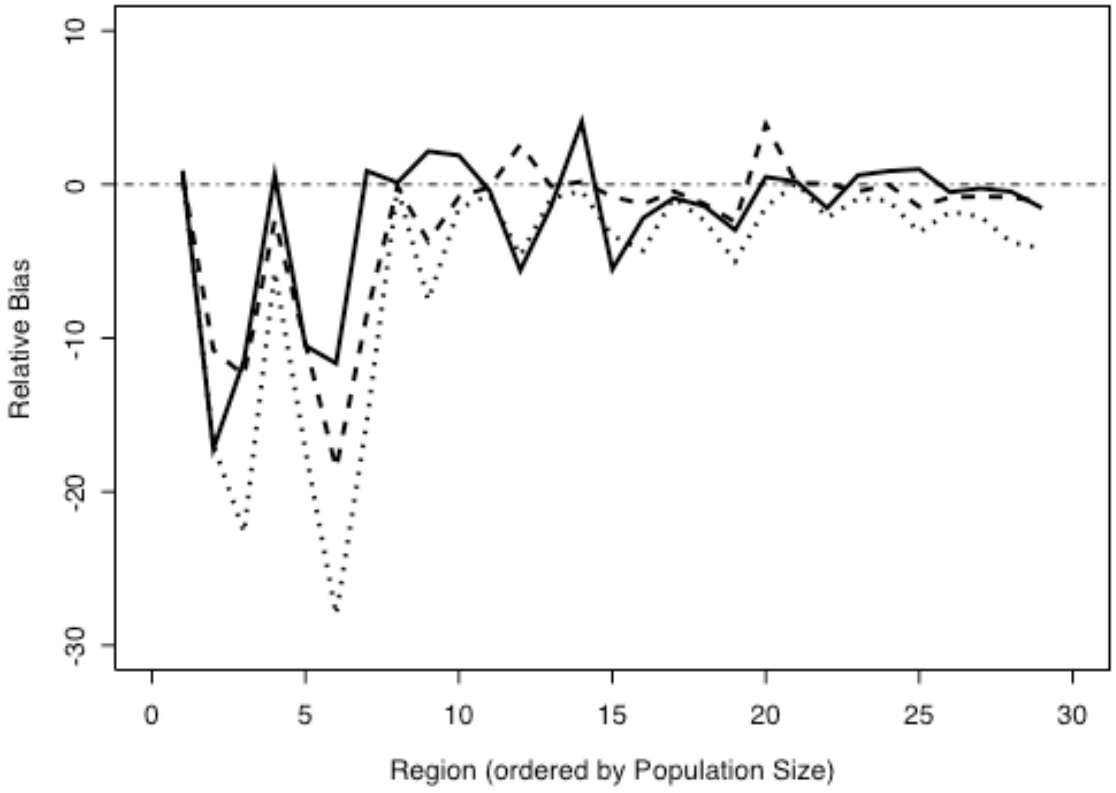


Figure 4 Relative biases and RMSEs by region for weighting based on SizeZone*Area specification. Dotted line is Fixed Effects model, dashed line is Random Intercepts model and solid line is Random Slopes model.



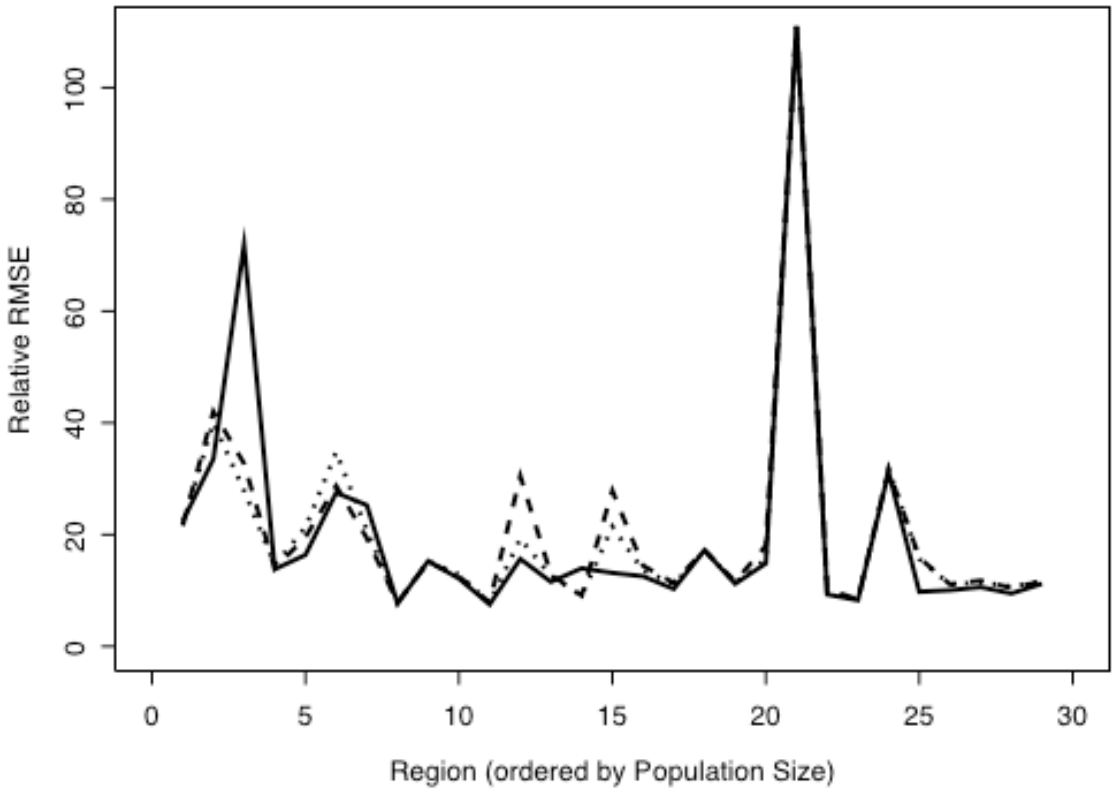
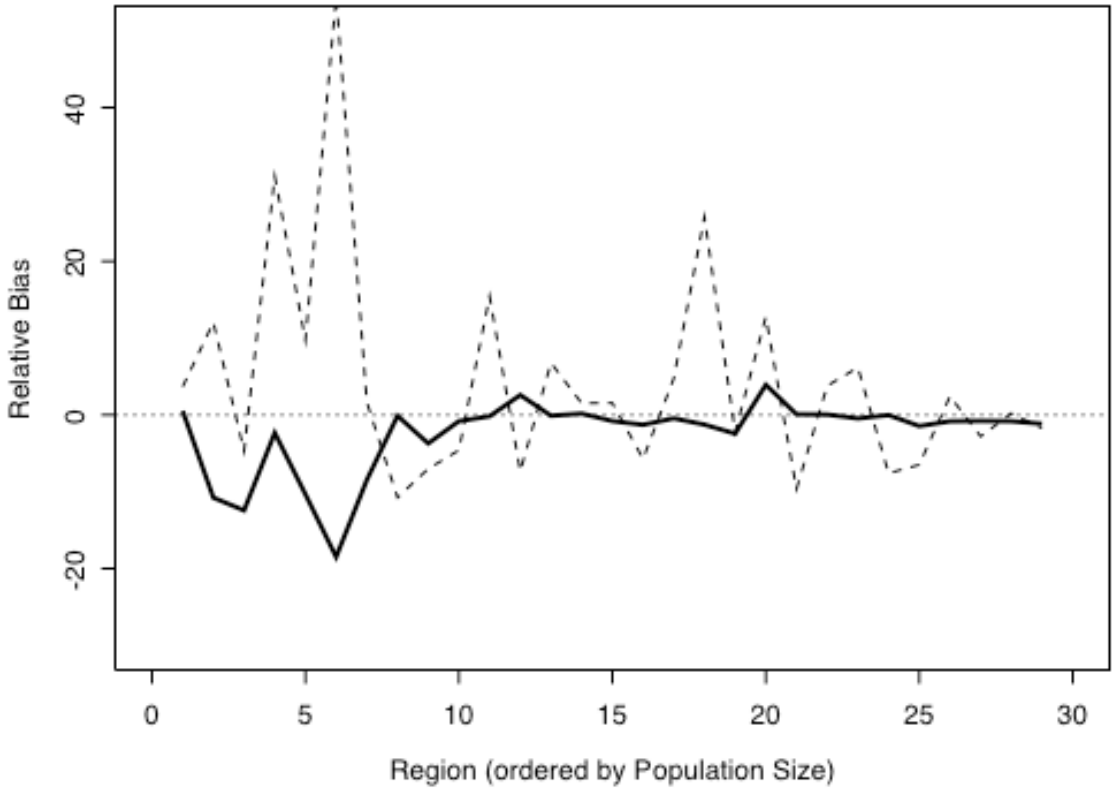


Figure 5 Relative biases and RMSEs by region for weighting based on SizeZone*Area specification. Dashed line is prediction estimation based on Random Intercepts model and solid line is weighting estimation based on the same model.



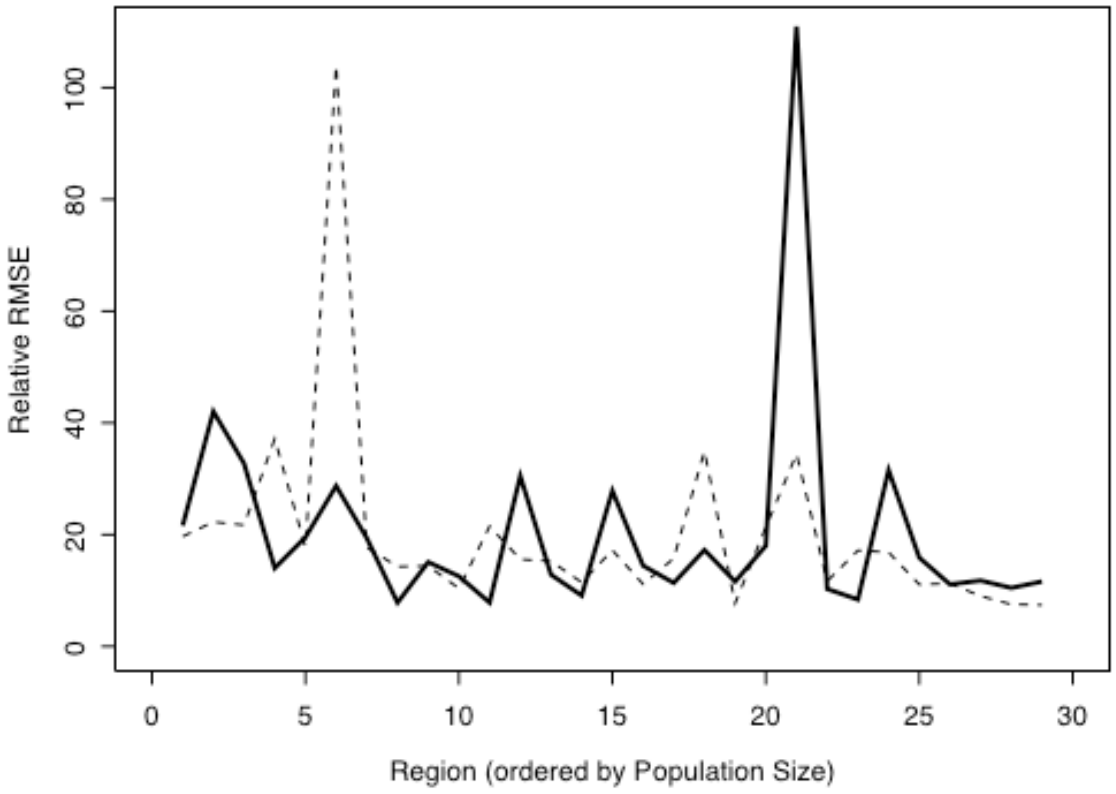


Figure 6 Relative biases and RMSEs by region for weighting based on SizeZone*Area specification. Dashed line is prediction estimation based on Random Slopes model and solid line is weighting estimation based on the same model.