Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 66(1) 2012 213-228

Small Area Estimation in Practice: An Application to Agricultural Business Survey Data

Nikos Tzavidis[1*], Ray Chambers[2], Nicola Salvati[3] and Hukum Chandra[4]
[1]*University of Southampton, United Kingdom*
[2]*University of Wollongong, Australia*
[3]*University of Pisa, Italy*
[4]*Indian Agricultural Statistics Research Institute, New Delhi, India*
Received 30 July 2011; Revised 12 September 2011; Accepted 12 September 2011

SUMMARY

This paper describes an application of small area estimation (SAE) to agricultural business survey data. Both well known small area estimators, such as the empirical best linear unbiased predictor (EBLUP), and more recently proposed small area estimators, for example, the M-quantile, the robust EBLUP and the Model Based Direct estimators are considered. Mean squared error estimation is discussed. Using a real agricultural business survey dataset, we place emphasis on model diagnostics for specifying the small area working model, on diagnostic measures for validating the reliability of direct and indirect (model-based) small area estimators and on providing practical guidelines to the prospective user of small area estimation techniques.

Keywords : Diagnostics, Direct estimator, Model-based estimation, Outlier robustness, MSE estimation.

1. INTRODUCTION

Sample surveys provide a cost-effective way of obtaining estimates for population characteristics of interest. On many occasions, however, the interest is in estimating parameters for domains that contain only a small number of data points. The term 'small areas' is used to describe domains whose sample sizes are not large enough to allow sufficiently precise direct estimation. When direct estimation is not possible, one has to rely on alternative, model-based methods for producing small area estimates. Such methods depend on the availability of population level auxiliary information related to the variable of interest, and are commonly referred to as indirect methods (Ghosh and Rao 1994, Rao 2003, Pfefferman 2002). The industry standard for small area estimation is to use unit or area level models that include random area effects to account for between area variation beyond that explained by the auxiliary information (Fay and Herriot 1979, Battese *et al*. 1988) and in this paper we solely focus on unit level small area models.

In recent years there has been a number of developments in the small area estimation literature. This involves both extensions of the conventional random effects small area model and the estimation of parameters other than averages and totals for example, quantities of the small area distribution function of the outcome of interest (Tzavidis *et al*. 2010) and complex indicators (Molina and Rao 2010, Marchetti *et al*. 2012). One research direction has focused on nonparametric versions of the random effects model (Opsomer *et al*. 2008) while a further research area that has attracted interest is in the specification of models that borrow strength over space either by specifying models with spatially correlated or nonstationary random effects (Salvati *et al*. 2011, Chandra *et al*. 2012). The issue of outlier robust small area estimation

* *Corresponding author :* Nikos Tzavidis
E-mail address : n.tzavidis@soton.ac.uk

has also attracted a fair amount of interest mainly due to the fact that in many real data applications the Gaussian assumptions of the conventional random effects model are not satisfied. Two main approaches to outlier robust small area estimation have been proposed. The first one is based on M-estimation of the until level random effects model (Sinha and Rao 2009) while the second is based on the use of an M-quantile model under which area effects are estimated using a semi-parametric approach (Chambers and Tzavidis 2006, Tzavidis *et al*. 2010, Chambers *et al*. 2009). The aim of this paper is to offer to the prospective user of small area estimation practical guidance on the implementation of a range of small area estimation methodologies. We illustrate how recently proposed small area methods can be used with a real agricultural business survey dataset, what model diagnostics can be used for specifying a small area working model and what diagnostics can be used for assessing the validity of small area estimates.

The paper is organised as follows. In Section 2 we review unit-level models for small area estimation. In particular, we review the Empirical Best Linear Unbiased Predictor (Rao 2003), and we then present the model-based direct estimator (MBDE) (Chandra and Chambers 2009) and a range of outlier robust small area estimators using either an outlier robust version of the unit level random effects model or an M-quantile small area model. In the same Section we further discuss Mean Squared Error (MSE) using both analytic and bootstrap-based approaches. Using real survey data, in Section 3 we present model diagnostics that assist us to build a model for performing small area estimation. In Section 4 we present the results from small area estimation and diagnostics for evaluating the reliability of the small area estimates. Finally, in Section 5 we provide some concluding remarks and inform the prospective user about the availability of software for implementing the estimation procedures.

2. UNIT-LEVEL MODELS FOR SMALL AREA ESTIMATION

In this Section we assume that unit level data are available at small area level. For the sampled units in the population this consists of small area identifiers, values of the outcome variable y_j , values $\mathbf{x}_j$ of a $p \times 1$ vector of individual level covariates, and values $\mathbf{z}_j$ of a vector of area level covariates. For the non-sampled population units we lack information about y_j but it is assumed that all areas are sampled and that we know the number of units in each small area and the corresponding small area averages of $\mathbf{x}_j$ and $\mathbf{z}_j$.

A popular approach to small area estimation is to assume a linear mixed model, with random area effects (see Rao 2003). Let $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{Z}$ denote the population level vector and matrices defined by y_j, $\mathbf{x}_j$ and $\mathbf{z}_j$ respectively,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (1)$$

where $\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_m^T)^T$ is a vector of random area effects with $\mathbf{u} \sim N(\mathbf{0}, \Sigma_u)$ and $\mathbf{e} \sim N(\mathbf{0}, \Sigma_e)$ is a vector of unit level random effects. It is also assumed that $\mathbf{u}$ is distributed independently of $\mathbf{e}$. We assume that the covariance matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ are defined in terms of a lower dimensional set of parameters $\theta = (\theta_1, \ldots, \theta_K)$, which are typically referred to as the variance components of (1), while $\boldsymbol{\beta}$ is usually referred to as its fixed effect. The covariance matrix of the vector $\mathbf{y}$ is given by $Var(\mathbf{y}) = \mathbf{V}$.

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ denote estimates of the fixed and random effects in (1). The EBLUP of the area i mean of the y_j under (1) is then

$$\hat{\bar{y}}_i^{EBLUP} = N_i^{-1}\left\{n_i \bar{y}_{si} + (N_i - n_i)\left(\bar{\mathbf{x}}_{ri}^T\hat{\boldsymbol{\beta}} + \bar{\mathbf{z}}_{ri}^T\hat{\mathbf{u}}\right)\right\}, \quad (2)$$

where $\hat{\mathbf{u}} = \left(\hat{\mathbf{u}}_1^T, \ldots, \hat{\mathbf{u}}_m^T\right)^T$ denotes the vector of the estimated area specific random effects, we use indices of s and r to denote sample and non-sample quantities, respectively, and $\bar{\mathbf{x}}_{ri}$ and $\bar{\mathbf{z}}_{ri}$ denote the vectors of average values of $\mathbf{x}_j$ and $\mathbf{z}_j$ respectively for the $N_i - n_i$ non-sampled units in the corresponding area.

Unbiased direct estimators for small area quantities are usually considered to be too variable to be of any practical use. Chandra and Chambers (2009) describe a class of model-based direct estimators (MBDE) for small area quantities that appear to overcome this limitation in the sense that these estimators are comparable in efficiency to indirect model-based small area estimators such as the EBLUP that is now widely used. Throughout this paper we assume that the sampling method used is uninformative for the population values of y given the corresponding values of the auxiliary variables and knowledge of the area affiliations of the population units. As a consequence,

model (1) represents our model for both sampled and non-sampled population units. It follows that we can partition **y**, **X**, **Z**, **e** into components defined by the n sampled and $N - n$ non-sampled population units. For example, $\mathbf{X}_s$ represents the matrix defined by the n sample values of the auxiliary variable vector, while $\mathbf{V}_{rr}$ is the matrix of covariances of the response variable among the $N - n$ non-sampled units. With this sample and non-sample partition, under the population level linear mixed model (1), the sample weights that define the EBLUP for the population total of y are

$$\mathbf{w}_{EBLUP} = (w_{j,EBLUP}) = \mathbf{1}_s + \hat{\mathbf{H}}'\left(\mathbf{X}^T\mathbf{1}_N - \mathbf{X}_s^T\mathbf{1}_s\right) + \left(\mathbf{I}_s - \hat{\mathbf{H}}^T\mathbf{X}_s^T\right)\hat{\mathbf{V}}_{ss}^{-1}\hat{\mathbf{V}}_{sr}\mathbf{1}_r \quad (3)$$

where $\hat{\mathbf{H}} = \left(\sum_i \mathbf{X}_{is}^T\hat{\mathbf{V}}_{iss}^{-1}\mathbf{X}_{is}\right)^{-1}\left(\sum_i \mathbf{X}_{is}^T\hat{\mathbf{V}}_{iss}^{-1}\right)$. See Royall (1976). The model-based direct estimator of the i^{th} small area mean is then defined as

$$\hat{\bar{y}}_i^{MBD} = \sum_{j\in s_i} w_{j,EBLUP}\, y_j \Big/ \sum_{j\in s_i} w_{j,EBLUP}. \quad (4)$$

MSE estimation of the MBDE estimator (4) is carried out by pseudo-linear MSE estimation approach described in Chandra and Chambers (2009). There are many practical advantages associated with the use of MBDE arising mainly from the fact that this estimator is computed as weighted linear combinations of the actual sample data from the small areas of interest. Perhaps the most important advantage is the simplicity of both the point and the MSE estimation. Furthermore, the MBDE estimator is easy to interpret and to build into a survey processing system. In contrast to design-based direct estimators, MBDE "borrows strength" from other areas via the linear mixed model used in defining the sample weights. That is, the MBDE of a small area mean improves upon the efficiency of the design-based direct estimator by using weights that define the EBLUP for the population total (see Royall 1976) under the same linear mixed model with random area effects that underpins the EBLUP for the small area mean.

A topic that has attracted interest in more recent small area applications is that of outlier robust small area estimation. The need for outlier robust methodologies has arisen mainly due to the fact that many applications of small area estimation involve business and economic survey data making the normality assumptions of (1) hard to satisfy. One approach for making estimator (2) insensitive to sample outliers is by replacing $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ with corresponding outlier robust quantities. In particular, denoting by ψ a bounded influence, Sinha and Rao (2009) discuss outlier robust small area estimation that is based on a robust version of (1). The Sinha and Rao (2009) robust alternative to (2) is then

$$\hat{\bar{y}}_i^{REBLUP} = N_i^{-1}\left\{n_i\bar{y}_{si} + \left(N_i - n_i\right)\left(\bar{\mathbf{x}}_{ri}^T\hat{\boldsymbol{\beta}}^{\psi} + \bar{z}_{ri}^T\hat{\mathbf{u}}^{\psi}\right)\right\}, \quad (5)$$

where the REBLUP stands for Robust EBLUP. An entirely different way of thinking about small area estimation and in particular about outlier robust SAE is the M-quantile regression-based method proposed by Chambers and Tzavidis (2006). This is based on a linear model for the Mquantile regression of **y** on **X**, i.e.

$$m_q(\mathbf{X}) = \mathbf{X}\hat{\boldsymbol{\beta}}_q^{\psi}, \quad (6)$$

where $m_q(\mathbf{X})$ denotes the M-quantile of order q of the conditional distribution of **y** given **X**. An estimate $\hat{\boldsymbol{\beta}}_q^{\psi}$ of $\boldsymbol{\beta}_q^{\psi}$ can be calculated for any value of q in the interval (0, 1), and for each unit in sample we define its unique M-quantile coefficient under this fitted model as the value q_j such that $y_j = \mathbf{x}_j^T\hat{\boldsymbol{\beta}}_{q_j}^{\psi}$, with the sample average of these coefficients in area i denoted by $\bar{q}_i$. The M-quantile estimate of the mean of y_j in area i is then

$$\hat{\bar{y}}_i^{MQ} = N_i^{-1}\left\{n_i\bar{y}_{si} + \left(N_i - n_i\right)\bar{\mathbf{x}}_{ri}^T\hat{\boldsymbol{\beta}}_{\bar{q}_i}^{\psi}\right\}. \quad (7)$$

Note that the regression M-quantile model (6) depends on the influence function ψ underpinning the M-quantile. When this function is bounded, sample outliers have limited impact on $\hat{\boldsymbol{\beta}}_q^{\psi}$. That is, (7) corresponds to assuming that all non-sample units in area i follow the working model (6) with $q = \bar{q}_i$, in the sense that one can write $y_j = \mathbf{x}_j^T\boldsymbol{\beta}_{\bar{q}_i}^{\psi}$ + noise for all such units.

A problem with the Robust EBLUP and M-quantile small area estimation approaches is the assumption that all non-sampled units follow the

working model that is, any deviations from this model cancel out on average. Methods for dealing with this assumption was developed by Chambers *et al*. (2009) leading to the so-called robust predictive approach. Under the linear mixed model (1) one can see that provided the individual errors of the non-sampled units are symmetrically distributed about zero, the REBLUP estimator (5) will perform well since it is based on the implicit assumption that the average of these errors over the non-sampled units in area *i* converges to zero. The M-quantile estimator (7) is no different since it assumes that the errors from the area specific M-quantile fit are 'noise' and hence also cancel out on average. Note that this does not mean that these non-sample units are not outliers. It is just that their behaviour is such that our best prediction of their corresponding average value is zero.

Starting with a working linear model linking the population values of y_j and $\mathbf{x}_j$, and sample data containing representative outliers with respect to this model, Welsh and Ronchetti (1998) extend the approach of Chambers (1986) to robust prediction of the empirical distribution function of the population values of y_j. Their argument immediately applies to robust prediction of the empirical distribution function of the area *i* values of y_j, and leads to a predictor of the form

$$\hat{F}_i^{\psi\phi}(t) = N_i^{-1}\left[\sum_{j\in s_i} I(y_j \le t) + n_i^{-1}\sum_{j\in s_i}\sum_{k\in r_i} I\left(\mathbf{x}_k^T\hat{\boldsymbol{\beta}}^{\psi} + \omega_{ij}^{\psi}\phi\left\{(y_j - \mathbf{x}_j^T\hat{\boldsymbol{\beta}}^{\psi})\big/\omega_{ij}^{\psi}\right\} \le t\right)\right]. \tag{8}$$

Here ω_{ij}^{ψ} is a robust estimator of the scale of the residual $y_j - \mathbf{x}_j^T\hat{\boldsymbol{\beta}}^{\psi}$ in area *i* and ϕ denotes a bounded influence function that satisfies $|\phi| \ge |\psi|$. Tzavidis *et al*. (2010) note that the robust estimator of the area *i* mean of the y_j defined by (8) is the expected value of the functional defined by it, which is

$$\hat{\bar{y}}_i^{\psi\phi} = \int t d\hat{F}_i^{\psi\phi}(t) = N_i^{-1}\left[n_i\bar{y}_{si} + (N_i - n_i)\left(\bar{\mathbf{x}}_{ri}^T\hat{\boldsymbol{\beta}}^{\psi} + n_i^{-1}\sum_{j\in s_i}\omega_{ij}^{\psi}\phi\left\{\left(y_j - \mathbf{x}_j^T\hat{\boldsymbol{\beta}}^{\psi}\right)\big/\omega_{ij}^{\psi}\right\}\right)\right]. \tag{9}$$

These authors therefore suggest an extension to the M-quantile estimator (7) by replacing $\hat{\boldsymbol{\beta}}^{\psi}$ in (8) by $\hat{\boldsymbol{\beta}}_{\bar{q}_i}$, which leads to a 'bias-corrected' version of (7), given by

$$\hat{\bar{y}}_i^{MQ-BC} = N_i^{-1}\left[n_i\bar{y}_{si} + (N_i - n_i)\left(\bar{\mathbf{x}}_{ri}^T\hat{\boldsymbol{\beta}}_{\bar{q}_i} + n_i^{-1}\sum_{j\in s_i}\omega_{ij}^{MQ}\phi\left\{\left(y_j - \mathbf{x}_j^T\hat{\boldsymbol{\beta}}_{\bar{q}_i}\right)\big/\omega_{ij}^{MQ}\right\}\right)\right], \tag{10}$$

and ω_{ij}^{MQ} is a robust estimator of the scale of the residual $y_j - \mathbf{x}_j^T\hat{\boldsymbol{\beta}}_{\bar{q}_i}$ in area *i*. If ϕ is an identity function then estimator (10) becomes a Chambers and Dunstan estimator (Tzavidis *et al*. 2010 - CD hereafter).

A similar argument can be used to modify the REBLUP estimator (5). In particular, a Robust Predictive version of this estimator, hereafter REBLUP-BC leads to

$$\hat{\bar{y}}_i^{REBLUP-BC} = \hat{\bar{y}}_i^{REBLUP} + \left(1 - n_i N_i^{-1}\right)n_i^{-1}\sum_{j\in s_i}\omega_{ij}^{\psi}\phi\left\{\left(y_j - \mathbf{x}_j^T\hat{\boldsymbol{\beta}}^{\psi} - \mathbf{z}_j^T\hat{\mathbf{u}}^{\psi}\right)\big/\omega_{ij}^{\psi}\right\}, \tag{11}$$

where the ω_{ij}^{ψ} are now robust estimates of the scale of the area *i* residuals $y_j - \mathbf{x}_j^T\hat{\boldsymbol{\beta}}^{\psi} - \mathbf{z}_j^T\hat{\mathbf{u}}^{\psi}$.

Estimating the mean squared error (MSE) of different small area estimators is both an important and challenging task. Starting with the EBLUP, the most popular analytic MSE estimator is the one proposed Prasad and Rao (1990). See also Rao (2003) for details on this estimator. An alternative, bias-robust approach to analytic MSE estimation that is based on an extension of the ideas by Royall and Cumberland (1978) has been proposed by Chambers *et al*. (2011). An appealing feature of this alternative MSE estimator is that it can be used with a wide range of small area predictors as long as they can be expressed as weighted sums of the sample values. Chambers *et al*. (2011) use the bias-robust MSE estimator for computing the MSEs of the EBLUP estimator (2), MBDE (4) and the M-quantile estimator (7). The overwhelming evidence from the results of this paper indicates that the bias-

robust MSE estimator comes at the price of higher variability and should not be used when the area-specific sample sizes are very small. On the other hand, if there is reasonable doubt about the validity of the assumptions of the linear mixed model (1), the bias-robust MSE estimator can be more efficient than alternative MSE estimators. More recently, the bias-robust MSE estimator has been used for estimating MSEs of the REBLUP estimator (5) and of the bias-corrected M-quantile and REBLUP estimators (10) and (11) (Chambers *et al*. 2009).

Analytic estimators offer only one approach to MSE estimation. As the complexity of small area models increases we tend to rely more often on computer intensive approaches to MSE estimation based for example on bootstrap. Sinha and Rao (2009) proposed the use of a parametric bootstrap MSE estimator for the REBLUP predictor that is similar in spirit to the bootstrap proposed by Hall and Maiti (2006). Tzavidis *et al*. (2010) proposed a non-parametric bootstrap MSE estimator for the M-quantile predictors (7) and (10). Marchetti *et al*. (2011) studied the properties of the nonparametric bootstrap MSE estimator and found it to be more stable than the analytic bias-robust MSE estimator that was proposed by Chambers *et al*. (2011).

3. THE DATASET

The data used in the analysis reported in this paper come from a sample of 1,652 Australian broadacre farms that participated in the annual Australian Agricultural and Grazing Industries Survey (AAGIS) organised by the Australian Bureau of Agricultural and Resource Economics in the late 1980s. Australian broadacre farms are spread across 29 regions of Australia. Sample sizes within these regions varied from a low of 6 to a high of 117. See Table 1 for distribution of regional sample sizes. These regions are the small areas of interest. Fig. 1 shows where these 29 farming regions (or small areas) and zones are located with the numbers shown in the map corresponding to region codes. The *Y*-variable of interest in this analysis is the

Table 1. Distribution of regional sample sizes.

Min	25%	50%	Mean	75%	Max	Total number of regions
6	32	55	57	79.5	117	29

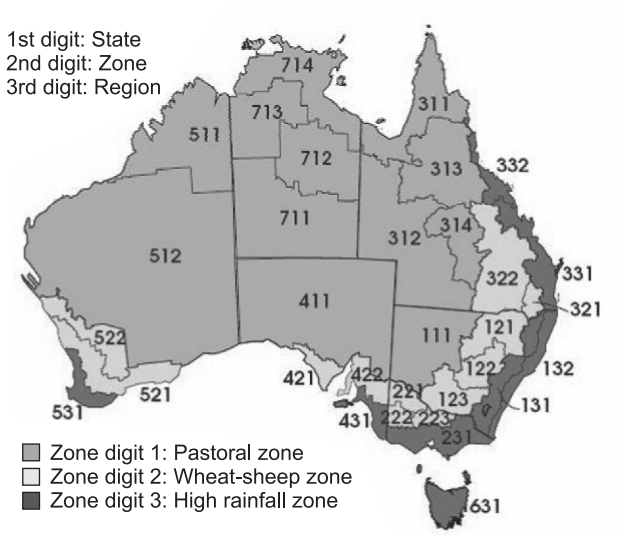


Fig. 1. Australian broadacre zones and farming regions.

Total Cash Costs (TCC) of the farm business in the reference year. An auxiliary size variable (Area) is available for each farm and is defined as the total area of the farm in hectares. The target is to estimate the average TCC in each small area.

3.1 Exploring the AAGIS Data

Before implementing small area estimation we must find a working model that we can use for this purpose. To start with, the overall linear relationship between TCC and Area is rather weak, however, this improves when separate linear models are fitted within the six strata defined by the interaction between area and region as shown in Table 2 and in Fig. 3. In particular, these six strata are defined by splitting each zone into small farms (farm area less than zone median) and large farms (farm area greater than or equal to zone median). The six Size by Zone Strata are then defined as follows: 1 = Pastoral zone and area of 50,000 hectares or less; 2 = Pastoral zone and area of more than

Table 2. Results showing the significance of the main and interaction effects.

Source	Parameters	DF	Sum of Squares	F Ratio	Prob > F
Area	1	1	6.262e+11	5.3970	0.0203
Stratum	5	5	5.625e+13	96.9683	<.0001
Area* Stratum	5	5	6.866e+13	118.3618	<.0001

50,000 hectares; 3 = Wheat-sheep zone and area of 1,500 hectares or less; 4 = Wheatsheep zone and area of more than 1,500 hectares; 5 = High rainfall zone and area of 750 hectares or less; 6 = High rainfall zone and area of more than 750 hectares. In Fig. 2 (top plot) we also notice the presence of two outlier data points. The linear relationship between TCC and farm size improves when these two points are excluded from the analysis. In particular, the values of R^2 (and root mean square error) increase (and decrease) from 0.05 (and 970358.4) to 0.236 (and 410043.8) (see also Fig. 2). The fit of the model between TCC and farm size within the six post-strata when the two outlier data points are excluded is shown in Fig. 3.

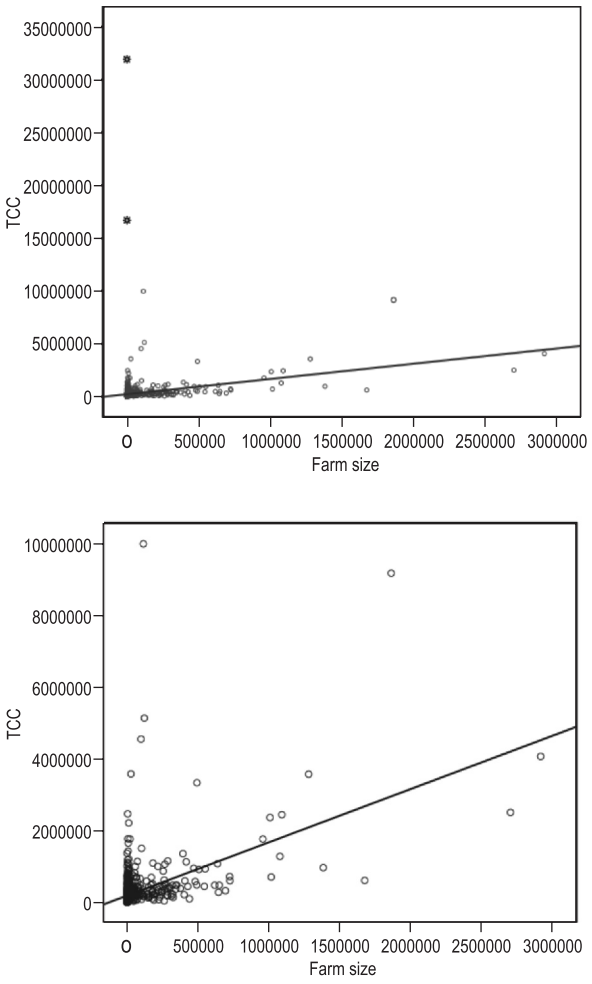


Fig. 2. Relationship between total cash costs (TTC) and farm sizes in AAGIS sample. The top plot shows the model fit with the two outlying data points included. The bottom plot shows the model fit after the two outlying data points have been removed.

3.2 Specifications of the Mixed and M-quantile Small Area Models

Many of the best known small area estimators are based on the unit level mixed model with area random effects. For specifying the fixed part of the model we used the model fitting results we reported in the previous Section. Hence, in the fixed part of the mixed model (1), the design matrix **X** of auxiliary variable values is defined to include an effect for Area, effects for the Size by Zone strata and effects for interactions between Area and the Size by Zone strata (see Fig. 3).

In the random part of the model (1) we need to test for the presence of random intercepts or random slopes which corresponds to comparing two different specifications for **Z** in (1). An empirical approach for testing for random intercepts or random slopes is by modelling the residuals from the fixed part of the model (see Section 3.1) as a function of the 29 regions. The results from this regression model indicated that there is a significant region effect, which indicates the presence of random intercepts. We then regressed the residuals from the fixed part of the model as a function of region, area and an interaction between area and region. Adding this interaction term between improved the fit of the model. Hence, the exploratory data analysis indicated that specifying a random slope on area should provide a better model fit. The presence of random intercepts or random slopes can be more formally tested by using the Akaike information criterion (AIC). The smaller the value of AIC implies a better model fit. In our test result the *p*-value of the random slopes model was slightly smaller than the random intercepts model. Fig. 4 shows normal probability plots of level 1 and level 2 residuals from the random intercepts model. It appears that the normality assumptions of the mixed model do not hold and hence we must also consider outlier robust small area estimators.

Until now we have only considered the specification of the mixed model. However, a number of recently proposed small area estimators are based on the M-quantile model. In the case of the M-quantile model the design matrix **X** of auxiliary variable values is just the design matrix for the fixed part of the mixed effect model (i.e. Area*SizeZone stratum). In this case we need to remember that there is no formal specification of a random part. Instead of specifying

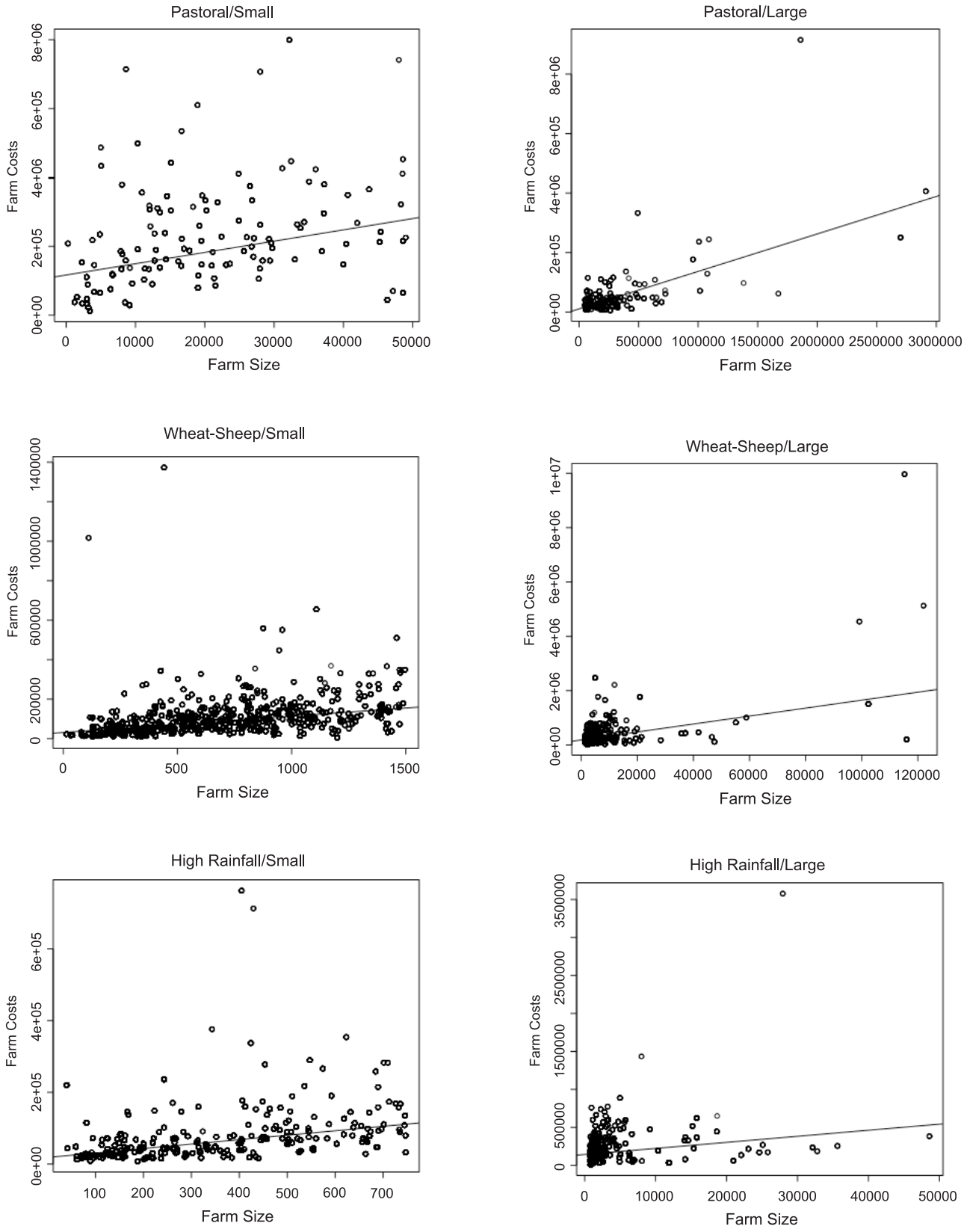


Fig. 3. Relationship between total cash costs and farm sizes in six post-strata when the two outliers are excluded from the AAGIS sample.

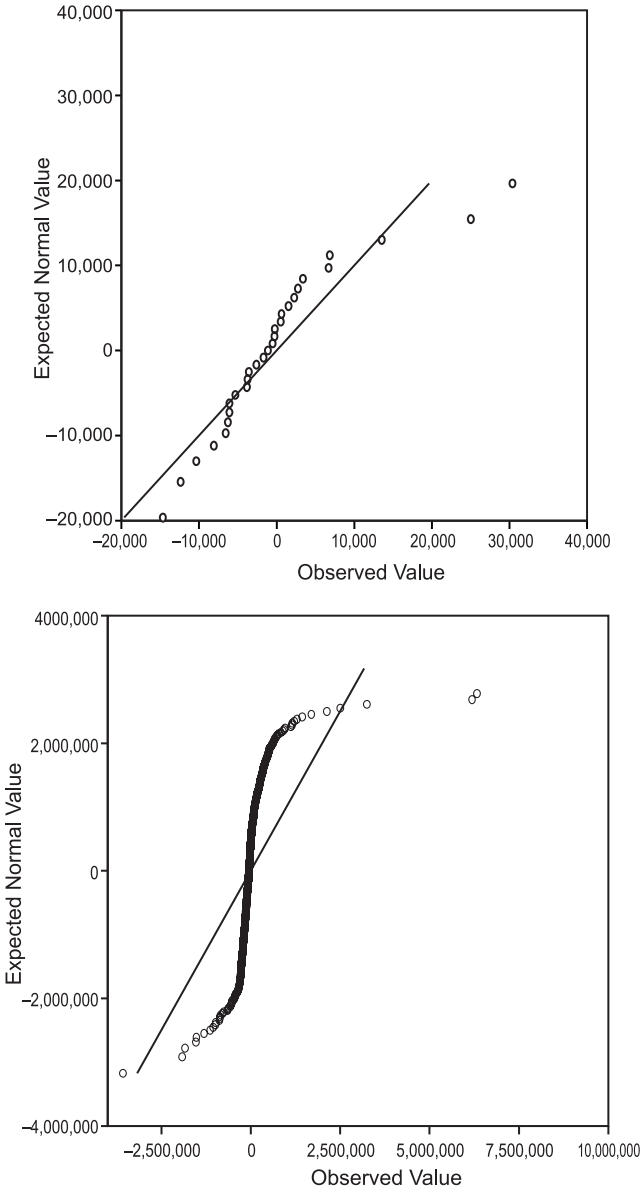


Fig. 4. Normal probability plots of level 1 (top) and level 2 (bottom) residuals from the random intercepts model.

random effects (i.e. random intercepts or random slopes) the between area variability is captured via the area specific M-quantile coefficients, which are estimated by using an empirical (moment-type) estimator.

4. APPLICATION

As we mentioned in Section 1 the purpose of this article is to demonstrate the application of different small area methods in practice. We started by presenting some approaches for specifying the working small area model. In what follows we use this working model for obtaining small area estimates and corresponding estimates of precision taking into account the characteristics of the sample data. For performing small area estimation we use the original AAGIS sample data described in Section 3 and we generate a synthetic population dataset. From this population we then draw one sample. We then assume that this is the real population and we carry out small area estimation using the single sample we selected from the population.

4.1 Setting the Scene

A population of N = 81982 farms generated by bootstrapping the original AAGIS sample. That is, the 1652 farms in the original AAGIS sample were themselves sampled with replacement 81982 times using selection probabilities proportional to a farm's AAGIS sample weight. An independent sample of n = 1652 farms taken from this population using stratified random sampling, with regions defining the strata and with strata sample allocations equal to those in the original AAGIS sample. Using the selected sample, we replicated the exploratory analyses of Section 3. The conclusions from the exploratory data analyses with the randomly selected sample from the synthetic population were the same as the ones we reported in Section 3 with original AAGIS sample. Consequently, the model specifications described in Section 3.2 were also valid in this case.

4.2 Computing and Evaluating the Small Area (Region) Estimates

Using the various small area estimation methods discussed in Section 2 and with the sample data and population information described in Section 4.1, we calculated small area estimates using a wide range of small area point and MSE estimators. In particular, regional estimates for average TCC and their respective mean squared error estimates were calculated using the statistical software R and the following estimators: (1) the direct estimator (regional sample mean), (2) the EBLUP based on random intercepts model (RI/EBLUP), (3) the EBLUP based on random slopes model (RS/EBLUP), (4) the model-based direct estimator based on the random intercepts model (RI/MBDE), (5) the model-based direct estimator based on the random slopes model (RS/MBDE), (6) the M-quantile naive estimator (MQ/Naive), (7) the M-quantile CD estimator (MQ/CD), (8) the M-quantile bias corrected estimator (MQ/BC), (9) the robust

Table 3. Correlations between the alternative regional estimates.

	RI/EBLUP	RI/MBDE	RS/EBLUP	RS/MBDE	MQ/Naive	MQ/CD	MQ/BC	REBLUP	REBLUP/BC
Direct	0.955	0.952	0.872	0.915	0.936	0.966	0.923	0.889	0.894
RI/EBLUP		0.978	0.958	0.964	0.973	0.981	0.971	0.968	0.974
RI/MBDE			0.951	0.988	0.980	0.989	0.983	0.940	0.956
RS/EBLUP				0.944	0.952	0.964	0.977	0.940	0.969
RS/MBDE					0.978	0.966	0.978	0.935	0.950
MQ-/Naive						0.978	0.992	0.946	0.956
MQ-/CD							0.986	0.932	0.954
MQ-/BC								0.946	0.967
REBLUP									0.991

Table 4. Goodness of Fit diagnostic. A smaller value (less than 42.5 in this case) indicates no statistically significant difference between model-based and direct estimates.

Model-based estimator	Goodness of Fit[1]
RI/EBLUP	12.60
RI/MBDE	3.36
RS/EBLUP	33.05
RS/MBDE	5.41
MQ/Naive	31.90
MQ/CD	9.13
MQ/BC	25.12
REBLUP	131.82
REBLUP/BC	132.10

EBLUP estimator based on the random intercepts model (REBLUP) and (10) the robust bias corrected EBLUP based on random intercepts model (REBLUP/BC). MSE estimation is performed using analytic and bootstrap (parametric and nonparametric) estimators (see Section 2).

Point estimates for each region are reported in Table 5. Region specific precision estimates are reported in Table 6 and region specific coefficients of variation (CV) are reported in Table 7. For assessing the different estimators we must use a set of diagnostics. Such diagnostics are suggested in Brown *et al*. (2001). Model-based estimates should be (a) consistent with unbiased direct estimates and (b) more precise than direct estimates. The results reported in Table 3 show that the correlation between the model-based estimates and the direct estimates are positive and high, which indicates that the model-based estimates are consistent with the direct estimates. Table 7 also shows that overall the MQ/CD, MQ/BC, REBLUP and REBLUP/BC provide advantages over the direct estimator as the CVs of these model-based estimates are smaller than the corresponding CVs of the direct estimates.

There is a number of additional diagnostics that can be used for evaluating small area estimates. Practitioners can use the following diagnostics which are described below.

Bias Diagnostic: Plot the direct estimates on Y-axis and model-based estimates on X-axis and look for divergence of the regression line from $Y = X$.

Goodness of Fit (GoF) Diagnostic: This diagnostic tests whether the direct and model-based estimates are statistically different. The null hypothesis is that the direct and model-based estimates are statistically equivalent. The alternative is that the direct and model-based estimates are statistically different. The GoF diagnostic is computed using the following Wald statistic for every model based estimator

$$W = \sum_i \left\{ \frac{(\text{Direct Estimate}_i - \text{Model-Based Estimate}_i)^2}{\hat{Var}(\text{Direct Estimate}_i) + \hat{MSE}(\text{Model-Based Estimate}_i)} \right\}.$$

Table 5. Region specific estimates of average TCC values.

Regions	Direct	RI/ EBLUP	RI/ MBDE	RS/ EBLUP	RS/ MBDE	MQ/ Naive	MQ/ CD	MQ/ BC	REBLUP	REBLUP/ BC
1	180290	174077	179061	174196	188204	148821	180277	162954	148179	164375
2	205239	183078	196181	179570	198814	134125	193713	152467	142998	157367
3	100898	100507	96901	107073	98145	90410	99788	96044	93015	96198
4	207130	162557	186705	125826	180715	105097	182036	130226	115119	134535
5	110376	94628	106294	97405	106914	83022	98005	94053	94206	94523
6	40613	44067	39779	60412	39835	38038	37660	33252	34753	33909
7	84807	98662	80856	105502	80392	82633	91657	88601	90655	91054
8	82864	79720	80567	87499	79259	68626	78781	74865	74118	76319
9	64440	69099	64789	76499	66226	57596	65725	63714	65863	66369
10	67969	64398	66949	61611	68835	52738	66597	59787	61205	63086
11	127259	189660	121562	174580	141347	127506	139963	137246	141426	133570
12	303359	256445	251148	196262	305792	285864	239154	283727	292193	291734
13	164618	164448	160356	206673	184034	140785	173849	165225	154253	167197
14	208315	197493	207425	201520	205846	185783	209757	209757	192451	209144
15	93374	100551	89186	99611	91355	88394	97591	95314	91674	95140
16	182096	155491	167449	148782	162479	153191	159449	152201	138900	148897
17	80385	95541	84684	109189	76521	74639	84874	68813	97018	91559
18	52439	158025	61787	112620	60865	24491	35525	23571	229406	225856
19	263139	242451	225699	232568	190347	202237	238604	221643	228906	231114
20	102408	127687	100507	133625	101394	112230	105444	105444	100373	106336
21	72775	84517	71905	89265	74906	75363	78845	77430	78825	78754
22	82765	93992	76897	94584	78742	76528	92212	76397	79698	78945
23	500274	538320	497536	790841	523119	490384	638057	638057	488525	638385
24	209831	256422	206096	224414	252305	211430	205616	204395	174612	197254
25	234962	210603	226588	229050	258555	155369	224931	192604	168612	197563
26	229565	268593	223413	323008	244159	227876	261781	261781	265456	260938
27	94254	84270	89735	74456	88162	64938	90155	77305	77907	85425
28	113202	110309	106858	97211	94528	79168	119367	103283	97576	106401
29	1043862	705424	594341	695434	558929	571135	812597	656148	534603	632967
Mean	**182880**	**176243**	**160733**	**183079**	**165542**	**145118**	**175931**	**162286**	**156984**	**170859**

Table 6. Region specific estimates of Root MSE values.

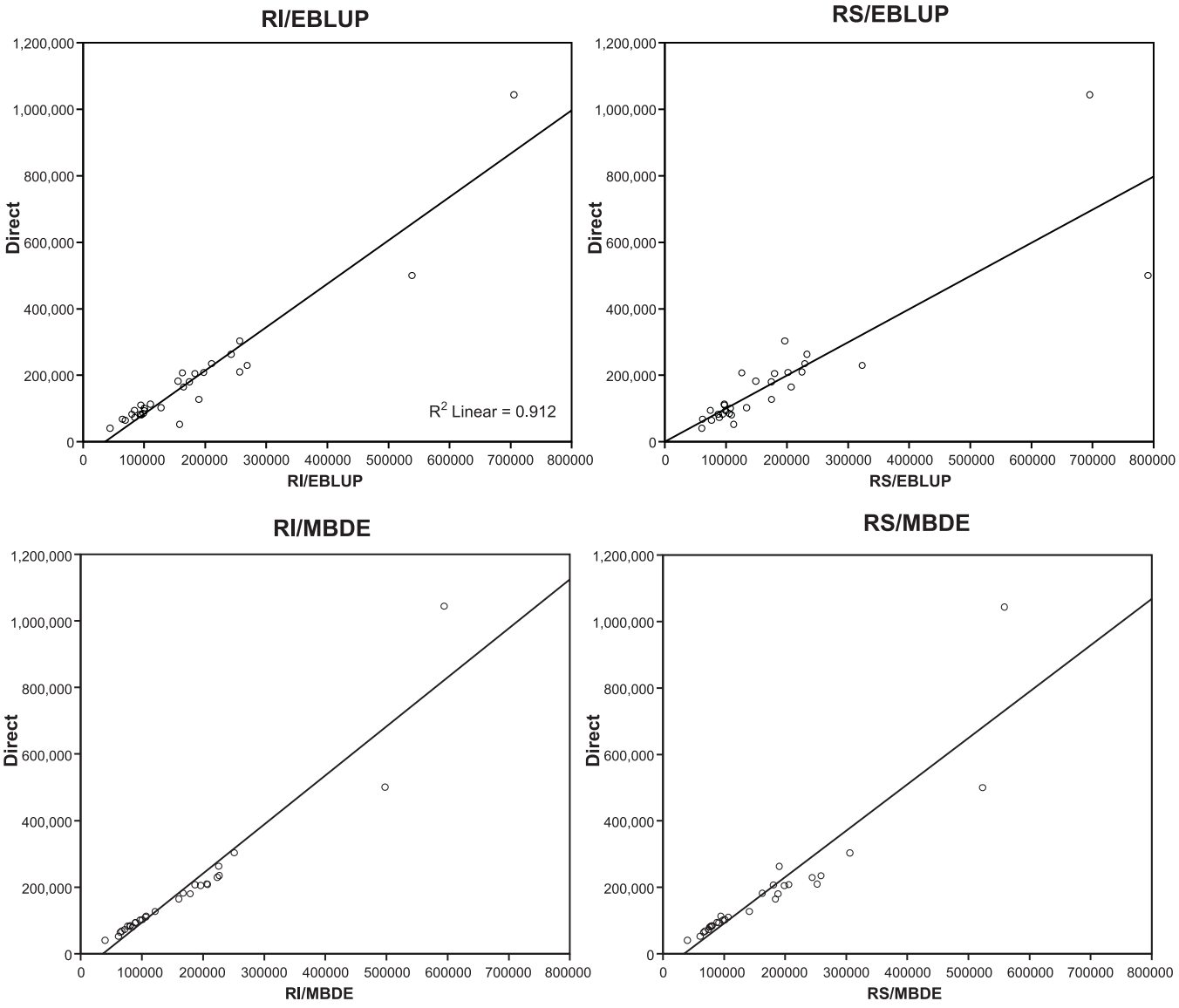
Regions	Direct	RI/ EBLUP	RI/ MBDE	RS/ EBLUP	RS/ MBDE	MQ/ Naive	MQ/ CD	MQ/ BC	REBLUP	REBLUP/ BC
1	25307	32374	25536	34647	22484	25090	25143	10384	13640	14067
2	29700	20514	44804	12128	27801	28801	27778	8143	7440	7411
3	8729	19846	6243	11727	6276	6851	6473	4598	6222	6033
4	49620	19512	80289	8248	29000	35537	33576	6514	7085	7000
5	10038	22264	73387	12036	13414	8357	7314	4909	8114	8068
6	8929	30960	104338	11438	9535	15230	7665	4158	10949	11202
7	9243	25677	17554	14140	14873	12707	6052	4014	9183	9382
8	8405	24820	5700	11605	5457	9267	5946	3504	9815	9646
9	6665	22934	10470	10765	5300	9940	5008	2497	7906	8063
10	7250	22705	103069	10843	5162	5271	5118	1975	8709	8497
11	46715	43390	56281	81517	68995	55523	40615	27322	21256	21250
12	65856	30089	43909	19958	177801	50662	46121	17994	11439	11854
13	23635	32213	22462	29821	30068	21351	21387	13119	11467	11784
14	13309	29180	14125	26708	14622	31999	12592	12929	11073	10924
15	9487	22095	11312	13791	14922	9239	6955	4836	8267	7968
16	21771	18688	200005	10454	28807	18126	12157	7996	5831	5620
17	13544	26773	60710	15554	17737	14634	12871	5828	9377	9550
18	13395	67459	33026	74818	173195	32708	34216	29242	21685	22184
19	49629	30036	19537	19909	53502	35132	24154	13019	11184	11181
20	9127	26334	16219	19284	19558	28346	7973	8071	8792	8644
21	5399	22075	13883	12784	9290	6616	4130	2697	7702	7432
22	11845	23011	74040	13775	18321	10042	10005	2879	8097	8037
23	128552	50191	198976	113101	232369	128847	116000	120198	26375	27901
24	39072	34545	36161	31448	92446	57854	24153	19118	12849	13105
25	25787	21752	27981	14079	33356	33640	22506	8412	7449	7402
26	16451	27360	39064	38535	69019	37257	12114	11801	11605	10807
27	10967	24992	93738	11541	7907	9740	8797	3511	8369	8193
28	17311	24560	84764	11979	18845	16877	13414	6483	9706	9701
29	301550	31466	139265	23755	203982	192003	191136	98097	12682	13146
Mean	**34044**	**28545**	**57133**	**24841**	**49105**	**32677**	**25909**	**16009**	**10837**	**10898**

Table 7. Region specific coefficients of variation %.

Regions	Direct	RI/ EBLUP	RI/ MBDE	RS/ EBLUP	RS/ MBDE	MQ/ Naive	MQ/ CD	MQ/ BC	REBLUP	REBLUP/ BC
1	14.04	18.60	14.26	19.89	11.95	16.86	13.95	6.37	9.20	8.56
2	14.47	11.21	22.84	6.75	13.98	21.47	14.34	5.34	5.20	4.71
3	8.65	19.75	6.44	10.95	6.39	7.58	6.49	4.79	6.69	6.27
4	23.96	12.00	43.00	6.56	16.05	33.81	18.44	5.00	6.15	5.20
5	9.09	23.53	69.04	12.36	12.55	10.07	7.46	5.22	8.61	8.54
6	21.99	70.26	262.29	18.93	23.94	40.04	20.35	12.51	31.51	33.04
7	10.90	26.03	21.71	13.40	18.50	15.38	6.60	4.53	10.13	10.30
8	10.14	31.13	7.07	13.26	6.88	13.50	7.55	4.68	13.24	12.64
9	10.34	33.19	16.16	14.07	8.00	17.26	7.62	3.92	12.00	12.15
10	10.67	35.26	153.95	17.60	7.50	10.00	7.69	3.30	14.23	13.47
11	36.71	22.88	46.30	46.69	48.81	43.55	29.02	19.91	15.03	15.91
12	21.71	11.73	17.48	10.17	58.14	17.72	19.29	6.34	3.91	4.06
13	14.36	19.59	14.01	14.43	16.34	15.17	12.30	7.94	7.43	7.05
14	6.39	14.78	6.81	13.25	7.10	17.22	6.00	6.16	5.75	5.22
15	10.16	21.97	12.68	13.84	16.33	10.45	7.13	5.07	9.02	8.38
16	11.96	12.02	119.44	7.03	17.73	11.83	7.62	5.25	4.20	3.77
17	16.85	28.02	71.69	14.24	23.18	19.61	15.17	8.47	9.67	10.43
18	25.54	42.69	53.45	66.43	284.55	133.55	96.32	124.06	9.45	9.82
19	18.86	12.39	8.66	8.56	28.11	17.37	10.12	5.87	4.89	4.84
20	8.91	20.62	16.14	14.43	19.29	25.26	7.56	7.65	8.76	8.13
21	7.42	26.12	19.31	14.32	12.40	8.78	5.24	3.48	9.77	9.44
22	14.31	24.48	96.28	14.56	23.27	13.12	10.85	3.77	10.16	10.18
23	25.70	9.32	39.99	14.30	44.42	26.27	18.18	18.84	5.40	4.37
24	18.62	13.47	17.55	14.01	36.64	27.36	11.75	9.35	7.36	6.64
25	10.98	10.33	12.35	6.15	12.90	21.65	10.01	4.37	4.42	3.75
26	7.17	10.19	17.49	11.93	28.27	16.35	4.63	4.51	4.37	4.14
27	11.64	29.66	104.46	15.50	8.97	15.00	9.76	4.54	10.74	9.59
28	15.29	22.26	79.32	12.32	19.94	21.32	11.24	6.28	9.95	9.12
29	28.89	4.46	23.43	3.42	36.50	33.62	23.52	14.95	2.37	2.08
Mean	**15.37**	**22.00**	**48.06**	**15.50**	**29.95**	**23.49**	**14.70**	**11.12**	**8.95**	**8.68**

The value from the test statistic is compared against the value from a chi square distribution with *D* degrees of freedom. In our case, this is the chi square value with *D=29* degrees of freedom which is 42.56 at 5% level of significance.

We now apply these two diagnostics to the estimates generated using the agricultural business survey data. Fig. 4 presents bias diagnostic plots. We note that all model-based estimators have similar consistency with the direct estimates. Overall, model-based estimates appear to be consistent with direct estimates, however, there are two regions for which the model-based and direct estimates are notably different. The GoF diagnostic results are presented in Table 4. These results indicate that all model-based estimates are not statistically different from the direct estimates apart from the REBLUP and the REBLUP/BC. The GoF diagnostic result for the REBLUP and the REBLUP/BC can be affected by the MSE estimation. For some regions the MSE estimates of the REBLUP and the REBLUP/BC appear to be very small compared to the MSE estimates of other model-based estimators. We think that in those cases we underestimate the MSE, which can be explained by the fact that parametric bootstrap MSE is used when the assumptions of the model may not hold (see Fig. 5). Taking into account the results from the coefficient of variation, the bias diagnostic plots and the GoF diagnostic we suggest that in this case the MQ/CD and MQ/BC have to be used. The EBLUP estimator that is based on the random slopes model appears to provide some efficiency gains over the direct estimator. Deciding to use the REBLUP or REBLUP/BC must be done bearing in mind that MSE estimation for these two estimators can be problematic in this case.



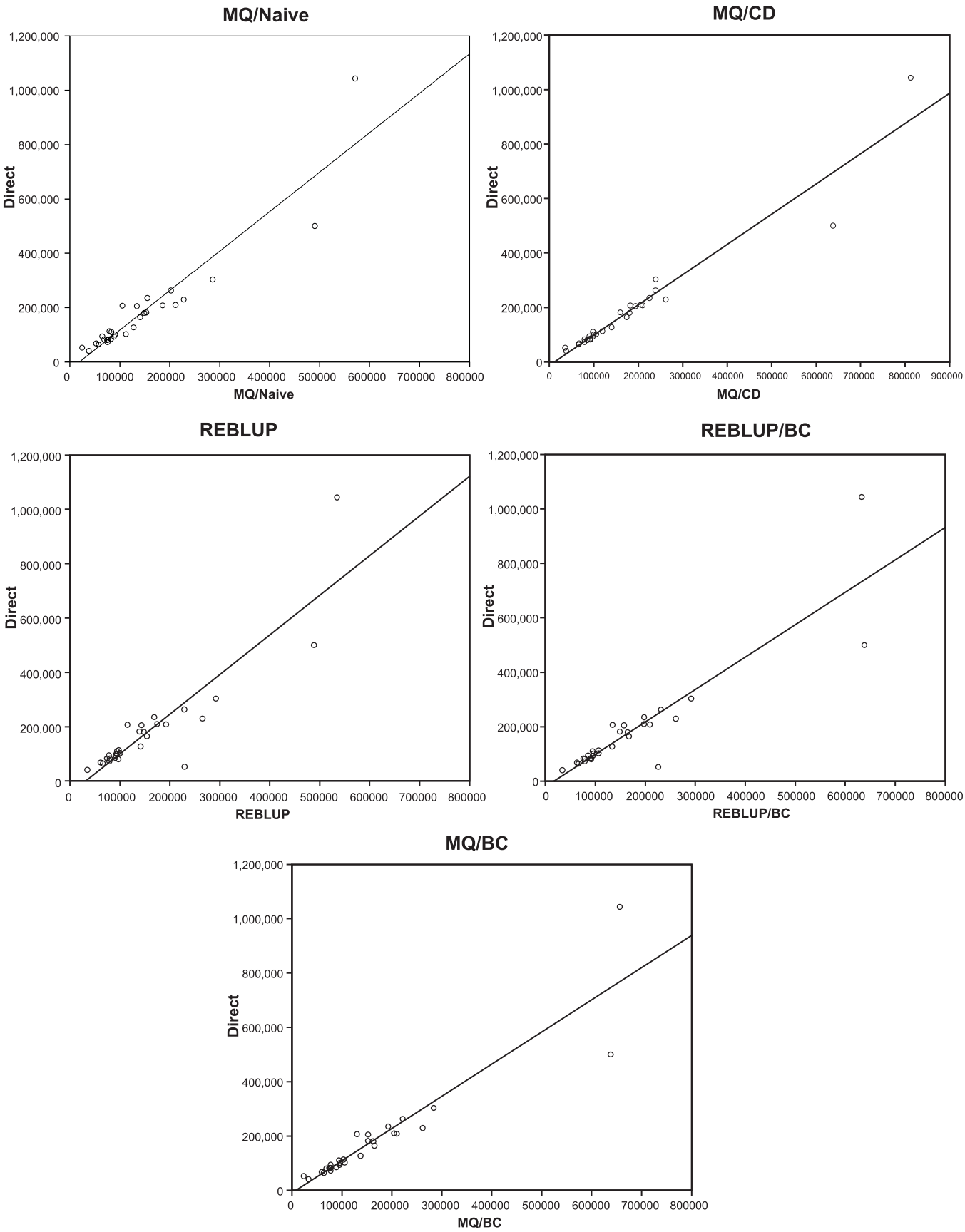


Fig. 5. Bias diagnostics plots

5. CONCLUSIONS

In this paper we present a small area application using an real agricultural business survey dataset. A range of well known and more recently proposed small area estimators are considered. These include the EBLUP estimator based on a random intercepts and random slopes mixed model, the MBDE estimator based on a random intercepts and random slopes mixed model and a range of outlier robust small area estimator using either an outlier robust version of the mixed model or the M-quantile small area model. Outlier robust estimators are considered in this paper due to departures from the assumptions of the mixed model (see Fig. 4). Although, not explored in this paper one can also examine the transformation based small area estimation when the model is not linear on raw scale, see Chandra and Chambers (2011). Emphasis is placed upon practical aspects of small area estimation. We start by presenting model diagnostics for deciding the best specification of the small area working model. This consists of deciding (a) what is the best specification for the fixed part of the model, (b) what is the best specification for the random part of the model and (c) whether outlier robust estimation is needed. Once the best possible small area working model has been found, we then have to use this model for producing small area point and MSE estimates. A number of diagnostics can be used for assessing the model-based estimates. These include (a) the consistency between the model-based and direct estimates which can be assessed by a GoF diagnostic, a bias diagnostic and by using the correlation coefficient between the model-based and direct estimates that has to be positive and high, and (b) the precision of model-based estimates, using for example the CV, which has to be smaller than that of direct estimates. However, the prospective user must be extremely careful when using these diagnostics. For example, the CV and the GoF diagnostics depend on the estimated MSE. For some model-based estimators, MSE estimation is performed by using parametric bootstrap which depends on the validity of the model assumptions. However, if the model assumptions do not hold or hold only approximately, the parametric bootstrap results may be misleading and hence the CV and GoF diagnostics may also be misleading. For this reason, the user must always be critical when using diagnostics for evaluating the results from small area estimation. All procedures for point and MSE estimation we presented in this paper can be implemented by using R functions. These are available upon request from the authors of this paper.

REFERENCES

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **83**, 401, 28-36.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001). Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS. *Proceedings of Statistics Canada Symposium 2001*. Achieving Data Quality in a Statistical Agency: A Methodological Perspective.

Chambers, R.L. (1986). Outlier robust finite population estimation. *J. Amer. Statist. Assoc.,* **81**, 1063-1069.

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika,* **93**, 255-268.

Chandra, H. and Chambers, R. (2009). Multipurpose weighting for small area estimation. *J. Off. Statist.*, **25**, 379-395.

Chandra, H. and Chambers, R. (2011). Small area estimation under transformation to linearity. *Survey Methodology,* **37(1)**, 39-51.

Chandra, H., Salvati, N., Chambers, R. and Tzavidis, N. (2012). Small area estimation under spatial nonstationarity. To appear

Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2009). Outlier Robust Small Area Estimation. Working Paper 16-09, Centre for Statistical and Survey Methodology, University of Wollongong. (Available from: http://cssm.uow.edu.au/publications).

Chambers, R., Chandra, H. and Tzavidis, N. (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, **37(2)**, 153-170.

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statist. Sci.*, **9**, 1, 55-93.

Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *J. Roy. Statist. Soc.,* **B68**, 221-238.

Marchetti, S., Tzavidis, N. and Pratesi, M. (2012). Non parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators. This will appear in *Compu. Statist. Data Anal.*

Molina, I. and Rao, J.N.K. (2010). Small area estimation of poverty indicators. *Canad. J. Statist.*, **38**, 369-385.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *J. Roy. Statist. Soc.,* **B70**, 265-286.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *Intern. Statist. Rev.*, **70**, 1, 125-143.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *J. Amer. Statist. Assoc.*, **85**, 163-171.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley, London.

Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *J. Amer. Statist. Assoc.,* **71**, 657-664.

Royall, R.M. and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *J. Amer. Statist. Assoc.*, **73**, 351-358.

Salvati, N., Tzavidis, N., Pratesi, M. and Chambers, R. (2011). Small area estimation via M-quantile geographically weighted regression, forthcoming in *TEST,* DOI 10.1007/s11749-010-0231-1.

Sinha, S.K. and Rao, J.N.K. (2009). Robust small area estimation. *Canad. J. Statist.*, **37**, 381-399.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010). Robust prediction of small area means and distributions. *Austr. & Newzealand J. Statist.*, **52**, 167-186.

Welsh, A.H. and Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *J. Roy. Statist. Soc.,* **B60**, 413-428.