Cathie Marsh Centre for Census and Survey Research

Robust Prediction of Small Area Means and Distributions

CCSR Working Paper 2007-08
Nikos Tzavidis and Ray Chambers
Nikos.tzavidis@manchester.ac.uk, ray@ouw.edu.au

Small area estimation techniques typically rely on mixed models containing random area effects to characterise between area variability. In contrast, Chambers and Tzavidis (2006) describe an approach to small area estimation based on regression M-quantiles. This approach avoids conventional Gaussian assumptions and problems associated with specification of random effects, allowing between area differences to be characterized by the variation of area-specific M-quantile coefficients. However, the resulting M-quantile predictors of small area means can be biased. In this paper we propose a general framework for robust bias adjusted small area prediction that corrects this problem, and is based on representing a small area predictor as a functional either of the Chambers and Dunstan (1986) or of the Rao-Kovar-Mantel (1990) predictor of the within area distribution of the target variable.

Robust Prediction of Small Area Means and Distributions

Nikos Tzavidis[1] and Ray Chambers[2]

1. Centre for Census and Survey Research
University of Manchester
Manchester M13 9PL, UK
nikos.tzavidis@manchester.ac.uk

2. Centre for Statistical and Survey Methodology
School of Mathematics and Applied Statistics
University of Wollongong
Wollongong, NSW 2522, Australia
ray@uow.edu.au

ABSTRACT

Small area estimation techniques typically rely on mixed models containing random area effects to characterise between area variability. In contrast, Chambers and Tzavidis (2006) describe an approach to small area estimation based on regression M-quantiles. This approach avoids conventional Gaussian assumptions and problems associated with specification of random effects, allowing between area differences to be characterized by the variation of area-specific M-quantile coefficients. However, the resulting M-quantile predictors of small area means can be biased. In this paper we propose a general framework for robust bias adjusted small area prediction that corrects this problem, and is based on representing a small area predictor as a functional either of the Chambers and Dunstan (1986) or of the Rao-Kovar-Mantel (1990) predictor of the within area distribution of the target variable. In doing so we note that use of this bias adjustment is not restricted to M-quantile models, but can be applied in conjunction with other models, e.g. mixed models, that are also suited to small area prediction. A further advantage of this framework is that it allows integrated prediction of small area means and quantiles. We demonstrate the usefulness of this framework through both model-based as well as design-based simulation, with the latter based on two realistic survey data sets containing small area information. The paper includes an application of the bias adjusted M-quantile approach to predicting key percentiles of district level distributions of per-capita household consumption expenditure in Albania in 2002.

KEY WORDS: Chambers-Dunstan estimator; Finite population distribution function; M-quantile regression; Rao-Kovar-Mantel estimator; Robust regression; Small area estimation; Smearing estimator; Weighted least squares

1. INTRODUCTION

Sample surveys provide a cost effective way of obtaining estimates for characteristics of interest at both population and sub-population (domain) level. In most practical applications domain sample sizes are not large enough to allow direct estimation. The term 'small areas' is typically used to describe such domains. When direct estimation is not possible, one has to rely upon alternative 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 or model-based methods.

The current industry standard for small area estimation is to use models that include random area effects to account for between area variation beyond that explained by the auxiliary information (Fay and Herriot 1979, Rao 2003). Traditionally, such models depend on Gaussian assumptions and require a formal specification of the random effects structure. In contrast, Chambers and Tzavidis (2006, hereafter referred to as CT) propose an alternative approach to small area estimation when unit level covariate information is available based on modelling the M-quantiles of the population-level conditional distribution of the target variable given the covariates. Such models avoid imposing strong distributional assumptions, and have the added benefit of not requiring formal specification of the random part of the model (i.e. those components of the model that capture unexplained heterogeneity caused by between area variability). Instead, between-area variability is captured by variation in so-called area-specific M-quantile coefficients. However, CT also observed that M-quantile predictors of small area means are biased. In this paper we therefore propose a bias adjustment to the M-quantile predictor of a small

area mean proposed by CT. This adjustment is based on representing this predictor as a functional of a corresponding predictor of the small area empirical distribution function. In particular, we consider use of the Chambers and Dunstan (1986, hereafter referred to as CD) smearing type predictor or the Rao-Kovar-Mantel predictor (1990, hereafter referred to RKM) of this distribution function. In doing so we note that use of this bias adjustment is not restricted to M-quantile models, but is generally applicable to any predictor that relies on substituting predicted values for non-sampled units in the small area, including predictors based on mixed models. An immediate consequence of posing the small area mean estimation problem within the context of prediction of the small area empirical distribution is that other small area distribution-related quantities, e.g. the small area quantiles, can also be predicted in a way that is consistent with prediction of the small area mean. This is especially useful if there are extreme values in the small area sample data, or the small area distribution is highly skewed.

The structure of the paper is as follows: In the following section we briefly review unit level mixed models with random area effects and M-quantile models for small area estimation. Then in section 3 we describe a general framework for small area estimation when unit level covariates are available, based on representing the small area target of inference as a functional either of the CD or the RKM predictors of the corresponding small area distribution. This naturally leads to a bias adjusted alternative to the M-quantile predictor for the small area mean proposed by CT, and, more generally, to any predictor of this mean based on predicted values for the non-sampled units within the small area. We also extend this idea to predicting the quantiles of the small area population distribution function. In section 4 we discuss mean squared error estimation

for the bias adjusted M-quantile predictor. In section 5 we present a series of model-based and design-based simulation studies for assessing the performance of the different small area predictors considered in this paper. In section 6, we use data from the 2002 Albanian Living Standards Measurement Study to predict within district distributions of per-capita consumption expenditure in this country. Finally, in section 7 we summarize our main findings.

2. UNIT LEVEL MODELS FOR SMALL AREA ESTIMATION

In what follows we assume that a vector x of p auxiliary variables is known for each of N units making up a population U and that values of the variable of interest y are available for each of n units making up a sample s from U. We also assume that U can be partitioned into d domains, which we refer to as areas, indexed by $j = 1, ..., d$, with area j containing N_j units, n_j of which comprise the sample s_j in the area, with the remaining unsampled $N_j - n_j$ units denoted by r_j. The target is to use the sample values for y and the population values for x to predict various area specific quantities, including (but not only) the area j mean m_j of y.

The most popular method used for this purpose employs linear mixed models. In the general case such a model specifies that for unit i in area j,

$$y_i = x_i^T \beta + z_i^T \gamma_j + \varepsilon_i \qquad (1)$$

where γ_j denotes a vector of random effects and z_i denotes a vector of auxiliary 'contextual' variables whose values are known for all units in the population. The role of the random effects in (1) is to characterise differences in the conditional distribution of y given x between the small areas. The parameters that characterise the joint distribution of

the γ_j and the ε_{ij} are usually referred to as the variance components associated with (1). Under this model m_j is typically predicted by

$$\hat{m}_j^{MX} = N_j^{-1}\left(\sum_{i\in s_j} y_i + \sum_{i\in r_j}\left\{x_i^T\hat{\beta} + z_i^T\hat{\gamma}_j\right\}\right) \tag{2}$$

where $\hat{\beta}$, $\hat{\gamma}_j$ are defined by 'plugging in' optimal (e.g. ML or REML) estimates of the variance components into the best linear unbiased estimator of β and the best linear unbiased predictor (BLUP) of γ_j respectively. Estimator (2) is often referred to as the empirical best linear unbiased predictor (EBLUP) of m_j (Henderson 1953).

An alternative approach to small area estimation is based on the use of quantile or M-quantile regression to characterise area effects. In the linear case, quantile regression leads to a family (or 'ensemble') of planes indexed by the value of the corresponding percentile coefficient $q \in (0,1)$ (Koenker and Bassett, 1978). For each value of q, the corresponding model shows how $Q_q(x)$, the q^{th} quantile of the conditional distribution of y given x, varies with x. A linear model for this conditional quantile is $Q_q(x) = x^T\beta_q$. The vector β_q in this model is estimated by minimising

$$\sum_{i=1}^{n}\left|y_i - x_i^T b\right|\left\{(1-q)I\left(y_i - x_i^T b \le 0\right) + qI\left(y_i - x_i^T b > 0\right)\right\}$$

with respect to b (Koenker and D'Orey, 1987). Quantile regression can be viewed as a generalisation of median regression. M-quantile regression (Breckling and Chambers, 1988) provides a generalisation of quantile regression based on influence functions, with the M-quantile of order q of the conditional density of y given x defined as the function $Q_q(x;\psi)$ that satisfies the estimating equation $\int \psi_q(y-Q) f(y\,|\,x)dy = 0$. A linear M-

quantile regression model is then one where we assume that $Q_q(x;\psi) = x^T \beta_\psi(q)$. That is, we allow a different set of regression parameters for each value of q. For specified q and ψ, an estimate $\hat{\beta}_\psi(q)$ of $\beta_\psi(q)$ can be obtained by solving

$$\sum_{i=1}^{n} \psi_q \left[\nu^{-1} \left\{ y_i - x_i^T \hat{\beta}_\psi(q) \right\} \right] x_i = 0$$

where $\psi_q(t) = 2\psi(t)\{qI(t>0) + (1-q)I(t \le 0)\}$ and ν is a suitable robust estimate of scale, e.g. the MAD estimate $\nu = median\left|r_{iq\psi}\right| / 0.6745$. Here ψ is an appropriately chosen influence function, which from now on is assumed to be the Huber function $\psi(t) = tI(-c \le t \le c) + c\,\mathrm{sgn}(t)I(|t| > c)$, with its default tuning constant c = 1.345.

CT extend the use of M-quantile regression models to small area estimation. Following their development (see also Kokic *et. al.*,1997; Aragon *et. al.*, 2005), we characterise the conditional variability across the population of interest by the M-quantile coefficients of the population units. For unit i with values y_i and x_i, this coefficient is the value θ_i such that $Q_{\theta_i}(x_i;\psi) = y_i$. Note that these M-quantile coefficients are determined at the population level. Consequently, if a hierarchical structure does explain part of the variability in the population data, then we expect units within clusters defined by this hierarchy to have similar M-quantile coefficients. When the conditional M-quantiles are assumed to follow a linear model, with $\beta_\psi(q)$ a sufficiently smooth function of q, this suggests a predictor of m_j of the form

$$\hat{m}_j^{MQ} = N_j^{-1} \left[\sum_{i \in s_j} y_i + \sum_{i \in r_j} \left\{ x_i^T \hat{\beta}_\psi(\hat{\theta}_j) \right\} \right]. \qquad (3)$$

Here $\hat{\theta}_j$ is an estimate of the average value of the M-quantile coefficients of the units in area j. However, alternative definitions of $\hat{\theta}_j$ are possible, for example the area j median of the unit M-quantile coefficients.

Note that the M-quantile approach to small area estimation is not restricted to continuous influence functions like the Huber function defined above. It can clearly also be implemented using quantile regression models, in which case the influence function underpinning the method is the discontinuous $\psi(t) = \text{sgn}(t)$. Here we use M-quantile regression models instead of 'standard' quantile regression models as the basis of our method for essentially practical reasons. Algorithms for fitting quantile regression models do not necessarily guarantee convergence and a unique solution. In contrast, the iteratively reweighted least squares algorithm used to fit an M-quantile regression converges to a unique solution (Kokic *et al.*, 1997) when a continuous monotone influence function is used. Finally, results from sensitivity analyses show that the choice of influence function does not impact upon the performance of the M-quantile-based small area estimators.

3. A GENERAL FRAMEWORK FOR SMALL AREA ESTIMATION

Given the finite population *U*, the area specific empirical distribution function of *y* for area *j* is

$$F_j(t) = N_j^{-1}\left\{\sum_{i\in s_j} I(y_i \le t) + \sum_{i\in r_j} I\left(y_i \le t\right)\right\}. \qquad (4)$$

The problem of predicting $F_j(t)$ essentially reduces to predicting the values y_i for the non-sampled units in small area j. One straightforward way of achieving this is to simply

replace the unknown non-sample values of y in (4) by their predicted values $\hat{y}_i$ under an appropriate model, leading to a predictor of (4) of the form

$$\hat{F}_j(t) = N_j^{-1}\left\{\sum_{i\in s_j} I(y_i \le t) + \sum_{i\in r_j} I\left(\hat{y}_i \le t\right)\right\}. \qquad (5)$$

A predictor of the mean m_j of y in area j is then defined by the value of the mean functional defined by (5). This leads to the usual plug-in predictor of the mean,

$$\hat{m}_j = \int_{-\infty}^{+\infty} t d\hat{F}_j(t) = N_j^{-1}\left(\sum_{i\in s_j} y_i + \sum_{i\in r_j} \hat{y}_i\right).$$

It immediately follows that the EBLUP (2) is the mean functional defined by (5) when $\hat{y}_i = x_i^T\hat{\beta} + z_i^T\hat{\gamma}_j$, while the M-quantile predictor (3) is also a mean functional corresponding to (5) but now with $\hat{y}_i = x_i^T\hat{\beta}_\psi(\hat{\theta}_j)$. In both cases the predicted value of a non-sample unit i in area j corresponds to an estimate $\hat{\mu}_i$ of its expected value given that it is located in area j.

We refer to small area predictors that can be expressed as functionals of (5), with non-sample predictions derived as estimates of expected values, as *naïve* predictors below. As we have already noted, CT observed that the naïve M-quantile predictor (3) can be biased. The reason for this is now clear. The distribution function (5) underlying (3) is not consistent in general. In particular, when the non-sample predicted values in (5) are estimated expectations $\hat{\mu}_i$ that converge in probability to the actual expected values μ_i, then

$$\sum_{i\in r_j} I\left(\hat{y}_i \le t\right) = \sum_{i\in r_j} I\left(\hat{\mu}_i \le t\right) = \sum_{i\in r_j} I\left\{y_i - (y_i - \hat{\mu}_i) \le t\right\} \approx \sum_{i\in r_j} I\left(y_i \le t + \varepsilon_i\right) \ne \sum_{i\in r_j} I\left(y_i \le t\right)$$

in general. Here the $\varepsilon_i = y_i - \mu_i$ are the actual regression errors. If these errors are independently and identically distributed symmetrically about zero we expect that the summation on the left hand side above will closely approximate the summation on the right for values of *t* near the median of the non-sampled area *j* values of *y* but not anywhere else. More generally, for heteroskedastic and/or asymmetric errors this correspondence will typically occur elsewhere in the support of *y*, although one would expect that in most reasonable situations it will be 'close' to the median of *y*. In other words, it is not advisable to use (5) to predict a quantile of the area *j* distribution of *y* other than the median.

By combining a smearing argument (Duan, 1983) with a model for the finite population distribution of *y*, CD develop a model-consistent predictor for a finite population distribution function. In the context of the small area distribution function (4), and assuming that the residuals $\varepsilon_i = y_i - \mu_i$ are homoskedastic within the small area of interest (an assumption satisfied by the linear mixed model), this is of the form

$$\hat{F}_j^{CD}(t) = N_j^{-1}\left[\sum_{i\in s_j} I(y_i \le t) + n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j} I\left\{\hat{\mu}_k + \left(y_i - \hat{\mu}_i\right) \le t\right\}\right]. \qquad (6)$$

In the Appendix we show that the mean functional defined by (6) takes the value

$$\hat{m}_j^{CD} = \int_{-\infty}^{\infty} t d\hat{F}_j^{CD}(t) = N_j^{-1}\left\{\sum_{i\in s_j} y_i + \sum_{i\in r_j}\hat{\mu}_i + (f_j^{-1}-1)\sum_{i\in s_j}(y_i - \hat{\mu}_i)\right\} \qquad (7)$$

where $f_j = n_j N_j^{-1}$ is the sampling fraction in area *j*. Under a linear M-quantile approach to small area estimation, (7) then defines a bias-adjusted predictor for m_j that is an alternative to (3) when we substitute $\hat{\mu}_{ij} = x_i^T \hat{\beta}_\psi(\hat{\theta}_j)$ in (7). Similarly, we obtain a bias-

adjusted alternative to the EBLUP (2) when we substitute $\hat{\mu}_i = x_i^T \hat{\beta} + z_i^T \hat{\gamma}_j$ in (7). In the former case we refer to this predictor as the CD-based M-quantile predictor, or M-quantile/CD predictor for short, while in the latter case we refer to it as the CD-based EBLUP predictor, or EBLUP/CD predictor for short. Corresponding predictors based on (5) will be denoted M-quantile/Naïve and EBLUP/Naïve respectively.

Outliers in the sample data can lead to large errors in estimation for the small areas in which they occur. Chambers (1986) considered the general problem of outlier robust prediction of finite population totals and means. Welsh and Ronchetti (1998) extended this approach to prediction of the finite population distribution function in the presence of outliers. In the context of robust prediction of an area *j* specific distribution function this leads to replacing the CD predictor (6) by

$$\hat{F}_j^{CD/rob}(t) = N_j^{-1}\left\{\sum_{i\in s_j} I(y_i \le t) + n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j} I\left[\hat{\mu}_k^{rob} + \nu_i \phi_{jt}\left\{\nu_i^{-1}(y_i - \hat{\mu}_i^{rob})\right\} \le t\right]\right\} \qquad (8)$$

where $\hat{\mu}_i^{rob}$ denotes an outlier robust estimate of the expected value μ_i of population unit *i* in area *j*, ν_i is a robust estimate of the scale of its residual $y_i - \mu_i$ and ϕ_{jt} is an outlier robust (i.e. bounded) influence function that can depend both on *j* and *t*. For the case $\phi_{jt} = \phi$ the corresponding predictor of the small area mean then takes the form

$$\hat{m}_j^{CD/rob} = N_j^{-1}\left[\sum_{i\in s_j} y_i + \sum_{i\in r_j} \hat{\mu}_i^{rob} + (f_j^{-1} - 1)\sum_{i\in s_j} \nu_i \phi\left\{\nu_i^{-1}(y_i - \hat{\mu}_i^{rob})\right\}\right]. \qquad (9)$$

Provided the influence function ψ used to define $\hat{\beta}_\psi(\hat{\theta}_j)$ in (3) is 'more' outlier robust than ϕ, e.g. $|\phi(t) - \psi(t)| \ge 0$, we can substitute $\hat{\mu}_i^{rob} = x_i^T \hat{\beta}_\psi(\hat{\theta}_j)$ in (9) to define an outlier robust predictor of the area *j* mean. In what follows, we denote (8), as well as functionals

derived from it, e.g. (9), by M-quantile/CDR. Note, however, that the cost of this robustness is inconsistency of (9), reflecting the usual bias-variance trade-off in outlier robust estimation. Similar robustification of the EBLUP (2) requires an outlier robust methodology for fitting the mixed model (1). There has been some theoretical development of this, see Richardson and Welsh (1996), but little practical application as far as we know. Consequently, we substitute the 'usual' EBLUP for $\hat{\mu}_i^{rob}$ in (8), and associated functionals, e.g. (9), denoting such predictors by EBLUP/CDR below.

Dorfman and Wang (1996) point out that the CD predictor (6) is model-consistent but design-inconsistent. An alternative to this predictor that is both design-consistent and model-consistent has been proposed by RKM. Under simple random sampling the RKM predictor of the finite population distribution function is

$$\begin{aligned}\hat{F}_j^{RKM}(t) = n_j^{-1}\sum_{i\in s_j} I(y_i \le t) + N_j^{-1}\sum_{k\in r_j} n^{-1}\sum_{i\in s_j} I(y_i - \hat{y}_i \le t - \hat{y}_k) \\ -(n_j^{-1} - N_j^{-1})\sum_{k\in s_j} n_j^{-1}\sum_{i\in s_j} I(y_i - \hat{y}_i \le t - \hat{y}_k).\end{aligned} \tag{10}$$

Chambers, Dorfman and Hall (1992) compared the large-sample mean squared errors of (6) and (10) and concluded that neither dominates the other. When the model is correctly specified we expect (6) to outperform (10). However RKM demonstrated that (6) can be substantially biased when model assumptions fail, while (10) is much less sensitive. Here we just note that the RKM predictor can be used to define a predictor of a small area characteristic, say θ_j, that can represented as a functional of the small area distribution function in exactly the same way as the CD-type predictors (6) and (8) can be used for this purpose. In general, the resulting predictors of θ_j will not be the same. An exception is the RKM-based predictor of the area *j* mean, which is the same as the CD-based

predictor of this mean under simple random sampling. See the Appendix for the proof of this result. Following the notation already introduced, predictors based on (10) will be denoted M-quantile/RKM if they define $\hat{y}_i$ via an M-quantile regression model, and by EBLUP/RKM if they use the linear mixed model (1) for this purpose.

Turning now to prediction of small area percentiles, we note that a predictor $\hat{m}_{pj}$ of the p^{th} quantile of the distribution of y in area j is straightforwardly defined as the solution to the estimating equation

$$\int_{-\infty}^{\hat{m}_{pj}} d\hat{F}_j(t) = p \tag{11}$$

given a suitable predictor $\hat{F}_j(t)$ of the area j distribution of y. CT discuss median estimation based on (11) when $\hat{F}_j(t)$ is defined by (5), with $\hat{y}_i = x_i^T \hat{\beta}_\psi(\hat{\theta}_j)$, i.e. naïve prediction. As the preceding discussion makes clear, we anticipate that a better approach for quantiles other than the median is to use either the CD-type specifications (6) and (8) or the RKM specification (10) for $\hat{F}_j(t)$, with $\hat{\mu}_i$ defined either by an M-quantile regression model or a by linear mixed model. Empirical results that address this issue are set out in Section 5.

4. MEAN SQUARED ERROR ESTIMATION

A robust mean squared error estimation method for the naïve M-quantile predictor $\hat{m}_j^{MQ}$ was described in CT. Here we extend this argument to define an estimator that is a first order approximation to the mean squared error of the mean predictor (7) when this is based on a M-quantile regression fit. A more detailed discussion of this approach to mean squared error estimation is set out in Chambers, Chandra and Tzavidis (2007).

To start, we note that since an iteratively reweighted least squares algorithm is used to calculate the M-quantile regression fit at $\hat{\theta}_j$, we have

$$\hat{\beta}_\psi(\hat{\theta}_j) = \left(X_s^T W_{sj} X_s\right)^{-1} X_s^T W_{sj} y_s$$

where X_s and y_s denote the matrix of sample x values and the vector of sample y values respectively, and W_{sj} denotes the diagonal weight matrix of order n that defines the estimator of the M-quantile regression coefficient with $q = \hat{\theta}_j$. It immediately follows that (7) can be written

$$\hat{m}_j^{MQ/CD} = w_{sj}^T y_s \tag{12}$$

where $w_{sj} = (w_{ij}) = n_j^{-1}\Delta_{sj} + (1 - N_j^{-1} n_j) W_j X_s (X_s^T W_j X_s)^{-1} \left\{\bar{x}_{rj} - \bar{x}_{sj}\right\}$ with Δ_{sj} denoting the n-vector that 'picks out' the sample units from area j. Here $\bar{x}_{sj}$ and $\bar{x}_{rj}$ denote the sample and non-sample means of x respectively in area j. Also, these weights are 'locally calibrated' on x since

$$\sum_{i \in s} w_{ij} x_i = \bar{x}_{sj} + (1 - f_j)(\bar{x}_{rj} - \bar{x}_{sj}) = \bar{x}_j .$$

Given the linear representation (12), standard methods of robust mean squared error estimation for linear predictors of population quantities (Royall and Cumberland, 1978) can be used. In particular, the prediction variance of $\hat{m}_j^{MQ/CD}$ is estimated by

$$v(\hat{m}_j^{MQ/CD}) = \sum_{g=1}^{d} \sum_{i \in s_g} \lambda_{ijg} \left\{y_i - x_i^T \hat{\beta}_\psi(\hat{\theta}_g)\right\}^2, \tag{13}$$

where $\lambda_{ijg} = \left\{(w_{ij} - 1)^2 + (n_j - 1)^{-1}(N_j - n_j)\right\} I(g = j) + w_{ig}^2 I(g \neq j)$. This prediction variance estimator implicitly assumes a model where the regression of y on x varies between areas, and that this variation is consistently estimated by the fit of the M-quantile

regression model in each area. Furthermore, since the weights defining $\hat{m}_j^{MQ/CD}$ are locally calibrated on x, it immediately follows that this predictor is unbiased under the same model and hence no correction for its bias is necessary when estimating its mean squared error. That is, we can use (13) as a first order approximation to the mean squared error of $\hat{m}_j^{MQ/CD}$. This can be compared with the estimator of the mean squared error of the naïve M-quantile predictor $\hat{m}_j^{MQ}$ described in CT, which includes a squared bias term. As an aside, we note that since the RKM-based predictor of the small area mean is the same the corresponding CD-based predictor under simple random sampling within areas, (13) also defines an estimator of the mean squared error of the RKM-based mean predictor in this case. Finally, we note that the issue of robust mean squared error estimation for predictors of characteristics of the small area distribution other than the mean (e.g. the quantiles) is not addressed here. In theory, such analytical estimators can be based on the Taylor series approximations described in Chambers, Dorfman and Hall (1992). However, we have so far not explored the behaviour of these estimators, or of more computationally intensive alternatives such as the jackknife procedure proposed by Wu and Sitter (2001), in the context of small area estimation.

5. SIMULATION STUDIES

In this section we present results from simulation studies that were used to compare the performance of the different small area estimators discussed in the preceding sections. The first was a model-based simulation in which small area population and sample data were simulated based on a two level linear mixed model with different parametric assumptions for the area and unit level random effects. We then present two design-based simulations in which actual sample data for a number of small areas were used to

construct two synthetic populations, which were then sampled repeatedly. In both cases the sample design used was stratified random sampling with strata corresponding to the small areas of interest and with stratum allocations set to the small area sample sizes in the original datasets.

5.1 MODEL-BASED SIMULATIONS

Two methods were used to simulate population data. In both, N = 232,500 population values of x and y in H = 30 small areas were generated with $N_h = 500h$ in area h. For each area h we selected a simple random sample (without replacement) of size $n_h = 30$, leading to an overall sample size of n = 900. The sample values of y and the population values of x were then used to estimate the small area target parameters, which were taken to be the small area means and selected quantiles of y. This process was repeated 1000 times.

The first method of population simulation (scenario 1) generated population values of x in small area h as $x_{ih} \sim N(\mu_h, \mu_h^2/36)$, where $\gamma_h \sim N(0,1)$, $\varepsilon_{ih} \sim N(0,64)$, and with $\mu_h \sim U[40,120]$ held fixed over the simulations. The second (scenario 2) generated these values as $x_{ih} \sim \chi^2(d_h)$, $\varepsilon_{ih} \sim \chi^2(3)-3$, $\gamma_h \sim \chi^2(1)-1$, with $d_h \sim U[1,200]$ held fixed over the simulations. The purpose of scenario 2 was to examine the effect of misspecification of the Gaussian assumptions of a mixed model. Population values of y in small area h in both scenarios were then generated from $y_{ih} = 5 + x_{ih} + \gamma_h + \varepsilon_{ih}$.

Two different types of small area linear models were fitted to the sample data obtained in the simulations. These were (a) a random intercepts specification of (1) and (b) a linear M-quantile regression specification. The random intercepts model used in (a) was based

on fitting a linear random effects model to the sample data using the default settings of the *lme* function (Venables and Ripley, 2002, section 10.3) in R. The M-quantile linear regression fit underpinning (b) was obtained using a modified version of the *rlm* function (Venables and Ripley, 2002, section 8.3) in R. Estimated model coefficients obtained from these fits were then used to compute naïve, CD and RKM-based versions of the EBLUP and M-quantile-based predictors of means and quantiles in the different areas.

Biases and mean squared errors over these simulations, averaged over the 30 areas, are set out in Table 1 (scenario 1) and in Table 2 (scenario 2). Under scenario 1 all approaches perform reasonably well. The differences between the naïve, CD and RKM versions of the EBLUP and M-quantile regression-based predictors are much more pronounced under scenario 2 (area effects distributed as chi-square). Here we see that use of naïve predictors leads to substantial biases as far as estimation of quantiles is concerned. In contrast, the CD and RKM-based predictors (both EBLUP and M-quantile) are essentially unbiased, even for extreme quantiles, with the CD-based predictors somewhat more efficient. On the basis of these results it would appear that predictors that are functionals of either the CD or the RKM distribution function predictors are preferable if there is concern about misspecification of the distribution of area effects.

Figure 1 shows the distribution across areas of coverage rates of nominal 95 per cent confidence intervals for the small area mean under both scenarios. These confidence intervals were constructed using the M-quantile/CD version of (12) together with its robust prediction variance estimator (13). We note that under scenario 1 (normal area effects) these intervals all achieve coverage close to the nominal 95 per cent across the different small areas. Under scenario 2 (chi square area effects) we also obtain coverage

rates close to nominal, although in this case there are some areas with relatively smaller coverage. Finally, in Table 3 we show key percentiles of the across areas distributions of the true and estimated mean squared errors of the predicted small area means calculated using the M-quantile/CD predictor (12). The estimated mean squared errors were calculated using (13) and are averaged over the Monte Carlo simulations. These results indicate that the mean squared error estimator (13) provides a good approximation to the true mean squared error of (12).

5.2 DESIGN-BASED SIMULATIONS

Case Study 1: Total cash costs by region for Australian broadacre farms

The population data on which these simulations are based are the same as those discussed in CT, and were obtained from a sample of 1652 broadacre farms spread across 29 agricultural regions of Australia. The y-variable of interest is the Total Cash Costs (TCC) of the farm business in the reference year. Auxiliary information available for each farm included the farm's sample weight, the total area of the farm in hectares (FarmArea) and the climatic zone in which the farm is situated. This information was used to classify the farms into six SizeZone strata on the basis of farm size and the climatic zone of the farm. The aim of this simulation study was to compare estimation of regional means of TCC under repeated sampling using both linear mixed models and linear M-quantile models. There were a total of N = 81982 farms generated for the study population, from which five hundred independent samples were selected. See CT for further details on how this was done and on the stratified sampling procedure, which replicated the regional distribution of sample farms in the original dataset. As in CT, all models used the same

set of x variables, defined by the main effects and interactions for the Farmarea and SizeZone variables.

Predicted values of regional means were obtained using both naïve and CD-based predictors assuming either a linear mixed model with random intercepts (EBLUP/Naïve and EBLUP/CD) or a linear M-quantile model (M-quantile/Naïve and M-quantile/CD). Note that the CD-based predictors are identical to RKM-based predictors in this case. These simulation results are set out in Table 4, which shows relative bias and relative root mean squared error (both expressed in percentage terms) averaged over the 29 regions. We immediately see that naïve M-quantile predictor of the mean is biased. However, this bias effectively disappears from the CD-based version of this predictor, which also records the lowest average RMSE value. As noted in CT, this population contains some extreme outliers, and this is reflected in the naïve EBLUP exhibiting some bias. Again we see that this bias is corrected by using the CD version of the EBLUP predictor, though in this case there is no corresponding reduction in RMSE.

Although we do not show these results here, we also evaluated the EBLUP and M-quantile versions of the outlier-robust CDR-based predictor (9), using 'huberised' residuals (based on a tuning constant of $c = 5$) to define the bias adjustment. As we expected, both of these further improved on the RMSE performance of their corresponding 'standard' versions (6), but at the cost of increased negative bias.

Figure 2 shows the regional distributions of coverage rates of nominal 95 per cent confidence intervals for regional means derived using the CD-based version (12) of the M-quantile predictor. In general intervals associated with this predictor display good coverage rates, with significant under-coverage only in one region that contained an

extremely large outlier. In Table 5 we further summarise the performance of (13) as an estimator of the mean squared error of (12) by comparing key percentiles of the distribution across areas of the average value of (12) over the simulations with the true (i.e. simulation-based) mean squared error of (12). These results indicate that (13) provides a good approximation to the true mean squared error of (12). In contrast, as reported in CT, the coverage rates of confidence intervals based on the naïve M-quantile predictor shows extensive undercoverage in this situation, which in this case is attributable to the bias of this predictor.

In addition to estimating regional means, we also predicted selected percentiles of distribution of TCC within the different regions by numerically solving (11), using naïve, CD, CDR and RKM predictors of the within region distribution function. Here we focus on the 10^{th} percentile, the median and the 90^{th} percentile. Our results are summarized in Figure 3, where we see that for both the 10^{th} and the 90^{th} percentile, the M-quantile and EBLUP versions of the naïve predictor (box plots 7 and 8) have larger absolute biases and root mean squared errors across the different regions than the corresponding CD and RKM-based predictors. As suggested in section 3, the situation is reversed at the median, where the M-quantile/Naïve predictor performs the best. Generally, these results indicate that for this population using a predictor based on an M-quantile model (box plots 1,2) is preferable to using one based on a linear mixed model (box plots 3,4), and that using a RKM-based predictor (box plots 2,4) is preferable to using a CD-based predictor (box plots 1,3). The outlier robust version of the CD predictor (box plots 5,6) seems to offer no worthwhile efficiency gains in this case.

Case Study 2: Per-capita consumption expenditure by district for Albania

This simulation is based on the 2002 Albanian Living Standards Measurement Study, and consists of data obtained from a sample of 3591 households spread across 36 districts of Albania. Unlike the farm survey data set, these data contain few outlying values and hence offer an opportunity for benchmarking the methodology under less extreme conditions. The variable of interest is the per-capita consumption expenditure of households, and the simulation study was implemented in two steps. A population of N = 724782 households was first created by sampling N times with replacement from the original sample of 3591 households and with probability proportional to a household's sample weight, then two hundred independent stratified random samples of the same size as the original sample were selected from this simulated population. District sample sizes for these simulated samples were fixed to be the same as in the original sample. Both the M-quantile and mixed (random intercepts) models used household level covariates defined by the presence of facilities in the dwelling (TV and parabolic dish antenna) and ownership of land.

The results set out in Table 6 are for prediction of district means. Again we see the bias in the naïve M-quantile predictor, which effectively disappears when the CD (or equivalently, RKM) version of this predictor is used. Also, for this population the bias of the naïve EBLUP predictor is small (essentially due to a small number of extreme values located in two districts), and so there is little to be gained from using its CD version. In Figure 4 we show the distribution over districts of the coverage rates of nominal 95 per cent confidence intervals for district means calculated using the M-quantile/CD predictor (12) and mean squared error estimator (13). With the exception of two districts these

coverages are quite close to the nominal 95 per cent level. Finally, in Table 7 we compare the distributions, across districts, of the average of the estimated mean squared errors of (12), computed via (13), and that of the actual Monte Carlo mean squared error of this predictor. Again, we see good agreement between these two distributions.

6. APPLICATION: ESTIMATING THE DISTRIBUTION OF DISTRICT LEVEL PER-CAPITA CONSUMPTION EXPENDITURE FOR ALBANIA

In this section we show how the methodology described in this paper may be practically employed for providing assistance and guidance to poverty alleviation programmes. In particular, we use the M-quantile/CD small area predictor to construct district level predictions of the mean and key percentiles of the distribution of per-capita household consumption expenditure in Albania in 2002. These predictions use data collected in the 2002 Albanian Living Standards Measurement Study (LSMS), which was funded as part of a programme by the World Bank aimed at improving the type and quality of household data collected by statistical offices in developing countries, and thereby fostering increased use of household data as a basis for policy decision making. The LSMS provides valuable information on a variety of issues related to living conditions of the people of Albania, including details on income and non-income dimensions of poverty in Albania, and forms the basis of poverty assessment in this country.

Albania consists of twelve prefectures, each composed of several districts, with a total of 36 districts. Our target is to use the LSMS data to predict the distribution of per-capita household consumption expenditure (PCHCE) in each of these districts, using an M-quantile modelling approach, with the objective of gaining a deeper insight into within district inequalities and how these are linked to district level estimates of poverty. The

selection of covariates for this model followed other studies on poverty assessment in Albania (Betti, 2003), taking into account the various non-income dimensions of poverty and deprivation that can potentially dominate the pure income dimension. The final model for PCHCE used therefore included the following household level covariates: the household size, which is a strong indicator of poverty; the presence of facilities in the household dwelling (TV, parabolic dish antenna, refrigerator, air conditioning, personal computer); ownership of the dwelling; ownership of land and ownership of a car.

Predicted values of average PHCE for each district were calculated using the M-quantile/CD predictor (12). Predicted values of other percentiles of the distribution of this variable in each district were obtained by numerical solution of the CD version of (11). Summary statistics showing the distribution of these predicted values across the 36 districts are set out in Table 8. An inspection of this Table, and particularly the inter-percentile ranges shown in its last two rows, indicates clearly that between district inequalities in PCHCE increase with increasing percentile of the within district distribution of this quantity. That is, districts are more unequal when compared in terms higher percentiles of PCHCE than they are when compared in terms of lower percentiles of this variable. In effect, between-district differences in the 2002 distribution of PCHCE cannot be explained by differences in the mean value of this variable. The interquartile and interdecile ranges could be potentially also used as district-level inequality measures. The results indicate that some of the wealthiest districts of Albania, mainly located in the costal (south west) and southern parts of the country, are also associated with high levels of inequality, while the poorer mountainous (north east) parts of the country are in many cases linked to lower levels of inequality. Finally, in Figure 5 we show maps of both the

predicted medians and means of PCHCE for Albania in 2002. It is worth noting how our view about the wealth of the different districts changes depending on whether we use a mean or a median for summarising PCHCE. More specifically means show a more wealthy central and northern part of the country than the medians. This illustrates the value of estimating small area distribution functions rather than just small area means.

7. SUMMARY AND EXTENSIONS

In this paper we outline an integrated and robust methodology for estimating small area means and distributions. The basis of our approach is the CD and RKM predictors of the small area distribution function, which can then be used to define a predictor of the small area mean as well as predictors of the small area quantiles. Our empirical results indicate that this approach shows promise when applied to unit level models for small area estimation, particularly when it is combined with the approach to small area estimation based on M-quantile regression modelling described in CT. However, the methodology described here has wider application, also leading to improvements in the efficiency of small area estimators based on mixed models.

Although we have not investigated them in any depth so far, extensions to the CD distribution function predictor that underpins our small area estimation framework are available, and lead to alternative predictors for small area characteristics. As we observed in section 3, Welsh and Ronchetti (1998) have proposed an outlier robust version (8) of the CD predictor (6). A slightly different modification to (6) uses local (i.e. nonparametric) weighting in the smearing process, leading to

$$\hat{F}_j^{CD/np}(t) = N_j^{-1}\left\{\sum_{i\in s_j} I(y_i \le t) + \sum_{i\in s_j}\sum_{k\in r_j} w_{ik} I\left[\hat{\mu}_k + \left(y_i - \hat{\mu}_i\right) \le t\right]\right\}. \qquad (14)$$

where the w_{ik} are 'local' weights that satisfy, for k in area j,

$$\sum_{i \in s_j} w_{ik} = 1 .$$

It is easy to show that the mean predictor implied by (13) is

$$\hat{m}_j^{np} = N_j^{-1} \left(\sum_{i \in s_j} y_i + \sum_{i \in r_j} \hat{y}_i + \sum_{i \in s_j} u_i (y_i - \hat{y}_i) \right)$$

where

$$u_i = \sum_{k \in r_j} w_{ik} .$$

We have not evaluated this option in the context of small area estimation, but previous experience with it for robust population level estimation (Chambers, Dorfman and Wehrly, 1993) indicates that it should also work well, particularly when there are significant non-linearities in the small area regression functions.

REFERENCES

Aragon, Y, Casanova, S., Chambers, R. and Leconte, E. (2005). Conditional ordering using nonparametric expectiles. *Journal of Official Statistics*, **21**, 617-633.

Betti, G. (2003). Poverty and inequality mapping in Albania: Final report (mimeo). World Bank and INSTAT: Washington DC and Tirana.

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, **75**, 761-771.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, **81**, 1063-1069.

Chambers, R. and Dunstan, R. (1986). Estimating distribution functions from survey data . *Biometrika*, **73**, 597-604.

Chambers, R.L., Dorfman, A.H. and Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika*, **79**, 577-82.

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, **88**, 268-277.

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

Chambers, Chandra and Tzavidis (2007). On robust mean squared error estimation for linear predictors for domains. *[Paper submitted for publication. A copy is available upon request]*.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, **78**, 605-610.

Fay, R.E. and Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, **9**, 226-252.

Kokic, P., Chambers, R., Breckling, J. and Beare, S. (1997). A measure of production performance. *Journal of Business and Economic Statistics*, **15**, 445-451

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33-50.

Koenker R. and D'Orey, V. (1987). Computing regression quantiles. *Applied Statistics*, **36**, 383-393.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, **77**, 365-75.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: Wiley.

Richardson, A.M. and Welsh, A.H. (1996). Covariate screening in mixed linear models. *Journal of Multivariate Analysis*, **58**, 27-54.

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

Venables,W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*. New York: Springer.

Wang, S. and Dorfman, A.H. (1996). A new estimator of the finite population distribution function. *Biometrika*, **83**, 639-52.

Welsh, A.H. and Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society B*, **60**, 413-428.

Wu, C. and Sitter, R.R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics*, **29**, 289-307.

APPENDIX

For notational simplicity, we drop the small area index *j*. The mean predictor defined by the CD predictor (6) is

$$\begin{aligned}
\hat{m}^{CD} &= N^{-1}\int_{-\infty}^{\infty} t d\hat{F}^{CD}(t) \\
&= N^{-1}\int_{-\infty}^{\infty} t d\left\{\sum_{i\in s} I(y_i \le t) + n^{-1}\sum_{i\in s}\sum_{j\in r} I\left(\hat{y}_j + y_i - \hat{y}_i \le t\right)\right\} \\
&= N^{-1}\left\{\sum_{i\in s}\int_{-\infty}^{\infty} t dI(y_i \le t) + n^{-1}\sum_{i\in s}\sum_{j\in r}\int_{-\infty}^{\infty} t dI\left(\hat{y}_j + y_i - \hat{y}_i \le t\right)\right\} \\
&= N^{-1}\left\{\sum_{i\in s} y_i + n^{-1}\sum_{i\in s}\sum_{j\in r}\left(\hat{y}_j + y_i - \hat{y}_i\right)\right\}
\end{aligned}$$

since $\int_{-\infty}^{\infty} t dI\left(y_i \le t\right) = y_i$. The identity (7) follows directly. Similarly, it is easy to see that under simple random sampling the RKM predictor of the mean satisfies

$$\begin{aligned}
\hat{m}^{RKM} &= n^{-1}\sum_{i\in s}\int_{-\infty}^{\infty} t dI(y_i \le t) + N^{-1}n^{-1}\sum_{j\in r}\sum_{i\in s}\int_{-\infty}^{\infty} t dI(\hat{y}_j + y_i - \hat{y}_i \le t) \\
&\quad -(n^{-1} - N^{-1})n^{-1}\sum_{j\in s}\sum_{i\in s}\int_{-\infty}^{\infty} I(\hat{y}_k + y_i - \hat{y}_i \le t) \\
&= n^{-1}\sum_{i\in s} y_i + N^{-1}n^{-1}\sum_{j\in r}\sum_{i\in s}\left(\hat{y}_j + y_i - \hat{y}_i\right) - (n^{-1} - N^{-1})n^{-1}\sum_{j\in s}\sum_{i\in s}\left(\hat{y}_j + y_i - \hat{y}_i\right) \\
&= n^{-1}\sum_{i\in s} y_i + N^{-1}n^{-1}\sum_{j\in r}\sum_{i\in s}\left(\hat{y}_j + y_i - \hat{y}_i\right) - (n^{-1} - N^{-1})\sum_{i\in s} y_i \\
&= N^{-1}\left\{\sum_{i\in s} y_i + n^{-1}\sum_{j\in r}\sum_{i\in s}\left(\hat{y}_j + y_i - \hat{y}_i\right)\right\} \\
&= \hat{m}^{CD}.
\end{aligned}$$

Table 1. Model-based simulation results for Scenario 1 (Gaussian random effects) averaged over 30 small areas.

Method	Target					
	10^{th}	25^{th}	50^{th}	Mean	75^{th}	90^{th}
	Relative Bias (%)					
EBLUP/Naïve	0.088	0.041	-0.002	-0.002	-0.036	-0.062
EBLUP/CD	0.096	0.046	0.051	-0.002	0.072	0.160
EBLUP/RKM	0.005	0.015	-0.024	-0.002	0.015	0.105
M-quantile/Naïve	0.090	0.044	0.003	0.003	-0.030	-0.055
M-quantile/CD	0.058	0.003	-0.003	-0.002	0.008	0.064
M-quantile/RKM	-0.011	0.002	0.008	-0.002	0.009	0.014
	Relative RMSE (%)					
EBLUP/Naïve	0.29	0.23	0.20	0.23	0.19	0.19
EBLUP/CD	0.34	0.25	0.22	0.24	0.21	0.26
EBLUP/RKM	0.31	0.25	0.21	0.24	0.20	0.20
M-quantile/Naïve	0.46	0.38	0.33	0.32	0.31	0.30
M-quantile/CD	0.34	0.25	0.21	0.24	0.21	0.24
M-quantile/RKM	0.32	0.25	0.22	0.24	0.21	0.22

Table 2. Model-based simulation results for Scenario 2 (Chi square random effects) averaged over 30 small areas.

Method	Target					
	10^{th}	25^{th}	50^{th}	Mean	75^{th}	90^{th}
	Relative Bias (%)					
EBLUP/Naïve	22.48	9.731	0.420	0.024	-4.708	-6.969
EBLUP/CD	0.373	0.205	0.079	-0.018	-0.073	-0.186
EBLUP/RKM	0.216	0.599	0.125	-0.018	-0.348	0.001
M-quantile/Naïve	17.24	5.653	-2.641	-1.794	-7.021	-8.787
M-quantile/CD	0.373	0.176	0.028	-0.018	-0.086	-0.188
M-quantile/RKM	0.211	0.596	0.124	-0.018	-0.348	0.003
	Relative RMSE (%)					
EBLUP/Naïve	22.56	9.99	2.86	1.97	4.93	7.03
EBLUP/CD	3.23	3.08	3.01	2.01	3.32	3.90
EBLUP/RKM	4.10	3.56	3.30	2.01	3.46	4.12
M-quantile/Naïve	17.60	6.70	3.30	2.49	7.04	8.80
M-quantile/CD	3.23	3.09	3.11	2.01	3.48	3.89
M-quantile/RKM	4.11	3.56	3.36	2.01	3.46	4.12

Table 3. Across areas distribution of true (i.e. Monte Carlo) mean squared errors and average over Monte Carlo simulations of estimated mean squared errors for the M-quantile/CD predictor (12). Estimated mean squared errors based on (13).

MSE Source	Percentiles of across areas distribution					
	10^{th}	25^{th}	50^{th}	Mean	75^{th}	90^{th}
	Gaussian area effects					
True	1.37	1.41	1.45	1.45	1.48	1.53
Estimated	1.42	1.43	1.44	1.44	1.47	1.47
	Chi square area effects					
True	0.41	0.43	0.44	0.44	0.46	0.47
Estimated	0.43	0.44	0.45	0.45	0.46	0.46

Table 4. Design-based simulation results for Australian broadacre farms data: Estimation of average TCC within regions. Entries show regional averages of Relative Bias % (RB) and Relative RMSE % (RRMSE) for different prediction methods.

Method	RB	RRMSE
EBLUP/Naïve	4.04	19.60
EBLUP/CD	1.43	20.84
M-quantile/Naïve	-16.17	20.41
M-quantile/CD	-0.20	18.23

Table 5. Australian farms data: Percentiles of the across regions distribution of the true (i.e. Monte Carlo) mean squared error of the M-quantile/CD predictor (12) of mean TCC within regions and the corresponding distribution of the average (over the Monte Carlo simulations) of its estimated mean squared error (13).

MSE Source	Percentiles of across regions distribution					
	10^{th}	25^{th}	50^{th}	Mean	75^{th}	90^{th}
True	7360	9847	17990	31550	30080	69454
Estimated	7281	9940	18290	30170	30300	66081

Table 6. Design-based simulation results for Albanian household expenditure data: Estimation of mean household consumption expenditure within districts. Entries show district averages of Relative Bias % (RB) and Relative RMSE % (RRMSE) for different prediction methods.

Method	RB	RRMSE
EBLUP/Naïve	0.61	5.46
EBLUP/CD	0.07	5.55
M-quantile/Naïve	-10.74	13.30
M-quantile/CD	0.07	5.55

Table 7. Albanian household expenditure data: Percentiles of the across districts distribution of the true (i.e. Monte Carlo) mean squared error of the M-quantile/CD predictor (12) of mean household consumption expenditure within districts and the corresponding distribution of the average (over the Monte Carlo simulations) of its estimated mean squared error (13).

MSE Source	Percentiles of across districts distribution					
	10^{th}	25^{th}	50^{th}	Mean	75^{th}	90^{th}
True	250	345	506	596	707	1163
Estimated	253	355	524	607	748	1122

Table 8. Albanian Living Standards Measurement Study: Across districts distribution of predicted within district mean and percentiles of per-capita consumption expenditure. M-quantile/CD predictors used.

Percentiles of across districts distribution	Within district target percentile					
	10^{th}	25^{th}	50^{th}	Mean	75^{th}	90^{th}
10^{th}	2772	4389	5999	6876	8153	11159
25^{th}	3129	5117	7185	7549	9697	12193
50^{th}	3824	5480	7870	8252	10309	13036
Mean	3918	5761	8009	8552	10676	13694
75^{th}	4462	6271	8702	9605	12274	15662
90^{th}	5245	7617	10416	11001	13547	17606
75^{th}-25^{th}	1333	1154	1517	2056	2577	3469
90^{th}-10^{th}	2473	3228	4417	4125	5394	6447

Figure 1. Distribution over areas of area-specific coverage rates of nominal 95 per cent confidence intervals for small area means in the model-based simulations. Intervals were defined as the CD-based M-quantile predictor (11) plus or minus twice its estimated standard error, calculated as the square root of (12).

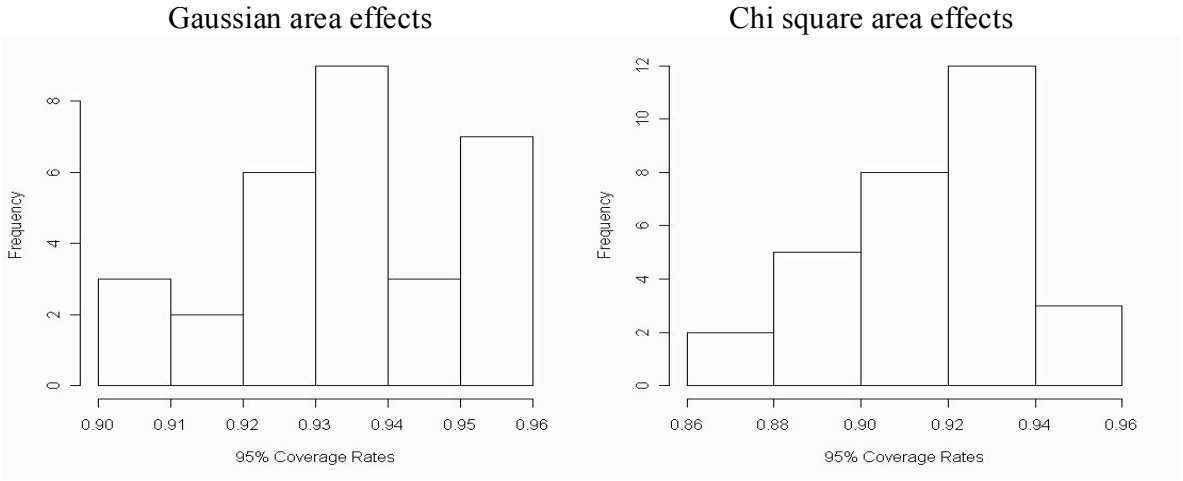

Figure 2. Australian farms data: Distribution over regions of region-specific coverage rates of nominal 95 per cent confidence intervals. Intervals were defined as the CD-based M-quantile predictor (11) plus or minus twice its estimated standard error, calculated as the square root of (12).

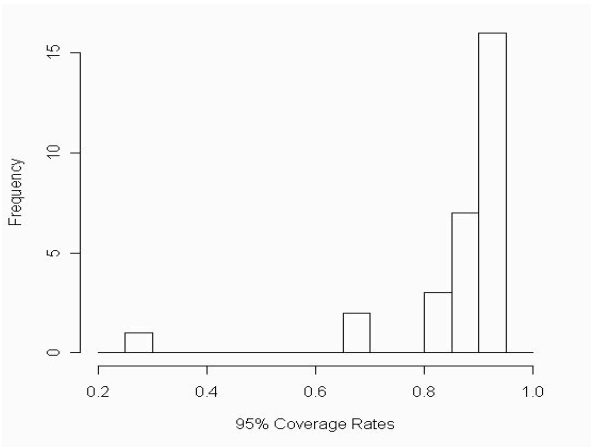

Figure 3. Australian farms data: Box plots showing across region distributions of average prediction error (left column) and root mean squared error (right column) for predicted percentiles (top = 10^{th} , middle = median, bottom = 90^{th}) of the within region distribution of TCC. Box plots correspond to different predictors: 1 = M-quantile/CD; 2 = M-quantile/RKM; 3 = EBLUP/CD; 4 = EBLUP/RKM; 5 = M-quantile /CDR (c = 5); 6 = EBLUP/CDR (c = 5); 7 = M-quantile/naïve; 8 = EBLUP/naïve.

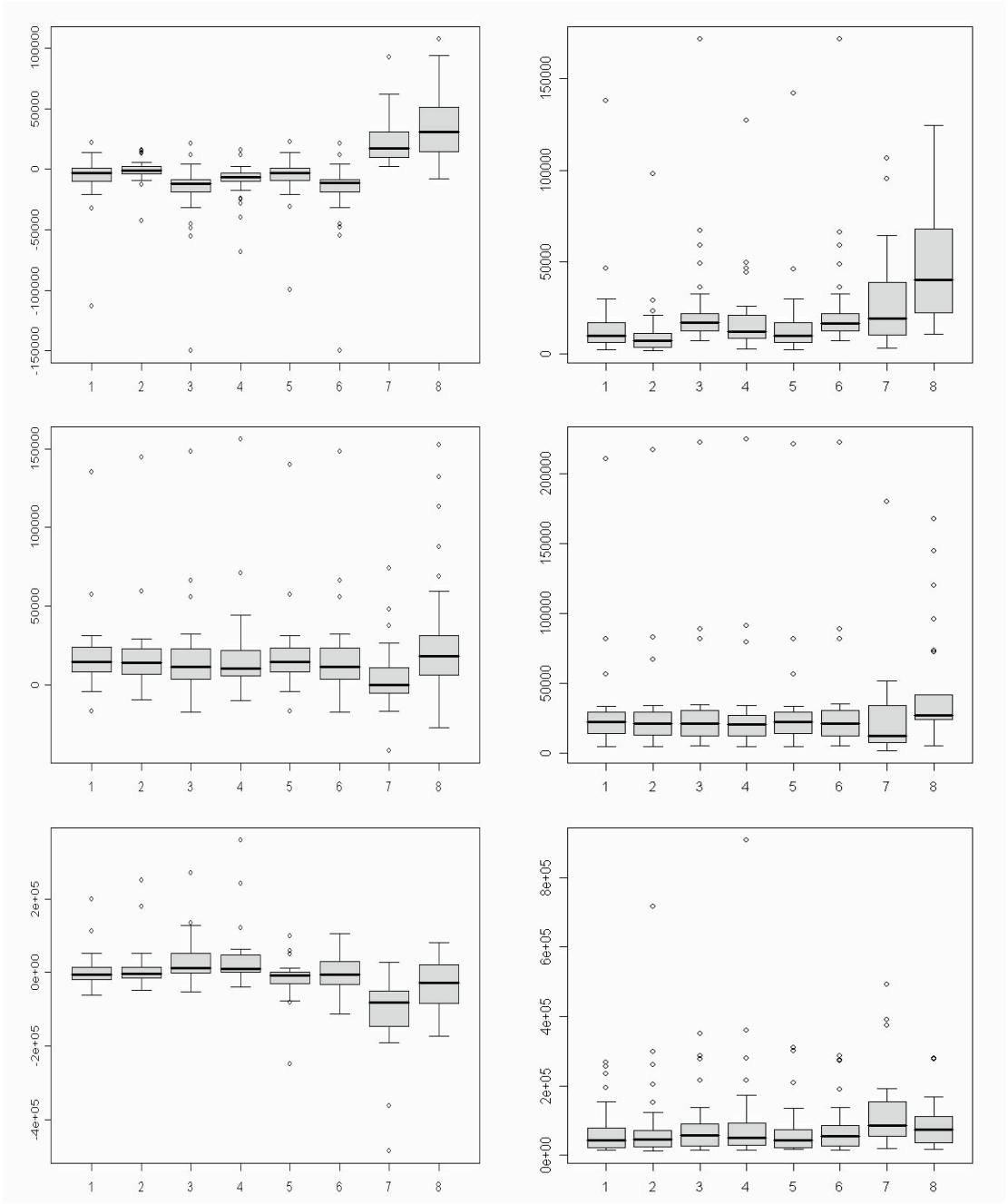

Figure 4. Albanian household expenditure data: Across district distribution of coverage rates of nominal 95 per cent confidence intervals. Intervals were defined as CD-based M-quantile predictors plus or minus twice their estimated standard errors, calculated as the square root of (12).

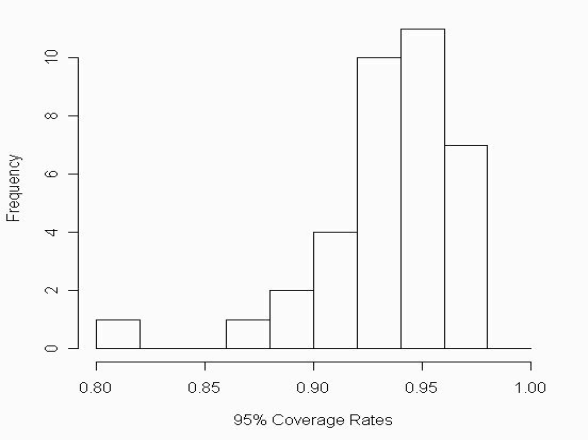

Figure 5. Albanian Living Standards Measurement Study: Maps showing estimated median consumption expenditure per district (left) and estimated average consumption expenditure per district (right).