Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

10-09

Robust Estimation of Small Area Means and Quantiles

(Short title: Robust Small Area Estimation)

Nikos Tzavidis, Stefano Marchetti and Ray Chambers

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

ROBUST ESTIMATION OF SMALL AREA MEANS AND QUANTILES

(Short Title: **ROBUST SMALL AREA ESTIMATION**)

Nikos Tzavidis[1], Stefano Marchetti[2] and Ray Chambers[3]
University of Manchester, University of Pisa and University of Wollongong

[1] Author to whom correspondence should be addressed:
Social Statistics and Centre for Census and Survey Research, University of Manchester, Manchester, M13 9PL, UK.
Email nikos.tzavidis@manchester.ac.uk; Telephone: +44 161 306 6953; Fax: +44 161 275 4722

[2] Department of Statistics and Mathematics applied to Economics, University of Pisa, Pisa, Via Ridolfi n. 10, 56124, Italy.

[3] Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW 2522, Australia.

Acknowledgements
The research described in this paper was supported in part by ARC Linkage Grant LP0776810 of the Australian Research Council and by a 7th European Framework Grant SSH-CT-2008-217565 FP7-SSH-2007-1 of the European Commission. We are grateful to the Technical Editor for suggestions that considerably improved the paper.

Summary

Small area estimation techniques have typically relied on plug-in estimation based on models containing random area effects. More recently, regression M-quantiles have been suggested for this purpose, thus avoiding conventional Gaussian assumptions, as well as problems associated with the specification of random effects. However, the plug-in M-quantile estimator for the small area mean can be shown to be the expected value of this mean with respect to a generally biased estimator of the small area cumulative distribution function of the characteristic of interest. To correct this problem, we propose a general framework for robust small area estimation, based on representing a small area estimator as a functional of a predictor of this small area cumulative distribution function. Key advantages of this framework are that it naturally leads to integrated estimation of small area means and quantiles and is not restricted to M-quantile models. We also discuss mean squared error estimation for the resulting estimators, and demonstrate the advantages of our approach through model-based and design-based simulations, with the latter using economic data collected in an Australian farm survey.

Key words: Australian farm data; Chambers-Dunstan estimator; finite population distribution function; M-quantile regression; Rao-Kovar-Mantel estimator; robust regression; small area estimation; smearing estimator.

1. Introduction

Sample surveys provide a cost-effective way of obtaining estimates for population characteristics of interest. However, this estimation may become problematic when these characteristics relate to a particular sub-population or domain for which the sample size is small. The term 'small areas' is typically used to describe domains whose sample sizes are not large enough to allow sufficiently precise direct estimation, i.e. estimation based only on the sample data for the domain. When direct estimation is not possible, one has to rely upon 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.

The standard approach to small area estimation uses regression models to predict the small area characteristics of interest, and incorporates random area effects to account for between-area variation beyond that explained by the model covariates (Fay & Herriot 1979, Rao 2003). Typically, these random effects are assumed to be Gaussian, and the models themselves require 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). In contrast, Chambers & Tzavidis (2006, hereafter referred to as CT) proposed an alternative approach to small area estimation when the target variable is measured on a continuous scale and unit level covariate information is available. This approach is based on modelling the regression M-quantiles of the population-level conditional distribution of the target variable. It avoids the strong distributional assumptions implicit in the mixed model approach, and has the added benefit of not requiring formal specification of random area effects. Instead, between-area variability is captured by variation in the area-specific M-quantile coefficients. However, the estimator of the small area mean suggested in CT is essentially a plug-in estimator and, as we show in Section 3,

corresponds to the expected value of this mean under a biased estimator of the small area cumulative distribution function (CDF). Consequently, we propose an alternative framework for small area estimation that is based on representing an estimator of a small area characteristic of interest as an appropriate functional of the Chambers & Dunstan (1986, hereafter referred to as CD) smearing-type estimator of this CDF. More generally, we note that our framework also allows small area estimates to be defined in terms of functionals of alternative smearing-type estimators of the small area CDF, e.g. the outlier resistant CDF estimator suggested by Welsh & Ronchetti (1998) or the design-consistent CDF estimator proposed by Rao, Kovar & Mantel (1990). The framework is generally applicable to any small area estimator that substitutes predicted values for non-sampled units in the small area, including those that use an M-quantile model or a mixed model for this purpose. An important consequence of formulating the small area mean estimation problem as an extension of the problem of estimation of the small area CDF is that other small area distribution-related quantities, e.g. the small area quantiles, can also be estimated in a way that is consistent with estimation of the small area mean. This is especially useful if there are extreme values in the small area sample data, or if the small area distribution of the characteristic of interest is highly skewed.

The structure of the paper is as follows: In the following section we briefly review the use of both unit-level linear models with random area effects and linear M-quantile models in 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 of the CD estimator of the corresponding small area CDF. This naturally leads to a bias-adjusted alternative to the M-quantile estimator of the small area mean proposed by CT, and, more generally, to any estimator of this mean that substitutes predicted values for the unknown non-sample values within the small area. We also extend this approach to estimation

of the corresponding small area quantiles. In Section 4 we describe linearization and bootstrap methods of mean squared error estimation for these bias adjusted estimators, and in Section 5 we assess the performance of the different small area estimators considered in this paper via model-based and design-based simulation studies. Finally, in Section 6 we summarise 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 *y* is measured on a continuous scale and that *U* can be partitioned into *d* mutually exclusive and exhaustive 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 estimate various area specific quantities, including (but not only) the area *j* mean m_j of *y*.

The most popular method used for this purpose is based on linear mixed models. In general, such a model specifies that for unit *i* in area *j*,

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \boldsymbol{\gamma}_j + \varepsilon_{ij}, \quad (1)$$

where $i = 1,\ldots,n_j$ and $j = 1,\ldots,d$. Here $\boldsymbol{\gamma}_j$ denotes a vector of random effects and $\mathbf{z}_{ij}$ 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 small area differences in the conditional distribution of *y* given **x**. The parameters that characterise the joint distribution of the area effects $\boldsymbol{\gamma}_j$ and the unit-level effects ε_{ij} are usually referred to as the variance components

associated with (1). Under this model, m_j is typically estimated by the mixed-model (MX) estimator

$$\hat{m}_j^{MX} = N_j^{-1}\left\{\sum_{i\in s_j} y_i + \sum_{i\in r_j}\left(\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T\hat{\boldsymbol{\gamma}}_j\right)\right\}, \quad (2)$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}_j$ are defined by 'plugging in' optimal (e.g. ML or REML) estimates of the variance components into the best linear unbiased estimator of $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) of $\boldsymbol{\gamma}_j$ 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 uses either 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 & Bassett 1978, Koenker 2005). For each value of *q*, the corresponding model shows how the *q*th quantile of the conditional distribution of *y* given **x**, denoted $Q_q(\mathbf{x})$, varies with **x**. A linear model for this conditional quantile is $Q_q(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}_q$. The vector $\boldsymbol{\beta}_q$ in this model is estimated by minimising

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

with respect to **b** (Koenker & D'Orey, 1987). Here $I(a)$ denotes the indicator function for the event *a*. M-quantile regression (Breckling & 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 $Q_q(\mathbf{x};\psi)$ that satisfies the estimating equation $\int \psi_q(y-Q)\,f(y\,|\,\mathbf{x})\;dy = 0$. Here $\psi_q(t) = 2\psi(t)\left\{qI(t>0) + (1-q)I(t\le 0)\right\}$ and ψ

is a user-specified influence function, e.g. the Huber Proposal 2 function $\psi(t) = tI(-c \le t \le c) + c\,\mathrm{sgn}(t)I(|t| > c)$, where c is a tuning constant. A linear M-quantile regression model is one where we assume that $Q_q(\mathbf{x};\psi) = \mathbf{x}^T\boldsymbol{\beta}_\psi(q)$. That is, we allow a different set of regression parameters for each value of q and for each choice of the influence function ψ. For specified q and ψ, an estimate $\hat{\boldsymbol{\beta}}_\psi(q)$ of $\boldsymbol{\beta}_\psi(q)$ can be obtained by using iteratively reweighted least squares to solve

$$\sum_{i=1}^{n} w_{iq\psi}\left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\psi(q)\right)\mathbf{x}_i = 0,$$

where $w_{iq\psi} = v_{q\psi}\left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\psi(q)\right)^{-1}\psi_q\left[v_{q\psi}^{-1}\left\{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\psi(q)\right\}\right]$. Here $v_{q\psi}$ is a suitable robust estimator of scale, e.g. the Median Absolute Deviation (MAD) estimator $v_{q\psi} = median|r_{iq\psi}| / 0.6745$, with $r_{iq\psi} = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\psi(q)$. In this paper we will always assume that ψ is the Huber Proposal 2 function, with its default tuning constant c = 1.345.

Following the development in CT (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 $\mathbf{x}_i$, this coefficient is the value θ_i such that $Q_{\theta_i}(\mathbf{x}_i;\psi) = y_i$. Note that these M-quantile coefficients are determined at the population level. If a hierarchical structure does explain part of the variability in the population data, we expect units within the clusters defined by this hierarchy to have similar M-quantile coefficients. Consequently, we characterise a cluster by the location of the distribution of its associated unit-level M-quantile coefficients. In particular, we define θ_j as the mean of the unit-level M-quantile coefficients within the jth cluster. When the conditional M-quantiles are

assumed to follow a linear model, with $\boldsymbol{\beta}_\psi(q)$ a sufficiently smooth function of q, CT suggested a plug-in M-quantile (MQ) estimator of m_j of the form

$$\hat{m}_j^{MQ} = N_j^{-1}\left[\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j}\left\{\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_j)\right\}\right]. \quad (3)$$

Here $\hat{\theta}_j$ is a suitable estimator of θ_j. Provided sampling is non-informative given **x** in area *j*, $\hat{\theta}_j$ can be calculated as the sample mean of the estimated unit-level M-quantile coefficients in that area.

Note that the M-quantile approach to small area estimation is not restricted to continuous influence functions like the Huber function defined above, since it can also be implemented using quantile regression models, in which case the influence function underpinning the method is the discontinuous function $\psi(t) = \text{sgn}(t)$. In this paper we use M-quantile regression models instead of 'standard' quantile regression models for essentially practical reasons. Algorithms for fitting quantile regression models do not necessarily guarantee convergence or a unique solution. In contrast, the iteratively reweighted least squares algorithm used to fit an M-quantile regression model based on a continuous and monotone influence function converges to a unique solution (Kokic *et al.* 1997). 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 CDF of *y* in area *j* is

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

The problem of estimating $F_j(t)$ given the sample data is therefore essentially one of predicting the non-sample sum of the zero-one values $I\left(y_{ij} \le t\right)$ 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 $\hat{y}_{ij}$ under an appropriate model, leading to a plug-in estimator of (4) of the form

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

An estimator 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 estimator of this mean,

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

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

We hereafter refer to small area estimators that can be expressed as functionals of (5), with non-sample predictions derived as estimates of expected values, as *naïve*. In particular, CT observed that the naïve M-quantile estimator (3) can be biased. The reason for this is now clear. The CDF estimator (5) underlying (3) is not consistent in general. When the non-sample predicted values in (5) are estimated expectations $\hat{\mu}_{ij}$ that converge in probability to the actual expected values μ_{ij}, we see that

$$\sum_{i\in r_j} I\left(\hat{y}_{ij} \le t\right) = \sum_{i\in r_j} I\left(\hat{\mu}_{ij} \le t\right) = \sum_{i\in r_j} I\left\{y_{ij} - (y_{ij} - \hat{\mu}_{ij}) \le t\right\} \approx \sum_{i\in r_j} I\left(y_{ij} \le t + \varepsilon_{ij}\right) \ne \sum_{i\in r_j} I\left(y_{ij} \le t\right).$$

Here $\varepsilon_{ij} = y_{ij} - \mu_{ij}$ is the actual regression error. 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* that is far from the median.

By combining a smearing argument (Duan, 1983) with a model for the finite population CDF of *y*, CD developed a model-consistent estimator for a finite population CDF. In the context of the small area CDF (4), and assuming that the residuals $\varepsilon_{ij} = y_{ij} - \mu_{ij}$ are homoskedastic within the small area of interest (an assumption satisfied by the linear mixed model), this is of the form

$$\hat{F}_j^{CD}(t) = N_j^{-1}\left[\sum_{i\in s_j} I(y_{ij} \le t) + n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j} I\left\{\hat{\mu}_{kj} + \left(y_{ij} - \hat{\mu}_{ij}\right) \le t\right\}\right]. \tag{6}$$

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

$$\hat{m}_j^{CD} = \int_{-\infty}^{\infty} t\, d\hat{F}_j^{CD}(t) = N_j^{-1}\left\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j} \hat{\mu}_{ij} + (f_j^{-1} - 1)\sum_{i\in s_j}(y_{ij} - \hat{\mu}_{ij})\right\}, \tag{7}$$

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 estimator of m_j that represents an alternative to (3) when we substitute $\hat{\mu}_{ij} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_j)$ in (7). We note that a corresponding bias-adjusted

alternative to the EBLUP (2) is obtained when we substitute $\hat{\mu}_{ij} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T\hat{\boldsymbol{\gamma}}_j$ in (7). In the former case we refer to this estimator as the CD-based M-quantile estimator, or M-quantile/CD estimator, while in the latter case we refer to it as the CD-based EBLUP estimator, or EBLUP/CD estimator. Corresponding estimators 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 & Ronchetti (1998) extended this approach to prediction of the finite population CDF in the presence of outliers. In the context of robust prediction of an area *j* specific CDF, these authors replace the CD estimator (6) by

$$\hat{F}_j^{WR}(t) = N_j^{-1}\left[\sum_{i\in s_j} I(y_{ij} \le t) + n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j} I\left[\hat{\mu}_{kj}^{rob} + v_{ij}\phi_{jt}\left\{v_{ij}^{-1}(y_{ij} - \hat{\mu}_{ij}^{rob})\right\} \le t\right]\right], \tag{8}$$

where $\hat{\mu}_{ij}^{rob}$ denotes an outlier-robust estimate of the expected value μ_{ij} of population unit *i* in area *j*, v_{ij} is a robust estimate of the scale of its residual $y_{ij} - \mu_{ij}$ 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 estimator of the small area mean based on (8) is then

$$\hat{m}_j^{WR} = N_j^{-1}\left[\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j}\hat{\mu}_{ij}^{rob} + (f_j^{-1} - 1)\sum_{i\in s_j} v_{ij}\phi\left\{v_{ij}^{-1}(y_{ij} - \hat{\mu}_{ij}^{rob})\right\}\right]. \tag{9}$$

Provided the influence function ψ used to define $\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_j)$ in (3) is 'more' outlier-robust than ϕ, i.e. $|\phi(t) - \psi(t)| \ge 0$, we can substitute $\hat{\mu}_{ij}^{rob} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_j)$ in (9) to define an outlier-robust estimator of the area *j* mean. In what follows, we denote (8), as well as functionals derived from it, e.g. (9), by M-quantile/WR. 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). One approach in this situation is to substitute the EBLUP $\hat{\mu}_{ij} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T\hat{\boldsymbol{\gamma}}_j$ for $\hat{\mu}_{ij}^{rob}$ in (8). We denote estimators, e.g. (9), based on (8) with this substitution by EBLUP/WR below. However, since the EBLUP $\hat{\mu}_{ij}$ is clearly not outlier-robust, this is not a satisfactory solution, and a better one would be to use a robustified version of $\hat{\mu}_{ij}$, building on the development in Richardson & Welsh (1995) and Richardson (1997). Although we do not pursue this idea further in this paper, a recent paper by Sinha & Rao (2009) is a step in this direction.

Wang & Dorfman (1996) pointed out that the CD estimator (6) is model-consistent but design-inconsistent. Rao, Kovar & Mantel (1990) proposed an alternative to this estimator that is both design-consistent and model-consistent. Under simple random sampling within the small areas, the estimator of the finite population CDF suggested by these authors is

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

Chambers, Dorfman & 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, Rao, Kovar & Mantel (1990) demonstrated that (6) can be substantially biased when model assumptions fail, while (10) is much less sensitive. Here we just note that, as with (6) and (8), (10) can be used to define an estimator of a small area characteristic that can be represented as a functional of the small area CDF. In general, the resulting estimators generated by (6) and (10) will not be the same. An exception is the estimator of the area *j* mean, which is the same under (6) and (10). See the Appendix for the proof of this

result. Following the notation already introduced, estimators based on (10) will be denoted M-quantile/RKM if they define $\hat{y}_{ij}$ via a linear M-quantile model, and by EBLUP/RKM if they use the linear mixed model (1) for this purpose.

Turning now to the small area quantiles, we note that an estimator of the *p*th quantile of the distribution of *y* in area *j* is straightforwardly defined as the solution to the estimating equation

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

where $\hat{F}_j(t)$ is an estimator of the area *j* CDF of *y*. CT discussed median estimation based on (11) when $\hat{F}_j(t)$ is defined by (5), with $\hat{y}_{ij} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_j)$, i.e. naïve estimation. As the preceding discussion makes clear, we anticipate that a better approach for estimating quantiles other than the median is to use smearing-type estimators like (6), (8) or (10) for $\hat{F}_j(t)$, with $\hat{\mu}_{ij}$ defined either by an M-quantile model or by a linear mixed model. Empirical results that address this issue are presented in Section 5.

4. Mean squared error estimation

4.1 Linearization-based MSE estimation for estimators of small area means

A robust estimator of the mean squared error of the naïve M-quantile estimator $\hat{m}_j^{MQ}$ is described in CT. Here we extend this argument to define an estimator that is a first order approximation to the mean squared error of (7) when this is based on an M-quantile regression fit. A more detailed discussion of this approach to mean squared error estimation is set out in Chambers, Chandra & Tzavidis (2008).

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 can write

$$\hat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_j) = \left(\mathbf{X}_s^T \mathbf{W}_{sj} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \mathbf{W}_{sj} \mathbf{y}_s$$

where $\mathbf{X}_s$ and $\mathbf{y}_s$ are the matrix of sample $\mathbf{x}$ values and the vector of sample y values respectively, and $\mathbf{W}_{sj}$ is 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 as the weighted sum

$$\hat{m}_j^{MQ/CD} = \mathbf{w}_{sj}^T \mathbf{y}_{sj}, \tag{12}$$

where $\mathbf{y}_{sj}$ denotes the vector of sample y values in area j and $\mathbf{w}_{sj} = (\mathbf{w}_{ij}) = n_j^{-1}\boldsymbol{\Delta}_{sj} + (1 - N_j^{-1} n_j)\mathbf{W}_{sj}\mathbf{X}_s(\mathbf{X}_s^T \mathbf{W}_{sj}\mathbf{X}_s)^{-1}\left(\bar{\mathbf{x}}_{rj} - \bar{\mathbf{x}}_{sj}\right)$ is a vector of implied area j specific sample weights. Here $\boldsymbol{\Delta}_{sj}$ denotes the n-vector that 'picks out' the sample units from area j and $\bar{\mathbf{x}}_{sj}$ and $\bar{\mathbf{x}}_{rj}$ denote the vectors of sample and non-sample means of $\mathbf{x}$ respectively in area j. Note that the weights in (12) are 'locally calibrated' on $\mathbf{x}$ since

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

A first order approximation to the mean squared error of (12) treats the weights defining this representation as fixed, and applies standard methods of robust mean squared error estimation for linear estimators of population quantities (Royall & Cumberland 1978). With this approach, the prediction variance of $\hat{m}_j^{MQ/CD}$ is estimated by

$$\sum_{g=1}^{d} \sum_{i \in s_g} \lambda_{ijg} \left\{ y_{ig} - \mathbf{x}_{ig}^T \hat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_g) \right\}^2, \tag{13}$$

where $\lambda_{ijg} = \left\{ (\mathbf{w}_{ij} - 1)^2 + (n_j - 1)^{-1}(N_j - n_j) \right\} I(g = j) + \mathbf{w}_{ig}^2 I(g \neq j)$. Note that this prediction variance estimator implicitly assumes a model where the regression of y on $\mathbf{x}$ varies between

areas. It also assumes 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 follows that (12) is unbiased for m_j under the same model, and so (13) can be used as an estimator of the MSE of $\hat{m}_j^{MQ/CD}$. This can be compared with the estimator of the mean squared error of the naïve M-quantile estimator $\hat{m}_j^{MQ}$ described in CT, which includes a squared bias term. As an aside, we note that, since the estimator of the small area mean defined by (10) is the same as that defined by (6), the expression (13) also defines an estimator of the mean squared error of the mean estimator defined by (10) when small area samples are obtained by simple random sampling.

4.2 Bootstrap MSE estimation for estimators of small area quantiles

The linearization-based prediction variance estimator (13) is defined only when the estimator of interest can be written as a weighted sum of sample values. Consequently, it cannot be used with quantile estimators defined by solving (11). In this section we describe an alternative non-parametric bootstrap approach to MSE estimation in this case, based on the approach of Lombardia, Gonzalez-Manteiga & Prada-Sanchez (2003). In particular, we define two bootstrap schemes that resample residuals from an M-quantile model fit. The first scheme draws samples from the empirical distribution of suitably re-centred residuals. The second scheme draws samples from a smoothed version of this empirical distribution. Using these two schemes, we generate a bootstrap population, from which we then draw bootstrap small area samples. In order to define the bootstrap population, we first calculate the M-quantile small area model residuals $e_{ij} = y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_j)$. A bootstrap finite population $U^* = \{y_{ij}^*, \mathbf{x}_{ij}\}, i \in U, j = 1,\ldots,d$ with $y_{ij}^* = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_j) + e_{ij}^*$ is then generated, where the bootstrap residuals e_{ij}^* are obtained by

sampling from an estimator of the CDF $\hat{G}(u)$ of the e_{ij}. In order to define $\hat{G}(u)$, we consider two approaches: (i) sampling from the empirical CDF of the residuals e_{ij} and (ii) sampling from a smoothed CDF of these residuals. In each case, sampling of the residuals can be done in two ways: (i) by sampling from the distribution of all residuals without conditioning on the small area (the unconditional approach); and (ii) by sampling from the conditional distribution of residuals within small area *j* (the conditional approach). The empirical CDF of the residuals is

$$\hat{G}(u) = n^{-1}\sum_{j=1}^{d}\sum_{i\in s_j} I(e_{ij} - \bar{e}_s \le u),$$

where $\bar{e}_s$ is the sample mean of the e_{ij}. Similarly, the empirical CDF of these residuals in area *j* is

$$\hat{G}_j(u) = n_j^{-1}\sum_{i\in s_j} I(e_{ij} - \bar{e}_{sj} \le u)$$

where $\bar{e}_{sj}$ is the sample mean of the e_{ij} in area *j*. A smoothed estimator of the unconditional CDF is

$$\hat{G}(u) = n^{-1}\sum_{j=1}^{d}\sum_{i\in s_j} K\left\{h^{-1}\left(u - e_{ij} + \bar{e}_s\right)\right\},$$

where $h > 0$ is a smoothing parameter and *K* is the CDF corresponding to a bounded symmetric kernel density *k*. Similarly a smoothed estimator of the conditional CDF in area *j* is

$$\hat{G}_j(u) = n_j^{-1}\sum_{i\in s_j} K\left\{h_j^{-1}\left(u - e_{ij} + \bar{e}_{sj}\right)\right\},$$

where $h_j > 0$ and *K* are the same as above. In the empirical studies reported in Section 5, we define *K* in terms of the Epanechnikov kernel, $k(u) = (3/4)\left(1-u^2\right)I\left(|u|<1\right)$, while the smoothing parameters h and h_j are chosen so that they minimize the cross-validation criterion

suggested by Bowman, Hall & Prvan (1998). That is, in the unconditional case, h is chosen in order to minimize

$$CV(h) = n^{-1}\sum_{j=1}^{d}\sum_{i\in s_j}\int\left[I\left\{(e_{ij} - \bar{e}_s) \le u\right\} - G_{-i}(u)\right]^2 du\,,$$

where $G_{-i}(u)$ is the version of $G(u)$ that omits sample unit *i*, with the extension to the conditional case being obvious. It can be shown (Li & Racine, 2007, section 1.5) that choosing h and h_j in this way is asymptotically equivalent to using the MSE optimal values of these parameters. In the simulation studies reported in the next section, we compute both the conditional and unconditional smoothed distribution functions of residuals using the *np* package in the R software environment (R Development Core Team 2008) that implements the above approach. In either case, bootstrap samples s_j^* are then drawn, using simple random sampling without replacement within the small areas. In what follows we denote by $F_j(t)$ the unknown true CDF of the finite population values in area *j*, by $\hat{F}_j^{CD}(t)$ the CD estimator of $F_j(t)$ based on sample s_j, by $F_j^*(t)$ the known true CDF of the bootstrap population U_j^* in area *j*, and by $\hat{F}_j^{CD*}(t)$ the CD estimator of $F_j^*(t)$ based on bootstrap sample s_j^*. Let $\tau_j = \tau(F_j)$ denote the functional defined by $F_j(t)$ that corresponds to the small area characteristic of interest, with associated CD-based estimator $\hat{\tau}_j^{CD} = \tau(\hat{F}_j^{CD})$. The bootstrap population value of this functional is then $\tau_j^* = \tau(F_j^*)$, with associated CD-based estimator $\hat{\tau}_j^{CD*} = \tau(\hat{F}_j^{CD*})$. We then estimate the MSE of the CD-based estimator $\hat{\tau}_j^{CD}$ as follows. Starting from the sample *s,* we generate *B* bootstrap populations, U^{*b}, using one of the four above-mentioned methods for estimating the

CDF of the residuals. From each bootstrap population, U^{*b}, we select L samples using simple random sampling without replacement within the small areas with $n_j^* = n_j$. The bootstrap estimator of the MSE of the CD-based estimator $\hat{\tau}_j^{CD}$ is then

$$B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}\left\{\hat{\tau}_j^{CD*bl} - \text{av}_L(\hat{\tau}_j^{CD*bl})\right\}^2 + \left\{B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}\left(\hat{\tau}_j^{CD*bl} - \tau_j^{*b}\right)\right\}^2. \quad (14)$$

Here τ_j^{*b} is the area j value of the characteristic of interest for the bth bootstrap population and $\text{av}_L(\hat{\tau}_j^{CD*bl}) = L^{-1}\sum_{l=1}^{L}\hat{\tau}_j^{CD*bl}$, where $\hat{\tau}_j^{CD*bl}$ is the CD-based estimator of this characteristic computed from the lth sample of the bth bootstrap population, (b = 1,…,B, l = 1,…,L). Note that this bootstrap procedure can also be used to construct confidence intervals for the value of τ_j by 'reading off' appropriate quantiles of the bootstrap distribution of $\hat{\tau}_j^{CD}$. Finally, we observe that this bootstrap approach is not restricted to functionals defined by the CD-based estimator (6), but can also be used to estimate the MSEs of functionals defined by the alternative smearing-type CDF estimators (8) and (10).

5. Simulation studies

In this section we present results from two simulation studies that were used to compare the performance of the different small area estimators defined 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. The second was a design-based simulation in which actual sample survey data for a number of small areas were used to construct a synthetic population, which was then sampled repeatedly. The sampling design used in this case was stratified random

sampling, with the strata corresponding to the small areas of interest, and with stratum allocations set to the small area sample sizes in the original survey.

5.1 Model-based simulations

Two methods were used to simulate bivariate population values (y, x) in $d = 30$ small areas. In both, $N = 232\ 500$ with $N_j = 500j$ in area j. For each area j, we selected a simple random sample (without replacement) of size $n_j = 30$, 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 simulation experiment (scenario 1) generated population values of y using $y_{ij} = 5 + x_{ij} + \gamma_j + \varepsilon_{ij}$, with the x_{ij} generated as independently and identically distributed realisations from $N(\xi_j, \xi_j^2 / 36)$. The small area x-means ξ_j were themselves drawn at random from the uniform distribution on the interval (40, 120), and held fixed over the simulations. Similarly, the random effects γ_j and ε_{ij} were independently and identically generated as $N(0,1)$ and $N(0,64)$ realisations respectively. The second simulation experiment (scenario 2) generated values of the target variable using the same linear model as in scenario 1, but in this case values of x_{ij} were generated as independently and identically distributed realisations from $\chi^2(d_j)$, with the d_j drawn at random from the integers 1 to 200, and held fixed over the simulations. Also, the random effects γ_j and ε_{ij} were independently and identically generated as mean-corrected $\chi^2(1)$ and $\chi^2(3)$ realisations respectively. The purpose of scenario 2 was to examine the effect of misspecification of the Gaussian assumptions of a mixed model.

Two different types of small area linear models were fitted to the sample data obtained in these Monte Carlo 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 fitted using the default settings of the *lme* function (Venables & Ripley, 2002, section 10.3) in the R software package. The M-quantile linear regression fit underpinning (b) was obtained using a modified version of the *rlm* function (Venables & Ripley, 2002, section 8.3) in R. Estimated model coefficients obtained from these fits were then used to compute a range of EBLUP and M-quantile-based estimators 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 estimators performed reasonably well. The differences between the estimators were much more pronounced under scenario 2 (area effects distributed as chi-squared). Here we see that the use of naïve estimators led to substantial biases as far as quantiles were concerned. In contrast, the estimators (both EBLUP and M-quantile) based on (6) and (10) were essentially unbiased, even for extreme quantiles, with the CD-based estimators somewhat more efficient. On the basis of these results it would appear that estimators that are defined as functionals of the CDF estimators (6) or (10) are preferable if there is concern about misspecification of the distribution of area effects.

TABLE 1 ABOUT HERE

TABLE 2 ABOUT HERE

In order to evaluate the performance of the linearization-based MSE estimator (13) and the bootstrap MSE estimator (14), we carried out a further model-based simulation study. In this study we focussed on MSE estimation for the 25th, 50th and 75th percentiles using the bootstrap estimator (14), and for the mean using either the linearization-based estimator (13) or the bootstrap estimator (14). A total of 200 Monte Carlo simulations were carried out for each

percentile and 100 Monte Carlo simulations for the mean, with the bootstrap MSE estimation implemented by generating a single bootstrap population at each Monte Carlo simulation and taking $L = 500$ bootstrap samples from this population. The bootstrap population was generated unconditionally, with bootstrap population values obtained by sampling from the smoothed residual distribution generated by the sample data obtained in each Monte Carlo simulation. Although it would have been theoretically preferable to have generated multiple bootstrap populations from each Monte Carlo sample, computing limitations restricted our investigation to $B = 1$. Since the estimates generated by the bootstrap procedure were then averaged over the 200 Monte Carlo simulations in our evaluation, this limitation is not as severe as it might appear to be, since the Monte Carlo simulations themselves serve as proxies for multiple bootstrap populations. Simulation results evaluating the resulting MSE estimators are set out in Tables 3 and 4 and in Figure 1. Focusing first on Table 3, we note that under both simulation scenarios, the linearization-based and the bootstrap MSE estimators tracked the true MSEs of the small area mean estimators very well, and provided coverage rates that were close to the nominal 95%.

TABLE 3 ABOUT HERE

Focusing next on Table 4 and Figure 1 we see that the bootstrap MSE estimator also performed well in terms of approximating the true MSEs of the small area quantile estimators. Again, coverage rates generated by 95% prediction intervals based on these MSE estimates were close to their nominal level.

TABLE 4 ABOUT HERE

FIGURE 1 ABOUT HERE

5.2 Design-based simulations

The synthetic population data on which these simulations were based are the same as those discussed in CT. They were obtained by nonparametrically bootstrapping an initial sample of

1652 Australian farms that responded to the Australian Agricultural and Grazing Industries Survey (AAGIS) up to a population of N = 81 982 farms spread across 29 agricultural regions of Australia, and referred to as the AAGIS dataset below. The variable of interest y is the Total Cash Costs (TCC) of the farm business in the reference year of the original AAGIS. Auxiliary information available for each farm in the population 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 from the (fixed) AAGIS dataset using both linear mixed models and linear M-quantile models. Five hundred independent samples were selected for the simulation. See CT for further details on how this was done and on the stratified sampling procedure, which replicated the regional distribution of the original sample farms. 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.

Estimated values of regional means were obtained using both naïve and CD-based estimators 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 estimators are identical to estimators based on the Rao, Kovar & Mantel (1990) CDF estimator (10) in this case. These simulation results are set out in Table 5, which shows relative bias and relative root mean squared error (both expressed in percentage terms) averaged over the 29 regions.

TABLE 5 ABOUT HERE

We immediately see that the naïve M-quantile estimator of the mean is biased. However, this bias effectively disappears from the CD-based version of this estimator, 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. This may be due to the violation of the mixed model assumptions. To illustrate this we present normal probability plots of level 1 (farm) and level 2 (Region) residuals that are based on fitting a two level linear mixed model to the original AAGIS sample data (Figure 2).

FIGURE 2 ABOUT HERE

These plots indicate that the model assumptions are not satisfied. Again we see that the bias of the naïve EBLUP estimator is corrected by using a CD version of the EBLUP estimator, though in this case there is no corresponding reduction in RMSE. Although we do not show these results, we also evaluated the EBLUP and M-quantile versions of the outlier-robust estimator (9), using 'huberised' residuals (based on a tuning constant of $c = 5$) to define the bias adjustment. As expected, both of these further improved on the RMSE performance of their corresponding 'standard' versions (7), but at the cost of increased negative bias.

Figure 3 shows the regional distributions of coverage rates of nominal 95% confidence intervals for regional means derived using the weighted version (12) of the CD-based M-quantile estimator and the linearization-based MSE estimator (13). In general, these intervals display good coverage rates, with significant under-coverage only in one region that contained an extremely large outlier. In Table 6 we further summarise the performance of (13) as an estimator of the MSE of (12) by comparing key percentiles of the distribution across areas of the Monte Carlo average value of (13) with the true (i.e. simulation-based) MSE of (12).

TABLE 6 ABOUT HERE

FIGURE 3 ABOUT HERE

These results indicate that (13) provides a good approximation to the true MSE of (12). In contrast, as reported in CT, the coverage rates of confidence intervals based on the naïve M-

quantile estimator show extensive undercoverage in this situation, which in this case is attributable to the bias of this estimator.

In addition to estimating regional means, we also estimated selected percentiles of the distribution of TCC within the different regions by numerically solving (11), using the estimators (5), (6), (8) and (10) of the within region CDF. Here we focus on the 10th percentile, the 50th percentile (the median) and the 90th percentile. Our results are summarized in Figure 4, where we see that, for both the 10th and the 90th percentile, the M-quantile and EBLUP versions of the naïve estimator (boxes 7 and 8) have larger absolute biases and root mean squared errors across the different regions than the corresponding estimators based on the smearing-type CDF estimators (6) and (10).

FIGURE 4 ABOUT HERE

As suggested in Section 3, the situation is reversed at the median, where the M-quantile/Naïve estimator performs the best. Generally, these results indicate that, for this population, using an estimator based on an M-quantile model (boxes 1 and 2) is preferable to using one based on a linear mixed model (boxes 3 and4), and that using the Rao, Kovar & Mantel (1990) estimator (10) (boxes 2 and 4) is preferable to using the CD-based estimator (boxes 1 and 3). The outlier-robust version (8) of the CD estimator (boxes 5 and 6) seems to offer no worthwhile efficiency gains in this case.

6. 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 use of smearing-type estimators of the small area CDF, which can then be used to define an estimator of the small area mean as well as estimators 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 estimator of the small area CDF that underpins our small area estimation framework are available, and lead to alternative estimators for small area characteristics. As we observed in Section 3, Welsh & Ronchetti (1998) have proposed an outlier robust version (8) of the CD estimator (6). A slightly different modification to (6) uses local (i.e. nonparametric) weighting in the smearing process, leading to

$$\hat{F}_j^{CD/np}(t) = N_j^{-1}\left[\sum_{i\in s_j} I(y_{ij} \le t) + \sum_{i\in s_j}\sum_{k\in r_j} w_{ik} I\left\{\hat{\mu}_{kj} + \left(y_{ij} - \hat{\mu}_{ij}\right) \le t\right\}\right], \tag{15}$$

where the w_{ik} are 'local' weights that satisfy, for non-sample unit *k* in area *j*,

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

It is easy to show that the mean estimator implied by (15) is

$$\hat{m}_j^{np} = N_j^{-1}\left\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j} \hat{y}_{ij} + \sum_{i\in s_j} u_{ij}(y_{ij} - \hat{y}_{ij})\right\}, \tag{16}$$

where

$$u_{ij} = \sum_{k\in r_j} w_{ik}\,.$$

We have not evaluated (16) in the context of small area estimation, but previous experience with it for robust population level estimation (Chambers, Dorfman & Wehrly 1993) indicates that it should also work well, particularly when there is significant non-linearity in the small area regression model.

Appendix

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

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

since $\int_{-\infty}^{\infty} t dI\left(y_i \le t\right) = y_i$. The expression (7) follows directly. Similarly, it is easy to see that, under simple random sampling, the estimator of the mean defined by the Rao, Kovar & Mantel CDF estimator (10) satisfies

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

References

ARAGON, Y., CASANOVA, S., CHAMBERS, R. & LECONTE, E. (2005). Conditional ordering using nonparametric expectiles. *J. Off. Statist.* **21**, 617-633.

BOWMAN, A.W., HALL, P. & PRVAN, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika* **85**, 799-808.

BRECKLING, J. & CHAMBERS, R.L. (1988). M-quantiles. *Biometrika* **75**, 761-771.

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

CHAMBERS, R.L., CHANDRA, H. & TZAVIDIS, N. (2008). On robust mean squared error estimation for linear predictors for domains. *Paper submitted for publication. A copy is available at* http://www.ccsr.ac.uk/publications/working/2007-10.pdf.

CHAMBERS, R.L., DORFMAN, A.H. & HALL, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika* **79**, 577-582.

CHAMBERS, R.L., DORFMAN, A.H. & WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *J. Amer. Statist. Assoc.* **88**, 268-277.

CHAMBERS, R.L. & DUNSTAN, R. (1986). Estimating distribution functions from survey data. *Biometrika* **73**, 597-604.

CHAMBERS, R.L. & TZAVIDIS, N. (2006). M-quantile models for small area estimation. *Biometrika*, **93**, 255-268.

DUAN, N. (1983). Smearing estimate: A nonparametric retransformation method. *J. Amer. Statist. Assoc.* **78**, 605-610.

FAY, R.E. & HERRIOT, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* **74**, 269-277.

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

KOKIC, P., CHAMBERS, R., BRECKLING, J. & BEARE, S. (1997). A measure of production performance. *J. Bus. Econ. Statist.* **15**, 445-451.

KOENKER, R. (2005). *Quantile Regression*. New York: Cambridge University Press.

KOENKER, R. & BASSETT, G. (1978). Regression quantiles. *Econometrica* **46**, 33-50.

KOENKER R. & D'OREY, V. (1987). Computing regression quantiles. *J. Roy. Statist. Soc. Ser. C* **36**, 383-393.

LI, Q. & RACINE, J.S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press.

LOMBARDÍA M.J., GONZÁLEZ-MANTEIGA W. & PRADA-SÁNCHEZ J.M. (2003). Bootstrapping the Dorfman-Hall-Chambers-Dunstan estimator of a finite population distribution function. *J. Nonpar. Statist.* **16***,* 63-90.

R DEVELOPMENT CORE TEAM (2008). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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

RAO, J.N.K., KOVAR, J.G. & MANTEL, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **77**, 365-375.

RICHARDSON, A.M. (1997). Bounded influence estimation in the mixed linear model. *J. Amer. Statist. Assoc.* **92**, 154-161.

RICHARDSON, A.M. & WELSH, A.H. (1995). Robust estimation in the mixed linear model. *Biometrics* **51**, 1429-1439.

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

SINHA, S.K. & RAO, J.N.K. (2009) Robust small area estimation. *Can. J. Statist.* **37**, 381-399.

VENABLES,W.N. & RIPLEY, B.D. (2002). *Modern Applied Statistics with S*. New York: Springer.

WANG, S. & DORFMAN, A.H. (1996). A new estimator of the finite population distribution function. *Biometrika* **83**, 639-652.

WELSH, A.H. & RONCHETTI, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *J. Roy. Stat. Soc. Ser. B* **60**, 413-428.

Table 1. Model-based simulation results for Scenario 1 (Gaussian area effects) averaged over 30 small areas. The target parameters are the small area means and selected percentiles of the small area distributions.

Method	Target Parameters					
	10th	25th	50th	Mean	75th	90th
	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-squared area effects) averaged over 30 small areas. The target parameters are the small area means and selected percentiles of the small area distributions.

Method	Target Parameters					
	10th	25th	50th	Mean	75th	90th
	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 error and average over Monte Carlo simulations of estimated mean squared error and coverage rates of nominal 95% confidence intervals for the M-quantile/CD estimator (12). Estimated mean squared errors based on (14) using the smoothed unconditional approach (Bootstrap) or (13) (Linearization). Intervals were defined as the M-quantile/CD estimator (12) plus or minus twice its estimated standard error, calculated as the square root of (13) or (14).

MSE	Percentiles of across areas distribution					
	Min	25th	50th	Mean	75th	Max
			Gaussian area effects			
True	0.271	0.331	0.411	0.419	0.481	0.783
Linearization	0.289	0.317	0.400	0.416	0.500	0.680
Bootstrap	0.282	0.319	0.401	0.418	0.504	0.715
Coverage Linearization	0.88	0.93	0.95	0.94	0.97	0.99
Coverage Bootstrap	0.88	0.94	0.96	0.96	0.97	0.99
			Chi-squared area effects			
True	0.344	0.453	0.549	0.589	0.736	1.051
Linearization	0.411	0.453	0.552	0.592	0.689	0.980
Bootstrap	0.398	0.444	0.559	0.589	0.706	1.003
Coverage Linearization	0.87	0.89	0.92	0.93	0.96	0.98
Coverage Bootstrap	0.92	0.95	0.96	0.96	0.97	1.00

Table 4. Across areas distribution of the true (i.e. Monte Carlo) mean squared error and average over Monte Carlo simulations of estimated mean squared error for the CD estimates of 0.25, 0.50 and 0.75 quantiles from (11). Estimated mean squared error for quantiles is based on (14) using smoothed unconditional approach.

MSE		Percentiles of across areas distribution					
		Min	25th	50th	Mean	75th	Max
				Gaussian area effects			
0.25 quantile	True	0.354	0.391	0.491	0.514	0.595	0.887
	Estimated	0.345	0.383	0.475	0.500	0.598	0.857
0.50 quantile	True	0.311	0.353	0.444	0.469	0.547	0.761
	Estimated	0.314	0.348	0.433	0.455	0.543	0.774
0.75 quantile	True	0.339	0.386	0.495	0.516	0.611	0.909
	Estimated	0.338	0.375	0.471	0.495	0.592	0.867
				Chi-squared area effects			
0.25 quantile	True	0.289	0.357	0.454	0.471	0.569	0.919
	Estimated	0.314	0.346	0.437	0.458	0.554	0.795
0.50 quantile	True	0.376	0.454	0.575	0.594	0.735	1.087
	Estimated	0.395	0.439	0.554	0.578	0.696	1.001
0.75 quantile	True	0.594	0.678	0.848	0.893	1.035	1.727
	Estimated	0.592	0.666	0.843	0.877	1.058	1.579

Table 5. Design-based simulation results for the AAGIS data: Estimation of average TCC within regions. Entries show regional averages of Relative Bias (RB) and Relative RMSE (RRMSE) for different small area estimators. Both RB and RRMSE are expressed in percentage terms.

Small Area Estimators	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 6. AAGIS data: Percentiles of the across regions distribution of the true (i.e. Monte Carlo) mean squared error of the M-quantile/CD estimator (12) of mean TCC within regions and the corresponding distribution of the average (over the Monte Carlo simulations) of its estimated MSE computed using the linearization estimator (13).

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

Figure 1. Distribution of area-specific coverage rates of nominal 95% confidence intervals for small area quantiles in the model-based simulations. Intervals were defined as the M-quantile/ CD estimator (11) plus or minus twice its estimated standard error, calculated as the square root of (14).

Figure 2. Normal probability plots of level 1 (left) and level 2 residuals (right) derived by fitting a two-level linear mixed model to the original AAGIS sample data.

Figure 3. AAGIS data: Distribution of region-specific coverage rates of nominal 95% confidence intervals. Intervals were defined as the CD-based M-quantile estimator (12) plus or minus twice its estimated standard error, calculated as the square root of (13).

Figure 4. AAGIS data: Box plots showing across-region distributions of average prediction error (left column) and root mean squared error (right column) for estimated percentiles (top = 10th, middle = median, bottom = 90th) of the within-region distribution of TCC. Boxes correspond to different estimators: 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 1

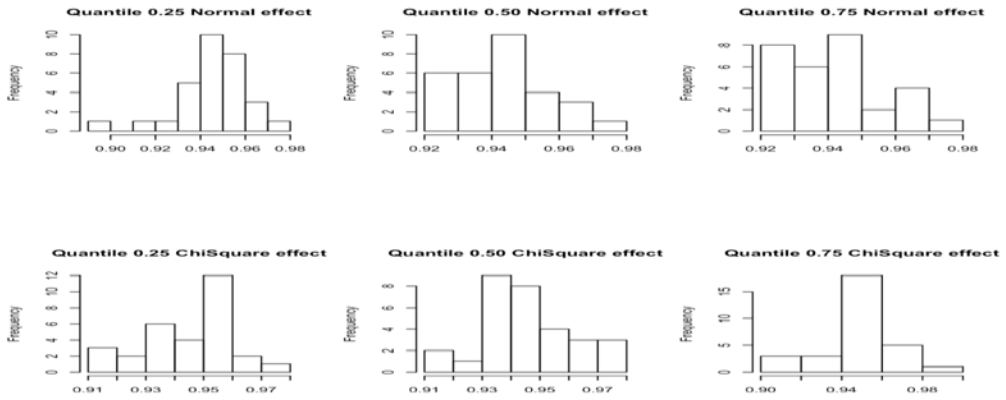

Quantile 0.25 Normal effect
Quantile 0.50 Normal effect
Quantile 0.75 Normal effect
Quantile 0.25 ChiSquare effect
Quantile 0.50 ChiSquare effect
Quantile 0.75 ChiSquare effect
Frequency

Figure 2

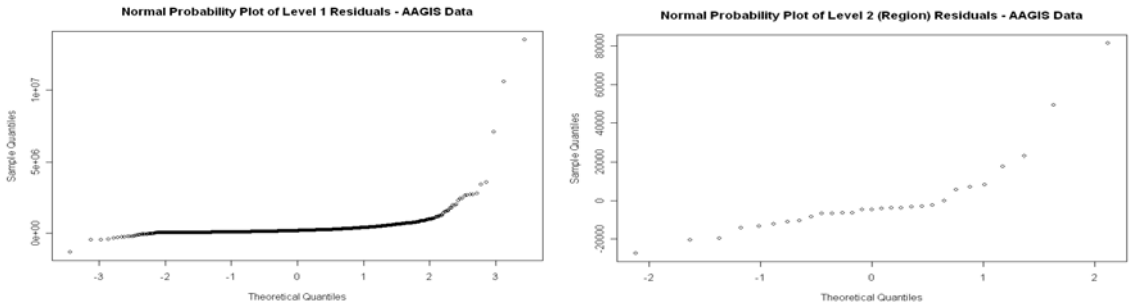

Normal Probability Plot of Level 1 Residuals - AAGIS Data
Sample Quantiles
0e+00
5e+06
1e+07
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Normal Probability Plot of Level 2 (Region) Residuals - AAGIS Data
Sample Quantiles
-20000
0
20000
40000
60000
80000
-2
-1
0
1
2
Theoretical Quantiles

Figure 3

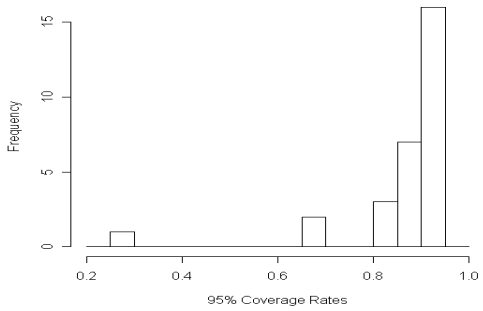
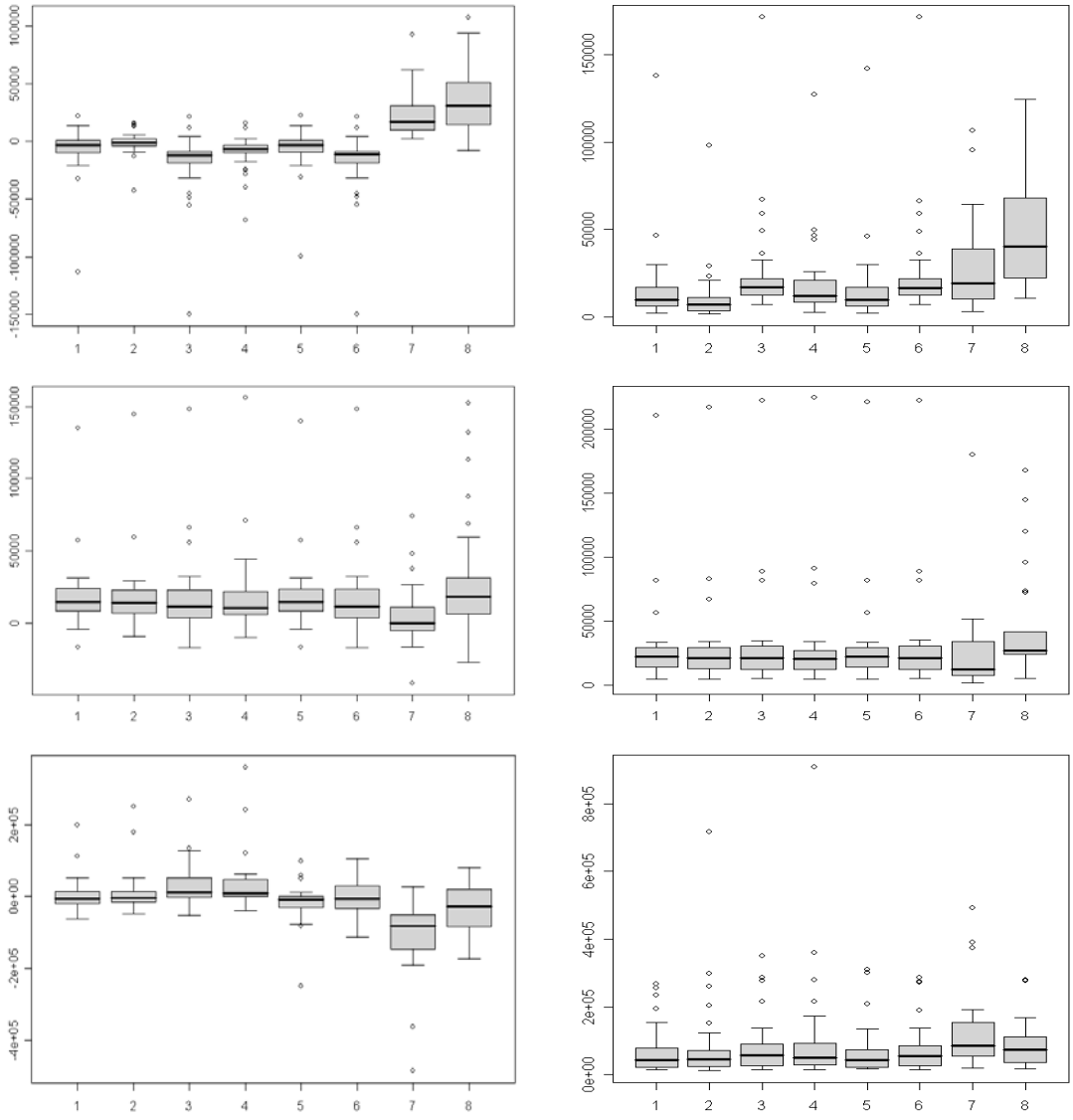


Figure 4


-150000 -100000 -50000 0 50000 100000
0 50000 100000 150000
0 50000 100000 150000
0 50000 100000 150000 200000
-4e+05 -2e+05 0e+00 2e+05
0e+00 2e+05 4e+05 6e+05 8e+05
1 2 3 4 5 6 7 8