# BIAS ADJUSTED ESTIMATION FOR SMALL AREAS WITH OUTLYING VALUES

**NIKOS TZAVIDIS, RAY CHAMBERS**

## ABSTRACT

Small area estimation techniques typically rely on regression models that use both covariates and random effects to explain between domain variation. Chambers and Tzavidis (2006) describe a novel approach to small area estimation that is based on modelling quantile-like parameters of the conditional distribution of the target variable given the covariates. This is an outlier robust approach that avoids conventional Gaussian assumptions and the problems associated with specification of random effects, allowing inter-domain differences to be characterized by the variation of area-specific M-quantile coefficients. These authors observed, however, that M-quantile estimates of small area means are biased with the magnitude of the bias being related to the presence of outliers in the data. In this paper we propose a bias adjustment to the M-quantile small area estimator of the mean that is based on representing this estimator as a functional of the small area distribution function. The method is then generalized for estimating other quantiles of the distribution function in a small area. The effect of this bias adjustment on small area estimation with random effects models in the presence of model misspecification is also examined..

Southampton Statistical Sciences Research Institute
Methodology Working Paper M06/09

# Bias Adjusted Estimation for Small Areas with Outlying Values

Nikos Tzavidis[1] and Ray Chambers[2]

1. *Centre for Longitudinal Studies, Institute of Education, University of London, UK*
2. *School of Mathematics and Applied Statistics, University of Wollongong, New South Wales, Australia*

## ABSTRACT

Small area estimation techniques typically rely on regression models that use both covariates and random effects to explain between domain variation. Chambers and Tzavidis (2006) describe a novel approach to small area estimation that is based on modelling quantile-like parameters of the conditional distribution of the target variable given the covariates. This is an outlier robust approach that avoids conventional Gaussian assumptions and the problems associated with specification of random effects, allowing inter-domain differences to be characterized by the variation of area-specific M-quantile coefficients. These authors observed, however, that M-quantile estimates of small area means are biased with the magnitude of the bias being related to the presence of outliers in the data. In this paper we propose a bias adjustment to the M-quantile small area estimator of the mean that is based on representing this estimator as a functional of the small area distribution function. The method is then generalized for estimating other quantiles of the distribution function in a small area. The effect of this bias adjustment on small area estimation with random effects models in the presence of model misspecification is also examined.

**Keywords**: Asymmetric data; Chambers-Dunstan estimator; Distribution estimation; Model misspecification; Quantile regression; Robust inference; 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 mixed (random) effects models that include area specific random effects to account for between area variation beyond that explained by the auxiliary information (Fay and Herriot 1979, Rao 2003). Such models depend on Gaussian assumptions and require a formal specification of the random effects structure. In a recent paper Chambers and Tzavidis (2006) proposed a new approach to small area estimation based on modelling quantile-like coefficients of the conditional distribution of the target variable given the covariates. With M-quantile models we avoid imposing strong distributional assumptions. Formal specification of the random part of the model is also not required. Instead, inter-domain variation is captured by variation in area-specific quantile coefficients. However, Chambers and Tzavidis (2006) also observed that M-quantile estimates of the small area means are biased, with the magnitude of the bias being related to the presence of outliers in the data. In this paper we propose a bias corrected M-quantile estimator of the small area mean. Our proposal is based on representing this estimator as a functional of the estimated

distribution function within the small area. The method is then generalized for estimating any quantile of the small area distribution function.

The structure of the paper is as follows: In Section 2 we review random effects models and M-quantile models for small area estimation. In Section 3 we propose a bias adjusted M-quantile estimator for the small area mean and extend this idea for estimating other quantiles of the small area population distribution function. In Section 4 we discuss approaches for estimating the mean squared error of the M-quantile-based small area estimators. In Section 5 we assess the performance of the different small area estimation methods using both simulated and real data. Finally, in Section 6 we summarize our main findings.

## 2. Models for Small Area Estimation

In what follows we assume that a vector of $p$ auxiliary variables $x_{ij}$ is known for each population unit $i$ in small area $j$ and that information for the variable of interest $y$ is available for units in the sample. The target is to use these data to estimate various area specific quantities, including (but not only) the small area mean $m_j$ of $y$.

The most popular method employs linear mixed effects models for this purpose. In the general case a linear mixed effects model has the following form

$$y_{ij} = x_{ij}^T \beta + z_{ij}^T \gamma_j + \varepsilon_{ij},\ i = 1, \ldots, n, j = 1, \ldots, d, \qquad (1)$$

where $\gamma_j$ denotes a vector of random effects and $z_{ij}$ denotes a vector of auxiliary variables whose values are known for all units in the population. Domain specific means are estimated by

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

where $s_j$ denotes $n_j$ sampled units in area $j$ and $r_j$ denotes the remaining $N_j - n_j$ units in the area. Estimator (2) is typically referred to as the Empirical Best Linear Unbiased Predictor (EBLUP) of $m_j$ (Henderson 1953). The role of the random effects in the model is to characterise differences in the conditional distribution of *y* given *x* between the small areas.

An alternative approach to small area estimation is based on the use of quantile or M-quantile regression models. 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 the $q^{th}$ conditional quantile *y* given *x* is $Q_q(\mathrm{x}) = x^T\beta_q$. The vector $\beta_q$ 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 "quantile-like" generalisation of regression based on influence functions (M-regression).

The M-quantile of order *q* for the conditional density of *y* given *x* is defined as the solution $Q_q(x;\psi)$ of the estimating equation $\int \psi_q(y-Q)\,f(y\mid x)dy = 0$, where $\psi$ denotes

the influence function associated with the M-quantile. A linear M-quantile regression model is 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 $\psi$, an estimate $\hat{\beta}_\psi(q)$ of $\beta_\psi(q)$ can be obtained by solving

$$\sum_{i=1}^{n} \psi_q(r_{iq\psi}) x_i = 0, \qquad (3)$$

where $r_{iq\psi} = y_i - x_i^T \hat{\beta}_\psi(q)$, $\psi_q(r_{iq\psi}) = 2\psi(s^{-1} r_{iq\psi}) \left\{ qI(r_{iq\psi} > 0) + (1-q) I(r_{iq\psi} \le 0) \right\}$ and *s* is a suitable robust estimate of scale for example, the MAD estimate $s = median\left| r_{iq\psi} \right| / 0.6745$.

Chambers and Tzavidis (2006) extended the use of M-quantile models to small area estimation. Following their development, we index population units by *i* and, following Kokic et.al (1997) and Aragon *et.al*. (2005), 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 $q_i$ such that $Q_{q_i}(x_i;\psi) = y_i$. 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. Consequently, if the conditional M-quantiles follow a linear model, with $\beta_\psi(q)$ a sufficiently smooth function of $q$, the following first order approximation holds

$$m_j = N_j^{-1} \left( \sum_{i \in s_j} y_i + \sum_{i \in r_j} x_i^T \beta_\psi(q_i) \right) \approx N_j^{-1} \left( \sum_{i \in s_j} y_i + \sum_{i \in r_j} x_i^T \beta_\psi(\theta_j) \right) + N_j^{-1} \sum_{i \in r_j} x_i^T \left( \frac{\partial \beta_\psi(\theta_j)}{\partial \theta_j} \right) \left( q_i - \theta_j \right).$$

Typically the first term on the right hand side of the above expression will dominate, suggesting a predictor of the form

$$\hat{m}_j = N_j^{-1}\left(\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T \beta_\psi(\hat{\theta}_j)\right), \qquad (4)$$

where a "hat" represents an estimator of the unknown quantity. Here $\hat{\theta}_j$ is 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.

## 3. A Bias Adjusted M-quantile Estimator for the Small Area Mean

We revisit small area estimation via mixed effects and M-quantile models using a unified estimation framework under which small area estimators are expressed as a functionals of the small area population distribution function.

Consider a finite population $P$ of $N$ units clustered within small areas of interest. For small area $j$ the area specific population distribution function is

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

The problem of estimating $F_j(t)$ essentially reduces to predicting the $y_{ij}$'s for the non-sampled units in small area $j$. This is achieved using a model suitable for small area estimation. Under a general linear model for small area estimation

$$\hat{F}_j(t) = N_j^{-1}\left(\sum_{i\in s_j} I(y_i \le t) + \sum_{i\in r_j} I\left(x_i^T \hat{\beta}_j \le t\right)\right), \qquad (6)$$

where $\hat{\beta}_j$ are the estimated model parameters for small area $j$. If we use an M-quantile model to predict $y$'s for out of sample units,

$$\hat{F}_j(t) = N_j^{-1}\left(\sum_{i\in s_j} I(y_i \le t) + \sum_{i\in r_j} I\left(x_i^T \hat{\beta}_\psi(\hat{\theta}_j) \le t\right)\right)$$

and the Chambers and Tzavidis (2006) estimator of the small area mean (4) is obtained as

$$\hat{m}_j = \int_{-\infty}^{\infty} t d\hat{F}_j(t) = N_j^{-1}\left(\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T \hat{\beta}_\psi(\hat{\theta}_j)\right). \qquad (7)$$

The same is true when a mixed effects model is used

$$\hat{F}_j(t) = N_j^{-1}\left(\sum_{i\in s_j} I(y_i \le t) + \sum_{i\in r_j} I\left(x_i^T \hat{\beta} + z_i^T \hat{\gamma}_j \le t\right)\right)$$

and the EBLUP estimator (2) is obtained as

$$\hat{m}_j = \int_{-\infty}^{\infty} t d\hat{F}_j(t) = N_j^{-1}\left(\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T \hat{\beta} + z_i^T \hat{\gamma}_j\right). \qquad (8)$$

Chambers and Tzavidis (2006) observed that (4) can produce biased estimates of small area means, particularly when small areas contain outliers. Hereafter, we refer to small area estimators derived under (6) as naïve estimators. Insight as to what might cause this bias is provided below.

Chambers (1986) considered the problem of estimating the finite population total, $T$, in the presence of representative outliers. The term representative, as opposed to non-representative, outliers is used to characterise observations that are correct but extreme relatively to the bulk of the data. It is well known that the Best Linear Unbiased Predictor of the finite population total of $y$ is (Royall 1970)

$$T_{LS} = \sum_{i \in s} y_i + \beta_{LS} \sum_{i \in r} x_i \,, \qquad (9)$$

where $\beta_{LS}$ is the generalised least squares estimator. However, it is also well known that $T_{LS}$ is sensitive to outliers. A first step in making (9) less sensitive to outliers might be to replace $\beta_{LS}$ by an outlier robust alternative. Although this approach stabilizes the variance of (9) in the presence of outliers, it does not address the problem of robust prediction of $T_{LS}$ in the presence of outliers leading to bias in the estimation of the total. Chambers (1986) proposes the use of an alternative estimator $T_n$ such that the distribution of the prediction error $T_n - T$ is unaffected by sample outliers. The general form of this estimator is

$$T_n = \sum_{i \in s} y_i + \beta \sum_{i \in r} x_i + \sum_{i \in s} \psi(y_i - \beta x_i) \,. \qquad (10)$$

Chambers's (1986) proposal suggests that $T_{LS}$ can be made more outlier robust by curtailing the influence of sample outliers based on the third term in the right hand side of (10). The robustness of (10) depends on the choice of $\beta$ and $\psi$.

Closely related to the work of Chambers (1986) is the work of Chambers and Dusntan (1986), hereafter denoted in formulae with subscript CD. These authors proposed an estimator of the distribution function, which under a general model and without any reference to the small area problem is of the following form

$$\hat{F}_{CD}(t) = N^{-1} \left\{ \sum_{i \in s} I(y_i \le t) + n^{-1} \sum_{i \in s} \sum_{k \in r} I\left[ x_k^T \hat{\beta} + (y_i - \hat{y}_i) \le t \right] \right\}, \qquad (11)$$

where $\hat{\beta}$ are the estimated model parameters and $\hat{y}_i = x_i^T \hat{\beta}$. The Chambers-Dunstan estimator of the distribution function is a bias adjusted version of (6). The adjustment is

by the residual $y_i - \hat{y}_i$. Welsh and Ronchetti (1998) considered the problem of estimating the population distribution function in the presence of outliers. To achieve this, they combine estimators of the form of (10) with the Chambers-Dunstan estimator of the distribution function.

Following these authors we propose a biased adjusted M-quantile estimator of the small area mean in the presence of outliers by combining the M-quantile small area model with the Chambers-Dunstan estimator. In this case an estimator of the population distribution function of small area *j* is

$$\hat{F}_{CD,j}(t) = N_j^{-1}\left\{\sum_{i\in s_j} I(y_i \le t) + n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j} I\left[x_k^T\hat{\beta}_\psi(\hat{\theta}_j) + (y_i - \hat{y}_i) \le t\right]\right\},$$

where $\hat{y}_i = x_i^T\hat{\beta}_\psi(\hat{\theta}_j)$. The proposed biased adjusted estimator of the small area mean is then

$$\hat{m}_j^{adj} = \int_{-\infty}^{\infty} t d\hat{F}_{CD,j}(t) = N_j^{-1}\left(\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T\hat{\beta}_\psi(\hat{\theta}_j) + \frac{N_j - n_j}{n_j}\sum_{i\in s_j}[y_i - \hat{y}_i]\right). \quad (12)$$

The derivation of (12) is given in the appendix. Estimator (4) adapts only the first step of the Chambers (1986) proposal i.e. the use of a robust $\beta$ such as the one estimated under an M-quantile model. However, this step does not protect us against the bias introduced when estimating the mean in the presence of outliers. In contrast, the proposed biased adjusted M-quantile estimator (12) is of the Chambers (1986) form where $\psi$ is the identity function. Using different definitions for the $\psi$ function, alternative bias-adjusted small area estimators of the small area mean are possible. Such estimators are considered in the empirical evaluations in section 5. An alternative, heuristic, approach to reducing the bias in the M-quantile estimate of the small area mean is to use expectile regression

(Newey and Powell 1987). To achieve this one can increase the tuning constant of the influence function $\psi$, $c \to \infty$. Expectile versions of (4) are also included in the empirical evaluations.

Although our main aim is to develop a bias adjusted M-quantile estimator of the small area mean, two further extensions are possible. Firstly, a modified version of the EBLUP estimator (2) is proposed by combining the mixed effects model (1) with the Chambers-Dunstan estimator

$$\hat{m}_j^{adj} = \int_{-\infty}^{\infty} t d\hat{F}_{CD,j}(t) = N_j^{-1}\left(\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T\hat{\beta} + z_i^T\hat{\gamma}_j + \frac{N_j - n_j}{n_j}\sum_{i\in s_j}[y_i - \hat{y}_i]\right), \qquad (13)$$

where $\hat{y}_i = x_i^T\hat{\beta} + z_i^T\hat{\gamma}_j$. It is well known that when the assumptions of the mixed model hold, (2) is the Empirical Best Linear Unbiased Predictor. It is of interest, however, to examine the usefulness of (13) when the model assumptions are wrongly specified. Secondly, the approach that we followed for defining the M-quantile bias adjusted estimator leads naturally to resolving the problem of estimating other quantiles, $q_j \in (0,1)$, of the population distribution function in a small area. This can be achieved using either an M-quantile or a mixed effects model and the Chambers-Dunstan estimator or estimator (6). In the former case a numerical solution to the following is required

$$\int_{-\infty}^{m_j} d\hat{F}_{CD,j}(t) = q_j . \qquad (14)$$

## 4. MSE Estimation

For fixed $q$, the estimator of the M-quantile regression coefficient $\beta_\psi(q)$ is $\hat{\beta}_\psi(q) = (X_s^T W_s(q) X_s)^{-1} X_s^T W_s(q) y_s$, where $X_s$ denotes the $n \times p$ matrix of sample

covariate values and $y_s$ is the *n*-vector of sample *y*-values. The diagonal matrix $W_s(q)$ contains the final set of weights produced by the iteratively reweighted least squares algorithm used to compute $\hat{\beta}_\psi(q)$. It immediately follows that $\hat{m}_j^{adj} = N_j^{-1} w_j^T y_s$, where $w_j$ is the vector of area *j* weights

$$w_j = \frac{N_j}{n_j} 1_{sj} + W_s(\hat{\theta}_j) X_s (X_s^T W_s(\hat{\theta}_j) X_s)^{-1} \left( t_{rj} - \frac{N_j - n_j}{n_j} t_{sj} \right). \qquad (15)$$

Here $1_{sj}$ is the *n*-vector with $i^{th}$ component equal to one whenever the corresponding sample unit is in area *j* and is zero otherwise, $t_{rj}$ is the sum of the non-sample covariate values in area *j* and $t_{sj}$ is the sum of the sample covariate values in area *j*.

We use the fact that the M-quantile estimator of $\beta_\psi(q)$ is linear in the sample values of *y* to develop an estimator of the mean squared error of the naïve M-quantile estimator. Note that our approach assumes $\hat{\theta}_j$ is constant, which leads to a first order approximation to the actual mean squared error. Mean squared error estimation of $\hat{m}_j^{adj}$ can be carried out using standard methods for robust estimation of the mean squared error of unbiased weighted linear estimators (Royall and Cumberland, 1978). That is, the prediction variance can be approximated by

$$var(\hat{m}_j^{adj} - m_j^{adj}) \approx N_j^{-2} \left( \sum_{i \in s_j} u_{ij}^2 var(y_i) + \sum_{i \in r_j} var(y_i) \right),$$

with $u_j = (u_{ij}) = \frac{N_j w_j - \sum w_j}{\sum w_j}$ (see also Chandra and Chambers 2005). We interpret $var(y_i)$ conditionally (i.e. specific to the area *j* from which $y_i$ is drawn) and hence

replace $var(y_i)$ in the first (sample) term on the right hand side above by $(y_i - x_i^T \hat{\beta}_\psi(\hat{\theta}_j))^2$ and the second term by

$$(N_j - n_j)(n_j - 1)^{-1} \sum_{i \in s_j} \{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_j)\}^2 .$$

Our estimator of the prediction variance in area *j* is therefore

$$\hat{V}_j = \sum_j \sum_{i \in s_j} \lambda_{ij} (y_i - x_i^T \hat{\beta}_\psi(\hat{\theta}_j))^2 , \tag{16}$$

where $\lambda_{ij} = N_j^{-2} \left( u_{ij}^2 + I(i \in j)(N_j - n_j)/(n_j - 1) \right)$. Next suppose $E(y_i \mid x_i, i \in j) = x_i^T \beta_j$. Then

$$E\left( N_j^{-1} \sum_{i \in s} w_{ij} y_i - m_j \right) \approx N_j^{-1} \left( \sum_j \sum_{i \in s_j} w_{ij} x_i^T \beta_j - \sum_{i \in j} x_i^T \beta_j \right),$$

where the summation over *k* on the right hand side above is over the set of areas represented in the sample. An estimate of the bias is

$$\hat{B}_j = N_j^{-1} \left( \sum_j \sum_{i \in s_j} w_{ij} x_i^T \hat{\beta}(\hat{\theta}_j) - \sum_{i \in j} x_i^T \hat{\beta}(\hat{\theta}_j) \right). \tag{17}$$

Our final estimator of the mean squared error of is therefore

$$\hat{M}_j = \hat{V}_j + \hat{B}_j^2 . \tag{18}$$

The proposed mean squared error estimator is similar to the mean squared estimator of the naïve M-quantile estimator proposed by Chambers and Tzavidis (2006). The difference is that for estimating the mean squared error of the naïve M-quantile estimator, instead of using the weights given by (15), we there use the weights

$$w_j = 1_{sj} + W_s(\hat{\theta}_j) X_s (X_s^T W_s(\hat{\theta}_j) X_s)^{-1} t_{rj} . \tag{19}$$

## 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 presented in section 3. The first is a model-based simulation in which small area population and sample data were simulated based on a two level hierarchical linear model with different parametric assumptions for the level one and level two variance components. The second is a design based simulation in which a fixed population containing a number of small areas was repeatedly sampled, holding the sample size in each small area fixed.

*5.1 Model –based Simulations*

In each simulation we generated $N$ = 232,500 population values of $x$ and $y$ in $H$ = 30 small areas with $N_h = 500h$ in area $h$. For each area $h$ we took 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 -small area means and other quantiles of the small area distribution function- of $y$ and the resulting estimation errors. This process was repeated 1000 times. Two scenarios for generating the data were used

**Scenario 1**: $x_{ih} \sim N(\mu_h, \mu_h^2/36)$, $\gamma_h \sim N(0,1)$, $\varepsilon_{ih} \sim N(0,64)$, with $\mu_h \sim U[40,120]$ held fixed over simulations.

**Scenario 2**: $x_{ih} \sim \chi^2(d_h)$, $\varepsilon_{ih} \sim \chi^2(3)$, $\gamma_h \sim \chi^2(1)$, with the $\varepsilon_{ih}$'$s$ and $\gamma_h$'$s$ centred around their means, and $d_h \sim U[1,200]$, held fixed over simulations. This second scenario is used to examine the effect of mis-specifying the Gaussian assumptions of a random effects model.

Population $y_{ij}$ values were then generated using $y_{ih} = 5 + x_{ih} + \gamma_h + \varepsilon_{ih}$. Two different methods of small area estimation were applied to the sample data obtained in the simulations, based on fitting linear models under (a) a random intercepts specification and (b) an M-quantile specification. The random intercepts model in (a) is based on fitting a linear mixed model to the sample data using the default settings of the *lme* function (Venables & Ripley, 2002, section 10.3) in R. The M-quantile regression fit underpinning (b) was obtained using a modified version of the *rlm* function (Venables & Ripley, 2002, section 8.3) in R.

For estimating small area means and other quantiles of the small area distribution function we employed either estimator (6) or the Chambers-Dunstan estimator (11). In general, we refer to estimators derived under (6) as the M-quantile naïve and the random intercepts naïve estimators and estimators derived under (11) as the M-quantile and the random intercepts bias adjusted estimators.

Biases and mean squared errors over these simulations, averaged over the 30 areas, are set out in Table 1 (under normality assumptions) and in Table 2 (under chi-square assumptions). When the model assumptions hold all approaches perform reasonably well. The Chambers-Dunstan adjustment offers bias correction mainly when estimating quantiles other than the median but the differences are not very pronounced. The differences between the naïve and the Chambers-Dunstan estimators are more pronounced when data are generated under chi-square assumptions. The use of the naïve estimator leads to biased estimates of the quantiles of the small area distribution function. In contrast, the Chambers-Dunstan estimator bias corrects the unadjusted estimators. This is true both for the M-quantile and the random intercepts model. In general, when the

model assumptions are not met the bias adjusted versions of the M-quantile and the random effects estimators perform well both in bias and mean squared error terms.

In Table 3 we report coverage rates of confidence intervals for the regional mean estimates based on the M-quantile naïve and the M-quantile bias adjusted estimators and the mean squared error estimator (18) with weights given by (19) or the mean squared error estimator (18) with weights given by (15) respectively. We conclude that the mean squared estimator of the bias adjusted M-quantile estimator of the mean exhibits good coverage properties.

*5.2 Design-Based Simulations*

The data on which these simulations were based were obtained from a sample of 1652 broadacre farms spread across 29 regions (Region) of Australia. This is the same dataset as the one employed by Chambers and Tzavidis (2006). We decided to use the same dataset in order to examine any potential gains from using the proposed biased adjusted small area estimators. 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 (1 = pastoral zone and a farm area of 50000 hectares or less; 2 = pastoral zone and a farm area of more than 50000 hectares; 3 = wheat-sheep zone and a farm area of 1500 hectares or less; 4 = wheat-sheep zone and a farm area of more than 1500 hectares; 5 = high rainfall zone and a farm area of 750 hectares or less; 6 = high rainfall zone and a farm area of more than 750 hectares). Individual (farm) level values for FarmArea, SizeZone and Region were assumed known at the population level.

The aim of this simulation study was to compare estimation of regional means of TCC under repeated sampling using both mixed effects models and M-quantile models. The study itself was implemented in two steps as follows: (1) A population of $N$ = 81982 farms was created by sampling $N$ times with replacement from the above sample of 1652 farms and with probability proportional to a farm's sample weight. Scatterplots of the distribution of $y$ and $x$ in this population show that it is highly heteroskedastic, with many outlying values. (2) Five hundred independently stratified random samples of the same size as the original sample were selected from this simulated population. Stratum (i.e. region) sample sizes were fixed to be the same as in the original sample. The same specification was used by all estimation methods, defined by the main effects and interactions for the Farmarea and SizeZone variables.

Small area mean estimates were obtained using a range of naïve and bias-adjusted estimators based both on the M-quantile and the random intercepts approaches. More specifically, under the M-quantile approach we can change the tuning constant of the influence function. For example when $c \to \infty$ we obtain expectile versions of the small area estimators. In addition, under the Chambers-Dunstan estimator of the distribution function one can use either an estimate of the raw residuals or an estimate of the huberized residuals for reducing the effect of large residuals. We consider both expectile versions -referred to as expectile estimators- of the M-quantile small area estimators as well as small area estimators that are derived using the Chambers-Dunstan estimator and either raw or huberized residuals -referred to as huberized estimators-.

The results set out in Table 4 focus on estimation of regional means under different M-quantile small area estimators. These show that the naïve M-quantile estimator of the

small area mean is severely biased (see also Chambers and Tzavidis 2006). The bias reduces significantly once we consider a set of alternative M-quantile estimators that are based on the use of the Chambers-Dunstan estimator of the distribution function. The adjusted M-quantile estimators exhibit good performance also in terms of relative root mean squared error. An interesting picture emerges for estimators based on mixed effects models (Table 5). Firstly, we expect that the naïve random intercepts estimator is not optimal any more because the model assumptions are violated due to the presence of outliers in the data. The use of adjusted random intercepts estimators offer a clear improvement.

Finally, in Table 6 we report coverage rates of confidence intervals for regional mean estimates based on the M-quantile bias adjusted estimator and the mean squared error estimator (18) with weights given by (15). We see that in general we derive good coverage rates, which can be attributed to the fact that we bias correct the naïve M-quantile estimator. Significant under-coverage still exists for 3 areas where we know that large outliers exist.

## 6. Summary

In the present paper we propose a bias adjustment to the naïve M-quantile estimator of the small area mean that is based on the Chambers-Dunstan estimator of the population distribution function. The bias-adjusted M-quantile estimator is more efficient than the naïve M-quantile estimator particularly in the presence of outliers. We further illustrate that the use of a Chambers-Dunstan adjustment may improve the estimation of small area means obtained from a random intercepts model when the assumptions of such models are violated. The problem of estimating other quantiles of the small area distribution

function is considered and results indicate that this can be achieved by employing the Chambers-Dunstan estimator of the distribution function either with M-quantile or with mixed effects models.

## Acknowledgements

The research reported in this paper was carried out with the support of grant H333250030 of the Economic and Social Research Council.

## References

Aragon, Y, Casanova, S., Chambers, R. and Leconte, E. (2005). Conditional ordering using nonparametric expectiles. *Journal of Official Statistics*, **21** (4), 617-33.

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika* **75**, 761-71.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association,* **81**, 1063-69.

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

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

Chandra, H. and Chambers, R. (2005). Comparing EBLUP and C-EBLUP for small area estimation, *[submitted to the Journal of Official Statistics]*.

Fay, R.E. and Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association,* **74**, 269-77.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226-52.

Kokic, P., Chambers, R., Breckling, J.and Beare, S. (1997). A measure of production performance. *Journal of Business and Economic Statistics,* **15**, 445-51

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**, 383-93.

Newey, W.K. & Powell, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, **55**, 819-47.

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

Royall, R.M. and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association,* **73**, 351 - 8.

Welsh, A.H. and Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers**,** *Journal of the Royal Statistical Society B,* **60**, 413-28.

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

## APPENDIX

### Derivation of the small area estimator of the mean under the Chambers-Dunstan estimator of the distribution function

Let $\hat{F}(t)$ denote an estimator of the sample empirical distribution.

$$\hat{F}(t) = n^{-1}\left(\sum_{i\in s} I(y_i \le t)\right)$$

An estimator of the mean is then given by

$$\hat{m} = \int_{-\infty}^{\infty} t d\hat{F}(t) = n^{-1}\sum_{i\in s} y_i \text{ since}$$

$$\int_{-\infty}^{\infty} t dI\left(y_i \le t\right) = y_i \tag{A1}$$

We now employ the Chambers-Dunstan estimator of the small area distribution function

$$\hat{F}_{CD,j}(t) = N_j^{-1}\left\{\sum_{i\in s_j} I(y_i \le t) + n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j} I\left[x_k^T\hat{\beta}_\psi(\hat{\theta}_j) + \left(y_i - \hat{y}_i\right) \le t\right]\right\}.$$

An estimator of the small area mean under the Chambers-Dunstan estimator is then given by

$$\hat{m}_j = \int_{-\infty}^{\infty} t d\hat{F}_{CD,j}(t) = N_j^{-1}\int_{-\infty}^{\infty} t d\left\{\sum_{i\in s_j} I(y_i \le t) + n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j} I\left[x_k^T\hat{\beta}_\psi(\hat{\theta}_j) + \left(y_i - \hat{y}_i\right) \le t\right]\right\}$$

which can be expressed as

$$\hat{m}_j = N_j^{-1}\int_{-\infty}^{\infty} t d\sum_{i\in s_j} I(y_i \le t) + n_j^{-1} t d\sum_{i\in s_j}\sum_{k\in r_j} I\left[x_k^T\hat{\beta}_\psi(\hat{\theta}_j) \le t\right] + n_j^{-1} t d\sum_{i\in s_j}\sum_{k\in r_j} I\left[\left(y_i - \hat{y}_i\right) \le t\right]$$

Applying (A1) to the previous expression we obtain the adjusted estimator of the small area mean (12)

$$\hat{m}_j^{adj} = N_j^{-1}\left(\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T\hat{\beta}_\psi(\hat{\theta}_j) + \frac{N_j - n_j}{n_j}\sum_{i\in s_j}[y_i - \hat{y}_i]\right)$$

## MODEL -BASED SIMULATIONS

*Table 1. Model-based simulation results of estimating a range of parameters of the small area distribution function when data are generated under Gaussian assumptions, averaged over 30 small areas.*

| Method | q10 | q25 | q50 | Mean | q75 | q 90 |
|---|---|---|---|---|---|---|
| | | | **Relative Bias (%)** | | | |
| Random Intercepts Naïve | 0.088 | 0.041 | -0.002 | -0.002 | -0.036 | -0.062 |
| Random Intercepts CD | 0.083 | 0.046 | 0.051 | -0.002 | 0.072 | 0.160 |
| 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 |
| | | | **RRMSE (%)** | | | |
| Random Intercepts Naïve | 0.36 | 0.29 | 0.25 | 0.23 | 0.24 | 0.24 |
| Random Intercepts CD | 0.42 | 0.31 | 0.27 | 0.24 | 0.26 | 0.31 |
| M-quantile Naïve | 0.57 | 0.48 | 0.42 | 0.41 | 0.39 | 0.38 |
| M-quantile CD | 0.43 | 0.32 | 0.27 | 0.24 | 0.26 | 0.30 |

*Table 2. Model-based simulation results of estimating a range of parameters of the small area distribution function when data are generated under chi-square assumptions, averaged over 30 small areas.*

| Method | q10 | q25 | q50 | Mean | q75 | q 90 |
|---|---|---|---|---|---|---|
| | | | **Relative Bias (%)** | | | |
| Random.Intercepts Naïve | 22.48 | 9.731 | 0.420 | 0.024 | -4.708 | -6.969 |
| Random .Intercepts CD | 0.374 | 0.205 | 0.079 | -0.018 | -0.073 | -0.186 |
| M-quantile Naïve | 17.24 | 5.653 | -2.641 | -1.794 | -7.021 | -8.787 |
| M-quantile CD | 0.373 | 0.176 | 0.027 | -0.018 | -0.085 | -0.188 |
| | | | **RRMSE (%)** | | | |
| Random.Intercepts Naïve | 23.11 | 10.81 | 3.62 | 1.97 | 5.60 | 7.44 |
| Random.Intercepts CD | 4.09 | 3.87 | 3.78 | 2.01 | 4.19 | 4.84 |
| M-quantile Naïve | 17.69 | 7.31 | 4.49 | 2.49 | 7.68 | 9.18 |
| M-quantile CD | 4.09 | 3.88 | 3.93 | 2.01 | 4.36 | 4.82 |

*Table 3. Simulated data under normality assumptions, coverage rates of 'two-sigma' confidence intervals. Intervals are defined by the M-quantile naïve and M-quantile bias-adjusted estimates plus or minus twice their corresponding standard errors using (18) with weights given by (19) or (18) with weights given by (15) respectively*

| Area | Coverage Rates for naïve M-quantile estimator | Coverage Rates for bias adjusted M-quantile estimator |
|---|---|---|
| 1 | 100 | 100 |
| 2 | 93 | 99 |
| 3 | 90 | 98 |
| 4 | 95 | 100 |
| 5 | 91 | 99 |
| 6 | 91 | 100 |
| 7 | 89 | 100 |
| 8 | 88 | 96 |
| 9 | 80 | 96 |
| 10 | 84 | 94 |
| 11 | 85 | 94 |
| 12 | 83 | 96 |
| 13 | 86 | 97 |
| 14 | 87 | 93 |
| 15 | 79 | 96 |
| 16 | 82 | 93 |
| 17 | 84 | 97 |
| 18 | 81 | 91 |
| 19 | 78 | 97 |
| 20 | 80 | 97 |
| 21 | 83 | 97 |
| 22 | 87 | 91 |
| 23 | 81 | 93 |
| 24 | 88 | 93 |
| 25 | 81 | 94 |
| 26 | 85 | 97 |
| 27 | 87 | 89 |
| 28 | 77 | 95 |
| 29 | 82 | 96 |
| 30 | 84 | 94 |
| **Mean** | **85.37** | **95.73** |

## DESIGN-BASED SIMULATIONS

*Table 4. Australian farms study, relative bias and relative root mean squared error estimates of regional means of Total Cash Costs under a range of M-quantile (Mq) small area estimators in design-based simulation study. Row entries correspond to a region with the last row reporting the column mean.*

| Mq Naïve | Mq Expectile Naive | Mq CD Huberized c=5 | Mq Expectile CD | Mq CD | Mq Naive | Mq Expectile Naive | Mq CD Huberized c=5 | Mq Expectile CD | Mq CD |
|---|---|---|---|---|---|---|---|---|---|
| **Relative Bias** | | | | | **Relative Root Mean Squared Error** | | | | |
| -14.20 | -10.95 | -1.29 | -0.21 | -0.43 | 15.88 | 13.72 | 11.49 | 12.31 | 11.96 |
| -31.29 | -25.13 | -17.36 | -0.18 | -0.16 | 31.50 | 25.42 | 18.71 | 31.26 | 31.32 |
| -15.17 | -9.09 | -6.84 | -0.38 | -0.33 | 15.58 | 9.83 | 9.02 | 9.73 | 9.83 |
| -23.76 | -18.5 | -6.22 | -0.66 | -0.50 | 24.07 | 19.05 | 8.81 | 9.03 | 9.46 |
| -16.18 | -9.11 | -6.31 | -0.18 | -0.15 | 16.81 | 10.31 | 9.27 | 9.14 | 9.28 |
| 10.87 | 18.93 | -1.71 | 0.54 | 0.46 | 14.21 | 21.58 | 15.94 | 18.15 | 17.30 |
| -12.27 | -5.31 | -4.35 | 0.26 | 0.23 | 13.19 | 7.60 | 7.92 | 10.08 | 10.34 |
| -14.89 | -8.70 | -2.66 | 0.01 | 0.00 | 15.56 | 10.51 | 8.28 | 9.21 | 9.21 |
| -26.39 | -20.14 | -23.53 | -1.42 | -1.46 | 26.74 | 21.35 | 24.62 | 104.31 | 104.02 |
| -19.05 | -13.48 | -5.84 | 0.05 | -0.02 | 19.41 | 14.04 | 8.94 | 9.23 | 9.58 |
| -22.20 | -12.16 | -7.28 | -2.99 | -2.53 | 31.89 | 25.64 | 38.82 | 37.84 | 38.64 |
| -10.71 | 2.95 | -6.96 | 3.21 | 0.38 | 13.05 | 16.10 | 12.35 | 23.71 | 16.68 |
| -22.90 | -21.38 | -4.45 | -0.17 | 0.04 | 23.92 | 24.09 | 14.48 | 17.04 | 14.92 |
| -16.61 | -15.85 | -0.25 | 0.11 | 0.14 | 17.67 | 16.86 | 7.59 | 7.77 | 7.46 |
| -14.56 | -7.18 | -6.29 | -0.93 | 0.02 | 15.44 | 9.33 | 10.97 | 16.08 | 18.78 |
| -15.87 | -3.06 | -3.61 | 0.73 | -0.10 | 17.18 | 12.12 | 8.91 | 9.93 | 9.00 |
| -4.43 | 12.34 | -2.62 | 0.27 | 0.24 | 12.65 | 18.27 | 13.98 | 14.32 | 14.20 |
| -21.31 | 0.72 | -15.68 | 10.80 | -3.76 | 37.38 | 40.10 | 34.14 | 41.95 | 34.44 |
| -12.90 | -0.20 | -3.10 | 1.09 | 0.45 | 14.81 | 10.03 | 9.45 | 13.77 | 11.32 |
| 2.84 | 12.30 | 0.53 | 0.46 | 0.53 | 7.02 | 13.74 | 7.01 | 7.36 | 7.01 |
| -8.20 | -1.59 | -1.34 | 0.34 | 0.17 | 9.11 | 4.75 | 6.54 | 7.65 | 7.05 |
| -21.29 | -14.96 | -8.98 | -0.41 | -0.45 | 21.78 | 15.74 | 11.97 | 8.85 | 9.08 |
| -14.12 | -0.13 | 1.18 | 1.67 | 1.76 | 24.32 | 24.36 | 25.96 | 32.22 | 25.48 |
| 6.81 | 15.91 | 0.22 | -0.22 | 0.25 | 15.51 | 20.98 | 11.43 | 12.85 | 11.41 |
| -29.46 | -24.31 | -11.01 | -1.71 | -0.31 | 29.62 | 24.54 | 12.52 | 11.79 | 15.86 |
| -10.06 | -3.77 | 3.85 | 3.12 | 2.05 | 13.74 | 11.43 | 11.80 | 15.42 | 12.54 |
| -30.97 | -25.09 | -10.39 | -0.57 | -0.59 | 31.23 | 25.43 | 13.34 | 10.67 | 11.12 |
| -31.51 | -24.87 | -8.60 | -1.06 | -0.43 | 31.90 | 25.39 | 13.22 | 15.23 | 17.06 |
| -29.13 | -14.65 | -13.98 | -1.23 | -1.37 | 30.88 | 23.47 | 19.15 | 23.13 | 24.32 |
| **-16.17** | **-7.81** | **-6.03** | **0.36** | **-0.20** | **20.41** | **17.78** | **14.02** | **18.97** | **18.23** |

*Table 5. Australian farms study, relative bias and relative root mean squared error estimates of regional means of Total Cash Costs under a range of random effects small area estimators in design-based simulation study. Row entries correspond to a region with the last row reporting the column mean.*

| Random Intercepts Naive | Random Intercepts CD | Random Intercepts CD Huberized c=5 | Random Intercepts Naive | Random Intercepts CD | Random Intercepts CD Huberized c=5 |
|---|---|---|---|---|---|
| **Relative Bias** | | | **Relative Root Mean Squared Error** | | |
| -4.59 | -0.21 | -0.86 | 10.37 | 12.24 | 12.17 |
| -7.78 | -0.16 | -13.98 | 17.05 | 31.25 | 16.37 |
| 2.44 | -0.41 | -5.27 | 11.65 | 9.60 | 8.46 |
| -2.72 | -0.65 | -3.90 | 8.98 | 9.08 | 8.05 |
| -0.07 | -0.21 | -4.02 | 7.68 | 9.09 | 8.58 |
| 25.62 | 0.30 | -0.70 | 34.91 | 22.12 | 21.53 |
| 7.08 | 0.34 | -2.79 | 15.51 | 9.96 | 7.43 |
| 4.83 | 0.04 | -1.31 | 16.24 | 9.25 | 8.51 |
| -10.49 | -1.42 | -19.08 | 29.85 | 104.64 | 21.15 |
| -1.60 | 0.15 | -3.51 | 7.46 | 9.00 | 7.88 |
| 12.12 | -1.58 | -1.64 | 22.13 | 33.79 | 33.71 |
| 1.94 | 2.75 | 3.03 | 17.84 | 22.77 | 20.14 |
| -7.10 | -0.25 | -1.64 | 14.43 | 16.26 | 16.17 |
| -10.51 | 0.20 | 0.00 | 14.21 | 7.79 | 7.89 |
| 2.10 | -1.07 | -3.57 | 18.07 | 15.39 | 10.87 |
| 4.55 | 0.91 | 0.65 | 12.20 | 11.03 | 11.87 |
| 12.35 | -0.01 | -0.44 | 20.55 | 15.21 | 15.06 |
| 50.54 | 42.85 | 39.03 | 98.51 | 91.09 | 92.68 |
| 9.15 | 1.17 | 5.37 | 17.43 | 16.76 | 15.60 |
| 15.36 | 0.36 | 0.75 | 21.21 | 8.25 | 8.00 |
| 6.26 | 0.44 | 0.73 | 17.54 | 8.22 | 7.42 |
| -2.55 | -0.34 | -4.01 | 7.72 | 8.73 | 9.45 |
| 3.57 | 0.08 | 0.13 | 19.63 | 33.05 | 32.99 |
| 30.66 | -0.52 | -0.50 | 36.66 | 13.89 | 13.90 |
| -6.33 | -1.48 | -5.58 | 11.13 | 12.16 | 9.62 |
| 1.08 | 3.16 | 6.35 | 11.45 | 15.06 | 14.05 |
| -6.21 | -0.44 | -3.79 | 11.27 | 10.43 | 10.53 |
| -7.87 | -0.99 | -5.72 | 15.24 | 15.17 | 12.28 |
| -4.77 | -1.53 | -6.84 | 21.52 | 22.93 | 19.81 |
| **4.04** | **1.43** | **-1.14** | **19.60** | **20.84** | **16.62** |

*Table 6. Australian farms study, coverage rates of 'two-sigma' confidence intervals for regional population means of Total Cash Costs. Intervals are defined by the M-quantile bias-adjusted estimates plus or minus twice their corresponding standard errors using (18) with weights given by (15)*

| Region | Coverage Rates |
|---|---|
| 1 | 0.955 |
| 2 | 0.705 |
| 3 | 0.900 |
| 4 | 0.895 |
| 5 | 0.923 |
| 6 | 0.968 |
| 7 | 0.953 |
| 8 | 0.958 |
| 9 | 0.308 |
| 10 | 0.913 |
| 11 | 0.980 |
| 12 | 0.888 |
| 13 | 0.948 |
| 14 | 0.985 |
| 15 | 0.875 |
| 16 | 0.953 |
| 17 | 0.930 |
| 18 | 0.923 |
| 19 | 0.985 |
| 20 | 0.973 |
| 21 | 0.965 |
| 22 | 0.933 |
| 23 | 0.983 |
| 24 | 1.000 |
| 25 | 0.878 |
| 26 | 0.935 |
| 27 | 0.948 |
| 28 | 0.933 |
| 29 | 0.770 |
| **Mean** | **0.906** |