2010

Model Based Direct Estimation of Small Area Distributions

Nicola Salvati
University of Pisa, Italy

Hukum Chandra
University of Wollongong, hchandra@uow.edu.au

Ray Chambers
University of Wollongong, ray@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

20-10

Model Based Direct Estimation of Small Area Distributions

Nicola Salvati, Hukum Chandra 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

Model Based Direct Estimation of Small Area Distributions

Nicola Salvati[1], Hukum Chandra[2] and Ray Chambers[3]

[1]Dipartimento di Statistica e Matematica Applicata all'Economia, University of Pisa, Italy,
E-mail: salvati@ec.unipi.it

[2]Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, Australia.
E-mail: hchandra@uow.edu.au

[3]Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, Australia.
Email: ray@uow.edu.au

Summary

Much of the small area estimation literature focuses on population totals and means. However, users of survey data are often interested in the finite population distribution of a survey variable, and the measures (e.g. medians, quartiles, percentiles) that characterise the shape of this distribution at small area level. In this paper we propose a model-based direct estimator (MBDE, see Chandra and Chambers, 2009) of the small area distribution function. The MBDE is defined as weighted sum of sample data from the area of interest, with weights derived from the calibrated spline-based estimate of the finite population distribution function introduced by Harms and Duchesne (2006), under an appropriately specified regression model with random area effects. We also discuss the mean squared error estimation of the MBDE. Monte Carlo simulations based on both simulated and real datasets show that the proposed MBDE and its associated mean squared error estimator perform well when compared with alternative estimators of the area-specific finite population distribution function.

Key words: Indicator function; Model-based direct estimator; Mean squared error estimator; Simulation experiments.

1. Introduction

Let $U = \{1,2,...,N\}$ be the finite population of size N and let y denote a variable of interest that takes values over this population. A common target of inference is then the proportion of values y_j that are bounded by a given constant (e.g. the proportion of households whose monthly per capita expenditure is below the poverty line). More generally, the target of inference is the value of the finite population distribution function for a variable y at a specified value t. This is $F_N(t) = N^{-1}\sum_{j=1}^{N} I(y_j \le t)$, i.e. the proportion of the population whose values for y are less than or equal to t, where $I(y_j \le t)$ is the indicator function that takes the value 1 if $y_j \le t$ and 0 otherwise and t is a specified constant. Clearly, once we obtain an estimator of the finite population distribution function, we can evaluate its inverse to obtain the associated estimator of the finite population quantile function. See Chambers and Dunstan (1986), Rao et al. (1990), Harms and Duchesne (2006) and Rueda et al. (2007, 2010).

Small area estimation (SAE) is an important objective of many surveys. Small areas or small domains are subsets of the population with small sample sizes, so standard survey estimation methods for these areas, which only use information from the small area samples, are unreliable. In this context SAE methods that 'borrow strength' via statistical models (Rao, 2003) can be used to produce reliable small area estimates. However, virtually all of these methods focus on estimation of linear parameters, e.g. small area means or totals. In this paper we focus on estimation of the small area distribution of a study variable and measures (e.g. medians, quartiles, percentiles) that characterise the shape of this distribution. This is especially useful if there are extreme values in the small area sample data, or if the small area distribution of the variable of interest is highly skewed (Tzavidis et al., 2010).

We propose a model based direct estimator (MBDE) for the small area distribution function, extending the MBDE approach (Chandra and Chambers, 2009) to the estimation of the small area distribution function. This MBDE estimator is a weighted sum of the sample data from the small area of interest, with weights that are derived from a spline-based calibrated estimator of the population distribution function (Harms and Duchesne, 2006) under a regression model with random area effects.

The rest of the article is organized as follows. The following Section describes SAE based on the linear mixed model and the nonparametric regression model based on penalized splines and then uses these models to motivate estimators of the small area distribution function. Section 3 introduces the concept of calibrated sample weights for a finite population distribution function and uses these to define the MBDE estimator for this function. A bias-robust estimator of the mean squared error of the MBDE is also developed, based on the approach of Chambers et al. (2009). The empirical performances of the proposed MBDE as well as alternative estimators of the small area distribution function are evaluated in Section 4, using both model-based and design-based simulations, with the design-based simulations based on two real data sets. Concluding remarks are set out in Section 5.

2. Estimation of the Small Area Distribution Function

We assume that a finite population U containing N units can be partitioned into A non-overlapping domains, referred to from now on as small areas, or simply areas, indexed by $i = 1,...,A$, with area i containing N_i units, so $N = \sum_{i=1}^{A} N_i$. Let y_{ij} denote the value of the variable of interest y for unit j $(j = 1,\ldots,N_i)$ in area i $(i = 1,\ldots,A)$. The area-specific distribution function of y for area i is

$$F_i(t) = N_i^{-1} \sum_{j=1}^{N_i} I(y_{ij} \le t). \qquad (1)$$

Let *s* denotes a sample of *n* units drawn from *U* by some specified sampling design, and assume that values of the variable of interest y are available for each of these *n* sample units. The non-sample component of *U*, containing *N* - *n* units, is denoted by *r*. In what follows, we use a subscript of *i* to denote quantities specific to area i $(i = 1,...,A)$. For example, s_i and r_i denote the n_i sample and $N_i - n_i$ non-sample units respectively for area *i*. With this notation, the conventional estimators of the area *i* distribution function, $F_i(t)$, are the Horvitz-Thompson (HT) estimator

$$\hat{F}_i^{HT}(t) = N_i^{-1} \sum_{j \in s_i} \pi_j^{-1} I(y_j \le t), \tag{2}$$

and the Hajek estimator

$$\hat{F}_i^{Hajek}(t) = \sum_{j \in s_i} \pi_j^{-1} I(y_j \le t) \Big/ \sum_{j \in s_i} \pi_j^{-1} . \tag{3}$$

Here π_j denotes the sample inclusion probability of unit *j*. Both (2) and (3) are area-specific design-based direct estimators and do not depend on an assumed model for their validity (Cochran, 1977). Unfortunately, empirical evidence presented in Rueda et al. (2007) shows that these estimators can be substantially biased, while the fact that they only use information from the area *i* sample makes them too unstable for SAE.

Model-based small area estimators based on the linear mixed model are widely used in SAE. However, if the functional form of the regression relationship between the variable of interest and the available auxiliary variables is unknown or has a complicated functional form, then SAE based on the use of a nonparametric regression model can offer significant advantages compared with one based on a linear model. In particular, a nonparametric regression model based on p-splines is attractive because it represents a relatively straightforward extension of a linear regression model (Eliers and Marx, 1996). Opsomer et al. (2008) describe the use of a spline-based nonparametric regression model for SAE. See also Salvati et al. (2010). In the rest of this Section we therefore summarize the model-based

approach to estimation of the small area distribution function under the linear mixed model and under a nonparametric regression model.

2.1 Estimation under the linear mixed model

SAE theory for this case is now well established, see Rao (2003). We briefly describe it below since this allows us to introduce notation that will be used elsewhere in the paper. To start, we note that throughout this paper we will assume that we have access to the population values of *p* auxiliary scalar variables that are, to a greater or lesser extent, correlated with *y*. Let $\mathbf{x}_{ij}$ denote the vector of values of these auxiliary variables that are associated with y_{ij} and let $\mathbf{z}_{ij}$ denote a vector of auxiliary 'contextual' variables whose values are known for all units in the population. Let $\mathbf{y}_U$, $\mathbf{X}_U$ and $\mathbf{Z}_U$ denote the population level vector and matrices defined by y_{ij}, $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$, respectively. Then the linear mixed model is

$$\mathbf{y}_U = \mathbf{X}_U\boldsymbol{\beta} + \mathbf{Z}_U\mathbf{u} + \mathbf{e}_U, \tag{4}$$

where $\boldsymbol{\beta}$ is a *p* vector of regression coefficients, $\mathbf{u}$ is a random vector of area effects and $\mathbf{e}_U$ is a population *N*-vector of random individual effects. In general, area effects are vector-valued, so $\mathbf{u}^T = \left(\mathbf{u}_1^T, \mathbf{u}_2^T, \cdots \mathbf{u}_A^T\right)$ and $\mathbf{Z}_U = diag\left\{\mathbf{Z}_i; i = 1, \ldots, A\right\}$, where *i* indexes the *A* areas that make up the population and $\mathbf{Z}_i$ is of dimension $N_i \times q$. The area effects $\left\{\mathbf{u}_i; i = 1, \ldots, A\right\}$ are assumed to be independent and identically distributed realisations of a random vector of dimension *q* with zero mean and covariance matrix Σ_u. Similarly, the scalar individual effects making up $\mathbf{e}_U$ are assumed to be independent and identically distributed realisations of a random variable with zero mean and variance σ_e^2, with area and individual effects mutually independent. The covariance matrix of the vector $\mathbf{y}_U$ is then $Var(\mathbf{y}_U) = \mathbf{V}_U = \mathbf{Z}_U\Sigma_u\mathbf{Z}_U^T + \sigma_e^2\mathbf{I}_N$, where $\mathbf{I}_k$ denotes the identity matrix of dimension *k*. The parameters $\theta = (\Sigma_u, \sigma_e^2)$ are typically referred to as the variance components of (4).

We also assume throughout this paper that the method of sampling is non-informative given the auxiliary variables, so the model (4) holds for both sampled and non-sampled population units. Consequently, we can partition $\mathbf{y}_U$, $\mathbf{X}_U$, $\mathbf{Z}_U$ and $\mathbf{e}_U$ into components defined by the *n* sampled and *N* – *n* non-sampled population units, denoted by subscripts of *s* and *r* respectively, and re-express (4) as follows:

$$\mathbf{y}_U = \begin{bmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{bmatrix} = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_s \\ \mathbf{Z}_r \end{bmatrix} \mathbf{u} + \begin{bmatrix} \mathbf{e}_s \\ \mathbf{e}_r \end{bmatrix},$$

with the variance of **y** similarly partitioned,

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

Thus $\mathbf{X}_s$ represents the matrix defined by the *n* sample values of the auxiliary variable vector, while

$$\mathbf{V}_{ss} = diag\left\{\mathbf{V}_{iss}; i = 1,\ldots,A\right\} = diag\left\{\mathbf{Z}_{is}\Sigma_u \mathbf{Z}_{is}^T + \sigma_e^2 \mathbf{I}_{is}; i = 1,\ldots,A\right\}$$

and

$$\mathbf{V}_{sr} = diag\left\{\mathbf{V}_{isr}; i = 1,\ldots,A\right\} = diag\left\{\mathbf{Z}_{is}\Sigma_u \mathbf{Z}_{ir}^T; i = 1,\ldots,A\right\}.$$

Here $\mathbf{Z}_{is}$ and $\mathbf{Z}_{ir}$ respectively denote the restriction of $\mathbf{Z}_i$ to sampled and non-sampled units in area *i*.

The distribution function for small area *i* given by (1) can be expressed as $F_i(t) = N_i^{-1}\left\{\sum_{j\in s_i} I(y_j \le t) + \sum_{j\in r_i} I(y_j \le t)\right\}$, where the first term on the left is known and the second is unknown. The problem of estimating $F_i(t)$ therefore reduces to predicting the values y_j for the non-sample units in area *i*. Given estimated values $\hat{\theta} = (\hat{\Sigma}_u, \hat{\sigma}_e^2)$ of the variance components we can define the estimated covariance matrix $\hat{\mathbf{V}}_U$, and the predicted values of y_j are $\hat{y}_j^{EBLUP} = \mathbf{x}_j^T \hat{\boldsymbol{\beta}}^{EBLUE} + \mathbf{z}_j^T \hat{\mathbf{u}}^{EBLUP}$, where $\hat{\boldsymbol{\beta}}^{EBLUE} = (\mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{y}_s$ is the

empirical best linear unbiased estimator (EBLUE) of $\boldsymbol{\beta}$ and $\hat{\mathbf{u}}^{EBLUP} = \hat{\Sigma}_u \mathbf{Z}_s^T \hat{\mathbf{V}}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})$ is the empirical best linear unbiased estimator (EBLUP) of $\mathbf{u}$. Substituting estimated values for the parameters of (4) then allows us to define an estimator for $F_i(t)$ of the form

$$\hat{F}_i^{EBP}(t) = N_i^{-1}\left\{\sum_{j\in s_i} I(y_j \le t) + \sum_{j\in r_i} I(\hat{y}_j^{EBLUP} \le t)\right\}. \qquad (5)$$

We refer to (5) as the empirical best predictor or EBP. An alternative way of predicting $F_i(t)$ is via the Chambers and Dunstan (hereafter CD) estimator. See Chambers and Dunstan (1986) for details. Since the within area residuals are homoskedastic under (4), the CD estimator of $F_i(t)$ can be written

$$\hat{F}_i^{CD}(t) = N_i^{-1}\left\{\sum_{j\in s_i} I(y_j \le t) + n_i^{-1}\sum_{j\in r_i}\sum_{k\in s_i} I\left\{\left[\hat{y}_j^{EBLUP} + \left(y_k - \hat{y}_k^{EBLUP}\right)\right] \le t\right\}\right\}. \qquad (6)$$

Note that the CD estimator is asymptotically unbiased if (4) is correctly specified.

2.2 Estimation under a nonparametric mixed model

The CD estimator (6) will be biased if the functional form of the relationship between the response variable and the auxiliary variables (i.e. the regression function) is not linear or the variance term in the regression model is misspecified (Tzavidis et al., 2010). This susceptibility of parametric model-based methods to misspecification bias provides motivation for the use of alternative non-parametric model-based methods. We now summarize application of the p-spline nonparametric regression model to SAE (Opsomer et al., 2008), and, for simplicity, consider the univariate case. The underlying regression model is then $y_j = m(x_j) + e_j$, where e_j are independent random variables with zero means. The function $m(x)$ is unknown and assumed to be approximated sufficiently well by

$$m(x, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \beta_0 + \beta_1 x + \cdots + \beta_p x^b + \sum_{k=1}^{K} \gamma_k (x - \kappa_k)_+^b, \qquad (7)$$

where b is the degree of the spline, $(c)_+^b = c^b I(c > b)$, κ_k is a set of fixed constants called knots for $k = 1, \ldots, K$, $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^T$ is the coefficient vector of the parametric part of the model and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_K)^T$ is the vector of spline coefficients. The approximating function $m(x, \boldsymbol{\beta}, \boldsymbol{\gamma})$ in (7) uses truncated polynomial basis functions for simplicity and, if the number of knots K is sufficiently large, can approximate most smooth functions. Ruppert et al. (2003, Chapter 5) suggest the use of a knot for every four observations, up to a maximum of about 40 knots for a univariate application. Using a large number of knots in (7) can lead to an unstable fit. In order to overcome this problem, an upper limit is usually imposed on the size of the spline coefficient vector $\boldsymbol{\gamma}$. Estimating $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by minimizing the squared deviations of model (7) from the actual data values subject to this constraint is equivalent to minimizing the penalized loss function

$$\sum_j \left(y_j - m(x_j, \boldsymbol{\beta}, \boldsymbol{\gamma}) \right)^2 + \lambda \boldsymbol{\gamma}^T \boldsymbol{\gamma} . \tag{8}$$

Here λ is a Lagrange multiplier that controls the level of smoothness of the resulting fit.

Wand (2003) and Ruppert et al. (2003, Chapter 4) note the equivalence between minimizing (8) and maximizing the likelihood of the response variable under the linear model (7) where the spline coefficients are treated as random effects. In particular, let $\mathbf{y}_U = (y_1, y_2, ..., y_N)^T$,

$$\mathbf{X}_U = \begin{bmatrix} 1 & x_1 & \cdots & x_1^b \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & \cdots & x_N^b \end{bmatrix} \text{ and } \boldsymbol{\Delta}_U = \begin{bmatrix} (x_1 - \kappa_1)_+^p & \cdots & (x_1 - \kappa_K)_+^b \\ \vdots & \ddots & \vdots \\ (x_N - \kappa_1)_+^p & \cdots & (x_N - \kappa_K)_+^b \end{bmatrix}.$$

The spline approximation (7) can then be written as the linear mixed model

$$\mathbf{y}_U = \mathbf{X}_U \boldsymbol{\beta} + \boldsymbol{\Delta}_U \boldsymbol{\gamma} + \mathbf{e}_U , \tag{9}$$

where $\boldsymbol{\gamma}$ and $\mathbf{e}$ are now assumed to be independent Gaussian random vectors of dimension K and N respectively. In particular, it is assumed that

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_K) \text{ and } \mathbf{e}_U \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_N).$$

Opsomer et al. (2008) adapt p-splines to the SAE context by adding area random effects to (9), which then becomes

$$\mathbf{y}_U = \mathbf{X}_U \boldsymbol{\beta} + \boldsymbol{\Delta}_U \boldsymbol{\gamma} + \mathbf{Z}_U \mathbf{u} + \mathbf{e}_U, \tag{10}$$

where, as in Section 2.1, $\mathbf{Z}_U = \left(\mathbf{Z}_1, \ldots, \mathbf{Z}_N\right)^T$ is a matrix of known covariates of dimension $N \times A$ characterising differences among the areas and $\mathbf{u}$ is the *A*-vector of random area effects. In the simplest case, $\mathbf{Z}_U$ is given by a matrix whose *i*-th column, for *i* = 1, …, *A*, is an indicator variable that takes the value 1 if a unit is in area *i* and is zero otherwise. It is assumed that the area effects are distributed independently of the spline effects $\boldsymbol{\gamma}$ and the individual effects $\mathbf{e}$, with $\mathbf{u} \sim N(\mathbf{0}, \Sigma_u)$, so that the covariance matrix of the vector $\mathbf{y}_U$ is $Var(\mathbf{y}_U) = \mathbf{V} = \sigma_\gamma^2 \boldsymbol{\Delta}_U \boldsymbol{\Delta}_U^T + \mathbf{Z}_U \Sigma_u \mathbf{Z}_U^T + \sigma_e^2 \mathbf{I}_N$. The variance components of (10) are then given by $\theta = \left(\sigma_\gamma^2, \Sigma_u, \sigma_e^2\right)$. Note that, as in previous Section, the use of non-informative sampling given the auxiliary variables means that (10) also holds at the sample level.

When the variance components are known, well-established theory (McCulloch and Searle, 2001, Chapter 9) leads to the generalised least squares estimator of $\boldsymbol{\beta}$, i.e. $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{y}_s$, and the best linear unbiased predictors (BLUPs) for $\boldsymbol{\gamma}$ and $\mathbf{u}$, i.e. $\hat{\boldsymbol{\gamma}} = \sigma_\gamma^2 \boldsymbol{\Delta}_s^T \mathbf{V}_{ss}^{-1} (\mathbf{y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}})$ and $\hat{\mathbf{u}} = \Sigma_u \mathbf{Z}_s^T \mathbf{V}_{ss}^{-1} (\mathbf{y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}})$. In practice, the variance components are unknown and must be estimated from sample data using methods such as maximum likelihood or restricted maximum likelihood; see Harville (1977). In what follows we use $\left(\hat{\sigma}_\gamma^2, \hat{\Sigma}_u, \hat{\sigma}_e^2\right)$ to denote such estimates, allowing us to define the plug-in estimator $\hat{\mathbf{V}}_{ss} = \hat{\sigma}_\gamma^2 \boldsymbol{\Delta}_s \boldsymbol{\Delta}_s^T + \mathbf{Z}_s \hat{\Sigma}_u \mathbf{Z}_s^T + \hat{\sigma}_e^2 \mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of order *n*. This leads to the nonparametric model-based EBLUE for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}^{NPEBLUE} = (\mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{y}_s$, and to the

corresponding nonparametric EBLUPs (NPEBLUPs) for the spline and area effects in (10), $\hat{\boldsymbol{\gamma}}^{NPEBLUP} = \hat{\sigma}_{\gamma}^2 \boldsymbol{\Delta}_s^T \hat{\mathbf{V}}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}}^{NPEBLUE})$ and $\hat{\mathbf{u}}^{NPEBLUP} = \hat{\Sigma}_u \mathbf{Z}_s^T \hat{\mathbf{V}}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}}^{NPEBLUE})$.

Under (10), the nonparametric empirical best predictor of the distribution function for area i (denoted by NPEBP) is

$$\hat{F}_i^{NPEBP}(t) = N_i^{-1}\left\{\sum_{j\in s_i} I(y_j \le t) + \sum_{j\in r_i} I(\hat{y}_j^{NPEBLUP} \le t)\right\}, \tag{11}$$

where $\hat{y}_j^{NPEBLUP} = \mathbf{x}_j^T \hat{\boldsymbol{\beta}}^{NPEBLUE} + \boldsymbol{\delta}_j^T \hat{\boldsymbol{\gamma}}^{NPEBLUP} + \mathbf{z}_j^T \hat{\mathbf{u}}^{NPEBLUP}$, and $\mathbf{x}_j^T$, $\boldsymbol{\delta}_j^T$ and $\mathbf{z}_j^T$ denote respectively the rows of $\mathbf{X}_U$, $\boldsymbol{\Delta}_U$ and $\mathbf{Z}_U$ that correspond to unit j in area i. Similarly, under (10), the nonparametric version of the CD estimator of the distribution function for area i is

$$\hat{F}_i^{NPCD}(t) = N_i^{-1}\left[\sum_{j\in s_i} I(y_j \le t) + n_i^{-1}\sum_{j\in r_i}\sum_{k\in s_i} I\left\{\left[\hat{y}_j^{NPEBLUP} + \left(y_k - \hat{y}_k^{NPEBLUP}\right)\right] \le t\right\}\right]. \tag{12}$$

3. The Model-Based Direct Estimator for the Small Area Distribution Function

A direct estimate for a small area is simple to interpret, since the estimated value of the variable of interest for the area is just a weighted average of the sample data from the same area. This is not true of an indirect estimator like the EBLUP, which is a weighted sum over the entire sample. Unfortunately, when weights are the inverses of sample inclusion probabilities, conventional direct estimators like (2) and (3) can be quite inefficient. The Model-Based Direct Estimator (MBDE) of a small area mean improves upon the efficiency of these conventional direct estimators by using the weights that define the EBLUP for the population total under a model with random area effects. See Chandra and Chambers (2009) and Salvati et al. (2010). MBDEs for the population mean of y using weights based on the linear model (4) as well as those based on the non-parametric model (10) are therefore possible. However, the finite population distribution function is the population mean of an indicator variable, which does not satisfy either (4) or (10). Consequently, 'standard' EBLUP

weights are not appropriate for defining the MBDE of this function. Instead, we use sample weights that are calibrated to the known finite population distribution of the auxiliary variables in **x** and are based on a model with random area effects.

For simplicity, we restrict our discussion below to a single scalar covariate *x*, noting that the extension to multiple scalar covariates is straightforward. The calibrated estimator of a finite population distribution function $F_N(t)$ was defined in Harms and Duchesne (2006) as a weighted empirical distribution function

$$\hat{F}_N^{HD}(t) = N^{-1}\sum\nolimits_{j\in s} w_j I(y_j \le t) \tag{13}$$

where the sample weights w_j in (13) are calibrated to the known finite population distribution of *x*. In particular, let $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_K < 1$ denote an ordered set of constants. Then the weights used in (13) sum to *N* and, for $k = 1,\ldots,K$, also satisfy

$$\sum\nolimits_{j\in s} w_j I\left\{x_j \le Q_x(\alpha_k)\right\} = N\alpha_k, \tag{14}$$

where $Q_x(\alpha_k)$ is the known α_k-quantile of the finite population distribution of x. That is, the weights used in (13) are calibrated to both the population size *N* and to the population totals of the auxiliary variables defined by the indicators $I\left\{x_j \le Q_x(\alpha_k)\right\}$.

Standard results from calibration theory (Deville and Särndal, 1992; Chambers, 1996) can be used to show that if these calibrated weights w_j are then chosen to minimise their chi-square distance from the weights used in Horvitz-Thompson estimator (2), as is commonly done, then (13) is a regression estimator of $F_N(t)$ under the linear model

$$I(y_j \le t) = \beta_{0t} + \sum_{k=1}^{K} \beta_{kt} I\left\{x_j \le Q_x(\alpha_k)\right\} + \varepsilon_{jt}, \tag{15}$$

where the ε_{jt} are uncorrelated errors with zero expectation and variance $\sigma_{\varepsilon t}^2$ (Chambers, 2005). However, (15) is also easily seen to be a p-spline model with knots at the α_k-th

quantiles of the finite population distribution of x. That is, $\hat{F}_N^{HD}(t)$ is actually a p-spline estimator of $F_N(t)$. Define $g_{jk} = I\left\{x_j \le Q_x(\alpha_k)\right\}$ and let $\mathbf{g}_{Uk} = \left(g_{jk}; j = 1,....,N\right)$ be the corresponding population N-vector, so $\mathbf{G}_U = \left[\mathbf{1}_N, \mathbf{g}_{U1}, \ldots, \mathbf{g}_{UK}\right]$ denotes the population level matrix of values of these variables, where $\mathbf{1}_N$ denotes a N-vector of ones. Also, define $d_{jt} = I(y_j \le t)$ and put $\mathbf{d}_{Ut}$ equal to the N-vector of population values of the d_{jt}. The population level version of model (15) is then

$$\mathbf{d}_{Ut} = \mathbf{G}_U \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_{Ut}. \tag{16}$$

Given the appropriate sample and non-sample components of $\mathbf{d}_{Ut}$, $\mathbf{G}_U$ and the covariance matrix $\mathbf{V}_{Ut} = \sigma_{\varepsilon t}^2 \mathbf{I}_U$ of $\boldsymbol{\varepsilon}_{Ut}$, the vector of sample weights w_{jt}^{DF} that define the EBLUP of the population total of the d_{jt} under (16) is then

$$\mathbf{w}_{st}^{DF} = \left(w_{jt}^{DF}; j \in s\right) = \mathbf{1}_n + \hat{\mathbf{H}}_{st}^T(\mathbf{G}_U^T \mathbf{1}_N - \mathbf{G}_s^T \mathbf{1}_n) + (\mathbf{I}_n - \hat{\mathbf{H}}_{st}^T \mathbf{g}_s^T)\hat{\mathbf{V}}_{sst}^{-1}\hat{\mathbf{V}}_{srt}\mathbf{1}_{N-n}, \tag{17}$$

where $\hat{\mathbf{H}}_{st} = \left(\mathbf{G}_s^T \hat{\mathbf{V}}_{sst}^{-1} \mathbf{G}_s\right)^{-1} \mathbf{G}_s^T \hat{\mathbf{V}}_{sst}^{-1}$. Under (16), $\hat{\mathbf{V}}_{sst} = \hat{\sigma}_{\varepsilon t}^2 \mathbf{I}_n$ and $\hat{\mathbf{V}}_{srt} = \mathbf{0}$, so these weights simplify to

$$\mathbf{w}_s^{DF} = \left(w_j^{DF}; j \in s\right) = \mathbf{1}_n + \mathbf{G}_s\left(\mathbf{G}_s^T \mathbf{G}_s\right)^{-1}(\mathbf{G}_U^T \mathbf{1}_N - \mathbf{G}_s^T \mathbf{1}_n) = \mathbf{1}_n + \mathbf{G}_s\left(\mathbf{G}_s^T \mathbf{G}_s\right)^{-1} \mathbf{G}_{N-n}^T \mathbf{1}_{N-n}.$$

The model (16) is easily adapted to small area estimation by including random area effects. That is, we replace (16) by

$$\mathbf{d}_{Ut} = \mathbf{G}_U \boldsymbol{\beta}_t + \mathbf{Z}_U \mathbf{u}_t + \boldsymbol{\varepsilon}_{Ut} \tag{18}$$

where $\mathbf{Z}_U$ was defined following (4) and $\mathbf{u}_t \sim N(\mathbf{0}, \boldsymbol{\Omega}_t)$ is an A-vector of random area effects. As usual, we assume that $\mathbf{u}_t$ and $\boldsymbol{\varepsilon}_{Ut}$ are independently distributed, so that $Var(\mathbf{d}_{Ut}) = \mathbf{V}_{Ut} = \mathbf{Z}_U \boldsymbol{\Omega}_t \mathbf{Z}_U^T + \sigma_{\varepsilon t}^2 \mathbf{I}_N$. The sample weights w_{jt}^{DF} that define the EBLUP of the population total of the d_{jt} under (18) are then still given by (17), but now with

$\hat{\mathbf{V}}_{sst} = \mathbf{Z}_s\hat{\boldsymbol{\Omega}}_t\mathbf{Z}_s^T + \hat{\sigma}_{\varepsilon t}^2\mathbf{I}_n$ and $\hat{\mathbf{V}}_{srt} = \mathbf{Z}_s\hat{\boldsymbol{\Omega}}_t\mathbf{Z}_r^T$, where $\hat{\boldsymbol{\Omega}}_t$ and $\hat{\sigma}_{\varepsilon t}^2$ are the estimated values of the variance components of (18).

In practice, one first needs to decide on the calibration constraints (14) before (18) can be fitted and (17) calculated. This in turn requires that one has chosen the values $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_K < 1$. We adapt the ordered half-sample cross validation procedure described in Chambers (2005) for this purpose. In particular, we fix $K = 1$ and then search for the value α_t^{opt} that maximises the concordance between the sample values of d_{jt} and the sample values of $g_j = I\left\{x_j \le Q_x(\alpha)\right\}$. The steps in this procedure are as follows:

1. Order the sample *x*-values: $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(n-1)}, x_{(n)}$;

2. Create two sets $E = \left\{x_{(1)}, x_{(3)}, ...\right\}$ and $V = \left\{x_{(2)}, x_{(4)}, ...\right\}$;

3. For given α and *t*, fit the model (18) and then compute the weights (17), treating *E* as the 'sample' and *V* as the 'nonsample'. Denote the corresponding value of (13) based on these weights by $\hat{F}_N^{HD(n)}(t,\alpha)$;

4. The optimal value α_t^{opt} then satisfies

$$\left\{\hat{F}_N^{HD(n)}(t,\alpha_t^{opt}) - n^{-1}\sum_{j\in s} I(y_j \le t)\right\}^2 = \min_{0<\alpha<1}\left\{\hat{F}_N^{HD(n)}(t,\alpha) - n^{-1}\sum_{j\in s} I(y_j \le t)\right\}^2.$$

We note that although this procedure only identifies a single 'most concordant' calibration constraint to use in (14), there is nothing to stop it being extended to identification of multiple calibration constraints. However, some care must then be taken to ensure that the resulting values of $Q_x(\alpha)$ are separated sufficiently in the interval spanned by the sample values of the auxiliary *x*. Failure to do this could result in the sample design matrix defined by (18) not being of full rank.

Finally, given the weights (17), we write down the MBDE for the area *i* distribution function $F_i(t)$ as

$$\hat{F}_i^{MBDE}(t) = \sum_{j \in s_i} w_{jt}^{DF} I(y_j \le t) \Big/ \sum_{j \in s_i} w_{jt}^{DF} \ . \qquad (19)$$

We refer (19) as a direct estimator because it is a weighted average of the sample data from the area of interest. However, this does not mean that it can be calculated from these data alone. The weights (17) are a function of the data from the entire sample. That is, they 'borrow strength' from other areas via the model (18).

It should also be pointed out that since the weights (17) depend on *t*, there is no guarantee that (19) defines a monotone function of *t*, i.e. one where $t_1 < t_2$ implies $\hat{F}_i^{MBDE}(t_1) \le \hat{F}_i^{MBDE}(t_2)$. This issue will usually not be relevant when one wishes to estimate the distribution of interest at points that are well separated, but can be a problem when the aim is to invert (19) as a function of *t* in order to estimate quantiles. In such a situation we recommend that (19) be first transformed to be monotone in *t*, e.g. using the approach described in He (1997).

3.1 Mean squared error estimation for the MBDE

A bias-robust estimator of the mean squared error (MSE) of the MBDE is described in Chandra and Chambers (2009), see also Chambers et al. (2009), and we use this approach here to define a corresponding MSE estimator for (19). This is the estimator

$$\hat{M}\left\{\hat{F}_i^{MBDE}(t)\right\} = \hat{V}_{it} + \hat{B}_{it}^2 \qquad (20)$$

where $\hat{V}_{it}$ is a heteroskedasticity-robust estimator of the conditional prediction variance of $\hat{F}_i^{MBDE}(t)$ (Royall and Cumberland, 1978), $\hat{B}_{it}$ is an estimator of the corresponding conditional prediction bias, and the conditioning is with respect to the value of the area effect. In particular, we use

$$\hat{V}_{it} = N_i^{-2} \sum_{j \in s_i} \left\{ \left(N_i w_{jt}^{DF(i)} - 1 \right)^2 + (N_i - n_i) n^{-1} \right\} (d_{jt} - \hat{\mu}_{jt})^2 \ , \qquad (21)$$

where $w_{jt}^{DF(i)} = w_{jt}^{DF} \Big/ \sum_{k \in s_i} w_{kt}^{DF}$ and $\hat{\mu}_{jt}$ is an unbiased linear estimator of the conditional expected value $\mu_{jt} = E(d_{jt} \big| g_j, \mathbf{u}_t)$. Chambers et al. (2009) recommend that $\hat{\mu}_{jt}$ be computed as the 'unshrunken' version of the EBLUP for μ_{jt}, i.e.

$$\hat{\mu}_{jt} = \hat{\beta}_{0t} + g_j \hat{\beta}_{1t} + \mathbf{z}_j^T \left(\mathbf{Z}_s^T \mathbf{Z}_s \right)^{-1} \mathbf{Z}_s^T \left(\mathbf{I}_s - \hat{\mathbf{H}}_{st}^T \mathbf{g}_s^T \right)^T \mathbf{l}_n .$$

For the conditional bias of the MBDE, we use a simple 'plug-in' estimator of the form

$$\hat{B}_{it} = \sum_{j \in s_i} w_{jt}^{DF(i)} \hat{\mu}_{jt} - N_i^{-1} \sum_{j \in U_i} \hat{\mu}_{jt} . \tag{22}$$

Note that the MSE estimator (20) ignores the extra variability associated with estimation of the variance components, and is therefore a heteroskedasticity-robust first order approximation to the actual conditional MSE of the MBDE. Also, (20) treats the weights (17) as fixed, i.e. it ignores the contribution to the MSE from the estimated variance components. Chambers et al. (2009) refer to this as a pseudo-linearization assumption since for large overall sample sizes the contribution to the overall MSE of (19) arising from the variability of variance components will be of smaller order of magnitude then the fixed weights prediction variance estimated by (21). However, the extent of this underestimation will depend on the small area sample sizes and the characteristics of the population of interest, particularly the strength of the small area effects. Finally, we note that (22) is a conservative estimator of the squared bias, since $E(\hat{B}_{it}^2) = Var(\hat{B}_{it}) + E^2(\hat{B}_{it})$. However, the extent of this overestimation is typically very small.

4. Empirical Evaluations

In this Section we report the results from model-based and design-based simulation studies that illustrate the performance of the different estimators of the small area distribution function defined in the preceding two Sections. These estimators are set out in Table 1. Their

performance in the simulation studies is evaluated by computing for each small area the absolute relative bias (*ARB*), the relative root mean squared error (*RRMSE*) and coverage rate (*CR*) of nominal 95 per cent confidence intervals defined as follows:

$$ARB_i = \left(R^{-1} \sum_{r=1}^{R} F_{ir} \right)^{-1} \left\{ R^{-1} \left| \sum_{r=1}^{R} \left(\hat{F}_{ir} - F_{ir} \right) \right| \right\} \times 100 ,$$

$$RRMSE_i = \left(R^{-1} \sum_{r=1}^{R} F_{ir} \right)^{-1} \left\{ \sqrt{R^{-1} \sum_{r=1}^{R} \left(\hat{F}_{ir} - F_{ir} \right)^2} \right\} \times 100 , \text{ and}$$

$$CR_i = \frac{1}{R} \sum_{r=1}^{R} I \left(\left| \hat{F}_{ir} - F_{ir} \right| \leq 2 \sqrt{\hat{M}_{ir}} \right) \times 100 .$$

Here R denotes the number of simulations, F_{ir} denotes the true value of the area i distribution function at simulation r, $\hat{F}_{ir}$ denotes an estimate of this value, and $\hat{M}_{ir}$ denotes an estimate of the MSE of $\hat{F}_{ir}$. The value of the true MSE for $\hat{F}_{ir}$ is calculated as $R^{-1} \sum_{r=1}^{R} \left(\hat{F}_{ir} - F_{ir} \right)^2$. Note that in the design-based simulations $F_{ir} = F_i$.

4.1 Model-based simulations

In the model-based simulations we set A = 30 and use two types of models to generate the population values of y. The first is a linear model, $y_{ij} = 500 + 1.5 x_{ij} + u_i + e_{ij}$, where $x_{ij} \sim \chi^2(20)$, $j = 1,....,N_i$ and $i = 1,...,A$, with random area effects u_i are generated as independent realizations from a $N(0, 23.52)$ distribution and e_{ij} distributed as $N(0, 94.09)$, corresponding to an intra-area correlation of $\sigma_u^2 / \left(\sigma_u^2 + \sigma_\varepsilon^2 \right) = 0.2$. Simulations based on this model are referred to as set 1 simulations. The second model is a multiplicative model, $y_{ij} = 5 x_{ij}^{\beta} u_i e_{ij}$, where the values of x_{ij} are independently drawn from the lognormal distribution $\log(x_{ij}) \sim N\left(6, \sigma_x^2\right)$, and the individual effects and area effects are independently drawn as $\log(e_{ij}) \sim N\left(0, \sigma_e^2\right)$ and $\log(u_i) \sim N\left(0, \sigma_u^2\right)$ respectively. We use two sets of parameters for this model, defined by β(1 or 2), σ_u(0.4 or 0.6), σ_e(0.7 or 1.0) and σ_x(2.25

or 1.20). These are referred to from now on as set 2a and set 2b. Data values for y generated under set 2a are almost linear in x while those generated under set 2b are quite non-linear in x. The small area population sizes N_i are randomly drawn from a uniform distribution on [450,550] and kept fixed over the simulations. The small area sample sizes n_i are determined by first selecting a simple random sample of size n =600 from the population and noting the resulting sample sizes in each small area. These area specific sample sizes n_i are then fixed in the simulations by treating the small areas as strata and carrying out stratified random sampling. A total of R = 1000 simulations are then carried out for each combination of model and individual error distribution, with each simulation corresponding to first generating the population values and then drawing a sample. The average ARB values and the average RRMSE values of the different small area distribution function estimators are shown in Table 2 and 3 respectively. These values are in percentage terms, and the averages are over the 30 small areas. All estimators are evaluated at the 0.1, 0.25, 0.5, 0.75 and 0.9 quantiles of y.

4.2 Design-based simulations

The design-based simulations are based on two real survey data sets. The first survey data set is based on data collected in the 1995-96 Australian Agricultural Grazing Industry Survey (AAGIS) conducted by the Australian Bureau of Agricultural and Resource Economics. In the original sample there were 759 farms from 12 regions (the small areas of interest), which make up the wheat-sheep zone for Australian broadacre agriculture. We used these sample data to generate a synthetic population of size N = 39,562 farms by re-sampling the original AAGIS sample of n = 759 farms with probability proportional to a farm's sample weight. This fixed population was then repeatedly sampled using stratified random sampling with regions corresponding to strata and with stratum sample sizes the same as in the original sample. The variable of interest is total cash costs (TCC) and the auxiliary variable is land area. Based on the original AAGIS sample data, the fit of the linear mixed model (AIC =

20012.32) and the fit of the nonparametric p-spline regression model (AIC = 19998.02) were essentially the same, indicating that addition of the nonparametric spline component does not improve the fit of the mixed model. We therefore do not expect to see much difference between the distribution function estimates generated by these two models. The aim is to estimate the values of the regional distribution functions at the 0.1, 0.25, 0.5, 0.75 and 0.9 quantiles of the finite population distribution of TCC.

The data for the second design-based simulation come from the Environmental Monitoring and Assessment Program (EMAP) survey carried out by the Space Time Aquatic Resources Modelling and Analysis Program (STARMAP) at Colorado State University, and we replicate the design-based simulation experiment carried out by Salvati et al. (2010). The background to this data set is that EMAP conducted a survey of lakes in the North-Eastern states of the United States of America between 1991 and 1996. The data collected in this survey included 551 measurements of Acid Neutralizing Capacity (ANC) - an indicator of the acidification risk of water bodies in water resource surveys - from a sample of 349 of the 21,028 lakes located in this area. Here we define lakes grouped by 6-digit Hydrologic Unit Code (HUC) as our small areas of interest. Since three HUCs have sample sizes of one, these are combined with adjacent HUCS, leading to a total of 23 small areas. Sample sizes in these 23 areas vary from 2 to 45. A (fixed) pseudo-population of N = 21,028 lakes is defined by sampling N times with replacement and with probability proportional to a lake's sample weight from the original sample of 349 lakes. A total of R = 1000 independent stratified random samples of the same size as the original sample are selected from this pseudo-population, with HUCs corresponding to strata and stratum sample sizes fixed to be the same as in the original sample. The survey variable of interest is the ANC value of a lake, with its elevation defining the auxiliary variable. Using the original EMAP data, the fit of the linear mixed model (AIC = 6714.31) is worse than that of the nonparametric regression model (AIC

= 6580.2). In this case, therefore, there are gains from including the spline component in the mixed model, and so we expect that estimates of the distribution function based on the nonparametric regression model will perform better than those based on the linear mixed model. Again, the aim is to estimate the values of the individual HUC distribution functions at the 0.1, 0.25, 0.5, 0.75 and 0.9 quantiles of the finite population distribution of ANC.

Tables 4 and 5 show the average over small areas of the ARB and RRMSE values of the different distribution function estimators based on the R = 1000 independent stratified samples taken from the AAGIS and EMAP populations respectively. Similarly, Table 6 shows the corresponding averages over the areas of the true RMSEs and estimated RMSEs, and the actual coverage rates of nominal 95 percent confidence intervals for the true area-specific distribution function values based on the MBDE estimator (19) and its associated MSE estimator (20). Figures 1 and 2 show the area-specific values of the true RMSE and estimated RMSE of the MBDE (19) for the design-based simulations of the AAGIS and EMAP data.

4.3 Discussion

Two things stand out in Tables 2 and 3. The first is that the MBDE offers substantial bias gains over the other DF estimators, at all quantiles, when the relationship between the study variable and the covariate is complicated and/or the usual mixed model distributional assumptions are invalid (sets 2a and 2b). If the underlying population structure is linear and the usual mixed model assumptions hold (set 1) the CD and NPCD estimators have slightly smaller absolute biases than the MBDE. The larger biases of the 'plug-in' EBP and NPEBP estimators are not unexpected in set 1 because these estimators ignore unit level variability in y. Second, the NPCD estimator generally records the lowest RRMSE among the alternatives to the MBDE, but when the relationship between y and x is complicated, as under sets 2a and 2b, the RRMSE values recorded by the MBDE are comparable, and sometimes lower, than

those recorded by the NPCD estimator. On the other hand, under the linear specification (set 1), the MBDE is clearly less efficient than its alternatives.

Design-based simulations serve to complement model-based simulations for SAE, providing evidence of comparative performance and robustness in realistic data scenarios. Table 4 shows the results for the design-based simulations using the AAGIS data. Here we see that the MBDE has lower bias and RMSE than the other predictors at all quantiles. As expected, given the linear relationship between y and x, the CD-based estimators of the DF based on the linear mixed model are generally more efficient than those based on the nonparametric spline regression model. However, the reverse is true for the EBP-based estimators, perhaps reflecting the lower (but still substantial) biases of the NPEBP. Table 5 reports the design-based simulation results for EMAP data. These again indicate that the MBDE dominates the other estimators in terms of bias. The results for RRMSE are not as clear-cut as in the AAGIS simulations, but still show that the performance of the MBDE is comparable with the performance of the NPCD estimator, which was consistently the best of the alternative estimators in terms of RRMSE.

We now turn to an examination of the performance of the MSE estimator (20) for the MBDE. Figures 1 and 2 show that this estimator accurately tracks the simulation (i.e. repeated sampling) area-specific MSEs of the MBDE at all five target quantiles for y. This good performance is confirmed by the results in Table 6, which shows that the area averages of the true RMSEs and the estimated RMSEs obtained using (20) are very close. Finally, we note that one can combine the MBDE estimator (19) with the MSE estimator (20) to generate 'normal theory' confidence intervals for the area-specific value of the distribution function, i.e. as the small area estimate plus or minus twice its corresponding estimated RMSE. Table 6 shows that the actual coverage rates achieved by these intervals, though generally less than 95 per cent, are still close enough to their target value to be practically useful.

Finally, we note that an alternative to the CD estimator that is both model-consistent and design-consistent, has been proposed by Rao et al. (1990). Although the relevant results are not reported here, we also explored the performance of both parametric and nonparametric versions of this estimator in our simulations. In all cases, this performance was almost identical to that of the parametric and nonparametric versions of the CD predictor.

5. Conclusions

This paper develops an MBDE estimator for the value of the area-specific finite population distribution of a response variable *y*. This estimator is based on sample weights that are calibrated to the finite population distribution of an auxiliary variable *x*, and also allow for random area effects. We then compare the performance of this MBDE estimator with two competing estimators based on either a linear mixed model or a nonparametric mixed model for *y*. Our results indicate that the proposed MBDE can sometimes be much better than these alternatives, particularly in realistic applications where fitted models are approximations at best. On the other hand, if the model assumptions are valid (e.g. set 1 in the model-based simulations), then area-specific distribution function estimators based on the CD representation are preferable. We also provide a method for estimating the MSE of the MBDE and demonstrate empirically that it performs well.

References

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

Chambers, R. (2005). Imputation vs. Estimation of Finite Population Distributions. *Southampton Statistical Sciences Research Paper*. S3RI Methodology Working Papers, M05/06.

Chambers, R., Chandra, H. and Tzavidis, N. (2009). On Bias-Robust Mean Squared Error Estimation for Linear Predictors for Domains. *Working Papers,* Centre for Statistical and Survey Methodology, The University of Wollongong, Australia. (Available from: http://cssm.uow.edu.au/publications).

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

Chandra, H. and Chambers, R. (2009). Multipurpose weighting for small area estimation. *Journal of Official Statistics*, **25**, 3, 379-395.

Cochran, W.G. *(*1977)*. Sampling Techniques*, 3rd edition. Wiley & Sons, NY.

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

Eilers, P. and Marx, B. (1996). Flexible Smoothing using B-splines and Penalized Likelihood (with comments and rejoinder). *Statistical Science*, **11**, 1200-1224.

Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, **32**, 37-52.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320-338.

He, X. (1997). Quantile curves without crossing. *American Statistician*, **51**, 186–192.

McCulloch, C.E., and Searle, S.R. (2001). *Generalized Linear and Mixed Models*. Wiley, New York.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B*, **70**, 265-286.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution fucntions and quantiles from survey data using auxiliary information. *Biometrika*, **77**, 365-375.

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

Rueda, M., Martínez, S., Martínez, H. and Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, **137**, 435-448.

Rueda, M., Sánchez-Borrego, I., Arcos, A. and Martínez, S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika*, **71**, 33-44.

Ruppert, D., Wand, M.P. and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Royall, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association,* **71**, 657-664.

Royall, R.M. and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, **71**, 351-358.

Salvati, N., Chandra, H., Ranalli, M.G. and Chambers, R. (2010). Small area estimation using a nonparametric model-based direct estimator. *Computational Statistics and Data Analysis*, **54**, 2159-2171.

Tzavidis, N., Marchetti, S., and Chambers, R. (2010). Robust prediction of small area means and quantiles. *Australian and New Zealand Journal of Statistics*, **52**, 167-186.

Wand, M.P. (2003). Smoothing and mixed models. *Computational Statistics*, **18**, 223-249.

Table 1. Description of the estimators considered in the simulation studies.

Estimator	Description
MBDE	MBDE (19) with sample weights (17) based on model (18)
EBP	EBLUP-based EBP estimator (5) under linear mixed model (4)
CD	EBLUP-based CD estimator (6) under linear mixed model (4)
NPEBP	NPEBLUP-based EBP estimator (11) under spline-based mixed model (10)
NPCD	NPEBLUP-based CD estimator (12) under spline-based mixed model (10)

Table 2. Area averages of absolute relative bias (*ARB*, %) generated by model-based simulations.

Set	Population quantile	MBDE	EBP	CD	NPEBP	NPCD
1	0.10	2.41	71.94	1.24	71.83	1.28
	0.25	1.29	30.92	0.61	30.83	0.62
	0.50	0.84	2.61	0.40	2.65	0.39
	0.75	0.52	9.17	0.26	9.14	0.25
	0.90	0.27	5.46	0.15	5.43	0.15
2a	0.10	2.40	127.28	141.01	114.80	160.20
	0.25	1.30	3.13	17.97	4.57	24.39
	0.50	0.80	39.42	10.49	16.33	8.94
	0.75	0.51	19.18	9.05	7.42	8.97
	0.90	0.28	1.12	4.00	1.35	3.92
2b	0.10	2.18	444.41	344.70	175.30	202.23
	0.25	1.38	120.62	80.84	21.72	33.14
	0.50	0.79	13.75	5.82	17.00	10.75
	0.75	0.53	17.62	29.28	12.20	11.48
	0.90	0.29	17.36	23.47	3.30	5.67

Table 3. Area averages of relative root mean squared error (*RRMSE*, %) generated by model-based simulations.

Set	Population quantile	MBDE	EBP	CD	NPEBP	NPCD
1	0.10	63.22	82.54	38.12	82.52	38.21
	0.25	36.55	39.05	22.35	39.21	22.40
	0.50	21.17	15.25	12.76	15.45	12.78
	0.75	12.38	11.23	6.93	11.22	6.92
	0.90	7.16	6.46	3.61	6.43	3.60
2a	0.10	65.17	314.20	179.08	242.08	180.82
	0.25	37.57	115.75	41.11	80.16	36.71
	0.50	21.66	71.40	16.70	39.96	14.19
	0.75	12.54	37.44	11.32	18.98	10.97
	0.90	7.23	6.43	6.04	5.81	5.67
2b	0.10	64.88	455.68	351.48	297.19	218.47
	0.25	37.30	128.20	86.91	92.85	43.29
	0.50	21.53	26.30	18.50	44.63	16.34
	0.75	12.43	21.17	30.80	26.43	13.67
	0.90	7.19	18.70	24.98	10.84	7.47

Table 4. Average values over 12 regions of absolute relative bias (*ARB*, %) and relative root mean squared error (*RRMSE*, %) for the AAGIS data.

Population quantile	MBDE	EBP	CD	NPEBP	NPCD
			ARB (%)		
0.10	1.51	97.14	87.92	95.03	143.74
0.25	0.92	94.18	50.74	64.10	53.45
0.50	0.35	67.99	13.96	38.27	16.34
0.75	0.30	24.39	3.97	15.34	10.70
0.90	0.15	10.26	1.83	6.76	2.69
			RRMSE (%)		
0.10	47.75	131.26	108.26	117.60	155.65
0.25	23.40	114.07	59.25	81.29	58.53
0.50	14.48	81.50	19.17	45.62	19.06
0.75	7.59	29.43	8.53	20.23	12.26
0.90	3.81	10.67	4.20	8.51	4.36

Table 5. Average values over 23 HUCs of absolute relative bias (*ARB*,%) and relative root mean squared error (*RRMSE*,%) for the EMAP data.

Population quantile	MBDE	EBP	CD	NPEBP	NPCD
			ARB (%)		
0.10	2.10	71.13	32.37	50.85	21.14
0.25	0.74	51.53	17.20	42.38	18.74
0.50	0.67	43.44	13.83	33.09	11.86
0.75	0.43	21.92	6.22	18.12	9.17
0.90	0.25	11.55	2.23	11.92	3.61
			RRMSE (%)		
0.10	46.76	72.02	47.71	58.38	43.91
0.25	28.41	58.93	32.64	47.92	29.02
0.50	30.51	52.17	25.18	36.83	21.60
0.75	14.76	27.94	16.04	21.70	15.21
0.90	5.30	14.13	6.29	13.57	6.06

Table 6. Average values of true RMSE and estimated RMSE and actual coverage rate (CR, %) of nominal 95 per cent confidence intervals generated by the MBDE (19) and associated MSE estimator (20) for the AAGIS and EMAP data. Averages are over regions.

Population quantile	AAGIS			EMAP		
	True RMSE	Estimated RMSE	CR	True RMSE	Estimated RMSE	CR
0.10	0.034	0.034	89	0.018	0.021	95
0.25	0.051	0.052	91	0.041	0.041	92
0.50	0.061	0.061	95	0.054	0.055	93
0.75	0.051	0.051	94	0.052	0.058	93
0.90	0.031	0.032	90	0.028	0.034	93

Figure 1. Region-specific values of actual repeated sampling RMSE (solid line) and average estimated RMSE (dashed line) of MBDE (19) for the AAGIS data.

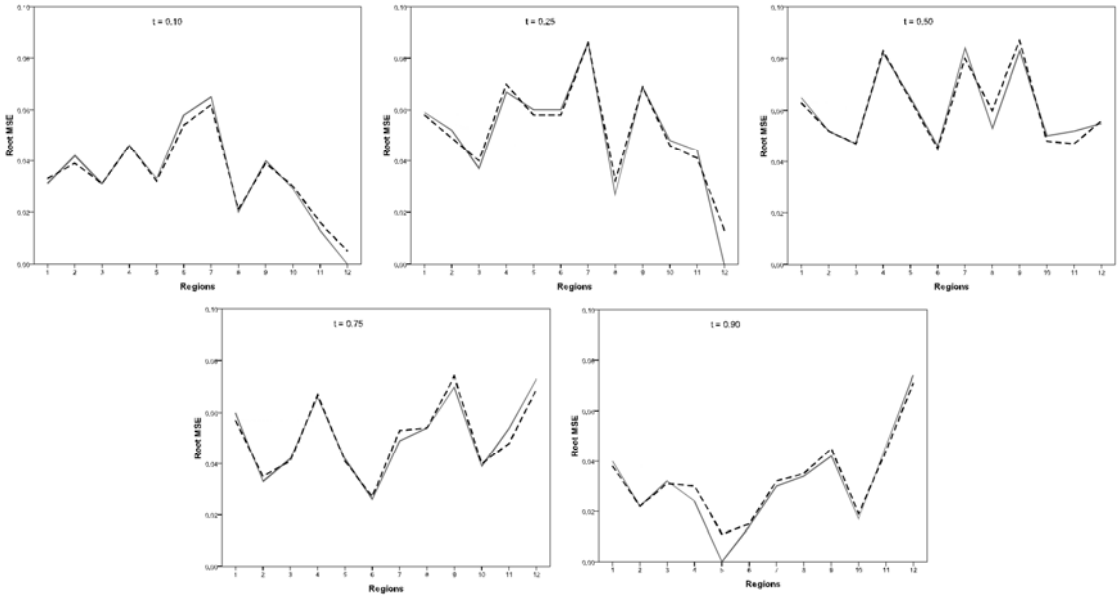

Figure 2. HUC-specific values of actual repeated sampling RMSE (solid line) and average estimated RMSE (dashed line) of MBDE (19) for the EMAP data.