Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

19-09

Small Area Estimation Using a Nonparametric Model Based Direct Estimator

Nicola Salvati, Hukm Chandra, M. Giovanna Ranalli and Ray Chambers

Copyright © 2008 by the Centre for Statistical & Survey Methodology, UOW. Work in progress, no part of this paper may be reproduced without permission from the Centre.

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

Small Area Estimation Using a Nonparametric Model Based Direct Estimator

Nicola Salvati[1], Hukum Chandra[2], M. Giovanna Ranalli[3] and Ray Chambers[4]

Abstract. Nonparametric regression is widely used as a method of characterising a non-linear relationship between a variable of interest and a set of covariates. Practical application of nonparametric regression methods in the field of small area estimation is fairly recent, and has so far focussed on the use of empirical best linear unbiased prediction under a model that combines a penalized spline (p-spline) fit and random area effects. In this paper, we propose an alternative approach to using nonparametric regression to estimate a small area mean, based on application of the recently introduced concept of model-based direct estimation. Under this approach, the estimator of the small area mean is a weighted average of the sample values from the area, with weights derived from an appropriately specified linear regression model with random area effects. Here we extend this model to incorporate a smooth, non-parametrically specified trend. Estimation of the mean squared error of the proposed small area estimator is also discussed. Monte Carlo simulations based on both simulated and real datasets show that the proposed model-based direct estimator and its associated mean squared error estimator perform well and are worth considering in small area estimation applications where the underlying population regression relationships are non-linear or have a complicated functional form.

Key words: Non-linear regression model; Empirical best linear unbiased prediction; Penalized splines; Mean squared error estimator; Unit level model.

[1]Corresponding Author: Dipartimento di Statistica e Matematica Applicata all'Economia, University of Pisa, Via Ridolfi, 10, 56124 - Pisa, Italy, phone: +39 (0)50 2216492, fax: +39 (0)50 2216375 E-mail: salvati@ec.unipi.it

[2]Indian Agricultural Statistics Research Institute, Library Avenue, PUSA, New Delhi-110012, India, Email: hchandra@iasri.res.in

[3]Dipartimento di Economia, Finanza e Statistica, Università degli Studi di Perugia, E-mail: giovanna@stat.unipg.it

[4]Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW, 2522, Australia. Email: ray@uow.edu.au

1. Introduction

Data collected in sample surveys are extensively used to provide reliable direct estimates of totals and means for target populations as well as for subpopulations defined by large areas or domains. However, using these data to obtain corresponding reliable estimates for smaller geographic areas and subpopulations can be problematic. An area is regarded as 'small' if the area sample size is not large enough to support a design-based direct estimator of adequate precision. Model-based small area estimation methods, e.g. those based on linear mixed models with area specific random effects, can lead to significant efficiency gains in such cases (for a review see Rao, 2003, or Jiang and Lahiri, 2006). These methods typically involve the use of indirect estimators, such as those based on empirical best linear unbiased prediction (EBLUP; Battese *et al*., 1988; Prasad and Rao, 1990).

However, indirect estimators are sensitive to model misspecification and so their high efficiency when the assumed model is true comes at a price - they can be quite biased under model misspecification. In contrast, direct estimators are generally less sensitive to model specification, but typically have large variability. Chandra and Chambers (2005, 2009) have recently proposed model-based direct estimation (MBDE) as a compromise between these two extremes. The MBDE for a small area mean is a weighted average of the sample values from the area, but with weights derived from a linear predictor of the corresponding population mean under a linear model with random area effects. The weights used in the MBDE are based on an extension of the approach by Royall (1976) to the case of a mixed effects model. These authors also show that estimation of the mean squared error (MSE) of the MBDE can be carried out using a simple pseudo-linearization type estimator.

When the functional form of the relationship between the response variable and the covariates is unknown or has a complicated functional form, an approach based on the use of a non-linear regression model can offer significant advantages compared with one based on a

linear model. In this paper we focus on non-parametric regression modelling based on a p-spline approximation to the true regression function (see Eilers and Marx, 1996 and Ruppert *et al.*, 2003). Wand (2003) shows how a p-spline model may be fitted by treating the spline coefficients as random effects in a linear mixed model. On the basis of this property, Opsomer *et al*. (2008) and Ugarte *et al.* (2009) have recently proposed a new approach to small area estimation that extends the unit level nested error regression model of Battese *et al.* (1988) by combining small area random effects with a p-spline regression model. In this paper we explore the extension of the MBDE to the case where the population level relationship between the variable of interest and a subset of the model covariates is non-linear. In particular, we use a p-spline to model this non-linearity, while at the same time including small area random effects in the model. The resulting nonparametric MBDE, hereafter NPMBDE, is then a weighted sum of the sample values from the small area of interest, with weights derived from the EBLUP of the population mean defined by this p-spline regression model with random area effects. MSE estimation for the NPMBDE follows using the same pseudo-linearization type estimator as used with the MBDE. As we show in Section 2.2, a spin-off of this approach to MSE estimation is an alternative MSE estimator for the nonparametric EBLUP, hereafter NPEBLUP, proposed by Opsomer *et al.* (2008).

The rest of the paper is organized as follows. Section 2 reviews the use of the p-spline model in small area estimation and develops the pseudo-linearization estimator of the MSE of the NPEBLUP. This model is then used to define the NPMBDE, along with an estimator of its MSE. The Section concludes with a discussion of the use of the nonparametric synthetic estimator, hereafter NPSYN, for areas without sample observations. The performances of these estimators are compared in Section 3 through two simulation studies: the first study uses model-based simulation of artificial populations; the second study is based on a real dataset from the Environmental Monitoring and Assessment Program (EMAP) survey of lakes in the

North-Eastern states of the United States. Concluding remarks are set out in Section 4, where the pros and cons of the use of the NPMBDE are summarized.

2. Small Area Estimation Based on Penalized Spline Regression

Nonparametric smoothing is a popular way of modelling a non-linear regression relationship, and smoothing models based on p-splines are particularly attractive because they represent a relatively straightforward extension of linear regression models (Eilers and Marx, 1996). In Section 2.1 we briefly summarize the p-spline regression model and, following Wand (2003) and Ruppert *et al*. (2003), show how it can be formulated in terms of a random effects model. We then explain its use in a small area estimation context following the proposal by Opsomer *et al.* (2008) and we use it in Section 2.2 to obtain an alternative mean squared error estimator. In Section 2.3 we extend the MBDE to define an estimator based on the p-spline regression model with small area effects and, finally, in Section 2.4 we discuss synthetic estimation based on this model.

2.1 The Nonparametric EBLUP

Let y_j denote the value of the variable of interest *y* and x_j the value of the auxiliary variable *x* associated with unit *j*. For simplicity, we focus on the univariate case here. Extension to bivariate smoothing is considered in Section 3.2. The underlying regression model is written as $y_j = m(x_j) + \varepsilon_j$, where ε_j are zero mean independent random variables. The function $m(x)$ is unknown and assumed to be approximated sufficiently well by

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

where p is the degree of the spline, $(t)_+^p = t^p$ if $t > 0$ and 0 otherwise, κ_k for $k = 1, \ldots, K$ is a set of fixed constants called knots, $\boldsymbol{\beta} = \left(\beta_0, ..., \beta_p\right)^T$ is the coefficient vector of the parametric portion of the model and $\boldsymbol{\gamma} = \left(\gamma_1, ..., \gamma_K\right)^T$ is the vector of spline coefficients. If the number of knots K is sufficiently large, the class of functions in expression (1) 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. Note that the approximating function $m(x, \boldsymbol{\beta}, \boldsymbol{\gamma})$ in equation (1) uses truncated polynomial basis functions for simplicity. Other basis functions, e.g. B-splines (Eilers and Marx, 1996) or radial functions, can be used; in particular we employ the latter option when smoothing in two dimensions in the design-based simulation study reported in Section 3.2.

Using a large number of knots in expression (1) 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 (1) from the actual data values subject to this constraint is equivalent to minimizing the following penalized loss function

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

Here λ is a Lagrange multiplier that controls the level of smoothness of the resulting fit and can be defined, for example, by (generalized) cross validation. Alternatively, Wand (2003) and sRuppert *et al*. (2003, Chap. 4) note the equivalence between minimizing expression (2) and maximizing the likelihood of the response variable and the spline coefficients under a random effects model, so that solution to expression (2) defines the BLUP for $m(x, \boldsymbol{\beta}, \boldsymbol{\gamma})$. In particular, let $\mathbf{y} = \left(y_1, y_2, ..., y_N\right)^T$, where N is the population size,

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & \cdots & x_N^p \end{bmatrix}$$

and

$$\mathbf{Z} = \begin{bmatrix} (x_1 - \kappa_1)_+^p & \cdots & (x_1 - \kappa_K)_+^p \\ \vdots & \ddots & \vdots \\ (x_N - \kappa_1)_+^p & \cdots & (x_N - \kappa_K)_+^p \end{bmatrix}.$$

The spline approximation of equation (1) can therefore be written as the linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}, \tag{3}$$

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

$$\boldsymbol{\gamma} \sim N_K(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_K) \text{ and } \mathbf{e} \sim N_N(\mathbf{0}, \sigma_e^2 \mathbf{I}_N), \tag{4}$$

where $\mathbf{I}_t$ denotes the identity matrix of dimension *t*. Opsomer *et al.* (2008) use p-splines in a small area estimation context by adding small area random effects to model (3) that capture the dissimilarities among small areas that are not explained by the covariates included in the model. Let *A* be the number of small areas. Then model (3) can be extended to

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{D}\mathbf{u} + \mathbf{e}, \tag{5}$$

where $\mathbf{D} = \left(\mathbf{d}_1, \ldots, \mathbf{d}_N\right)^T$ is a matrix of known covariates of dimension $N \times A$ characterising differences among the small areas and $\mathbf{u}$ is the *A*-vector of random area effects. In the simplest case, $\mathbf{D}$ 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 $\mathbf{u}$ are distributed independently of the spline effects $\boldsymbol{\gamma}$ and the individual effects $\mathbf{e}$, with $\mathbf{u} \sim N_A(\mathbf{0}, \sigma_u^2 \mathbf{I}_A)$, so that the covariance matrix of the vector $\mathbf{y}$ is given by

$$Var(\mathbf{y}) = \mathbf{V} = \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}^T + \sigma_u^2 \mathbf{D}\mathbf{D}^T + \sigma_e^2 \mathbf{I}_N.$$

The parameters σ_γ^2, σ_u^2 and σ_e^2 are typically referred to as the variance components of (5).

Throughout this paper we assume that the sampling method used is non-informative for the population values of *y* given the corresponding values of the auxiliary variables and knowledge of the area affiliations of the population units. As a consequence, the nonparametric model (5) represents our model for both sampled and non-sampled population units. It follows that we can partition **y**, **X**, **Z**, **D** and **e** into components defined by the *n* sampled and $N - n$ non-sampled population units, denoted by subscripts of *s* and *r* respectively. We can therefore write (5) as follows:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{bmatrix} = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_s \\ \mathbf{Z}_r \end{bmatrix} \boldsymbol{\gamma} + \begin{bmatrix} \mathbf{D}_s \\ \mathbf{D}_r \end{bmatrix} \mathbf{u} + \begin{bmatrix} \mathbf{e}_s \\ \mathbf{e}_r \end{bmatrix}, \tag{6}$$

with variance matrix given by

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}. \tag{7}$$

Thus, $\mathbf{X}_s$ in (6) represents the matrix defined by the *n* sample values of the auxiliary variable vector, while $\mathbf{V}_{rr}$ in (7) is the matrix of covariances of the response variable among the $N - n$ non-sampled units. We use a subscript of *i* to denote restriction to small area *i*, so that N_i (n_i) denotes the population (sample) size in small area *i*. The overall population size is $N = \sum_{i=1}^{A} N_i$, and the total sample size is $n = \sum_{i=1}^{A} n_i$. Similarly, s_i (r_i) denotes the set of sample (non-sample) population units from area *i*, and $U_i = s_i \cup r_i$ denotes the set of population units making up small area *i*.

When the variance components are known, well-established theory (see McCulloch and Searle, 2001, Chapter 9) suggests the following generalised least squares estimator of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{y}_s, \tag{8}$$

and the following best linear unbiased predictors (BLUPs) for $\boldsymbol{\gamma}$ and **u**

$$\hat{\boldsymbol{\gamma}} = \sigma_\gamma^2 \mathbf{Z}_s^T \mathbf{V}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}) \text{ and } \hat{\mathbf{u}} = \sigma_u^2 \mathbf{D}_s^T \mathbf{V}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}). \quad (9)$$

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 $\left(\hat{\sigma}_\gamma^2, \hat{\sigma}_u^2, \hat{\sigma}_e^2\right)$ to denote such estimates, allowing us to define the plug-in estimator $\hat{\mathbf{V}}_{ss} = \hat{\sigma}_\gamma^2 \mathbf{Z}_s \mathbf{Z}_s^T + \hat{\sigma}_u^2 \mathbf{D}_s \mathbf{D}_s^T + \hat{\sigma}_e^2 \mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of order *n*. This leads to the empirical best linear unbiased estimator $\hat{\boldsymbol{\beta}}^{EBLUE} = (\mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{y}_s$, and to the empirical BLUPs $\hat{\boldsymbol{\gamma}}^{EBLUP} = \hat{\sigma}_\gamma^2 \mathbf{Z}_s^T \hat{\mathbf{V}}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}^{EBLUE})$ and $\hat{\mathbf{u}}^{EBLUP} = \left(\hat{u}_i^{EBLUP}\right) = \hat{\sigma}_u^2 \mathbf{D}_s^T \hat{\mathbf{V}}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}^{EBLUE})$.

Under (5), the EBLUP for the mean $\bar{y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_j$ of *y* in small area *i* is

$$\hat{\bar{y}}_i^{NPEBLUP} = N_i^{-1}\left[\sum\nolimits_{j \in s_i} y_j + \sum\nolimits_{j \in r_i} \hat{y}_j^{NPEBLUP}\right], \quad (10)$$

where $\hat{y}_j^{NPEBLUP} = \mathbf{x}_j^T \hat{\boldsymbol{\beta}}^{EBLUE} + \mathbf{z}_j^T \hat{\boldsymbol{\gamma}}^{EBLUP} + \mathbf{d}_j^T \hat{\mathbf{u}}^{EBLUP}$, and $\mathbf{x}_j^T$, $\mathbf{z}_j^T$ and $\mathbf{d}_j^T$ denote respectively the rows of **X**, **Z** and **D** that correspond to unit *j* in area *i*. The EBLUP (10) is referred to as the nonparametric EBLUP or NPEBLUP by Opsomer *et al.* (2008), who study its theoretical properties and propose both an analytical estimator and a more computationally intensive nonparametric bootstrap-based estimator for its MSE. In particular, these authors provide a second order approximation to the mean squared error of the NPEBLUP. This is

$$\begin{aligned} MSE\left(\hat{\bar{y}}_i^{NPEBLUP}\right) = {} & \mathbf{c}_{is}\left(\mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{X}_s\right)^{-1} \mathbf{c}_{is}^T + \left(\sigma_\gamma^2 \bar{\mathbf{z}}_{is}^T \bar{\mathbf{z}}_{is} + \sigma_u^2 \bar{\mathbf{d}}_{is}^T \bar{\mathbf{d}}_{is}\right) \\ & - \left(\sigma_\gamma^2 \mathbf{Z}_s \bar{\mathbf{z}}_{is} + \sigma_u^2 \mathbf{D}_s \bar{\mathbf{d}}_{is}\right)^T \mathbf{V}_{ss}^{-1}\left(\sigma_\gamma^2 \mathbf{Z}_s \bar{\mathbf{z}}_{is} + \sigma_u^2 \mathbf{D}_s \bar{\mathbf{d}}_{is}\right) + tr\left(\mathbf{Q}_s \mathbf{V}_{ss} \mathbf{Q}_s^T \mathfrak{I}^{-1}\right)^T, \end{aligned} \quad (11)$$

where $\mathbf{c}_{is} = \bar{\mathbf{x}}_{is}^T - \left(\sigma_\gamma^2 \mathbf{Z}_s \bar{\mathbf{z}}_{is} + \sigma_u^2 \mathbf{D}_s \bar{\mathbf{d}}_{is}\right)^T \mathbf{V}_{ss}^{-1} \mathbf{X}_s$, and $\bar{\mathbf{x}}_{is}$, $\bar{\mathbf{z}}_{is}$ and $\bar{\mathbf{d}}_{is}$ are the column vectors defined by averaging the columns of the matrices $\mathbf{X}_{is}$, $\mathbf{Z}_{is}$ and $\mathbf{D}_{is}$ respectively. The matrix $\mathbf{Q}_s$ is of order $3 \times n$, with *f*-th row given by

$$\mathbf{Q}_{sf} = \overline{\mathbf{z}}_{is}^T \frac{\partial \sigma_\gamma^2 \mathbf{I}_K}{\partial (\sigma^2)_f} \mathbf{Z}_s^T \mathbf{V}_{ss}^{-1} + \overline{\mathbf{d}}_{is}^T \frac{\partial \sigma_u^2 \mathbf{I}_A}{\partial (\sigma^2)_f} \mathbf{D}_s^T \mathbf{V}_{ss}^{-1} + \sigma_\gamma^2 \overline{\mathbf{z}}_{is}^T \mathbf{Z}_s^T \frac{\partial \mathbf{V}_{ss}^{-1}}{\partial (\sigma^2)_f} + \sigma_u^2 \overline{\mathbf{d}}_{is}^T \mathbf{D}_s^T \frac{\partial \mathbf{V}_{ss}^{-1}}{\partial (\sigma^2)_f},$$

for f = 1, 2, 3 and $\sigma^2 = \left(\sigma_\gamma^2, \sigma_u^2, \sigma_e^2\right)$, and the 3×3 matrix $\Im$ is the Fisher information matrix with respect to the vector σ^2 of variance components. If we use a 'hat' to denote substitution of estimators for unknown parameters in an expression, then the analytical estimator of the MSE of the NPEBLUP proposed in Opsomer et al. (2008) is

$$\begin{aligned} \widehat{MSE}\left(\hat{\bar{y}}_i^{NPEBLUP}\right) = {} & \hat{\mathbf{c}}_{is}\left(\mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{X}_s\right)^{-1} \hat{\mathbf{c}}_{is}^T + \left(\hat{\sigma}_\gamma^2 \overline{\mathbf{z}}_{is}^T \overline{\mathbf{z}}_{is} + \hat{\sigma}_u^2 \overline{\mathbf{d}}_{is}^T \overline{\mathbf{d}}_{is}\right) \\ & - \left(\hat{\sigma}_\gamma^2 \mathbf{Z}_s \overline{\mathbf{z}}_{is} + \hat{\sigma}_u^2 \mathbf{D}_s \overline{\mathbf{d}}_{is}\right)^T \hat{\mathbf{V}}_{ss}^{-1} \left(\hat{\sigma}_\gamma^2 \mathbf{Z}_s \overline{\mathbf{z}}_{is} + \hat{\sigma}_u^2 \mathbf{D}_s \overline{\mathbf{d}}_{is}\right) \\ & + 2\left(\mathbf{y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}}\right)^T \left(\hat{\mathbf{Q}}_s^T \hat{\Im}^{-1} \hat{\mathbf{Q}}_s\right)^T \left(\mathbf{y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}}\right). \end{aligned} \tag{12}$$

These authors also propose testing the significance of the small area effects and the spline components of model (5) using a restricted likelihood ratio test and a nonparametric bootstrap-based test. Similarly, Lombardia and Sperlich (2008) propose a bootstrap procedure to test for non-linearity when fitting a generalized linear mixed model where the nonparametric trend is approximated using kernel-based methods.

2.2 An Estimator of the Conditional MSE of the NPEBLUP

Standard methods of MSE estimation for small area EBLUPs (e.g. Prasad and Rao, 1990) focus on estimation of its unconditional MSE, averaging over the distribution of the random area effects. An alternative measure of uncertainty can be obtained by conditioning on the realised values of the area effects (see Longford, 2007). In what follows we propose an estimator of the conditional MSE of the NPEBLUP that is much less computationally demanding than the unconditional MSE estimators suggested by Opsomer *et al*. (2008). The proposed estimator is based on the pseudo-linearization approach to MSE estimation described by Chambers *et al*. (2007). See also Chandra and Chambers (2005, 2009) and

Chandra *et al*. (2007). It is motivated by first re-expressing the NPEBLUP (10) in a pseudo-linear form, i.e. as a weighted sum of the sample values of *y*, and then applying heteroskedasticity-robust prediction variance estimation methods that treat these weights (which typically depend on estimated variance components) as known. More precisely, we note that the BLUP of $\bar{y}_i$ under (5) (i.e. the NPBLUP) can be expressed as

$$\hat{\bar{y}}_i^{NPBLUP} = \sum_{j \in s} w_{ij}^{NPBLUP} y_j = \left(\mathbf{w}_{is}^{NPBLUP}\right)^T \mathbf{y}_s,\ i \in 1 \ldots A, \tag{13}$$

where

$$\left(\mathbf{w}_{is}^{NPBLUP}\right)^T = N_i^{-1}\left\{\mathbf{d}_{is}^T + (N_i - n_i)\left[\bar{\mathbf{x}}_{ir}^T \mathbf{H}_s + \left(\sigma_\gamma^2 \mathbf{Z}_s \bar{\mathbf{z}}_{ir} + \sigma_u^2 \mathbf{D}_s \bar{\mathbf{d}}_{ir}\right)^T \mathbf{V}_{ss}^{-1}\left(\mathbf{I}_n - \mathbf{X}_s \mathbf{H}_s\right)\right]\right\}.$$

Here $\mathbf{d}_{is}$ is the *i*-th column of $\mathbf{D}_s$, $\mathbf{H}_s = \left(\mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1}$ and $\bar{\mathbf{x}}_{ir}$, $\bar{\mathbf{z}}_{ir}$ and $\bar{\mathbf{d}}_{ir}$ are the vectors defined by averaging the columns of the matrices $\mathbf{X}_{ir}$, $\mathbf{Z}_{ir}$ and $\mathbf{D}_{ir}$ respectively. The Appendix provides details on the computation of such weights. Note that the NPEBLUP (10) can be expressed in exactly the same way, except that all values in the vector $\mathbf{w}_{is}^{NPBLUP}$ that depend on (unknown) variance components now need a 'hat'. Given this 'pseudo-linear' representation for the NPEBLUP, we show how well-known results in model-based survey sampling theory can be used to develop a simple first order approximation to its MSE that is robust to misspecification of the variances of the individual effects e_j in (5).

In what follows, we assume the conditional version of the nonparametric model (5), with random effects (both the spline effect and the area effect) considered as fixed. In this case expression (5) corresponds to a fixed effects linear model, with regression parameters that vary from area to area, and we can apply the approach described by Royall and Cumberland (1978) to estimate the prediction variance of the NPBLUP for $\bar{y}_i$. Let $I(j \in i)$ denote the indicator for whether unit *j* is in area *i*. Then

$$Var\left(\hat{\bar{y}}_i^{NPBLUP} - \bar{y}_i \middle| \mathbf{X},\boldsymbol{\gamma},\mathbf{u}\right) = N_i^{-2}\left\{\sum\nolimits_{j\in s}\left(N_i w_{ij}^{NPBLUP} - I(j\in i)\right)^2 Var(y_j \middle| x_j,\boldsymbol{\gamma},\mathbf{u}) + \sum\nolimits_{j\in r_i} Var(y_j \middle| x_j,\boldsymbol{\gamma},\mathbf{u})\right\}.$$

The first term on the right hand side above is the leading term, and is estimated by replacing $Var(y_j \middle| x_j,\boldsymbol{\gamma},\mathbf{u})$ by $\lambda_j^{-1}(y_j - \hat{\mu}_j)^2$, where $\hat{\mu}_j = \sum_{k\in s}\phi_{kj} y_k$ is an unbiased linear estimator of the conditional expected value $\mu_j = E(y_j \middle| x_j,\boldsymbol{\gamma},\mathbf{u})$, the ϕ_{kj} are weights that are defined implicitly by the expression for $\hat{\mu}_j$ in (15) below, and $\lambda_j = \left\{1 - 2\phi_{jj} + \sum_{k\in s}\phi_{kj}^2\right\}$ is a scaling constant that ensures

$$\lambda_j^{-1} E\left\{(y_j - \hat{\mu}_j)^2 \middle| x_j,\boldsymbol{\gamma},\mathbf{u}\right\} = Var\left(y_j \middle| x_j,\boldsymbol{\gamma},\mathbf{u}\right)$$

under the more restrictive assumption that $Var(e_j) = \sigma_e^2$ in model (5). This homoskedasticity assumption can be invoked once again to allow unbiased estimation of the second order term in the prediction variance above using the sample average of these scaled residuals. The final result is an estimator of the conditional prediction variance of the NPBLUP of the form

$$\hat{V}(\hat{\bar{y}}_i^{NPBLUP}) = N_i^{-2}\sum\nolimits_{j\in s}\left\{a_{ij}^2 + (N_i - n_i)n^{-1}\right\}\hat{\lambda}_j^{-1}(y_j - \hat{\mu}_j)^2, \qquad (14)$$

where $a_{ij} = N_i w_{ij}^{NPBLUP} - I(j\in i)$ and $\hat{\lambda}_j$ is defined by substituting estimated variance components in the expression for λ_j above.

Since area sample sizes are usually small, obtaining a stable and conditionally unbiased linear estimator of μ_j under the conditional version of model (5) is problematic in the small area context. Therefore, an initial specification for $\hat{\mu}_j$ is to replace it by the BLUP of μ_j under (5), i.e.

$$\hat{\mu}_j = \mathbf{x}_j^T\mathbf{H}_s\mathbf{y}_s + \left\{\sigma_\gamma^2\mathbf{z}_j^T\mathbf{Z}_s^T + \sigma_u^2\mathbf{d}_j^T\mathbf{D}_s^T\right\}\mathbf{V}_{ss}^{-1}\left(\mathbf{y}_s - \mathbf{X}_s\mathbf{H}_s\mathbf{y}_s\right).$$

However, because of the well-known shrinkage effect associated with BLUPs, this specification leads to biased estimation of the prediction variance under the conditional model. Consequently, Chambers *et al.* (2007) recommend that $\hat{\mu}_j$ be computed as the 'unshrunken' version of the BLUP for μ_j. However, before applying this idea, we note that expression (5) also includes random spline coefficients that are inappropriate to 'unshrink', and so we suggest that any such 'unshrinking' be restricted to the BLUP for the area effect that is implicit in $\hat{\mu}_j$ above. This corresponds to setting

$$\hat{\mu}_j = \mathbf{x}_j^T \mathbf{H}_s \mathbf{y}_s + \sigma_\gamma^2 \mathbf{z}_j^T \mathbf{Z}_s^T \mathbf{V}_{ss}^{-1} (\mathbf{y}_s - \mathbf{X}_s \mathbf{H}_s \mathbf{y}_s) + \mathbf{d}_j^T \tilde{\mathbf{u}}, \tag{15}$$

where $\tilde{\mathbf{u}} = \left(\mathbf{D}_s^T \mathbf{D}_s\right)^{-1} \mathbf{D}_s^T \left(\mathbf{y}_s - \mathbf{X}_s \mathbf{H}_s \mathbf{y}_s\right)$ replaces the BLUP $\hat{\mathbf{u}}^{BLUP} = \sigma_u^2 \mathbf{D}_s^T \mathbf{V}_{ss}^{-1} \left(\mathbf{y}_s - \mathbf{X}_s \mathbf{H}_s \mathbf{y}_s\right)$ of the vector of area effects. Note that $\hat{\lambda}_j = 1 + O(n^{-1})$ in this case, so that $\hat{\lambda}_j$ will be very close to one in most practical applications. This suggests that there is little to be gained by not setting $\hat{\lambda}_j \equiv 1$ when calculating the conditional prediction variance (14). This simplification is used in the simulation studies described in the next Section.

Next, we note that the conditional bias of the NPBLUP (13) under (5) is given by

$$E\left(\hat{\bar{y}}_i^{NPBLUP} - \bar{y}_i \middle| \mathbf{X}, \boldsymbol{\gamma}, \mathbf{u}\right) = \sum\nolimits_{j \in s} w_{ij}^{NPBLUP} \mu_j - N_i^{-1} \sum\nolimits_{j \in U_i} \mu_j ,$$

which has the simple 'plug-in' estimator

$$\hat{B}(\hat{\bar{y}}_i^{NPBLUP}) = \sum\nolimits_{j \in s} w_{ij}^{NPBLUP} \hat{\mu}_j - N_i^{-1} \sum\nolimits_{j \in U_i} \hat{\mu}_j ,$$

with $\hat{\mu}_j$ defined by expression (15). Collecting terms, the estimator of the conditional MSE of the NPBLUP (13) is then

$$\hat{M}(\hat{\bar{y}}_i^{NPBLUP}) = \hat{V}(\hat{\bar{y}}_i^{NPBLUP}) + \left\{\hat{B}(\hat{\bar{y}}_i^{NPBLUP})\right\}^2 . \tag{16}$$

Finally, we estimate the conditional MSE of the NPEBLUP (10) by replacing all unknown variance components in (16) by their estimated values. Note that this MSE estimator 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 NPEBLUP. Since use of the NPEBLUP (10) will typically require a large overall sample size, we expect that any consequent underestimation of the conditional MSE of the NPEBLUP will be small. 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. However, in a realistic small area estimation application where use of a linear mixed model is appropriate, Chambers *et al*. (2007) report median relative biases of 0.1% and -0.8% when this pseudo-linearization approach is used to estimate the conditional MSE of the EBLUP and the MBDE respectively.

2.3 The Nonparametric MBDE

Direct estimation for small areas is simple to implement and to interpret, since the estimated value of the mean for area *i* is just a weighted average of the sample data from this area. The same is not true of indirect estimators like the EBLUP, which can only be represented as weighted sums over the entire sample. Unfortunately, when these weights are the inverses of sample inclusion probabilities, the direct estimator can be quite inefficient. The Model-Based Direct Estimator (MBDE) of a small area mean improves upon the efficiency of the design-based direct estimator by using weights that define the EBLUP for the population total (see Royall, 1976) under the same linear mixed model with random area effects that underpins the EBLUP for the small area mean. That is, if the weights w_j^{EBLUP} define the EBLUP for the population total of the y_j, then the MBDE (Chandra and Chambers, 2005) of the area *i* mean of these values is

$$\hat{\bar{y}}_i^{MBDE} = \sum_{j \in s_i} w_j^{EBLUP} y_j \Big/ \sum_{j \in s_i} w_j^{EBLUP} .$$

Given that the nonparametric model (5) holds, the vector of sample weights that defines the corresponding EBLUP (i.e. the NPEBLUP) of the population total of the y_j is

$$\mathbf{w}_s^{NPEBLUP} = \left(w_j^{NPEBLUP} \right) = \mathbf{1}_n + \hat{\mathbf{H}}_s^T (\mathbf{X}^T \mathbf{1}_N - \mathbf{X}_s^T \mathbf{1}_n) + (\mathbf{I}_n - \hat{\mathbf{H}}_s^T \mathbf{X}_s^T) \hat{\mathbf{V}}_{ss}^{-1} \hat{\mathbf{V}}_{sr} \mathbf{1}_{N-n}, \qquad (17)$$

where, as usual, a 'hat' denotes substitution of estimated variance components, and $\hat{\mathbf{H}}_s = \left(\mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{X}_s \right)^{-1} \mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1}$, $\hat{\mathbf{V}}_{ss} = \hat{\sigma}_\gamma^2 \mathbf{Z}_s \mathbf{Z}_s^T + \hat{\sigma}_u^2 \mathbf{D}_s \mathbf{D}_s^T + \hat{\sigma}_e^2 \mathbf{I}_n$ and $\hat{\mathbf{V}}_{sr} = \hat{\sigma}_\gamma^2 \mathbf{Z}_s \mathbf{Z}_r^T + \hat{\sigma}_u^2 \mathbf{D}_s \mathbf{D}_r^T$. The nonparametric model-based direct estimator (NPMBDE) of small area *i* mean of *y* is therefore

$$\hat{\bar{y}}_i^{NPMBDE} = \sum\nolimits_{j \in s_i} w_j^{NPEBLUP} y_j \Big/ \sum\nolimits_{j \in s_i} w_j^{NPEBLUP} . \qquad (18)$$

Note that we refer to the NPMBDE (18) as a direct estimator because it is a weighted average of the sample data from the small area of interest. However, this does not mean that it can be calculated just using these data, since the weights (17) are a function of the data from the entire sample. That is, they 'borrow strength' from other areas through the model (5).

If the assumed linear mixed model underpinning the EBLUP actually holds, then this indirect estimator will be superior to the MBDE (as it should be). However, the MBDE is typically more efficient than the design-based direct estimator, and, as the simulation results described by Chandra and Chambers (2005, 2009) and Chandra *et al*. (2007) attest, it is often at least as, efficient as the EBLUP in situations where the assumed linear model is misspecified.

MSE estimation for the NPMBDE is carried out using the same heteroskedasticity-robust pseudo-linearization approach as that leading to the MSE estimator (16) for the NPEBLUP. The only difference from the development in Section 2.2 is that the weights w_j^{NPBLUP} used in expression (14) are now replaced by corresponding NPMBDE weights

$$w_j^{NPMBDE} = I(j \in s_i) w_j^{NPEBLUP} \Big/ \sum\nolimits_{k \in s_i} w_k^{NPEBLUP} .$$

2.4 Synthetic Nonparametric Prediction

In some situations we are interested in estimating the characteristics of small areas containing no sample observations. The conventional approach to estimating the area mean in this case is synthetic estimation (Rao, 2003, page 46), based on a suitable mixed model fitted to the data from the sampled areas. This is equivalent to setting the area effect for the non-sampled area to zero. Exactly the same approach can be taken with the spline-based small area model (5). When geo-referenced population location data are available, and spline smoothing is over these locations (e.g. using radial basis functions), the nonparametric model (5) is effectively accounting for spatial correlation in the population values of *y* over and above that 'explained' by the random area effects. In this case, model (5) has the potential to improve conventional synthetic estimation for out of sample areas. We therefore define the nonparametric synthetic (NPSYN) predictor for area *i* as

$$\hat{\bar{y}}_i^{NPSYN} = N_i^{-1}\left[\sum\nolimits_{j\in U_i}\left(\mathbf{x}_j^T\hat{\boldsymbol{\beta}} + \mathbf{z}_j^T\hat{\boldsymbol{\gamma}}\right)\right]. \tag{19}$$

That is, NPSYN is NPEBLUP with $\hat{u}_i^{EBLUP} = 0$. Since MBDE-type estimators cannot be computed for out-of-sample areas, we propose using NPSYN for those areas, even though, like the traditional synthetic estimator, this estimator can be biased. MSE estimation of NPSYN is also implemented using (16), but now based on weights defined by expressing (19) as a weighted sum of the sample values of *y*, i.e. we write

$$\hat{\bar{y}}_i^{NPSYN} = \left(\mathbf{w}_{is}^{NPSYN}\right)^T \mathbf{y}_s,$$

with

$$\left(\mathbf{w}_{is}^{NPSYN}\right)^T = (w_{ij}^{NPSYN}) = \bar{\mathbf{x}}_i^T\hat{\mathbf{H}}_s + \hat{\sigma}_\gamma^2\bar{\mathbf{z}}_i^T\mathbf{Z}_s^T\hat{\mathbf{V}}_{ss}^{-1}\left(\mathbf{I}_n - \mathbf{X}_s\hat{\mathbf{H}}_s\right).$$

3. Simulation Studies

In this Section we investigate the performance of different estimators for small area means through two different types of Monte Carlo simulation experiments. In Section 3.1 we report results from model-based simulations with a single covariate and in Section 3.2 we report on a design-based simulation study with a single covariate and geo-referenced data. In the model-based simulations we generate synthetic populations under a variety of models, then draw samples from them. In contrast, in the design-based simulation study we assess the performance of the estimators described in the previous Section in the context of a real population and realistic sampling methods, using the Environmental Monitoring and Assessment Program (EMAP) survey data provided by the Space Time Aquatic Resources Modelling and Analysis Program (STARMAP) at Colorado State University.

The three estimators considered in the simulations are the NPEBLUP (10), the NPSYN (19) and the NPMBDE (18). The MSEs for the NPEBLUP and the NPSYN are estimated using the analytical estimator (12) as well as the heteroskedasticity-robust pseudo-linearization MSE estimator (16). MSE estimation for the NPMBDE uses expression (16) only. The performance of the estimators is evaluated by computing for each small area the Average Relative Bias (*ARBias*) and the Average Relative Root MSE (*ARRMSE*) defined as follows:

$$ARBias_i = \left(T^{-1} \sum_{t=1}^{T} \bar{y}_{it} \right)^{-1} \left\{ T^{-1} \sum_{t=1}^{T} \left(\hat{\bar{y}}_{it} - \bar{y}_{it} \right) \right\}$$

and

$$ARRMSE_i = \left(T^{-1} \sum_{t=1}^{T} \bar{y}_{it} \right)^{-1} \left\{ \sqrt{T^{-1} \sum_{t=1}^{T} \left(\hat{\bar{y}}_{it} - \bar{y}_{it} \right)^2} \right\}.$$

Here $\bar{y}_{it}$ denotes the actual area *i* mean at simulation *t*, with predicted value $\hat{\bar{y}}_{it}$. Note that in the design-based simulation study $\bar{y}_{it} = \bar{y}_i$.

3.1 Model-Based Simulations

Model-based simulations are a common way of illustrating the sensitivity of an estimation procedure to variation in assumptions about the structure of the population of interest. Here we fix the number of small areas *at A* = 30 and use the following three types of models to generate the population values of y_j. In particular, for unit *j* in area *i* we generate values $y_j = m(x_j) + u_i + \varepsilon_j$, $j = 1,....,N_i; i = 1,...,A$, with $m(x)$ specified as follows:

Specification	$m(x)$
Linear	$1+2\left(x-0.5\right)$
Cycle	$2+100\sin(2\pi x)$
Exponential	$\exp(6x)/400$

In all three cases *x* values are independently drawn from a Uniform distribution on [0, 1] and the random area effects u_i are generated as *A* independent realizations from a $N\left(0,0.04\right)$ distribution. Note that the Linear specification above corresponds to a situation in which the NPEBLUP, NPSYN and NPMBDE estimators may not be appropriate. In contrast, the Cycle and Exponential specifications represent more complicated relationships between *y* and *x.* Three different settings are used when generating the individual error terms ε_j: (a) 'regularly' noisy data with ε_j distributed as $N\left(0,0.16\right)$, corresponding to an intra area correlation $\sigma_u^2 \big/ \left(\sigma_u^2 + \sigma_\varepsilon^2\right) = 0.2$; (b) more noisy data with the likely presence of extreme and outlying observations, with the ε_j distributed as Cauchy with location parameter 0 and scale parameter 0.05; and (c) skewed individual errors with $\varepsilon_j \sim \chi^2(3) - 3$, i.e. mean corrected chi-squared variates with 3 degrees of freedom. In (b) the intra-area correlation is not defined, while in (c) it is rather low, with $\sigma_u^2 \big/ \left(\sigma_u^2 + \sigma_\varepsilon^2\right) = 0.0066$. This provides a 3×3 design for the

simulations. The small area population sizes N_i are randomly drawn from a uniform distribution on [100,300] 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 =389 from the population and noting the resulting sample sizes in each small area. These area specific sample sizes n_i are fixed in the simulations by treating the small areas as strata and carrying out stratified random sampling. A total of T = 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. For each sample drawn, the mean of each small area is estimated by the NPEBLUP, the NPSYN and the NPMBDE. Estimates of the corresponding MSEs of these estimators are also calculated.

Table 1 shows the absolute values of *ARBias* and Table 2 shows the values of *ARRMSE* that are obtained in the simulations. Both tables show the mean and the five-point summary (minimum, first quartile, median, third quartile and maximum) of the distribution of these ratios over the small areas.

Two things stand out in Tables 1 and 2. The first is that the NPMBDE offers substantial gains over the NPEBLUP and NPSYN in terms of lower absolute bias irrespective of whether the underlying population structure is linear and the usual mixed model assumptions hold or when the relationship between y and x is complicated and/or the usual mixed model distributional assumptions are invalid. Second, under Gaussian and Chi-squared errors the NPMBDE records lower median values of ARRMSE than both the NPEBLUP and the NPSYN when the relationship between y and x is complicated (i.e. under the Cycle and Exponential specifications). On the other hand, under the Linear specification, the NPMBDE tends to be less efficient than these two alternatives. One would typically not use a nonparametric small area regression model like (5) if the relationship between y and x is linear, employing instead a more standard linear mixed model. Consequently, this lack of

efficiency is of limited consequence. Although these results are not presented in this paper, we also calculated the values of the standard linear mixed model-based EBLUP and MBDE in this simulation study and observed that these estimators, unsurprisingly, perform substantially better than the nonparametric model-based predictors under the Linear specification. However, they (again, as one would expect) perform extremely poorly under the Cycle and Exponential specifications.

Table 1. Across areas distribution of *ARBias* generated by model-based simulations.

Error	Estimator	Min	Q1	Median	Mean	Q3	Max
			Linear				
Gaussian	NPSYN	-53.31	-2.49	0.77	3.24	4.54	110.80
	NPEBLUP	-18.46	-0.71	-0.11	0.98	1.05	41.75
	NPMBDE	-2.63	-0.30	-0.07	-0.19	0.01	0.68
Cauchy	NPSYN	-74.22	-4.77	-1.02	1.91	3.85	114.67
	NPEBLUP	-72.95	-3.90	-1.09	2.84	2.31	110.05
	NPMBDE	-9.36	-0.75	-0.02	1.32	1.32	18.43
Chi-squared	NPSYN	-84.56	-3.61	0.46	2.11	2.91	113.37
	NPEBLUP	-78.93	-3.16	0.35	1.79	2.80	97.70
	NPMBDE	-13.00	-2.39	-0.14	-0.69	1.11	9.43
			Cycle				
Gaussian	NPSYN	-1.78	-0.44	-0.09	0.12	0.32	3.26
	NPEBLUP	-1.13	-0.34	0.00	-0.12	0.18	0.37
	NPMBDE	-1.01	-0.03	0.02	0.03	0.04	1.26
Cauchy	NPSYN	-2.40	-0.50	-0.14	0.23	0.27	5.93
	NPEBLUP	-0.66	-0.27	-0.03	0.10	0.22	2.59
	NPMBDE	-4.79	-0.09	0.01	-0.08	0.12	2.37
Chi-squared	NPSYN	-9.29	-0.84	-0.06	-0.24	0.33	6.79
	NPEBLUP	-2.68	-0.31	-0.07	0.06	0.35	2.86
	NPMBDE	-5.79	-0.13	-0.06	-0.06	0.11	3.54
			Exponential				
Gaussian	NPSYN	-677.24	-0.17	0.22	-4.31	1.60	429.44
	NPEBLUP	-256.80	0.00	0.25	3.53	1.27	320.90
	NPMBDE	-32.20	-0.17	0.00	-1.22	0.05	15.21
Cauchy	NPSYN	-179.12	-0.88	0.09	5.33	0.55	307.07
	NPEBLUP	-98.94	-0.86	0.01	0.35	0.16	72.33
	NPMBDE	-43.23	-0.18	0.03	4.53	0.28	70.97
Chi-squared	NPSYN	-209.35	-0.64	0.20	-7.43	1.54	141.86
	NPEBLUP	-155.93	-14.85	-0.24	-20.76	0.22	44.40
	NPMBDE	-96.14	-1.46	-0.10	-1.52	0.05	123.05

Table 2. Across areas distribution of *ARRMSE* generated by model-based simulations.

Error	Estimator	Min	Q1	Median	Mean	Q3	Max
			Linear				
Gaussian	NPSYN	0.82	2.44	4.50	11.93	9.80	113.80
	NPEBLUP	1.38	2.03	2.86	7.69	5.58	49.57
	NPMBDE	1.67	2.20	3.53	8.40	5.93	39.32
Cauchy	NPSYN	3.08	4.42	6.48	20.78	12.51	125.84
	NPEBLUP	4.00	7.62	10.28	37.64	30.05	164.03
	NPMBDE	2.26	9.21	19.81	66.12	63.68	509.80
Chi-squared	NPSYN	4.30	5.83	7.18	20.54	14.20	122.65
	NPEBLUP	4.92	6.16	8.06	21.66	15.39	115.08
	NPMBDE	11.72	17.09	21.15	50.37	49.06	223.66
			Cycle				
Gaussian	NPSYN	0.27	0.59	0.79	1.43	1.65	9.50
	NPEBLUP	0.24	0.44	0.68	1.04	0.98	9.03
	NPMBDE	0.20	0.30	0.56	0.84	0.76	7.88
Cauchy	NPSYN	0.28	0.63	1.14	2.56	2.43	23.17
	NPEBLUP	0.24	0.76	1.37	3.82	2.71	36.11
	NPMBDE	0.22	0.65	1.38	4.68	2.96	42.86
Chi-squared	NPSYN	0.58	0.98	1.50	4.02	3.08	43.55
	NPEBLUP	0.60	0.93	1.44	3.81	2.53	43.99
	NPMBDE	0.61	0.99	1.35	4.15	2.64	48.11
			Exponential				
Gaussian	NPSYN	0.25	0.54	1.44	135.11	19.17	1913.55
	NPEBLUP	0.24	0.67	2.03	173.72	41.49	2005.00
	NPMBDE	0.14	0.35	0.73	151.98	21.28	2091.47
Cauchy	NPSYN	0.22	0.76	2.13	170.20	24.19	2873.03
	NPEBLUP	0.15	0.54	3.64	134.31	60.94	1346.19
	NPMBDE	0.13	0.37	3.19	143.09	46.69	1550.79
Chi-squared	NPSYN	0.41	1.38	3.84	185.43	120.43	1752.44
	NPEBLUP	0.44	1.45	11.21	338.39	172.52	2529.45
	NPMBDE	0.23	0.52	2.36	247.27	165.27	2327.64

3.2 Design-Based Simulation Study

Design-based simulations complement model-based simulations for small area estimation. In the environment in which small area estimation is usually applied, the underlying models, no matter how sophisticated, are invariably approximations. Consequently, basing a simulation on repeated sampling from a realistic finite population can be used to assess the robustness of model-based estimation methods to model misspecification. From a practical perspective, they are also more interesting than model-based simulations since, by effectively fixing the differences between the small areas, they constitute a more realistic representation of the small area estimation problem.

In this Section we replicate the design-based simulation experiment carried out by Chandra *et al*. (2007), using a population based on a data set obtained under the Environmental Monitoring and Assessment Program (EMAP) survey and provided by the Space-Time Aquatic Resources Modelling and Analysis Program (STARMAP) at Colorado State University. The only change from that simulation experiment is that we now also use information on the geographical coordinates (in the UTM coordinate system) of each population unit. 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 1000 independent

stratified random samples of the same size as the original sample are selected from this pseudo-population, with HUC sample sizes fixed to be the same as in the original sample. The survey variable *y* is taken to be the ANC value of a lake, with its elevation defining the auxiliary variable *x*. For each sample we have calculated the values of the NPEBLUP, the NPSYN and the NPMBDE for mean ANC for each of the 23 HUCs. In all cases the nonparametric spline regression model (5) fitted to the sample data employs bivariate spline components based on the UTM coordinates of the sampled lakes, defined using transformed radial basis functions (Ruppert *et al*., 2003).

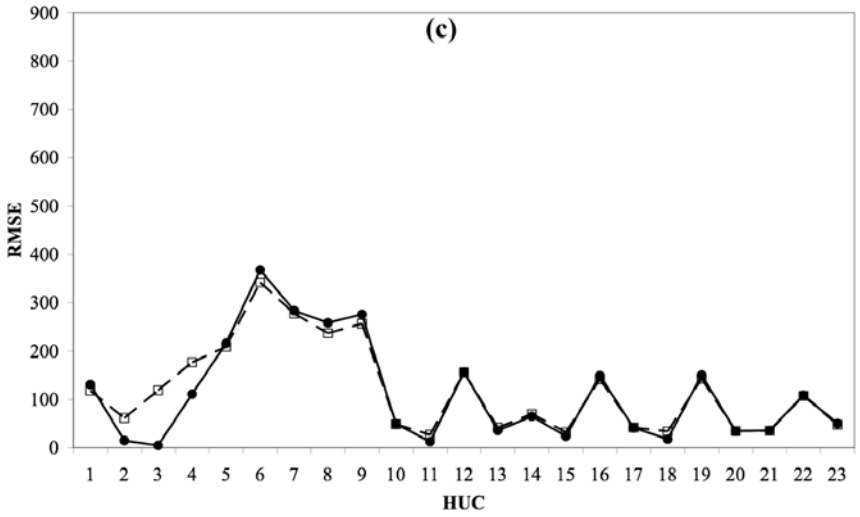
Table 3 shows the region specific values of $ARBias_i$ and $ARRMSE_i$ generated by the simulations. In the context of design-based simulation, these measures are referred to below as relative bias (RB) and relative root mean squared error (RRMSE). It can be noted that in area 1 (sample size equal to 2) all estimators are unstable, while in areas 2 and 3 (both with sample size 3) only the NPSYN and the NPEBLUP are unstable, possibly because there is little or no variability in the population values of *y* in these areas. In contrast, the NPMBDE appears unaffected, recording RB values that are consistently low. This is in contrast to the NPSYN, which is quite biased in a number of areas. Overall, the NPMBDE records the lowest value of RRMSE in 10 of the 23 areas, with the NPSYN next (7 out of 23) and finally the NPEBLUP (6 out of 23). Across all 23 areas, the NPMBDE records the lowest average RRMSE value, followed by the NPEBLUP and then the NPSYN. When one considers medians, rather than means, the NPEBLUP is marginally more efficient than the NPMBDE. Overall, taking both RB and RRMSE performance into account, it appears that the NPMBDE is preferable to both the NPEBLUP and the NPSYN in this simulation study.

Table 3. Relative bias (RB) and relative root mean squared error (RRMSE) for the EMAP data. Areas are arranged in order of increasing sample size.

			Relative bias (%)			Relative RMSE (%)		
Area	n_i	N_i	NPSYN	NPEBLUP	NPMBDE	NPSYN	NPEBLUP	NPMBDE
1	2	58	1027.38	237.39	-7.26	257.96	314.82	253.62
2	3	151	273.07	-25.88	0.56	34.53	237.78	26.68
3	3	236	3922.00	528.43	1.73	37.20	773.58	40.74
4	3	378	15.68	-0.95	-0.24	11.52	16.40	10.01
5	3	515	-19.56	7.10	-1.18	33.29	44.75	25.80
6	4	695	-36.46	-3.52	-0.09	16.28	11.39	12.74
7	6	782	-42.50	-3.69	2.29	34.69	33.52	69.60
8	7	504	19.19	-2.46	3.20	52.53	18.95	34.00
9	8	769	-50.73	1.47	-0.54	25.59	14.68	22.06
10	11	1155	47.52	-3.79	-0.10	38.73	45.13	51.09
11	12	336	-4.14	-0.71	0.01	9.89	21.57	11.25
12	12	709	1.67	2.38	0.37	26.80	22.67	19.57
13	13	1253	-12.97	2.24	0.26	7.27	11.80	9.26
14	14	756	-7.51	1.87	1.43	30.42	30.48	36.53
15	18	1243	17.06	2.03	0.61	10.73	14.09	8.56
16	18	1527	-25.93	1.31	0.47	16.64	13.18	11.75
17	18	1749	2.40	0.16	-0.30	14.01	16.35	15.94
18	19	448	355.39	2.21	-0.11	32.29	77.67	34.03
19	25	1141	-33.24	0.46	0.23	12.75	8.20	7.39
20	30	980	97.95	0.46	-0.55	21.95	21.25	21.89
21	34	1633	73.18	0.63	0.99	15.07	12.93	12.18
22	41	2508	-0.81	-0.08	0.00	15.25	9.32	9.19
23	45	1502	57.39	-2.28	2.09	21.24	9.97	10.33
Mean			246.78	32.38	0.17	281.93	77.41	32.79
Median			2.40	0.46	0.23	34.71	18.95	19.57

We now turn to an examination of the performance of the two methods of MSE estimation investigated in the simulation. MSE estimation for the NPMBDE is implemented via the pseudo-linearization MSE estimator (16), while for the NPEBLUP and the NPSYN both (16) and the analytical MSE estimator (12) are calculated. The bootstrap procedure proposed by Opsomer *et al*. (2008, Section 3.3) was also investigated. However, this failed because the nonparametric model fit to the EMAP data is not good enough to ensure comparability between the bootstrap resamples and the original sample. The behaviour of the empirical true root MSE and its estimator for each area and for each approach is shown in Figure 1. It can be seen that the pseudo-linearization MSE estimator (16) for the NPMBDE tracks the irregular profile of the area-specific empirical MSE (see Figure 1(c)) very well, while the analytic MSE estimator (12) for the NPEBLUP and the NPSYN produces somewhat over-smoothed estimates of area-specific empirical MSE. In contrast, the pseudo-linearization MSE estimator (16) applied to the NPEBLUP works well whereas the same estimator applied to the NPSYN tends to underestimate the true area-specific MSE, mainly because its squared bias component underestimates the actual squared bias of this predictor. These results also show that the area-specific MSE estimator of the NPMBDE tends to overestimate in a few areas (e.g. areas 2 and 3), mainly because there is little or no variability in ANC values in these areas.

Figure 1. HUC values of actual RMSE (solid line) and average estimated RMSE (dashed line and dotted line) obtained in the design-based simulations. Values for the MSE estimator (12) are indicated by the dotted line and by △ while those for the pseudo-linearization MSE estimator (16) are indicated by the dashed line and by □. The plots show the results for (a) the NPEBLUP, (b) the NPSYN and (c) the NPMBDE predictors. HUCs are ordered by increasing sample size.



4. Final Remarks

Most of the research in small area estimation focuses on regression structures that can be modelled using a linear mixed model. This can be a limitation when data do not follow a linear model or have an analytically intractable and/or complicated regression structure. In this article we consider using a nonparametric model for such data, but still restrict our attention to estimators for small area means that are pseudo-linear combinations of the sample data values. The "pseudo" is added since weights indeed depend on estimated values, and therefore, on the variable of interest. The way we have proposed to achieve this is to extend the Model Based Direct Estimator of Chandra and Chambers (2005) – which has this pseudo-linear structure and is based on a linear mixed model – to the case in which we have a nonparametric regression model based on p-splines that also incorporates random area effects. The simulation results show that this proposal can sometimes be more efficient than using the nonparametric EBLUP approach if the relationship between the response variable and the covariates is clearly non-linear and, is therefore, a good alternative to keep in mind when considering a nonparametric approach to small area estimation. On the other hand, MSE estimation for both the NPEBLUP and the NPMBDE using the heteroskedasticity-robust pseudo-linearization approach is straightforward to implement and seems to work promisingly well. Finally, the design-based simulation results reported in Section 3.2 indicate that the NPMBDE can be more robust than the NPEBLUP when the nonparametric spline regression model is locally misspecified. See, for example, the results in Table 3 for areas 2 and 3, where there is little variation in ANC values (most are zero) and the p-spline model (5) does not fit well.

Acknowledgments
The authors would like to acknowledge the valuable comments and suggestions of the Associate Editor and three anonymous referees. These led to a considerable improvement in the paper. The work of Salvati and Ranalli is supported by the project PRIN 2007 '*Efficient*

use of auxiliary information at the design and at the estimation stage of complex surveys: methodological aspects and applications for producing official statistics' awarded by the Italian Government to the Universities of Perugia, Cassino, Florence, Pisa and Trieste. The authors are grateful to Space-Time Aquatic Resources Modeling and Analysis Program (STARMAP) for data availability.

Appendix

Under the nonparametric regression model (5) the NPBLUP is

$$\begin{aligned}\hat{\bar{y}}_i^{NPBLUP} &= N_i^{-1}\left[\sum\nolimits_{j\in s_i} y_j + \sum\nolimits_{j\in r_i}\left(\mathbf{x}_j^T\hat{\boldsymbol{\beta}} + \mathbf{z}_j^T\boldsymbol{\gamma} + \mathbf{d}_j^T\mathbf{u}\right)\right] \\ &= N_i^{-1}\left[\sum\nolimits_{j\in s_i} y_j + \sum\nolimits_{j\in r_i}\left(\mathbf{x}_j^T\hat{\boldsymbol{\beta}}\right) + \sum\nolimits_{j\in r_i}\left(\mathbf{z}_j^T\boldsymbol{\gamma}\right) + \sum\nolimits_{j\in r_i}\left(\mathbf{d}_j^T\mathbf{u}\right)\right] \\ &= N_i^{-1}\left[\sum\nolimits_{j\in s_i} y_j + \left(\sum\nolimits_{j\in r_i}\mathbf{x}_j^T\right)\hat{\boldsymbol{\beta}} + \left(\sum\nolimits_{j\in r_i}\mathbf{z}_j^T\right)\boldsymbol{\gamma} + \left(\sum\nolimits_{j\in r_i}\mathbf{d}_j^T\right)\mathbf{u}\right].\end{aligned}$$

The components $\left(\sum_{j\in r_i}\mathbf{x}_j^T\right)$, $\left(\sum_{j\in r_i}\mathbf{z}_j^T\right)$ and $\left(\sum_{j\in r_i}\mathbf{d}_j^T\right)$ can be written as $\left(N_i - n_i\right)\bar{\mathbf{x}}_{ir}^T$, $\left(N_i - n_i\right)\bar{\mathbf{z}}_{ir}^T$ and $\left(N_i - n_i\right)\bar{\mathbf{d}}_{ir}^T$, respectively, where $\bar{\mathbf{x}}_{ir}$, $\bar{\mathbf{z}}_{ir}$ and $\bar{\mathbf{d}}_{ir}$ are the means for non sampled units from area *i* for the three sets of covariates. It immediately follows that the NPBLUP can be expressed as

$$\hat{\bar{y}}_i^{NPBLUP} = N_i^{-1}\left[\sum\nolimits_{j\in s_i} y_j + \left(N_i - n_i\right)\bar{\mathbf{x}}_{ir}^T\hat{\boldsymbol{\beta}} + \left(N_i - n_i\right)\bar{\mathbf{z}}_{ir}^T\boldsymbol{\gamma} + \left(N_i - n_i\right)\bar{\mathbf{d}}_{ir}^T\mathbf{u}\right].$$

Moreover, since $\hat{\boldsymbol{\beta}} = \mathbf{H}_s\mathbf{y}_s$, $\boldsymbol{\gamma} = \sigma_\gamma^2\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}\left(\mathbf{y}_s - \mathbf{X}_s\mathbf{H}_s\mathbf{y}_s\right) = \left(\mathbf{I}_n - \mathbf{X}_s\mathbf{H}_s\right)\mathbf{y}_s$ and $\mathbf{u} = \sigma_u^2\mathbf{D}_s^T\mathbf{V}_{ss}^{-1}\left(\mathbf{y}_s - \mathbf{X}_s\mathbf{H}_s\mathbf{y}_s\right) = \sigma_u^2\mathbf{D}_s^T\mathbf{V}_{ss}^{-1}\left(\mathbf{I}_n - \mathbf{X}_s\mathbf{H}_s\right)\mathbf{y}_s$, the NPBLUP can be written as

$$\begin{aligned}\hat{\bar{y}}_i^{NPBLUP} &= N_i^{-1}\left\{\mathbf{d}_{is}^T + (N_i - n_i)\left[\bar{\mathbf{x}}_{ir}^T\mathbf{H}_s + \left(\sigma_\gamma^2\mathbf{Z}_s\bar{\mathbf{z}}_{ir} + \sigma_u^2\mathbf{D}_s\bar{\mathbf{d}}_{ir}\right)^T\mathbf{V}_{ss}^{-1}\left(\mathbf{I}_n - \mathbf{X}_s\mathbf{H}_s\right)\right]\right\}\mathbf{y}_s \\ &= \left(\mathbf{w}_{is}^{NPBLUP}\right)^T\mathbf{y}_s\end{aligned}$$

where the weights are given by

$$\left(\mathbf{w}_{is}^{NPBLUP}\right)^T = N_i^{-1}\left\{\mathbf{d}_{is}^T + (N_i - n_i)\left[\bar{\mathbf{x}}_{ir}^T\mathbf{H}_s + \left(\sigma_\gamma^2\mathbf{Z}_s\bar{\mathbf{z}}_{ir} + \sigma_u^2\mathbf{D}_s\bar{\mathbf{d}}_{ir}\right)^T\mathbf{V}_{ss}^{-1}\left(\mathbf{I}_n - \mathbf{X}_s\mathbf{H}_s\right)\right]\right\}.$$

References

Battese, G., Harter, R. and Fuller, W. (1988). An Error-Components Model for Prediction of County Crop Areas using Survey and Satellite Data. *Journal of the American Statistical Association*, **83**, 28-36.

Chambers, R., Chandra, H. and Tzavidis, N. (2007). On Bias-Robust Mean Squared Error Estimation for Linear Predictors for Domains. *Working Papers, 09-08*, Centre for Statistical and Survey Methodology, The University of Wollongong, Australia. (Available from: http://cssm.uow.edu.au/publications).

Chandra, H. and Chambers, R. (2005). Comparing EBLUP and C-EBLUP for Small Area Estimation. *Statistics in Transition*, **7**, 637-648.

Chandra, H., Salvati, N. and Chambers, R. (2007). Small Area Estimation for Spatially Correlated Populations. A Comparison of Direct and Indirect Model-Based Methods. *Statistics in Transition*, **8**, 887-906.

Chandra, H. and Chambers, R. (2009). Multipurpose Weighting for Small Area Estimation. *Journal of Official Statistics*, **25**, 3, 379-395.

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

Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, **72**, 320-338.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussion). *Test*, **15**, 1–96.

Lombardía, M.J. and Sperlich, S. (2008). Semiparametric Inference in Generalized Mixed Effects Models. *Journal of the Royal Statistical Society, Series B*, **70**, 913-930.

Longford, N.T. (2007). On Standard Errors of Model-Based Small-Area Estimators. *Survey Methodology*, **33**, 69-79.

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.

Prasad, N.G.N. and Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small Area Estimators. *Journal of the American Statistical Association,* **85**, 163-171.

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

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

Royall, R.M. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. *Journal of the American Statistical Association*, 71, 351-358.

Ugarte, M.D., Goicoa, T., Militino, A.F. and Durbán, M. (2009). Spline Smoothing in Small Area Trend Estimation and Forecasting.. *Computational Statistics and Data Analysis, 53*, 3616-3629.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley & Sons, New York.

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