Nonparametric Small Area Estimation via M-quantile Regression using Penalized Splines

Monica Pratesi*

10 August 2008

Abstract

The demand of reliable statistics for small areas, when only reduced sizes of the samples are available, has promoted the development of small area estimation methods. In particular, an approach that is now widely used is based on linear mixed models. Chambers & Tzavidis (2006) have recently proposed an approach for small area estimation that is based on M-quantile models. However, when the functional form of the relationship between the q^{th} quantile and the covariates is not linear, it can lead to biased estimators of the small area parameters. In this paper a small area mean estimator and its mean squared error estimator are proposed allowing non linearities in the relationship between the quantiles of the distribution of the study variable and the auxiliary covariates by using a nonparametric specification of the conditional M-quantile of the response variable given the covariates (Pratesi et al., 2006). Simulation studies are presented that show the finite sample properties of the proposed estimation technique.

Keywords: Simulation study; Robust regression; Smoothing.

1 Introduction

Sample survey data are extensively used to provide reliable direct estimates of totals and means for the whole population and large areas or domains. Also estimation of population characteristics for sub-national domains (or smaller regions) is an important objective for statistical surveys. However, sample sizes may not be large enough within the domains/areas of interest to support direct estimates of adequate precision. The demand of reliable statistics for small areas, when only reduced sizes of the samples are available, has promoted the development of small area estimation methods: in particular, the model-based approach is now widely used. It is based on linear mixed models that include random area effects to account for between area variations. Under this class of models the Best Linear Unbiased Predictor (BLUP) is obtained. Details about this predictor, and its empirical version (EBLUP) for

*Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa, m.pratesi@ec.unipi.it

small area parameters are in Rao (2003). However, linear mixed models depend on strong distributional assumptions, require a formal specification of the random part of the model and do not easily allow for outlier robust inference.

Chambers & Tzavidis (2006) have recently proposed an approach for small area estimation that is based on M-quantile models. M-quantile regression provides a "quantile-like" generalization of regression based on influence functions. M-quantile small area models do not depend on strong distributional assumptions nor on a predefined hierarchical structure, and outlier robust inference is automatically performed when these models are fitted. However, M-quantile regression assumes that the quantiles of the distribution are some known parametric function of the covariates. When the functional form of the relationship between the q^{th} quantile and the covariates deviates from the assumed one, the traditional M-quantile regression can lead to biased estimators of the small area parameters. Pratesi et al. (2006) extended M-quantile regression to nonparametric modeling via penalized splines. Penalized splines (p-splines in the proceeding of the paper) regression is a flexible smoothing technique popularized by Eilers & Marx (1996). Ruppert et al. (2003) provide a thorough treatment of p-splines and their applications. Bollaerts et al. (2006) introduce quantile regression based on p-splines to estimate quantile growth curves and quantile antibody levels as a function of age. Lee & Oh (2007), independently of Pratesi et al. (2006), use M-regression to make p-splines robust against outliers. Using p-splines for M-quantile regression, beyond having the properties of M-quantile models, allows for dealing with an undefined functional relationship that can be estimated from the data. When the relationship between the q^{th} quantile and the covariates is not linear, a p-splines M-quantile regression model may have significant advantages compared to the linear M-quantile model.

In this paper a small area mean estimator and its mean squared error estimator are proposed allowing non linearities in the relationship between the quantiles of the distribution of the study variable and the auxiliary covariates. The nonparametric specification of the conditional M-quantile of y given $\boldsymbol{x}$ is described in Pratesi et al. (2006), and an application of the semiparametric version of the model is in Pratesi et al. (2008); here, in Section 2, we summarize the main steps of the method. In Section 3 the small area mean estimator and its mean squared error estimator are shown. In Section 4 a simulation study illustrates the finite-sample performance of the small area estimator based on p-splines M-quantile regression in comparison with linear M-quantile small area estimator and Empirical Best Linear Unbiased Prediction estimators based on Battese et al. (1988) model and on nonparametric regression model (Opsomer et al., 2008). Some final remarks are drawn in Section 5.

2 Nonparametric M-quantile regression

Given an influence function ψ, a nonparametric model with one covariate x_1 for the q^{th} quantile can be written as $Q_q(x_1, \psi) = \tilde{m}_{\psi,q}(x_1)$, where the function $\tilde{m}_{\psi,q}(\cdot)$ is unknown, but assumed to be approximated sufficiently well by the following function

$$m_{\psi,q}[x_1; \boldsymbol{\beta}_\psi(q), \boldsymbol{\gamma}_\psi(q)] = \beta_{0\psi}(q) + \beta_{1\psi}(q)x_1 + \ldots + \beta_{p\psi}(q)x_1^p + \sum_{k=1}^{K} \gamma_{k\psi}(q)(x_1 - \kappa_k)_+^p, \quad (1)$$

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 knots, $\boldsymbol{\beta}_\psi(q) = (\beta_{0\psi}(q), \beta_{1\psi}(q), \ldots, \beta_{p\psi}(q))^T$ is the coefficient vector of the parametric portion of the model and $\boldsymbol{\gamma}_\psi(q) = (\gamma_{1\psi}(q), \ldots, \gamma_{K\psi}(q))^T$ is the coefficient vector for the spline one. The latter portion of the model allows for handling nonlinearities in the structure of the relationship. If the number of knots K is sufficiently large, the class of functions in (1) is very large and can approximate most smooth functions. In particular, in the p-splines context, a knot is placed every 4 or 5 observations at uniformly spread quantiles of the unique values of x_1. The spline model (1) uses a truncated polynomial spline basis to approximate the function $\tilde{m}_{\psi,q}(\cdot)$. Other bases can be used; in particular we will later use radial basis functions to handle bivariate smoothing. More details on bases and knots choice can be found in Ruppert et al. (2003, Chapters 3 and 5).

The influence of the knots is limited by putting a constraint on the size of the spline coefficients: typically $\sum_{k=1}^{K} \gamma_{k\psi}^2(q)$ is bounded by some constant, while the parametric coefficients $\boldsymbol{\beta}_\psi(q)$ are left unconstrained. Therefore, estimation can be accommodated by mimicking penalization of an objective function and solving the following set of estimating equations

$$\sum_{i=1}^{n} \psi_q(y_i - \boldsymbol{x}_i\boldsymbol{\beta}_\psi(q) - \boldsymbol{z}_i\boldsymbol{\gamma}_\psi(q))(\boldsymbol{x}_i, \boldsymbol{z}_i)^T + \lambda \begin{bmatrix} \mathbf{0}_{(1+p)} \\ \boldsymbol{\gamma}_\psi(q) \end{bmatrix} = \mathbf{0}_{(1+p+K)}, \tag{2}$$

where $\boldsymbol{x}_i$ here is the i-th row of the $n \times (1+p)$ matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{11}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{1n}^p \end{bmatrix},$$

while $\boldsymbol{z}_i$ is the i-th row of the $n \times K$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} (x_{11} - \kappa_1)_+^p & \cdots & (x_{11} - \kappa_K)_+^p \\ \vdots & \ddots & \vdots \\ (x_{1n} - \kappa_1)_+^p & \cdots & (x_{1n} - \kappa_K)_+^p \end{bmatrix},$$

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

An algorithm based on iteratively reweighted penalized least squares is proposed in Pratesi et al. (2006) to effectively compute the parameter estimates. Once those estimates are obtained, $\hat{m}_{\psi,q}[x_1] = m_{\psi,q}[x_1; \hat{\boldsymbol{\beta}}_\psi(q), \hat{\boldsymbol{\gamma}}_\psi(q)]$ can be computed as an estimate for $Q_q(x_1, \psi)$. The approximation ability of this final estimate will heavily depend on the value of the smoothing parameter λ. Generalized Cross Validation (GCV) has been usefully applied in the context of smoothing splines (Craven & Wahba, 1979) and is used also in Pratesi et al. (2006) too.

Extension to bivariate smoothing can be handled by assuming $Q_q(x_1, x_2, \psi) = \tilde{m}_{\psi,q}(x_1, x_2)$. This is of central interest in a number of application areas as environment and public health. It has particular relevance when referenced responses need to be converted to maps. In particular, the following model is assumed at quantile q for unit i:

$$m_{\psi,q}[x_{1i}, x_{2i}; \boldsymbol{\beta}_\psi(q), \boldsymbol{\gamma}_\psi(q)] = \beta_{0\psi}(q) + \beta_{1\psi}(q)x_{1i} + \beta_{2\psi}(q)x_{2i} + \boldsymbol{z}_i\boldsymbol{\gamma}_\psi(q). \tag{3}$$

Here $\boldsymbol{z}_i$ is the i-th row of the following $n \times K$ matrix

$$\boldsymbol{Z} = [C(\tilde{\boldsymbol{x}}_i - \boldsymbol{\kappa}_k)]_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant k \leqslant K}} [C(\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'})]^{-1/2}_{1 \leqslant k, k' \leqslant K}, \tag{4}$$

where $C(\boldsymbol{t}) = ||\boldsymbol{t}||^2 \log ||\boldsymbol{t}||$, $\tilde{\boldsymbol{x}}_i = (x_{1i}, x_{2i})$ and $\boldsymbol{\kappa}_k$, $k = 1, \ldots, K$ are knots. See Pratesi et al. (2006) and for details on this. Here, it is enough to note that the estimation procedure can again be pursued with (2) where $\boldsymbol{x}_i = (1, \tilde{\boldsymbol{x}}_i)$.

It should be noted, then, that the estimating equations in (2) can be used to handle univariate smoothing and bivariate smoothing by suitably changing the parametric and the spline part of the model, i.e. once the $\boldsymbol{X}$ and the $\boldsymbol{Z}$ matrices are set up. Finally, other continuous or categorical variables can be easily inserted parametrically in the model by adding columns to the $\boldsymbol{X}$ matrix. This allows for semiparametric modeling, as intended in Ruppert et al. (2003), to be inherited and applied to M-quantile regression.

3 The methodology

P-splines M-quantile regression is applied to the estimation of a small area mean as follows. The first step is to estimate the M-quantile coefficients q_i for each unit i in the probabilistic sample s of size n without reference to the m small areas of interest. This is done defining a fine grid of values on the interval $(0, 1)$ and using the sample data to fit the p-splines M-quantile regression functions at each value q on this grid, as explained in the previous section. If a data point lies exactly on the qth fitted curve, then the coefficient of the corresponding sample unit is equal to q. Otherwise, to obtain q_i, a linear interpolation over the grid is used.

If a hierarchical structure does explain part of the variability in the population data, we expect units within clusters defined by this hierarchy to have similar M-quantile coefficients. Therefore, an estimate of the mean quantile for area j, $\bar{q}_j$, is obtained by taking the corresponding average value of the sample M-quantile coefficient of each unit in area j, $\hat{\bar{q}}_j = \sum_{i=1}^{n_j} q_i$. The small area estimator of the mean $\bar{y}_j$ is then

$$\hat{\bar{y}}_j = \frac{1}{N_j} \Big\{ \sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \hat{y}_{ij} \Big\}, \tag{5}$$

where s_j and r_j denote the sampled and non sampled units in area j, respectively, with $U_j = s_j \cup r_j$, and N_j is the known population size of area j. Note that the unobserved value for population unit $i \in r_j$ is predicted using

$$\hat{y}_{ij} = \boldsymbol{x}_{ij} \hat{\boldsymbol{\beta}}_\psi(\hat{\bar{q}}_j) + \boldsymbol{z}_{ij} \hat{\boldsymbol{\gamma}}_\psi(\hat{\bar{q}}_j),$$

where $\hat{\boldsymbol{\beta}}_\psi(\hat{\bar{q}}_j)$ and $\hat{\boldsymbol{\gamma}}_\psi(\hat{\bar{q}}_j)$ are the coefficient vectors of the parametric and spline portion, respectively, of the fitted p-splines M-quantile regression function at $\hat{\bar{q}}_j$.

The estimator of the small area mean can be biased for small areas containing outliers. This has already been noted in Tzavidis & Chambers (2006) for the estimator under the a linear M-quantile regression model. They propose an adjustment for bias based on the Chambers & Dunstan (1986) estimator of the small area distribution function. This adjustment can be used also in case of p-splines M-quantile regression models. The bias-adjusted estimator for the mean is given by

$$\hat{\bar{y}}_j = \frac{1}{N_j}\Big\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j}\hat{y}_{ij} + \frac{N_j - n_j}{n_j}\sum_{i\in s_j}(y_{ij} - \hat{y}_{ij})\Big\}, \tag{6}$$

where $\hat{y}_{ij}$ denotes the predicted values for the population units in s_j and in r_j. This estimator will be here denoted with PSPL when using a p-splines M-quantile regression model, and with LIN when using a linear one.

In many instances we are interested in estimating parameters for out of sample areas, that is areas where there are not sampled units even if in those areas there are population units with the characteristic of interest. In this case no area effects can be computed and the small area characteristic is estimated by using synthetic estimation. We can note that with synthetic estimation all variation in the area-specific predictions comes from the area-specific auxiliary information. One approach to improving estimation for out of sample areas is by borrowing strength over space (Saei & Chambers, 2005). In case of P-splines M-quantile regression, this can be achieved using model (3) and setting $\hat{\bar{q}}_j = 0.5$. A synthetic type mean predictor for out of sample area j is given by

$$\hat{\bar{y}}_j = \frac{1}{N_j}\Big\{\sum_{i\in r_j}\boldsymbol{x}_{ij}\hat{\beta}_\psi(0.5) + \boldsymbol{z}_{ij}\hat{\boldsymbol{\gamma}}_\psi(0.5)\Big\}. \tag{7}$$

We expect that when a truly spatially process is present, (7) will improve the efficiency of the other traditional synthetic estimators.

Following the approach described in Chandra & Chambers (2005) and Chambers & Tzavidis (2006), for fixed q and λ, the $\hat{\bar{y}}_j$ in (6) can be written as the following linear combination of the observed y_i,

$$\hat{\bar{y}}_j = \frac{1}{N_j}\sum_{i\in s} w_{ij} y_i, \tag{8}$$

where the n-vector of weights $\boldsymbol{w}_j = (w_{1j}, \ldots, w_{nj})^T$ is given by

$$\boldsymbol{w}_j = \frac{N_j}{n_j}\mathbf{1}_{s_j} + \boldsymbol{W}(\hat{\bar{q}}_j)[\boldsymbol{X}\ \boldsymbol{Z}]\Big([\boldsymbol{X}\ \boldsymbol{Z}]^T\boldsymbol{W}(\hat{\bar{q}}_j)[\boldsymbol{X}\ \boldsymbol{Z}] + \lambda\boldsymbol{G}\Big)^{-1}\Big(\boldsymbol{T}_{r_j} - \frac{N_j - n_j}{n_j}\boldsymbol{T}_{s_j}\Big) \tag{9}$$

with $\mathbf{1}_{s_j}$ the n-vector with i^{th} component equal to one whenever the corresponding sample unit is in area j and to zero otherwise, $\boldsymbol{W}(\hat{\bar{q}}_j)$ a diagonal $n\times n$ matrix that contains the final set of weights produced by the iteratively reweighted penalized least squares algorithm used to estimate the regression coefficients, $\boldsymbol{G} = \text{diag}\{\mathbf{0}_P, \mathbf{1}_K\}$ with P the number of columns of $\boldsymbol{X}$ and K the number of columns of $\boldsymbol{Z}$, and with $\boldsymbol{T}_{r_j}$ and $\boldsymbol{T}_{s_j}$ the totals of the covariates for the non-sampled and the sampled units in area j, respectively.

The weights derived from (9) are treated as fixed and a 'plug in' estimator of the mean squared error of estimator (8) given by

$$MSE(\hat{\bar{y}}_j) = var(\hat{\bar{y}}_j - \bar{y}_j) + [bias(\hat{\bar{y}}_j)]^2 \tag{10}$$

can be proposed by using standard methods for robust estimation of the variance of unbiased weighted linear estimators (Royall & Cumberland, 1978) and by following the results due to

Tzavidis & Chambers (2007). The prediction variance of (8) can be approximated by

$$var(\hat{\bar{y}}_j - \bar{y}_j) \approx \frac{1}{N_j^2}\Big(\sum_{i\in s_j}\Big\{d_{ij}^2 + \frac{N_j - n_j}{n_j - 1}\Big\}var(y_{ij}) + \sum_{i\in s\backslash s_j} d_{ij}^2 var(y_{ij})\Big) \tag{11}$$

with $d_{ij} = w_{ij} - 1$ if $i \in s_j$ and $d_{ij} = w_{ij}$ otherwise, and $s\backslash s_j$ the set of sampled units outside area j. The bias can be written as

$$bias(\hat{\bar{y}}_j) \approx \frac{1}{N_j}\Big(\sum_{k=1}^{m}\sum_{i\in s_k} w_{ij}\tilde{y}_{ik} - \sum_{i\in U_j}\tilde{y}_{ij}\Big) \tag{12}$$

where $\tilde{y}_{ik} = \boldsymbol{x}_{ik}\beta_\psi(\hat{\bar{q}}_k) + \boldsymbol{z}_{ik}\gamma_\psi(\hat{\bar{q}}_k)$ are the study variable values under the p-splines M-quantile regression model. Following the area level residual approach (Tzavidis & Chambers, 2006), we can interpret $var(y_{ij})$ conditionally to the specific area j from which y_i is drawn and hence replace $var(y_{ij})$ in (11) by $(y_{ij} - \hat{y}_{ij})^2$. An estimate of the bias is obtained replacing $\tilde{y}_{ik}$ by $\hat{y}_{ik}$ in (12). A robust estimator of the mean squared error of (8) is given by the sum of the estimator of the variance

$$\widehat{var}(\hat{\bar{y}}_j) = \frac{1}{N_j^2}\left[\sum_{i\in s_j}\left\{d_{ij}^2 + \frac{N_j - n_j}{n_j - 1}\right\}(y_{ij} - \hat{y}_{ij})^2 + \sum_{i\in s\backslash s_j} d_{ij}^2 (y_{ij} - \hat{y}_{ij})^2\right] \tag{13}$$

and the squared estimate of the bias

$$\hat{b}^2(\hat{\bar{y}}_j) = \frac{1}{N_j^2}\left(\sum_{k=1}^{m}\sum_{i\in s_k} w_{ij}\hat{y}_{ik} - \sum_{i\in U_j}\hat{y}_{ij}\right)^2. \tag{14}$$

Since the bias-adjusted nonparametric M-quantile estimator is an approximately unbiased estimator of the small area mean, the squared bias term will not impact significantly the mean squared error estimator. The main limitation of the MSE estimator is that it does not account for the variability introduced in estimating the area specific q's and λ. We note also that we can obtain an estimate only for areas where there are at least two sampled units. For all these reasons, we are currently investigating the use of bootstrap as an alternative approach for estimating the MSE.

4 Simulation study

In this section we use simulation studies to illustrate the finite-sample performance of the small area mean estimator based on p-splines M-quantile regression (6) – PSPL. It is compared with the estimator computed by standard linear M-quantile regression – LIN – and with the Empirical Best Linear Unbiased Prediction estimators based on Battese et al. (1988) model – EBLUP – and on nonparametric regression model (Opsomer et al., 2008) – NPEBLUP. The properties of the estimators have been assessed by Monte Carlo experiments using models with a single covariate.

Given the number of small areas $m = 30$, three synthetic populations of size $N = 10,550$ are generated using the following models for creating the true underlying relationship between the covariate x and the expected value of the response variable y $E(y|x) = m(x)$:

Linear. $m(x) = 3 + 2(x - 0.5)$;
Cycle. $m(x) = 2\sin(2\pi x)$;
Jump. $m(x) = 1 + 2(x - 0.5)I(x \leqslant 0.5) + 0.5I(x > 0.5)$.

Population values of y in small area j were generated under the random intercepts model

$$y_{ij} = m(x) + \gamma_j + \epsilon_{ij}$$

with x drawn from a Uniform distribution $[0, 1]$, area effects γ_j and individual effects ϵ_{ij} were independently drawn from $N(0, 0.04)$ and $N(0, 0.16)$ distributions respectively, with $\rho = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\epsilon^2} = 0.2$. A sample of size $n = 600$ was then selected from the simulated population, by simple random sampling. Each population was kept fixed for all simulation runs. A total of $T = 500$ simulations were carried out.

The linear case represents a situation in which LIN and EBLUP are based on a good representation of the true model and PSPL and NPEBLUP may be too complex and over-parametrized. The cycle model defines an increasingly more complicated structure of the relationship between y and x, while the jump one is a discontinuous function for which LIN, PSPL, EBLUP and NPEBLUP are based on a misspecified model. Gaussian errors provide a situation in which EBLUP and NPEBLUP consider the correct distributional assumptions for the distribution of the areas.

For each sample LIN, PSPL, EBLUP and NPEBLUP have been used to estimate the small area means. For each estimator and for each small area we computed the Monte Carlo estimate of the Bias

$$\mathrm{B}_{MC} = \frac{1}{T}\sum_{t=1}^{T}(\hat{\bar{y}}_{jt} - \bar{y}_j) \tag{15}$$

and with it the percentage relative bias

$$\mathrm{RB\%} = \frac{\mathrm{B}_{MC}}{\bar{y}_j}100; \tag{16}$$

the root Mean Squared Error

$$\mathrm{RMSE}_{MC} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{\bar{y}}_{jt} - \bar{y}_j)^2}, \tag{17}$$

and the corresponding percentage relative root Mean Squared Error

$$\mathrm{RRMSE\%} = \frac{\mathrm{RMSE}_{MC}}{\bar{y}_j}100. \tag{18}$$

Tables 1, 2 and 3 report the RB%, RRMSE% values obtained for this study under a linear, cycle and jump signal, respectively. MSE estimation was monitored comparing MSE estimates and Monte Carlo MSEs, and by checking 95% confidence intervals coverage rates CR%. For MSE estimation of the PSPL estimator we used expression (13), whereas the MSE estimation of LIN predictor was carried out following the method suggested in Tzavidis &

Chambers (2007). MSE estimation of the EBLUP and NPEBLUP comes from methods introduced in Prasad & Rao (1990) and Opsomer et al. (2008), respectively. Intervals are defined by the small area mean estimate plus or minus twice their corresponding estimated root mean squared error. Areas are arranged in order of increasing population size.

There is a promising result from this simulation study: the gain in terms of unbiasedness of PSPL and LIN estimators is relevant especially when the structure of the relationship between y and x is more complicated than the linear one and in case of misspecified models. The average value of the RB%, in the simulation experiments, for PSPL and LIN varies from 0.12% to 0.91%, whereas for EBLUP and NPEBLUP is between 0.66% and 3.15%.

Looking at the efficiency of the estimators, their performance changes, as expected, under different population models. It is evident that, under the assumption of linearity and Gaussian errors (Table 1), EBLUP and NPEBLUP outperform PSPL, even if PSPL is still competitive. In the Cycle Gaussian population (Table 2) PSPL is more efficient than LIN and EBLUP: the ratio of the average values of the percentage relative root Mean Squared Error (RRMSE%[LIN or EBLUP]/RRMSE%[PSPL]) is greater than 1.4 pointing out that the PSPL is at least 1.4 times closer to the true values of the target parameter. This gain is due to the better fitting of the model upon which PSPL is based. In addition, given that it is based on the right distributional assumptions, NPEBLUP is the best estimator in this case. The Jump signal is linear for most of its support. This explains the very similar behavior shown by the four estimators in this case (Table 3). It is noticeable the difference in bias between the M-quantile based estimators and the mixed models ones; this is particularly true for the estimates in Area 19. However, the overall RRMSE% does not seem to benefit from such reduction in bias.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

Figures (1), (2) and (3) show how different root mean squared estimators track the true root mean squared error of the different estimator under linear, cycle and jump signal. Each figure has the same structure. Top left is the PSPL predictor (6) with RMSE estimated using (13). Top right is the LIN predictor with RMSE estimator suggested by Tzavidis & Chambers (2007). Bottom left is EBLUP predictor with RMSE estimator suggested by Prasad & Rao (1990) and bottom right is the NPEBLUP estimator with RMSE estimator suggested by Opsomer et al. (2008).

Figure (1) shows the area-specific values of RMSE and average estimated RMSE in case of linear signal. The estimator (13) performs well, showing only a small amount of undercoverage both for PSPL and LIN estimators. Given that all its underlying assumptions are met, the Prasad & Rao (1990) and Opsomer et al. (2008) estimators of RMSE works very well in terms of empirical coverage. However, we note that they have a smoothing effect on the estimated variability of the small areas.

In case of cycle signal (Figure 2) the PSPL and the LIN MSE estimators have the best performance in tracking the true variability. EBLUP and NPEBLUP MSE estimators smooth the behavior across the areas. Under the jump signal (Figure 3) estimator (13) for PSPL and

LIN estimators tracks the true behavior of RMSE. Both Prasad & Rao (1990) and Opsomer et al. (2008) estimators confirm their smoothing effect.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

5 Conclusions

The PSPL model can be widely used in many important application areas, such as financial and economic statistics and environmental and public health modeling. In this work the PSPL models are used for small area estimation. Also in this case they appear to be a useful tool when the functional form of the relationship between the variable of interest and the covariates is left unspecified and the data are characterized by complex patterns of dependence. The method proposed for small area mean estimation relies on PSPL model: it takes advantages from the properties of M-quantile models in small area estimation and on the versatility of the penalized splines as a tool for capturing non linearities in the data. The method is outlier robust, it does not require strong assumptions on the data distribution and a pre-specified hierarchical structure of the data. In addition it allows for a sample based estimator of the MSE.

The finite-sample performance of small area mean estimator based on p-splines M-quantile regression – PSPL – method has been compared to the linear M-quantile – LIN – method, the EBLUP and the NPEBLUP estimators by Monte Carlo experiments carried out for a single covariate case. Finite populations were simulated using a mixed model with different trends with respect to x. Methods based on nonparametric regression techniques outperform those based on linear models when such trends were not linear. On the other side, the loss in efficiency in the case of a linear trend was not noticeable. In addition, the mixed model based simulation was set up to investigate the performance of the proposed estimator when the mixed models based estimators use the correct error specification. The performance was comparable in all cases. MSE estimators provide a good tracking of real MSE in most simulation studies. However, in some cases a poor coverage rate is shown, so that the issue of MSE estimation requires further work. Next step is to explore the consistency of parametric and/or semiparametric bootstrap estimators of the MSE with the analytical solution.

Acknowledgements

The work reported here has been developed under the support of the project PRIN *Metodologie di stima e problemi non campionari nelle indagini in campo agricolo-ambientale* awarded by the Italian Government to the Universities of Cassino, Florence, Perugia, Pisa and Trieste. The author would like to thank Ray Chambers, Giovanna Ranalli and Nicola Salvati for their help and support. The views expressed here are solely those of the author.

References

Battese, G., Harter, R. & 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.

Bollaerts, K., Eilers, P. H. C. & Aerts, M. (2006). Quantile regression with monotonicity restrictions using p-splines and the l-1-norm. *Statistical Modelling* **6 (3)**, 189–207.

Chambers, R. & Dunstan, P. (1986). Estimating distribution function from survey data. *Biometrika* **73**, 597–604.

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

Chandra, H. & Chambers, R. (2005). Comparing eblup and c-eblup for small area estimation. *Statistics in Transition* **7**, 637–648.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.

Lee, T. C. & Oh, H.-S. (2007). Robust penalized regression spline fitting with application to additive mixed modeling. *Computational Statistics* **22**, 159–171.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. & 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. & Rao, J. (1990). The estimation of mean squared error of small-area estimators. *Journal of the American Statistical Association* **85**, 163–171.

Pratesi, M., Ranalli, M. G. & Salvati, N. (2006). Nonparametric M-quantile regression via penalized splines. In *ASA Proceedings on Survey Research Methods, Alexandria, VA*.

Pratesi, M., Ranalli, M. G. & Salvati, N. (2008). Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US. *Environmetrics* **19**, 1–15.

Rao, J. N. K. (2003). *Small Area Estimation.* John Wiley & Sons.

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

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

Saei, A. & Chambers, R. (2005). Empirical best linear unbiased prediction for out of sample areas. In *S3RI Methodology Working Papers.* Southampton: ed. Southampton Statistical Sciences Research Institute, pp. 1–15.

Tzavidis, N. & Chambers, R. (2006). Bias adjusted distribution estimation for small areas with outlying values. In *S3RI Methodology Working Papers.* Southampton: ed. Southampton Statistical Sciences Research Institute, pp. 1–30.

Tzavidis, N. & Chambers, R. (2007). Robust prediction of small area means and distributions. In *CCSR Working Paper 2007-08.* Manchester: University of Manchester, pp. 1–34.

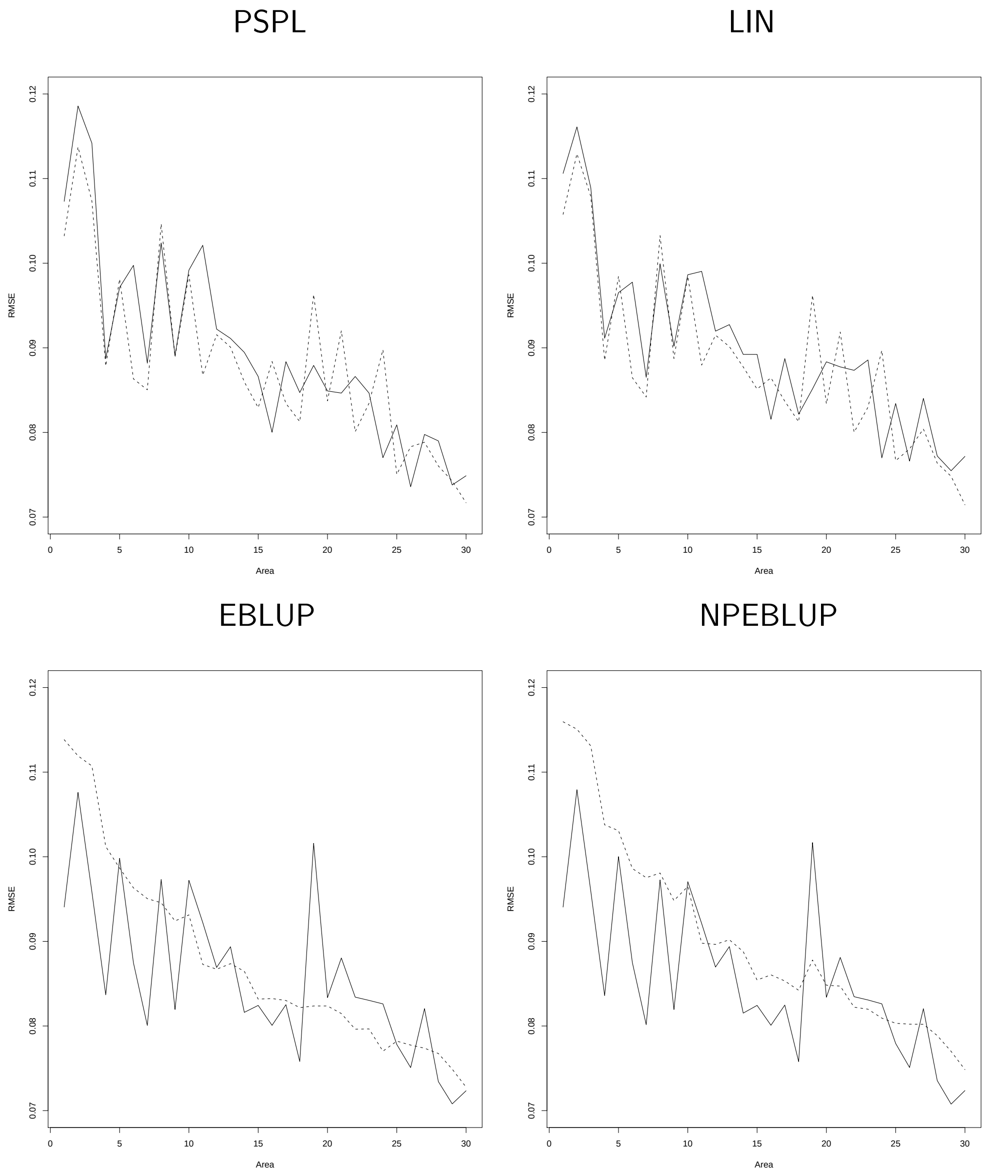

Figure 1: Area-specific values of RMSE (solid line) and average estimated RMSE (dashed line) in case of Linear signal. Areas are arranged in order of increasing population size.

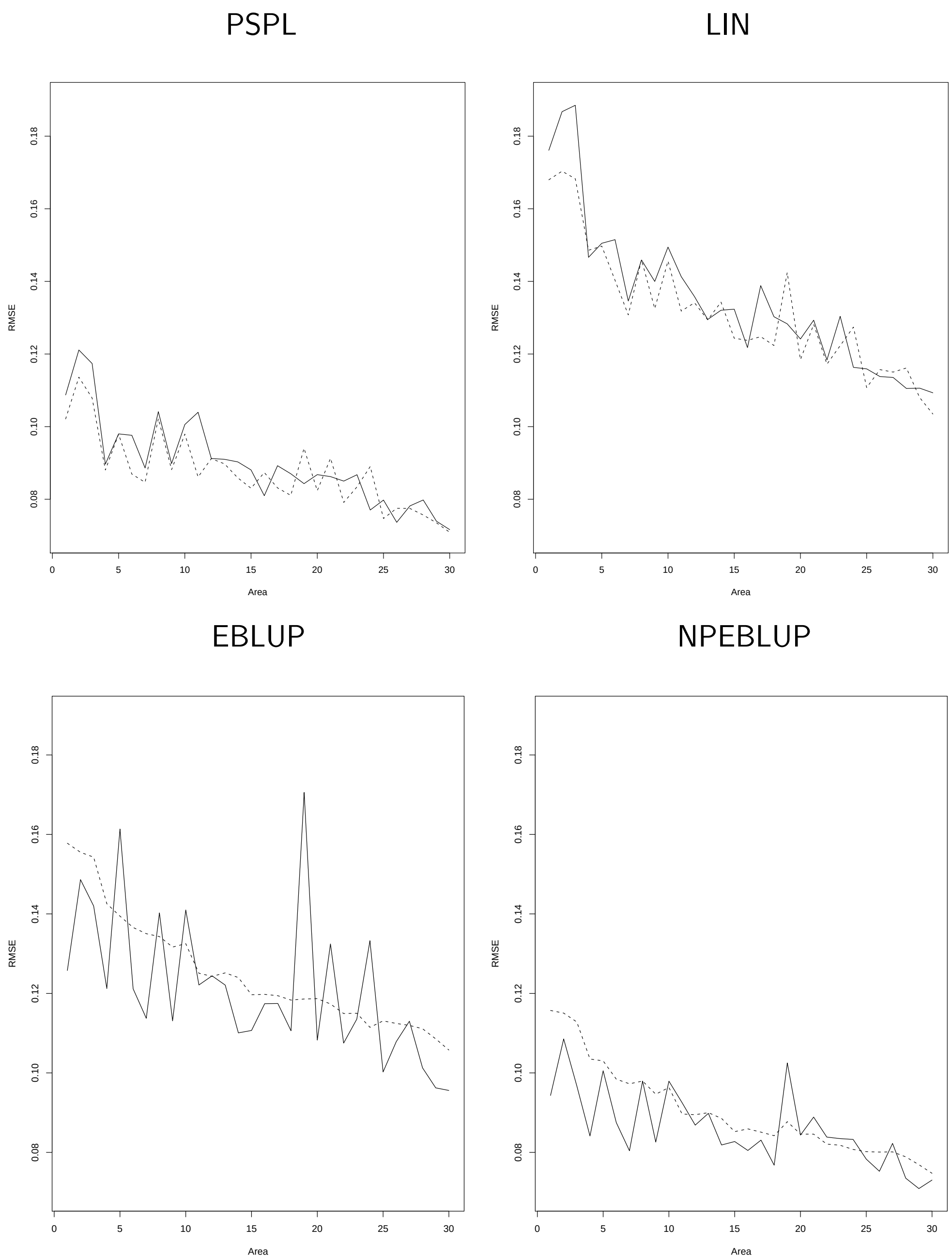

Figure 2: Area-specific values of RMSE (solid line) and average estimated RMSE (dashed line) in case of Cycle signal. Areas are arranged in order of increasing population size.

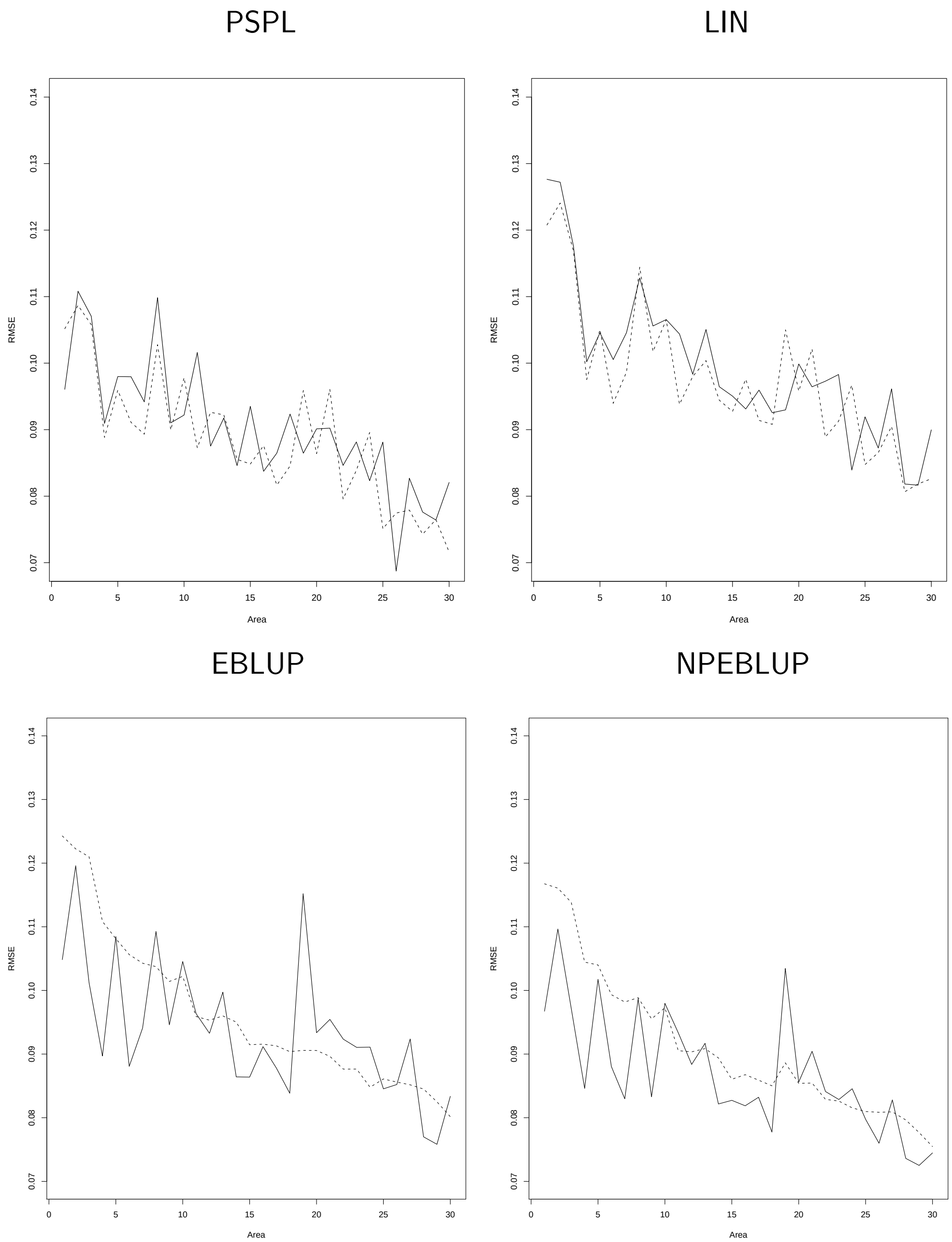

Figure 3: Area-specific values of RMSE (solid line) and average estimated RMSE (dashed line) in case of Jump signal. Areas are arranged in order of increasing population size.

Area	PSPL			LIN			EBLUP			NPEBLUP		
	RB%	RRMSE%	CR%	RB%	RRMSE%	CR%	RB%	RRMSE%	CR%	RB%	RRMSE%	CR%
1	0.30	3.64	89.5	0.10	3.75	89.8	0.19	3.19	98.4	0.20	3.19	98.4
2	-0.02	3.63	91.9	0.11	3.55	91.6	-1.20	3.29	96.6	-1.21	3.30	97.4
3	0.17	3.99	91.5	0.19	3.80	91.0	0.67	3.35	96.4	0.68	3.35	96.4
4	0.20	3.14	93.2	0.27	3.22	92.4	0.82	2.96	98.2	0.81	2.95	98.6
5	0.13	2.81	93.9	0.17	2.80	94.2	-1.37	2.89	95.4	-1.38	2.90	96.0
6	-0.53	3.31	89.5	-0.42	3.24	91.2	-0.48	2.90	97.6	-0.48	2.90	98.2
7	-0.24	2.75	89.2	-0.21	2.70	91.6	-0.60	2.50	98.4	-0.60	2.50	98.4
8	-0.33	3.96	93.6	-0.21	3.87	94.2	1.30	3.76	94.0	1.30	3.76	94.8
9	0.05	3.09	95.3	-0.10	3.13	93.6	0.31	2.84	96.6	0.31	2.84	97.0
10	-0.16	2.93	92.5	-0.05	2.91	93.4	-1.13	2.87	94.0	-1.12	2.86	94.8
11	-0.16	3.67	90.2	-0.05	3.56	91.0	0.59	3.32	94.2	0.58	3.31	94.8
12	-0.48	3.49	94.2	-0.28	3.48	94.4	0.73	3.29	95.6	0.72	3.29	96.6
13	-0.12	2.77	92.2	-0.07	2.82	92.0	-0.83	2.72	94.0	-0.84	2.72	94.6
14	0.03	2.90	93.6	0.16	2.89	93.6	0.01	2.65	96.8	0.00	2.64	97.6
15	-0.07	2.91	92.5	-0.10	3.00	92.2	-0.04	2.77	95.8	-0.04	2.77	95.8
16	0.08	2.40	94.2	0.02	2.45	94.2	-0.73	2.40	96.2	-0.73	2.40	96.6
17	-0.12	2.82	93.2	-0.07	2.83	91.8	-0.21	2.63	95.4	-0.21	2.63	96.2
18	0.12	2.84	91.9	-0.02	2.75	93.0	-0.03	2.54	96.8	-0.03	2.54	97.4
19	0.07	3.94	96.6	0.21	3.81	97.2	2.69	4.55	90.4	2.69	4.55	92.6
20	0.06	2.99	94.2	0.08	3.11	92.6	0.62	2.93	95.6	0.62	2.94	96.2
21	-0.08	2.44	95.6	0.03	2.53	94.2	-0.89	2.54	94.0	-0.89	2.54	94.8
22	0.15	2.62	93.2	0.03	2.64	92.8	-0.56	2.52	94.6	-0.56	2.53	95.0
23	-0.30	3.01	92.2	0.01	3.15	91.8	0.41	2.95	92.8	0.41	2.95	93.4
24	-0.02	2.16	97.6	0.01	2.16	97.8	-1.05	2.32	95.4	-1.05	2.32	96.2
25	-0.06	2.63	92.9	-0.01	2.71	92.8	-0.14	2.53	95.8	-0.14	2.53	96.4
26	0.10	2.72	94.2	0.21	2.83	93.6	0.84	2.77	96.8	0.85	2.78	97.0
27	-0.26	3.04	93.2	-0.17	3.20	92.6	0.81	3.12	94.0	0.81	3.12	94.4
28	0.07	2.83	94.2	0.06	2.77	94.6	0.49	2.63	96.0	0.49	2.64	97.0
29	-0.21	2.53	95.6	-0.13	2.59	95.6	0.02	2.43	97.2	0.02	2.42	97.2
30	-0.01	2.56	93.6	-0.04	2.64	92.4	0.06	2.47	96.4	0.06	2.48	97.0
Mean (abs. values)	0.16	3.02	93.2	0.12	3.03	93.1	0.66	2.89	95.6	0.66	2.89	96.2
Median (abs. values)	0.12	2.90	93.2	0.10	2.90	92.8	0.61	2.81	95.8	0.61	2.81	96.4

Table 1: Relative Bias (RB%), Relative Root Mean Squared Errors (RRMSE%) and Coverage Rate (CR%) in case of Linear signal. Areas are arranged in order of increasing population size.

Area	PSPL			LIN			EBLUP			NPEBLUP		
	RB%	RRMSE%	CR%	RB%	RRMSE%	CR%	RB%	RRMSE%	CR%	RB%	RRMSE%	CR%
1	0.48	5.42	89.6	-0.16	8.77	91.2	-0.07	6.27	98.6	0.29	4.70	98.4
2	-0.26	5.34	91.1	0.45	8.24	90.0	-2.50	6.56	96.4	-1.76	4.79	97.4
3	0.28	6.18	90.0	0.19	9.93	89.0	1.60	7.48	96.0	1.06	5.09	96.4
4	0.38	4.80	92.6	0.80	7.84	92.0	2.21	6.48	97.2	1.27	4.50	99.0
5	0.03	3.91	94.4	0.19	6.00	93.8	-4.10	6.44	90.4	-1.89	4.01	96.0
6	-0.81	4.82	91.1	-0.64	7.47	90.6	-0.82	5.98	97.6	-0.66	4.32	98.2
7	-0.35	4.25	90.4	0.00	6.45	93.4	-1.61	5.45	97.4	-0.95	3.85	98.8
8	-0.52	6.53	92.6	-0.81	9.15	93.8	4.55	8.80	94.4	2.16	6.14	94.8
9	0.09	4.68	94.1	-0.30	7.29	91.6	0.39	5.89	97.6	0.46	4.30	97.0
10	-0.44	4.34	92.6	-0.07	6.46	92.4	-2.93	6.09	93.0	-1.67	4.23	93.8
11	-0.26	5.80	88.5	-0.09	7.89	93.4	1.90	6.82	96.6	0.88	5.16	95.0
12	-0.71	5.25	94.1	-0.31	7.82	93.8	2.66	7.17	96.2	1.10	5.00	97.2
13	-0.20	3.96	92.6	-0.14	5.64	94.2	-2.40	5.31	96.2	-1.19	3.91	94.8
14	0.07	4.47	92.6	0.25	6.54	94.6	-0.36	5.45	97.0	0.01	4.05	97.2
15	-0.08	4.53	92.6	-0.38	6.81	91.8	0.03	5.70	97.0	-0.05	4.26	96.0
16	0.02	3.52	93.0	-0.04	5.29	94.2	-2.26	5.10	96.6	-1.08	3.50	96.8
17	-0.42	4.54	92.6	-0.20	7.06	91.0	-0.60	5.97	95.4	-0.35	4.22	95.8
18	0.14	4.45	89.6	0.17	6.65	92.0	0.44	5.65	96.6	-0.03	3.92	96.8
19	0.79	6.81	96.3	0.12	10.37	97.0	9.80	13.79	82.0	4.92	8.29	91.8
20	0.15	4.89	93.7	-0.05	7.00	92.6	1.47	6.10	96.6	1.02	4.75	96.0
21	-0.23	3.64	94.8	-0.24	5.47	93.8	-2.89	5.60	92.8	-1.31	3.76	94.2
22	0.11	3.82	92.6	0.23	5.33	94.8	-1.61	4.84	96.6	-0.86	3.77	94.8
23	-0.47	4.67	91.9	-0.24	7.02	92.6	0.98	6.11	95.0	0.64	4.50	93.6
24	-0.12	2.99	97.4	0.01	4.51	95.4	-3.13	5.17	92.2	-1.45	3.23	96.2
25	0.06	3.90	93.7	0.06	5.67	92.8	-0.51	4.90	97.8	-0.24	3.83	96.2
26	0.17	4.19	94.8	0.55	6.48	95.0	2.50	6.14	96.6	1.32	4.28	97.2
27	-0.29	4.88	92.6	-0.07	7.09	94.6	3.22	7.06	94.4	1.38	5.14	94.4
28	0.19	4.38	93.0	0.34	6.06	96.2	1.78	5.55	97.4	0.73	4.03	97.0
29	-0.34	3.80	94.8	-0.35	5.69	93.0	0.04	4.95	97.6	0.01	3.64	97.2
30	0.08	3.64	94.4	0.03	5.55	94.0	0.32	4.86	97.4	0.11	3.71	96.8
Mean (abs. values)	0.28	4.61	92.8	0.25	6.92	93.2	1.99	6.26	95.6	1.03	4.43	96.2
Median (abs. values)	0.24	4.50	92.6	0.19	6.73	93.4	1.69	5.97	96.6	0.99	4.24	96.3

Table 2: Relative Bias (RB%), Relative Root Mean Squared Errors (RRMSE%) and Coverage Rate (CR%) in case of Cycle signal. Areas are arranged in order of increasing population size.

Area	PSPL			LIN			EBLUP			NPEBLUP		
	RB%	RRMSE%	CR%	RB%	RRMSE%	CR%	RB%	RRMSE%	CR%	RB%	RRMSE%	CR%
1	0.57	9.87	94.6	0.24	13.12	91.0	0.18	10.77	97.6	0.43	9.94	97.8
2	-0.80	8.60	91.3	0.00	9.87	91.8	-4.11	9.28	95.6	-3.15	8.51	96.8
3	-0.21	12.32	92.4	0.57	13.53	92.2	2.47	11.65	97.8	2.32	11.18	96.0
4	-0.03	11.07	93.5	0.41	12.20	91.6	2.91	10.92	99.0	2.80	10.30	99.0
5	-0.81	6.76	93.5	0.46	7.21	94.4	-3.85	7.48	95.6	-3.30	7.02	96.2
6	-1.09	9.74	94.6	-1.00	10.00	92.2	-1.15	8.76	97.8	-1.37	8.75	98.0
7	-1.01	8.04	87.0	-0.85	8.93	92.2	-1.90	8.03	97.4	-1.83	7.08	98.2
8	-0.36	18.08	92.4	-0.45	18.57	93.8	6.76	17.98	94.4	5.59	16.24	95.4
9	0.50	10.43	91.3	-0.20	12.10	92.4	1.53	10.84	95.8	1.03	9.55	97.8
10	-0.47	6.64	93.5	-0.07	7.68	93.2	-3.21	7.53	94.0	-2.83	7.06	94.0
11	-1.06	12.76	92.4	0.18	13.11	91.6	2.64	12.10	95.2	2.11	11.72	94.4
12	-2.41	13.88	94.6	-1.22	15.60	94.8	4.06	14.79	96.2	3.28	14.02	96.6
13	0.57	7.11	94.6	-0.04	8.14	92.8	-2.40	7.73	94.4	-2.19	7.10	94.2
14	0.08	7.93	92.4	0.43	9.05	93.2	0.06	8.10	96.6	0.02	7.71	97.0
15	-0.39	9.69	93.5	-0.28	9.86	93.0	0.03	8.96	96.6	-0.22	8.58	96.8
16	0.75	6.26	93.5	-0.12	6.96	94.4	-2.42	6.81	95.6	-2.01	6.12	96.8
17	-1.17	7.76	90.2	-0.26	8.61	92.6	-0.64	7.88	95.8	-0.65	7.47	95.8
18	0.72	9.19	91.3	0.21	9.21	93.2	0.00	8.34	95.8	0.02	7.74	96.8
19	6.01	37.36	94.6	2.86	40.18	95.8	31.78	49.78	88.2	26.97	44.70	91.0
20	1.34	10.86	91.3	0.33	12.03	93.8	2.59	11.25	96.2	2.21	10.31	96.6
21	-0.04	6.13	95.7	0.27	6.56	94.8	-2.42	6.49	94.8	-2.19	6.15	94.6
22	-0.11	6.49	94.6	0.06	7.47	92.2	-1.75	7.09	93.0	-1.52	6.46	94.6
23	-2.72	10.80	90.2	0.06	12.04	92.8	1.73	11.16	93.4	1.45	10.15	93.6
24	-0.03	5.26	97.8	0.01	5.36	96.6	-2.90	5.82	94.6	-2.44	5.40	95.8
25	-0.61	8.14	85.9	-0.08	8.49	93.2	-0.59	7.80	96.6	-0.45	7.36	96.6
26	-0.83	9.94	94.6	0.78	12.62	93.0	4.02	12.33	95.8	3.34	10.99	97.2
27	-0.54	13.23	91.3	-1.08	15.38	93.6	3.88	14.78	93.8	3.61	13.25	94.2
28	-0.81	9.76	91.3	0.06	10.29	94.4	1.85	9.68	96.6	1.67	9.26	97.0
29	0.65	8.29	94.6	-0.09	8.86	94.8	0.43	8.22	97.2	0.13	7.86	97.4
30	0.53	8.77	93.5	0.03	9.62	93.0	0.32	8.91	95.8	0.24	7.96	96.0
Mean (abs. values)	0.91	10.37	92.7	0.42	11.42	93.3	3.15	11.04	95.6	2.71	10.20	96.1
Median (abs. values)	0.63	9.44	93.5	0.25	9.86	93.1	2.41	8.93	95.8	2.06	8.54	96.6

Table 3: Relative Bias (RB%), Relative Root Mean Squared Errors (RRMSE%) and Coverage Rate (CR%) in case of Jump signal. Areas are arranged in order of increasing population size.