

Model-assisted estimation of a spatial population mean

Giuseppe Cicchitelli and Giorgio E. Montanari

Abstract This paper deals with the estimation of the mean of a finite spatial population defined on a fine grid superimposed over a continuous domain. Under a design-based perspective, an estimator assisted by low-rank spline smoothers of the population data is proposed and studied. Proof is given that the estimator is design-consistent and has a normal limiting distribution. A simulation study is conducted to investigate the performance of the new estimator and of the estimator of its variance, in terms of relative bias, efficiency and confidence interval coverage rate.

Keywords: low-rank smoother; regression spline model; Horvitz-Thompson estimator; model-assisted estimator, spatial data analysis; sample design and estimation.

1 Introduction

In environmental studies dealing with the soil characterization, monitoring of natural resources or estimation of pollution concentration, it is often of interest to make an inference on the mean of a survey variable defined continuously over a region.

Let A be the domain of interest and let $y(\mathbf{x})$ be the survey variable defined at each location $\mathbf{x} = (x_1, x_2)'$ in A , where x_1 and x_2 are the geographical coordinates of the location. A common problem in this context is to estimate the mean of the survey variable. We will focus on the case of a spatially discrete population obtained by superimposing a very fine grid over the continuous study region. Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be the locations corresponding to the nodes of this grid, where $\mathbf{x}_i = [x_{i1}, x_{i2}]'$, $i = 1, \dots, N$, is the vector of the geographical coordinates. The set of points defined in this way constitutes the finite

population which will be denoted by $U = \{1, \dots, N\}$. Let $y(\mathbf{x}_1), \dots, y(\mathbf{x}_N)$ be the values taken by the survey variable $y(\mathbf{x})$ in such a population. Therefore, the population mean is given by $\bar{Y} = \sum_{i \in U} y(\mathbf{x}_i) / N$.

Given the vector $\mathbf{y}_s = [y(\mathbf{x}_1), \dots, y(\mathbf{x}_n)]'$ of n sample observations at locations $\mathbf{x}_1, \dots, \mathbf{x}_n$, the inference on $\bar{Y}$ can be conducted under two paradigms: (i) the model-based approach, (ii) the design-based approach. In the former approach $y(\mathbf{x})$ is assumed to be a random variable and the sample $\mathbf{y}_s$ a random vector. As a result, the source of randomness is given by the model, which describes the probability distribution of $\mathbf{y}_s$. On the contrary, in the design-based approach, $y(\mathbf{x})$ is considered a fixed quantity for every $\mathbf{x} \in A$ and the randomness arises from the chance mechanism used for selecting the sample locations $\mathbf{x}_1, \dots, \mathbf{x}_n$.

In this paper we present a new estimator of the spatial mean $\bar{Y}$ under the design-based paradigm. It is a model-assisted estimator in the following sense. A penalized regression spline model, capable of capturing the correlation structure underlying the population data, is estimated on the basis of the sample observations. Then the resulting fitted values are employed in a model-assisted estimator.

2 Design-based population low-rank smoothers

Under a model-based perspective, assume that the population values $y(\mathbf{x}_1), \dots, y(\mathbf{x}_N)$ associated with the locations $\mathbf{x}_1, \dots, \mathbf{x}_N$ are the realization of an N -dimensional random vector according to the following superpopulation model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}, \quad (1)$$

where $\mathbf{y} = [y(\mathbf{x}_1), \dots, y(\mathbf{x}_N)]'$, $\mathbf{X}$ is an $N \times 3$ matrix having $[1, x_{i1}, x_{i2}]$ as i -th row, for $i = 1, \dots, N$, $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]'$ is a vector of unknown constants, $\boldsymbol{\delta} = [\delta(\mathbf{x}_1), \dots, \delta(\mathbf{x}_N)]'$ and $\boldsymbol{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_N]'$ are N -dimensional random vectors such that:

$$E_{\xi}(\boldsymbol{\delta}) = E_{\xi}(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \text{Cov}_{\xi}(\boldsymbol{\delta}, \boldsymbol{\varepsilon}) = \mathbf{0}, \quad \text{Var}_{\xi}(\boldsymbol{\delta}) = \sigma_x^2 \boldsymbol{\Sigma}, \quad \text{Var}_{\xi}(\boldsymbol{\varepsilon}) = \sigma_{\varepsilon}^2 \mathbf{I}_N$$

(the suffix ξ means that the expected values, variances and covariance are taken with respect to the probability distribution of the random vectors considered). In model (1), $\mathbf{X}\boldsymbol{\beta}$ represents the large-scale deterministic structure that links the survey variable to the geographical covariates; $\boldsymbol{\delta}$ is the small-scale stochastic structure that models the spatial dependence; $\boldsymbol{\varepsilon}$ accounts for measurement error.

The covariance matrix of $\boldsymbol{\delta}$, $\sigma_x^2 \boldsymbol{\Sigma}$, is generally expressed in a parametric form using parsimonious models, under the assumption of second-order stationarity and isotropy. An important class of isotropic covariance functions is the Matérn family, which involves a three parameter vector $\boldsymbol{\theta} = (\sigma_x^2, \rho, \nu)$, where σ_x^2 is the variance of $\delta(\mathbf{x}_i)$, ρ is the range parameter (it controls how fast correlations decay with increasing distance) and ν is the smoothing parameter (it controls the smoothness of the resulting interpolating surface). For the special case $\nu = 1.5$, the covariance between observations at two locations separated by the vector $\mathbf{h}$ takes the form

Model-assisted estimation of a spatial mean

$$C(\mathbf{h}) = \sigma_x^2 (1 + \|\mathbf{h}\| / \rho) \exp(-\|\mathbf{h}\| / \rho),$$

where $\|\cdot\|$ is the Euclidean norm. Hereafter, without losing any generality, we assume $C(\mathbf{h})$ as covariance function.

Now, equation (1) can be written in the form of the following linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad (2)$$

where $\mathbf{Z} = \boldsymbol{\Sigma}^{1/2}$ and $\text{Var}_{\boldsymbol{\varepsilon}}(\mathbf{u}) = \sigma_x^2 \mathbf{I}_N$. Subsequently, a smoother for the entire population can be obtained by invoking Henderson's mixed model equations to derive the estimator of $\boldsymbol{\beta}$ and the predictor of the random effects vector $\mathbf{u}$ (Ruppert *et al.*, 2003, p. 99). The resulting smoother is the kriging predictor

$$\tilde{\mathbf{y}}_{KR} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{Z}\tilde{\mathbf{u}}. \quad (3)$$

Here $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, $\tilde{\mathbf{u}} = \sigma_x^2 \mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$, $\mathbf{V} = \sigma_x^2(\mathbf{Z}\mathbf{Z}' + \lambda\mathbf{I})$, where λ is the variance ratio $\lambda = \sigma_{\varepsilon}^2 / \sigma_x^2$. Note that $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ are the solutions of the following minimum problem

$$\min_{\boldsymbol{\beta}, \mathbf{u}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2 + \lambda \mathbf{u}'\mathbf{u}). \quad (4)$$

Then $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ correspond to fitting model (2) to population data by penalized least squares. Penalties are imposed to restrict the variation of the coefficients $u_1, \dots, u_N$. The larger the value of the smoothing parameter λ , the more the fit shrinks towards a polynomial fit, whereas small values of λ result in wiggly "overfitted" predictions.

Given λ , it is easy to show from (4) that $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ can also be written as

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}, \quad (5)$$

where $\mathbf{D} = \text{blockdiag}(\mathbf{0}_{3 \times 3}, \mathbf{I}_N)$. The smoother matrix is given by

$$\mathbf{S}_{\lambda} = [\mathbf{X}, \mathbf{Z}] \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{bmatrix}. \quad (6)$$

The trace of $\mathbf{S}_{\lambda}$ can be interpreted as the number of degrees of freedom, r , of the fit, i.e. as the equivalent number of fitting parameters. It obeys the inequalities $3 \leq r = \text{tr}(\mathbf{S}_{\lambda}) \leq N + 3$. The lower limit is reached as $\lambda \rightarrow \infty$, the upper limit is attained as $\lambda \rightarrow 0$. Formula (3) involves the covariance between all pairs of locations $\mathbf{x}_1, \dots, \mathbf{x}_N$; in this sense, it represents a *full-rank* smoother. To reduce the dimensionality of the problem, it has been suggested (see Nychka *et al.*, 1998) to use *low-rank* smoothers.

Consider a representative subset of the locations $\mathbf{x}_1, \dots, \mathbf{x}_N$; let $\{\boldsymbol{\kappa}_1, \dots, \boldsymbol{\kappa}_K\}$ be such a subset, whose elements are called *knots*. Let

$$c_{ik} = C(\|\mathbf{x}_i - \boldsymbol{\kappa}_k\| / \rho), \quad i = 1, \dots, N, \quad k = 1, \dots, K, \quad d_{kl} = C(\|\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_l\| / \rho), \quad k, l = 1, \dots, K,$$

where $C(r) = (1 + |r|) \exp(-|r|)$. Define the $N \times K$ matrix $\tilde{\mathbf{Z}}_L$ and the $K \times K$ matrix $\boldsymbol{\Omega}_L$ which have the typical elements c_{ik} and d_{kl} , respectively. Finally, define the $N \times K$ matrix $\mathbf{Z}_L = \tilde{\mathbf{Z}}_L \boldsymbol{\Omega}_L^{-1/2}$ and consider the following linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_L \mathbf{u} + \boldsymbol{\varepsilon}, \quad (7)$$

obtained from (2) replacing $\mathbf{Z}$ with $\mathbf{Z}_L$ and $\mathbf{u}$ with an analogous vector of dimension K . From (7), following the same procedure described above, we obtain the smoother

$$\tilde{\mathbf{y}}_L = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{Z}_L \tilde{\mathbf{u}},$$

where $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ can be obtained from equation (5) replacing $\mathbf{Z}$ with

The smoother matrix of $\tilde{\mathbf{y}}_L$ is obtained from equation (6) replacing $\mathbf{Z}$ with $\mathbf{Z}_L$ and putting $\mathbf{D} = \text{blockdiag}(\mathbf{0}_{3 \times 3}, \mathbf{I}_K)$. Obviously, the number of degrees of freedom of $\tilde{\mathbf{y}}_L$ obeys the inequalities $3 \leq r \leq K + 3$. The low-rank smoother just introduced is an approximation of the kriging smoother $\tilde{\mathbf{y}}_{KR}$. The number of knots controls the trade-off between accuracy of the approximation and numerical simplification: when $K = N$, the two smoothers coincide.

A third way to construct a smoother is connected with the so-called *thin plate* smoothing spline. Now, let

$$\dot{c}_{ik} = (\|\mathbf{x}_i - \boldsymbol{\kappa}_k\|)^2 \log(\|\mathbf{x}_i - \boldsymbol{\kappa}_k\|), \quad i = 1, \dots, N, \quad k = 1, \dots, K,$$

$$\dot{d}_{kl} = (\|\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_l\|)^2 \log(\|\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_l\|), \quad k, l = 1, \dots, K.$$

Form the $N \times K$ matrix $\tilde{\mathbf{Z}}_T$ and the $K \times K$ matrix $\boldsymbol{\Omega}_T$ which have the typical elements $\dot{c}_{ik}$ and $\dot{d}_{kl}$, respectively. Define the $N \times K$ matrix $\mathbf{Z}_T = \tilde{\mathbf{Z}}_T \boldsymbol{\Omega}_T^{-1/2}$ and consider the following linear mixed model (Ruppert *et al.*, 2003, p. 257).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_T \mathbf{u} + \boldsymbol{\varepsilon}, \quad (8)$$

obtained from (7) replacing $\mathbf{Z}_L$ with $\mathbf{Z}_T$. Then, from (8) we obtain the new smoother

$$\tilde{\mathbf{y}}_T = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{Z}_T \tilde{\mathbf{u}}, \quad (9)$$

where $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ can be derived from equation (5) replacing on the right-hand side $\mathbf{Z}$ with $\mathbf{Z}_T$ and putting $\mathbf{D} = \text{blockdiag}(\mathbf{0}_{3 \times 3}, \mathbf{I}_K)$. With the same procedure, the smoother matrix of (9) can be obtained from equation (6) by following the same procedure. Formula (9)

Model-assisted estimation of a spatial mean

represents a penalized spline regression smoother which has the rows of matrix $\mathbf{Z}_T$ as radial basis functions.

Now, consider a random sample of locations $s \subset U$, of expected size n , drawn from U by a sampling design $p(s)$ that assigns the inclusion probability

$$\pi(\mathbf{x}_i) = \sum_{s: i \in s} p(s)$$

to location $\mathbf{x}_i$, $i = 1, \dots, N$. Referring to model (8), which does not require the estimation of the covariance parameter as model (7), we want to find a design-based estimator of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ for smoother (9). For this purpose, define the diagonal matrix $\mathbf{\Pi}_U = \text{diag}(1/\pi(\mathbf{x}_i))_{i \in U}$. Let $\mathbf{\Pi}_s = \text{diag}(1/\pi(\mathbf{x}_i))_{i \in s}$ be the sample submatrix of $\mathbf{\Pi}_U$. Similarly, let $\mathbf{X}_s$ and $\tilde{\mathbf{Z}}_{Ts}$ be the submatrices of $\mathbf{X}$ and $\tilde{\mathbf{Z}}_T$ consisting of the rows for which $i \in s$. Then, under mild conditions, a design-consistent estimator of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ is given by

$$\begin{bmatrix} \hat{\tilde{\boldsymbol{\beta}}} \\ \hat{\tilde{\mathbf{u}}} \end{bmatrix} = \left[\begin{bmatrix} \mathbf{X}_s' \mathbf{\Pi}_s \mathbf{X}_s & \mathbf{X}_s' \mathbf{\Pi}_s \tilde{\mathbf{Z}}_{Ts} \mathbf{\Omega}^{-1/2} \\ \mathbf{\Omega}^{-1/2} \tilde{\mathbf{Z}}_{Ts}' \mathbf{\Pi}_s \mathbf{X}_s & \mathbf{\Omega}^{-1/2} \tilde{\mathbf{Z}}_{Ts}' \mathbf{\Pi}_s \tilde{\mathbf{Z}}_{Ts} \mathbf{\Omega}^{-1/2} \end{bmatrix} + \lambda \mathbf{D} \right]^{-1} \begin{bmatrix} \mathbf{X}_s' \mathbf{\Pi}_s \mathbf{y}_s \\ \mathbf{\Omega}^{-1/2} \tilde{\mathbf{Z}}_{Ts}' \mathbf{\Pi}_s \mathbf{y}_s \end{bmatrix}.$$

Hence, a design-based estimator of $\tilde{\mathbf{y}}_T = [\tilde{y}_T(\mathbf{x}_1), \dots, \tilde{y}_T(\mathbf{x}_N)]'$ is provided by

$$\hat{\tilde{\mathbf{y}}}_T = \mathbf{X} \hat{\tilde{\boldsymbol{\beta}}} + \mathbf{Z}_T \hat{\tilde{\mathbf{u}}}, \quad (10)$$

where $\hat{\tilde{\mathbf{y}}}_T = [\hat{\tilde{y}}_T(\mathbf{x}_1), \dots, \hat{\tilde{y}}_T(\mathbf{x}_N)]'$. A similar approach was adopted by Breidt *et al.* (2005) in a different context.

3 Penalized spline model-assisted estimation of the spatial mean

Using $\hat{\tilde{\mathbf{y}}}$ as a predictor of $\mathbf{y}$, following Särndal *et al.* (1992), a model-assisted estimator of the population mean is given by

$$\hat{\tilde{Y}}_{bspl} = \frac{1}{N} \sum_{i \in U} \hat{\tilde{y}}(\mathbf{x}_i) + \frac{1}{N} \sum_{i \in s} \frac{e(\mathbf{x}_i)}{\pi(\mathbf{x}_i)},$$

where $e(\mathbf{x}_i) = y(\mathbf{x}_i) - \hat{\tilde{y}}(\mathbf{x}_i)$.

Ignoring that $\hat{\tilde{\mathbf{y}}}$ is estimated from the sample, an estimator of the design variance of $\hat{\tilde{Y}}_{bspl}$ is given by

$$\hat{V}_p(\hat{\tilde{Y}}_{bspl}) = \frac{1}{N^2} \sum_{i \in s} \sum_{j \in s} \frac{\pi(\mathbf{x}_i, \mathbf{x}_j) - \pi(\mathbf{x}_i)\pi(\mathbf{x}_j)}{\pi(\mathbf{x}_i, \mathbf{x}_j)} \frac{e(\mathbf{x}_i)}{\pi(\mathbf{x}_i)} \frac{e(\mathbf{x}_j)}{\pi(\mathbf{x}_j)}, \quad (11)$$

where $\pi(\mathbf{x}_i, \mathbf{x}_j)$ is the joint inclusion probability of locations $\mathbf{x}_i, \mathbf{x}_j$, for $i, j \in s$, and $\pi(\mathbf{x}_i, \mathbf{x}_i) = \pi(\mathbf{x}_i)$ (the suffix p on the left-hand member of equation (11) means that here we are operating in the design-based framework, i.e. expectations are based on the sampling design). Estimator (11) is design-consistent as far as $\hat{\tilde{y}}$ is consistent for $\tilde{y}$. However it tends to be downward biased, since it does not take into account the sampling variability of $\hat{\tilde{y}}$.

Coming to the properties of estimator $\hat{\tilde{Y}}_{bspl}$, first of all we note that (using the same knots and penalty term for all survey variables), it is a linear estimator, i.e. it can be written as a weighted sum of the observations with weights that are the same for all survey variables. This is a feature which is much appreciated by survey samplers as regards coherence issues.

To establish the large-sample properties of the proposed estimator, let us introduce an appropriate asymptotic framework. Let k and qk be the number of rows and columns of the grid that generates the finite population. Consider the sequence of finite populations in domain A obtained by increasing the number of rows k , while maintaining q fixed as well as the region covered by the grid. Then, as far as $k \rightarrow \infty$, the grid becomes increasingly dense and the population size $N_k = qk^2$ tends to infinity. Assume that for any value of k the values of the survey variable $y(\mathbf{x}_1), \dots, y(\mathbf{x}_{N_k})$ are limited in the interval $[m, M]$. Let us denote with k the population defined by k rows. Now, for each population k , let $p_k(s)$ be the sampling design used to select a sample of expected size n_k , such that the inclusion probabilities are positive for all units of the population k . In such a case the Horvitz-Thompson estimator of the mean is unbiased. Assume that n_k goes to infinity as $k \rightarrow \infty$. In this framework and under some regularity conditions, it can be shown that the proposed estimator $\hat{\tilde{Y}}_{bspl}$ is design $\sqrt{n_k}$ -consistent and that

$$\hat{\mathbf{V}}_p(\hat{\tilde{Y}}_{bspl})^{-1/2}(\hat{\tilde{Y}}_{bspl} - \bar{Y}) \rightarrow N(0, 1).$$

4 A simulation study

In this section, we report on some simulation experiments carried out to study the performance of the proposed estimator. The simulations compare the behaviour of $\hat{\tilde{Y}}_{bspl}$ with that of the Horvitz-Thompson estimator $\hat{\tilde{Y}}_{HT}$ under a stratified random sampling with a proportional allocation. We considered four populations defined on a grid of $60 \times 60 = 3,600$ spatial locations, $\mathbf{x} = (x_{l1}, x_{l'2})'$, with $l, l' = 1, \dots, 60$, regularly located in the square $[0, 1] \times [0, 1]$, where $x_{lk} = (2l - 1)/120$, $l = 1, \dots, 60$, $k = 1, 2$. The following four finite populations were considered:

- A: this was obtained by the Cholesky triangularization technique as a realization of the random variables $y(\mathbf{x}_1), \dots, y(\mathbf{x}_{3,600})$, whose covariance matrix $\sigma_x^2 \Sigma$ has as typical element $\sigma_x^2 \rho^{d_{ij}}$, with $\sigma_x^2 = 100$, $\rho = 0.5$ and $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$.
- B: this was obtained with the same procedure as for population A, but with $\rho = 0.8$.
- C: this was obtained from the function $y(x_{l1}, x_{l'2}) = 5[\sin(x_{l1})]^2 + 5[\cos(x_{l'2})]^2 + 5x_{l1}$.

Model-assisted estimation of a spatial mean

- D: this was obtained from the function $y(x_{I1}, x_{I2}) = 100[(x_{I1} - 0.5)^2 + (x_{I2} - 0.5)^2]$.

The 60×60 grid was divided into nine squares of 20×20 locations, to obtain a spatial stratification of the population units. Then, we selected the knots using the software due to Nychka *et al.* (1998), and computed the matrix $\mathbf{Z}_T$, which appears in equation (10).

From each population, 2,000 stratified random samples were selected with a proportional allocation among the strata. The simulation study was carried out with two different sample sizes: $n=90$ and $n=360$. For each selected sample, the vector of the fitted values was calculated, assigning to λ a value such that the number of degrees of freedom of the model is equal to a given value of r , fixed in advance. Then the proposed estimator was calculated together with the Horvitz-Thompson estimator. To estimate the variance of $\hat{Y}_{bspl}$ we used a slightly different formula, $\hat{V}_p(\hat{Y}_{bspl})$, obtained by multiplying $\hat{V}_p(\hat{Y}_{bspl})$ by the ratio $(n-1)/(n-r)$ to compensate for the negative bias that might arise with an higher number of degrees of freedom. Distinct simulations were performed giving parameters r and K the values 5, 10, 20, 30, 40, 60 and 10, 30, 60, respectively. Of course, r and K were combined in such a manner that inequalities $3 \leq r \leq K+3$ hold. Part of the simulation results is presented in Table 1 where Monte Carlo evaluations of the values of the following indexes are reported: relative efficiency $\text{Eff}_{MC}(\hat{Y}_{bspl}) = \text{mse}_{MC}(\hat{Y}_{bspl}) / \text{mse}_{MC}(\hat{Y}_{HT})$; relative bias $R_{b/sd} = [E_p(\hat{Y}_{bspl}) - \bar{Y}] / [E_p(\hat{V}_p(\hat{Y}_{bspl}))]^{1/2}$; variance to mean squared error ratio $R_{var/mse} = E_p(\hat{V}_p(\hat{Y}_{bspl})) / \text{mse}_{MC}(\hat{Y}_{bspl})$; confidence Interval coverage (percentage).

Table 1 Monte Carlo estimate of the: (i) Relative efficiency; (ii) Relative bias; (iii) Variance to mse ratio; (iv) Confidence interval coverage.

r	K	$n = 90$				$n = 360$			
		Eff_{MC}	$R_{b/sd}$	$R_{var/mse}$	Coverage	Eff_{MC}	$R_{b/sd}$	$R_{var/mse}$	Coverage
Population A									
5	10	0.74	-0.02	0.98	94.0	0.74	-0.02	1.07	95.7
10	10	0.53	-0.02	0.96	93.6	0.53	-0.01	1.03	95.3
5	30	0.74	-0.02	0.97	94.0	0.73	-0.02	1.07	95.6
10	30	0.56	-0.01	0.92	93.4	0.55	-0.01	1.03	95.1
20	30	0.41	0.11	0.81	91.3	0.38	0.04	0.99	95.0
30	30	0.45	0.10	0.64	87.9	0.33	0.03	0.94	94.0
Population B									
5	10	0.67	-0.04	1.03	95.5	0.68	-0.05	1.00	95.0
10	10	0.56	0.02	1.00	94.9	0.55	-0.02	1.00	95.0
5	30	0.67	-0.03	1.04	95.6	0.68	-0.05	1.01	95.2
10	30	0.50	0.01	1.01	94.6	0.50	-0.03	1.00	95.4
20	30	0.46	0.11	0.81	91.9	0.40	0.01	0.95	94.9
30	30	0.59	0.14	0.55	84.2	0.36	0.02	0.90	93.6
Population C									
5	10	0.67	0.00	1.09	95.8	0.67	-0.01	1.02	94.4
10	10	0.22	0.01	1.04	95.6	0.22	-0.01	0.97	94.6
5	30	0.64	0.01	1.09	95.9	0.64	-0.01	1.03	94.3
10	30	0.29	0.01	1.04	95.3	0.29	-0.01	0.98	94.4
20	30	0.12	0.00	0.80	90.7	0.10	-0.03	0.91	94.0
30	30	0.11	-0.02	0.44	80.1	0.06	-0.05	0.85	92.7

Table 1 (cont.)

r	K	$n = 90$				$n = 360$			
		Eff_{MC}	$R_{b/sd}$	$R_{var/mse}$	Coverage	Eff_{MC}	$R_{b/sd}$	$R_{var/mse}$	Coverage
Population D									
5	10	0.51	-0.14	0.93	92.7	0.50	-0.07	1.00	94.7
10	10	0.11	-0.31	0.79	91.4	0.10	-0.14	0.92	94.0
5	30	0.48	-0.15	0.91	93.1	0.47	-0.07	1.02	94.9
10	30	0.08	-0.44	0.67	88.9	0.07	-0.19	0.91	93.4
20	30	0.01	-0.97	0.32	75.5	0.01	-0.34	0.67	90.1
30	30	0.00	-1.10	0.20	69.6	0.00	-0.35	0.70	90.3

For the proposed estimator, the following main indications can be drawn from the simulation study:

- it behaves as approximately unbiased even for moderate sample size;
- it appears to be far more efficient than the Horvitz-Thompson estimator based on a stratified sample (in practice the map is forced to be a step function);
- its variance estimator does work until the number of degrees of freedom of the smoother is not high: as a rule of thumb we found that the smoother dimension should be not greater than $n/6$ or, even better, then $n/10$;
- both the smoother dimension and the number of knots have an impact on the performance of the proposed estimator, but the first quantity plays a far more important role;
- the coverage rate of confidence intervals is fairly close to the nominal level, unless the variance estimator heavily underestimates the mean squared error, which seems to occur when the smoother dimension exceeds the threshold indicated above.

Acknowledgments Work 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”.

References

1. Breidt F.J., Claeskens G., Opsomer J.D.: Model-assisted estimation for complex surveys using penalized splines. *Biometrika* **92**, 831-846 (2005).
2. Nychka D., Haaland P., O’Connell M., Ellner S.: FUNFITS. Data Analysis and Statistical Tools for Estimating Functions. In: Nychka D., Piegorsch W.W., Cox L.H. (Eds.) *Case Study in Environmental Statistics* (Lecture Notes in Statistics, **132**, pp. 159-179). Springer-Verlag, New York (1998).
3. Ruppert D., Wand M.P., Carroll R.J.: *Semiparametric Regression*. Cambridge University Press, Cambridge (2003).
4. Särndal C.E., Swensson B., Wretman J. *Model Assisted Survey Sampling*. Springer-Verlag, New York (1992).