International Statistical Review (2012), 80, 1, 111–126 doi:10.1111/j.1751-5823.2011.00164.x

Model-Assisted Estimation of a Spatial Population Mean

Giuseppe Cicchitelli and Giorgio E. Montanari

Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via A. Pascoli, 06124 Perugia, Italy
E-mail: giorgioeduardo.montanari@unipg.it

Summary

This paper deals with the estimation of the mean of a spatial population. Under a design-based approach to inference, an estimator assisted by a penalized spline regression model is proposed and studied. Proof that the estimator is design-consistent and has a normal limiting distribution is provided. A simulation study is carried out to investigate the performance of the new estimator and its variance estimator, in terms of relative bias, efficiency, and confidence interval coverage rate. The results show that gains in efficiency over standard estimators in classical sampling theory may be impressive.

Key words: Spline regression model; Horvitz-Thompson estimation; sampling design; model-assisted estimator.

1 Introduction

Spatial populations arise in a number of disciplines, including geology, ecology, and environmental science, in connection with the study of natural phenomena in two-dimensional spaces. Examples of the latter are mineral resources, vegetation cover, soil chemical composition, pollution concentration in soil, abundance of fish in a lake, all involving areas or land plots. A spatial population can be formed by distinct entities and spatial units, such as small- to medium-sized lakes in a lake district, or can consist of a continuum, i.e. of an infinite number of points. In the first case, the population is discrete; in the latter case, the population is continuous. For discrete populations, the value taken by the response variable in a unit, for example, the measure of biodiversity, is referred to the entire unit, although it may be associated with a given point, i.e. the reference point that identifies the location of the unit. In such a case, the reference points are irregularly distributed across the study region. Also, a continuous spatial population can be discretized if we subdivide the entire region under study in appropriate small areas or plots that act as units. This can be done, for example, assuming as units the areal cells formed by superimposing a regular grid on the region of interest and taking as reference point the central location of the cell. In this case, the reference points are regularly disseminated across the study region.

In the discrete case, there exist a set A of location points, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, identifying the population units, where the response variable takes the values $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N)$. In the continuous case, we assume that the response variable is described by an integrable function $y(\mathbf{x})$ defined over a domain A.

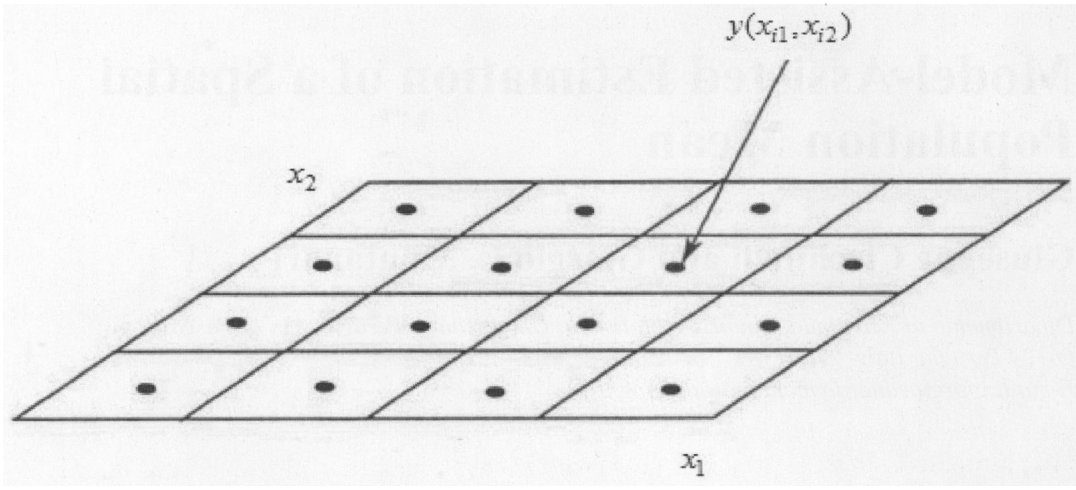

Figure 1. *Example of a rectangular grid.*

The population parameter of main interest is the total of the response variable, that is $T = \sum_{i=1}^{N} y(\mathbf{x}_i)$, in the discrete case, and $T = \int_A y(\mathbf{x})d\mathbf{x}$, in the continuous case. This is a quite general parameter, because estimates of mean values, variances, proportions, and distribution functions can be rewritten as functions of estimated totals, and therefore, formulated as estimates of sums or integrals over A.

In this paper, we deal with the estimation of the mean of a spatial population, given by $\overline{Y} = T/N$ in the discrete case, or by $\overline{Y} = T/|A|$ (where $|A|$ denotes the area of domain A) in the continuous case. To this end, a sample $s = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of n location points is drawn and the value of the survey variable is observed at each sampled location.

We assume that the location points are identified by their geographical coordinates x_{i1} and x_{i2}. Then, the point $\mathbf{x}_i$ of a discrete population must be interpreted as the vector $\mathbf{x}_i = [x_{i1}, x_{i2}]'$, and similarly, the point $\mathbf{x}$ of a continuous population as the vector $\mathbf{x} = [x_1, x_2]'$. In the discrete case, we further assume that population units are identified by the reference location points; in the case of a discretized continuous population, we assume that the reference points are placed at the centre of grid cells covering the domain of interest (see Figure 1).

Inference on $\overline{Y}$ can be conducted either according to the design-based approach or under the model-based approach. Under the first paradigm, also called probability sampling method, $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N)$ are considered as fixed but unknown quantities (in the continuous case, $y(\mathbf{x})$ is considered as fixed but unknown function) and the randomness arises from the chance mechanism used for selecting the sample of locations $s = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Under the second paradigm, $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N)$ are assumed to be spatially correlated random variables (in the continuous case, $y(\mathbf{x}), x \in A$, is assumed to be a random field) whose distribution is described by a model, which represents the stochastic mechanism that generates the data.

There has been a long-standing debate in the statistical literature on the relative merits of design-based and model-based approaches to spatial population surveys (see, for example, Brus & de Gruijter, 1997; Cox *et al.*, 1997). A wide-spread opinion is that a design-based approach is the best option if inference focuses on global quantities, such as means or totals, and besides, validity of the result is more important than efficiency (the word "validity" refers to the fact that the design-based approach warrants an objective assessment of the uncertainty of the estimator and produces confidence intervals with almost correct coverage, provided that the sample size is large enough to assume approximately both the normality of the estimator and the unbiasedness

of the variance estimator). A model-based approach is the best choice if we are interested in constructing a map, or in predicting the values of the survey variable for many small areas in the most efficient way. In this case, the efficiency is increased by postulating a model describing the autocorrelation of data, but this may weaken the validity of the resulting inference (efficiency more important than validity).

The design-based approach is largely adopted to assess natural resource condition (see de Gruijter *et al.*, 2006; Gregoire & Valentine, 2008). For example, the Environmental Monitoring and Assessment Program (EMAP) relies entirely on probability samples (see Stevens, 1994). The model based-approach is the most used framework in geostatistics (mining, soil studies, and air pollution monitoring), where modelling the spatial correlation of data is conceptually more appropriate.

One of the principal objections against the design-based approach to survey spatial populations is that it pays little care to ancillary information provided by the sample labels (space coordinates), using it mainly as the basis for stratification. One possible answer to this problem is a more intense use, at the design stage, of prior knowledge on spatial pattern in the response variable for achieving more efficient designs (see Quenouille, 1949; Bellhouse, 1977; Iachan, 1985). In the continuous case, a similar goal has been pursued by the proposal of spatially balanced sampling designs. In this context, we mention the random tessellation design (Overton & Stehman, 1993) and the generalized random tessellation stratified design (Stevens, 1997; Stevens & Olsen, 2004).

Less attention has been devoted to techniques aimed at improving efficiency at the estimation stage, using models for capturing insight into spatial pattern under the model-assisted setting (Särndal *et al.*, 1992). To our knowledge, only a few studies have appeared, which adopt this perspective. Brus (2000) has advocated the use of auxiliary variables in the form of a regression estimator within the model-assisted framework. Brus & Te Riele (2001) have dealt with the same problem in a two-phase sampling design where the first phase is used to estimate the unknown means of the auxiliary variables. For continuous spatial populations, Barabesi & Marcheselli (2005) have used the control-variate Monte Carlo integration method to increase the regression estimator accuracy.

In this paper, we want to go a step further, assuming as ancillary variable the space itself in a penalized spline regression model able to describe the spatial pattern in the data. More precisely, given a probability sample of locations, $s = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, we fit a spline regression model to population data in a design-based setting, and then, we employ the resulting population fitted values in a difference estimator. The idea is to mimic, in some way, what happens in geostatistics, where the autocorrelation of data is modelled by estimating the covariance function, and then, the population mean or total is predicted by a weighted average of the sample observations (the so-called kriging predictor) with weights depending on the estimated autocorrelation of data. We feel that the same bulk of prior information and knowledge should give results not very dissimilar, irrespective of the theoretical framework adopted.

The rest of this paper is organized as follows. In Section 2, we present the new estimator for a discrete population and establish its design properties. In Section 3, we extend the results to the case of continuous populations. Section 4 reports on a simulation study aimed at evaluating the performance of the proposed estimator. Section 5 contains some final comments.

2 Spline-Model-Assisted Estimator of a Finite Spatial Population Mean

In environmental populations, nearby locations tend to be influenced by the same set of factors. As a result, locations close together are more similar, in the response variable, than

locations that are far apart. In other words, data may be characterized by the presence of a spatial pattern. Thus, it makes sense to model the response variable assuming as predictors the spatial coordinates, which provide ancillary information. In this study, we focus on the penalized spline regression model, according to the lines established in Ruppert *et al.* (2003, Chapter 13) for bivariate smoothing.

We consider the case of a discretized continuous population whose units are given by the locations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ corresponding to the N grid cells into which the study area is subdivided by the superimposed regular grid. Following the procedure suggested in Ruppert *et al.* (2003, pp. 256–258), let $\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_K$ be a representative subset of the locations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, whose elements will be called *knots*, such that the $K \times K$ symmetric matrix $\boldsymbol{\Omega} = \{(||\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_l||)^2 \log(||\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_l||)\}_{k,l=1,\ldots,K}$, where $||\cdot||$ is the Euclidean norm, is nonsingular. For each location $\mathbf{x}_i$, define the following pseudocovariate values $[z_1(\mathbf{x}_i), \ldots, z_K(\mathbf{x}_i)] = [\breve{z}_1(\mathbf{x}_i), \ldots, \breve{z}_K(\mathbf{x}_i)]\boldsymbol{\Omega}^{-1/2}$, where $\breve{z}_k(\mathbf{x}_i) = (||\mathbf{x}_i - \boldsymbol{\kappa}_k||)^2 \log(||\mathbf{x}_i - \boldsymbol{\kappa}_k||)$; $i = 1, \ldots, N$; $k = 1, \ldots, K$.

Under a model-assisted approach to the randomization-based inference, assume that the population values $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N)$ are the realization of an N-dimensional random vector according to the following spline regression model, called ξ,

$$\begin{cases} \mathrm{E}_\xi[y(\mathbf{x}_i)] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_1 z_1(\mathbf{x}_i) + \cdots + u_K z_K(\mathbf{x}_i), & i = 1, \ldots, N, \\ \mathrm{V}_\xi[y(\mathbf{x}_i)] = \sigma_\varepsilon^2, & i = 1, \ldots, N, \end{cases} \tag{1}$$

where E_ξ and V_ξ denote the expected value and variance with respect to the model ξ; $\beta_0, \beta_1, \beta_2, u_1, \ldots, u_K$, and σ_ε^2 are the model parameters. We note that the pseudocovariates, $z_1(\mathbf{x}_i), \ldots, z_K(\mathbf{x}_i)$, depending on the spatial configuration of data and knots, contain local spatial information. As a result, coefficients $u_1, \ldots, u_K$ account for the local behaviour of the response variable.

If, hypothetically, the response variable was known for all population units, we would estimate model (1) by means of the penalized least-square criterion, minimizing the function (with respect to $\beta_0, \beta_1, \beta_2, u_1, \ldots, u_K$)

$$\sum_{i=1}^{N} [y(\mathbf{x}_i) - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - u_1 z_1(\mathbf{x}_i) - \cdots - u_K z_K(\mathbf{x}_i)]^2 + \lambda \sum_{k=1}^{K} u_k^2. \tag{2}$$

Penalties are imposed to regulate the variation of the coefficients $u_1, \ldots, u_K$ in model (1). By observing function (2), we can see that the larger the value of the penalty parameter $\lambda \geq 0$, the more the fit of population values shrinks towards the linear fit with respect to x_1 and x_2; in contrast, the smaller the values of λ the more wiggly the fit.

The minimization of function (2), for a fixed value of λ, gives the following estimator of $\beta_0, \beta_1, \beta_2, u_1, \ldots, u_K$:

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \left[\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix} + \lambda \mathbf{D} \right]^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix},$$

where: $\tilde{\boldsymbol{\beta}} = [\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2]'$; $\tilde{\mathbf{u}} = [\tilde{u}_1, \ldots, \tilde{u}_K]'$; $\mathbf{X}$ is the $N \times 3$-matrix with i-th row $[1, x_{i1}, x_{i2}]$, for $i = 1, \ldots, N$; $\mathbf{Z}$ is the $N \times K$ matrix whose ik-th entry is $z_k(\mathbf{x}_i)$, for $k = 1, \ldots, K$, $i = 1, \ldots, N$; $\mathbf{D} = \text{blockdiag}\,[\mathbf{0}_{3\times 3}, \mathbf{I}_K]$, $\mathbf{I}_K$ being a $K \times K$ identity matrix; $\mathbf{y} = [y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N)]'$.

The fitted population values $\tilde{\mathbf{y}} = [\tilde{y}(\mathbf{x}_1), \ldots, \tilde{y}(\mathbf{x}_N)]'$ are given by $\tilde{\mathbf{y}} = \tilde{\mathbf{S}}_\lambda \mathbf{y}$, where $\tilde{\mathbf{S}}_\lambda$ is the smoothing matrix

$$\tilde{\mathbf{S}}_\lambda = [\mathbf{X}, \mathbf{Z}] \left[\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix} + \lambda \mathbf{D} \right]^{-1} \begin{bmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{bmatrix}.$$

The trace of $\tilde{\mathbf{S}}_\lambda$ 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 $3 \leq r = \mathrm{tr}(\tilde{\mathbf{S}}_\lambda) \leq K + 3$. The lower limit is reached as $\lambda \to \infty$, and the upper limit is attained as $\lambda \to 0$.

We note that model (1) can be interpreted as a linear mixed model and it can be written in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\mathrm{E}_\xi(\mathbf{u}) = \mathbf{0}$, $\mathrm{E}_\xi(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{Cov}_\xi(\mathbf{u}, \boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{Var}_\xi(\mathbf{u}) = \sigma_u^2 \mathbf{I}_K$, $\mathrm{Var}_\xi(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}_N$. In this context, the least-square criterion consists in minimizing the function

$$\frac{1}{\sigma_\varepsilon^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \frac{1}{\sigma_u^2}\mathbf{u}'\mathbf{u}. \tag{4}$$

By comparing (2) and (4), we see that fitting data with model (1), for a fixed value of λ, is equivalent to fitting data with model (3) with $\sigma_\varepsilon^2/\sigma_u^2 = \lambda$. This may give some guidelines for choosing the value of the penalty parameter.

Model (3) is a low-rank radial smoother in contrast with the full-rank radial smoother that we would obtain if we assumed $K = N$ in the definition of matrix $\mathbf{Z}$, that is if we took as knots the entire set of locations $\mathbf{x}_1, \ldots, \mathbf{x}_N$. Notice that this smoother belongs to the class of thin-plate splines (see Huang & Chen, 2007, p. 1011). An alternative full-rank radial smoother is represented by kriging (see Cressie, 1993) that can be expressed in the form of model (3) if we put $K = N$ and define $\mathbf{Z} = \boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{\Sigma}$ is the covariance matrix of $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N)$ divided by σ_u^2. As concerns the prediction ability of the full-rank smoothers mentioned above, several comparative studies have highlighted that neither the smoothing spline method nor the kriging method dominates (see, for example, Hutchinson & Gessler, 1994; Laslett, 1994). Kriging assumes that $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N)$ come from a realization of a stochastic process with a spatial dependence structure that is estimated from the data, while smoothing splines use a particular generalized covariance function (see Cressie, 1993, Section 5.4).

It is clear that the low-rank smoother we use in this study can be seen as an approximation of the corresponding full-rank smoother. As will be clear later, this choice allows us to obtain the design-based consistent estimator of the vector of population fitted values $\tilde{\mathbf{y}} = [\tilde{y}(\mathbf{x}_1), \ldots, \tilde{y}(\mathbf{x}_N)]'$ defined above.

For the purpose of introducing the new estimator, we give a sketch of inclusion probabilities of the first and second order, considering, for simplicity, a sampling design of fixed size n. The inclusion probability of the first order is the probability that unit $\mathbf{x}_i$, $i = 1, \ldots, N$, is included in the random sample $s = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. If the sampling is without replacement, this probability is given by

$$\pi(\mathbf{x}_i) = \Pr[(\mathbf{X}_1 = \mathbf{x}_i) \cup \cdots \cup (\mathbf{X}_n = \mathbf{x}_i)] = \sum_{j=1}^{n} \Pr(\mathbf{X}_j = \mathbf{x}_i), \tag{5}$$

where random coordinates are in upper case and fixed coordinates are in lower case. For example, in simple random sampling without replacement, $\pi(\mathbf{x}_i) = n/N$, $i = 1, \ldots, N$. The inclusion probability of the second order, that is, the probability that units i and i' are simultaneously in the sample, is given by

$$\pi(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{n} \sum_{j' \neq j}^{n} \Pr(\mathbf{X}_j = \mathbf{x}_i, \mathbf{X}_{j'} = \mathbf{x}_{i'}). \tag{6}$$

For example, in simple random sampling without replacement, $\Pr(\mathbf{X}_j = \mathbf{x}_i, \mathbf{X}_{j'} = \mathbf{x}_{i'}) = 1/N(N-1)$, then $\pi(\mathbf{x}_i, \mathbf{x}_{i'}) = n(n-1)/N(N-1)$.

In the sequel, we will repeatedly make use of the Horvitz-Thompson estimator of the population total, which is given by $\hat{T}_{HT} = \sum_{j=1}^{n} y(\mathbf{x}_j)/\pi(\mathbf{x}_j)$.

Now, note that the vector $[\tilde{\boldsymbol{\beta}}', \tilde{\mathbf{u}}']'$ is an unknown population characteristic that can be estimated using a random sample. Now, consider a sample of locations $s = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ drawn from the target population by a sampling design that assigns the inclusion probability $\pi(\mathbf{x}_i)$ to location $\mathbf{x}_i, i = 1, \ldots, N$. It can be shown that a consistent estimator of $[\tilde{\boldsymbol{\beta}}', \tilde{\mathbf{u}}']'$ is given by (see Appendix I)

$$\begin{bmatrix} \hat{\tilde{\boldsymbol{\beta}}} \\ \hat{\tilde{\mathbf{u}}} \end{bmatrix} = \left[\begin{bmatrix} \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{X}_s & \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{Z}_s \\ \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{X}_s & \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{Z}_s \end{bmatrix} + \lambda \mathbf{D} \right]^{-1} \begin{bmatrix} \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{y}_s \\ \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{y}_s \end{bmatrix}, \tag{7}$$

where $\mathbf{X}_s$ and $\mathbf{Z}_s$ are the submatrices of $\mathbf{X}$ and $\mathbf{Z}$ consisting of the rows for which $\mathbf{x}_i \in s$; $\boldsymbol{\Pi}_s = \text{diag}(1/\pi(\mathbf{x}_i))_{\mathbf{x}_i \in s}$ is the submatrix of $\boldsymbol{\Pi}_U = \text{diag}(1/\pi(\mathbf{x}_i))_{i=1,\ldots,N}$; $\mathbf{y}_s$ is the subvector of $\mathbf{y}$ for $\mathbf{x}_i \in s$. The smoother matrix is now replaced by

$$\tilde{\mathbf{S}}_{\lambda s} = [\mathbf{X}, \mathbf{Z}] \left[\begin{bmatrix} \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{X}_s & \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{Z}_s \\ \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{X}_s & \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{Z}_s \end{bmatrix} + \lambda \mathbf{D} \right]^{-1} \begin{bmatrix} \mathbf{X}_s' \boldsymbol{\Pi}_s \\ \mathbf{Z}_s' \boldsymbol{\Pi}_s \end{bmatrix}, \tag{8}$$

and using (8), a design-based consistent estimator of $\tilde{\mathbf{y}} = [\tilde{y}(\mathbf{x}_1), \ldots, \tilde{y}(\mathbf{x}_N)]'$ is provided by

$$\hat{\tilde{\mathbf{y}}} = \tilde{\mathbf{S}}_{\lambda s} \mathbf{y}_s. \tag{9}$$

So far, we have obtained the fitted population values according to model (1), $\tilde{\mathbf{y}} = [\tilde{y}(\mathbf{x}_1), \ldots, \tilde{y}(\mathbf{x}_N)]'$, and the corresponding consistent estimator in (9). The next step is to estimate the population mean $\overline{Y}$ using $\hat{\tilde{\mathbf{y}}} = [\hat{\tilde{y}}(\mathbf{x}_1), \ldots, \hat{\tilde{y}}(\mathbf{x}_N)]'$ as predictor of $\mathbf{y}$ in a difference estimator.

When a variable $a(\mathbf{x}_i)$ is a good predictor of the study variable $y(\mathbf{x}_i)$ and it is known for all units in the population, an efficient design unbiased estimator for the population mean is the difference estimator

$$\hat{\overline{Y}}_d = \frac{1}{N} \sum_{i=1}^{N} a(\mathbf{x}_i) + \frac{1}{N} \sum_{j=1}^{n} \frac{y(\mathbf{x}_j) - a(\mathbf{x}_j)}{\pi(\mathbf{x}_j)}.$$

The more the predicted values are near to the true values, the more the design variance of this estimator approaches zero.

Hence, using $\hat{\tilde{\mathbf{y}}}$ in (9) as a predictor of $\mathbf{y}$, a model-assisted estimator of the population mean is given by

$$\hat{\overline{Y}}_{spl} = \frac{1}{N} \sum_{i=1}^{N} \hat{\tilde{y}}(\mathbf{x}_i) + \frac{1}{N} \sum_{j=1}^{n} \frac{e(\mathbf{x}_j)}{\pi(\mathbf{x}_j)}, \tag{10}$$

where $e(\mathbf{x}_j) = y(\mathbf{x}_j) - \hat{\tilde{y}}(\mathbf{x}_j)$. The estimator is no longer design unbiased since $\hat{\tilde{\mathbf{y}}}$ is not fixed, but estimated from the sample.

Ignoring that $\hat{\tilde{\mathbf{y}}}$ is estimated from the sample and using the standard variance estimator of $\hat{\overline{Y}}_d$, an estimator of the variance of $\hat{\overline{Y}}_{spl}$ is given by (see Särndal *et al.*, 1992, p. 294)

$$\hat{\mathrm{V}}_p(\hat{\overline{Y}}_{spl}) = \frac{1}{N^2} \left\{ \sum_{j=1}^{n} [1 - \pi(\mathbf{x}_j)] \frac{e^2(\mathbf{x}_j)}{\pi^2(\mathbf{x}_j)} + \sum_{j=1}^{n} \sum_{j' \neq j}^{n} \frac{\pi(\mathbf{x}_j, \mathbf{x}_{j'}) - \pi(\mathbf{x}_j)\pi(\mathbf{x}_{j'})}{\pi(\mathbf{x}_j, \mathbf{x}_{j'})} \frac{e(\mathbf{x}_j)}{\pi(\mathbf{x}_j)} \frac{e(\mathbf{x}_{j'})}{\pi(\mathbf{x}_{j'})} \right\} \tag{11}$$

(the suffix p on the left-hand side of equation (11) indicates that here, we are operating in the design-based framework, i.e. expectations are taken with respect to the sampling design). Estimator (11) is design-consistent as far as $\hat{\tilde{\mathbf{y}}}$ is consistent for $\tilde{\mathbf{y}}$. It tends to be downward biased since it does not take into account the sample variability of $\hat{\tilde{\mathbf{y}}}$. Although this problem is overcome when the sample size is sufficiently large (see the simulation study below), alternative variance estimators, based on resampling techniques, could be studied.

Note that estimator (10) can also be written as $\hat{\bar{Y}}_{spl} = N^{-1}\sum_{j=1}^{n} g_j(\lambda)y(\mathbf{x}_j)/\pi(\mathbf{x}_j)$ where

$$g_j(\lambda) = 1 + (\mathbf{T}_{\mathbf{v}} - \hat{\mathbf{T}}_{\mathbf{v}HT})' \left(\sum_{l=1}^{n} \frac{\mathbf{v}_l\mathbf{v}_l'}{\pi(\mathbf{x}_l)} + \lambda\mathbf{D}\right)^{-1} \mathbf{v}_j,$$

$\mathbf{v}_j$ is the vector $[1, x_{j1}, x_{j2}, z_1(\mathbf{x}_j), \ldots, z_K(\mathbf{x}_j)]'$, $\mathbf{T}_{\mathbf{v}} = \sum_{i=1}^{N}\mathbf{v}_i$ and $\hat{\mathbf{T}}_{\mathbf{v}HT} = \sum_{j=1}^{n}\mathbf{v}_j/\pi(\mathbf{x}_j)$. Quantities $g_j(\lambda), j = 1, \ldots, n$, are called g-weights, and following Särndal *et al.* (1992, equation (6.6.4)), an alternative variance estimator of (10) is defined as

$$\hat{\mathrm{V}}_{pg}(\hat{\bar{Y}}_{spl}) = \frac{1}{N^2}\left\{\sum_{j=1}^{n}[1-\pi(\mathbf{x}_j)]\frac{[g_j(\lambda)e(\mathbf{x}_j)]^2}{\pi^2(\mathbf{x}_j)} + \sum_{j=1}^{n}\sum_{j'\neq j}^{n}\frac{\pi(\mathbf{x}_j,\mathbf{x}_{j'}) - \pi(\mathbf{x}_j)\pi(\mathbf{x}_{j'})}{\pi(\mathbf{x}_j,\mathbf{x}_{j'})}\frac{g_j(\lambda)e(\mathbf{x}_j)}{\pi(\mathbf{x}_j)}\frac{g_{j'}(\lambda)e(\mathbf{x}_{j'})}{\pi(\mathbf{x}_{j'})}\right\}.$$

This estimator is the g-weighted version of equation (11) and it is known to have somewhat better conditional properties and reduced downward bias.

2.1 Properties of the Proposed Estimator

First, note that using the same knots and the same penalty term λ (not necessarily optimal) for all survey variables, $\hat{\bar{Y}}_{spl}$ is a linear estimator, i.e. it can be written as a weighted sum of observations with weights $g_j(\lambda)/N\pi(\mathbf{x}_j)$, for $j = 1, \ldots, n$, that are the same for all estimated survey variables means. This is a feature much appreciated by survey practitioners for coherence issues.

To establish the large-sample properties of the proposed estimator, let us introduce an asymptotic framework. In general, the grid consists of two sets of parallel segments intersecting at some angle with grid points at the intersections of the segments. A grid is regular if the spacing between segments is constant within each set, although it may differ between sets. So, a rectangular grid consists of two sets of segments that meet at right angles. Segments intersecting at 60° form a triangular grid. The reference point is the central point of the grid cell.

Let t and qt be, respectively, the number of segments in the first and second sets that generate the finite population. Consider the sequence of finite populations in domain A obtained increasing the number of segments in the first set t, while maintaining q fixed as well as the region covered by the grid. Then, as far as $t \to \infty$, the grid becomes more and more dense and the population size $N_t = (t-1)(qt-1)$ tends to infinity. Assume that for any value of t, the values of the survey variable $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_{Nt})$ are limited in some interval $[m, M]$. Let us denote with t the population with t segments in the first set. Now, for each population t, let $p_t(s)$ be the sampling design used to select a sample of size n_t such that the inclusion probabilities are positive for all units of the population. In such a case, the Horvitz-Thompson estimator of the mean is unbiased. Assume that n_t goes to infinity as $t \to \infty$.

In using the proposed estimator $\hat{\overline{Y}}_{spl}$, we assume that: *a*) the number K and the locations of the knots are fixed; *b*) the smoothing parameter increases with N_t, i.e. $\lambda_t = N_t\delta$ for some chosen δ. In this way, the number of degrees of freedom of the smoother is approximately the same for all populations. Let $o_p(.)$ and $O_p(.)$ denote infinitesimal terms for $t \to \infty$ of order, respectively, smaller than or equal to that in parenthesis. In order to prove our results, we make the following assumptions on the sequence of populations and sampling designs when $t \to \infty$.

ASSUMPTION I. *For any vector* $\mathbf{t}$ *of survey variables, the Horvitz-Thompson estimator* $\hat{\hat{\mathbf{t}}}_{HT}$ *of the mean of* $\mathbf{t}$ *satisfies a central limit theorem with a covariance matrix* $\mathrm{V}_p(\hat{\hat{\mathbf{t}}}_{HT})$ *such that the limit of* $n_t \mathrm{V}_p(\hat{\hat{\mathbf{t}}}_{HT})$ *exists and is positive definite.*

ASSUMPTION II. *The estimated covariance matrix* $\hat{\mathrm{V}}_p(\hat{\hat{\mathbf{t}}}_{HT})$ *of* $\hat{\hat{\mathbf{t}}}_{HT}$, *whose elements are, for example, of type equation* (11), *is design-consistent, i.e.* $n_t[\hat{\mathrm{V}}_p(\hat{\hat{\mathbf{t}}}_{HT}) - \mathrm{V}_p(\hat{\hat{\mathbf{t}}}_{HT})] = o_p(1)$.

ASSUMPTION III. *The limit* $\lim_{t\to\infty} \begin{bmatrix} \tilde{\boldsymbol{\beta}}_t \\ \tilde{\mathbf{u}}_t \end{bmatrix}$ *exists and* $\begin{bmatrix} \hat{\tilde{\boldsymbol{\beta}}}_t \\ \hat{\tilde{\mathbf{u}}}_t \end{bmatrix} - \begin{bmatrix} \tilde{\boldsymbol{\beta}}_t \\ \tilde{\mathbf{u}}_t \end{bmatrix} = o_p(1)$.

The above assumptions are common in finite population asymptotic literature and discussed throughout (see, for example, Breidt *et al.*, 2005, pp. 836–837). Assumptions I and II are satisfied for common sampling designs in reasonably behaved populations. Assumption II requires that the design is measurable (i.e. the second-order inclusion probabilities are all positive). Assumption III ensures that the sample fit $\hat{\tilde{\mathbf{y}}}$ and the population fit $\tilde{\mathbf{y}}$ share a common limit. This assumption is weaker than assuming that model (1) is true.

The following theorem establishes the design consistency and asymptotic normality of $\hat{\overline{Y}}_{spl}$, along with the design consistency of its variance estimator.

THEOREM 1. *Under Assumptions* I–III, *the proposed estimator* $\hat{\overline{Y}}_{spl}$ *is design* $\sqrt{n_t}$*-consistent, in the sense that* $\hat{\overline{Y}}_{spl} - \overline{Y} = O_p(n_t^{-1/2})$. *Furthermore,*

$$\frac{\hat{\overline{Y}}_{spl} - \overline{Y}}{\hat{\mathrm{V}}_p(\hat{\overline{Y}}_{spl})} \to \mathrm{N}(0, 1),$$

where $N(0,1)$ *is the standard normal distribution.*

The proof is deferred to Appendix II.

3 The Case of Continuous Populations

When the spatial population is continuous, we can define the analogues of formulas (5) and (6), called inclusion density function of the first and second order, replacing the marginal probability $\Pr(\mathbf{X}_j = \mathbf{x}_i)$ in (5) with the marginal density function $f_j(\mathbf{x})$, and the joint probability $\Pr(\mathbf{X}_j = \mathbf{x}_i, \mathbf{X}_{j'} = \mathbf{x}_{i'})$ in (6) with the joint marginal density $f_{jj'}(\mathbf{x}, \mathbf{x}')$, respectively. Hence, the first- and second-order inclusion density functions are defined as $\pi(\mathbf{x}) = \sum_{j=1}^{n} f_j(\mathbf{x})$ and $\pi(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{n} \sum_{j'\neq j}^{n} f_{jj'}(\mathbf{x}, \mathbf{x}')$, respectively. In this case, the Horvitz-Thompson estimator of the population total is formally identical to the formula given for the discrete case, where the inclusion probabilities must be replaced with the inclusion probability density functions (see Cordy, 1993).

Now, we want to find the analogue of formula (10) for the case of a continuous population. We assume that a probability sample of locations, $s = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, is available, selected according

to a given sampling design with inclusion density probability functions of the first and second order represented by $\pi(\mathbf{x}_j)$ and $\pi(\mathbf{x}_j, \mathbf{x}_{j'})$; $j, j' = 1, \ldots, n$.

With the same procedure followed in the discrete case, we begin with the choice of K knots, $\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_K$, in A, and with the definition of the pseudocovariate values $[z_1(\mathbf{x}), \ldots, z_K(\mathbf{x})] = [\breve{z}_1(\mathbf{x}), \ldots, \breve{z}_K(\mathbf{x})]\boldsymbol{\Omega}^{-1/2}$, $\mathbf{x} \in A$ (the matrix $\boldsymbol{\Omega}$ has been defined in Section 2), where $\breve{z}_k(\mathbf{x}) = (||\mathbf{x} - \boldsymbol{\kappa}_k||)^2 \log(||\mathbf{x} - \boldsymbol{\kappa}_k||)$, $\mathbf{x} \in A$, $k = 1, \ldots, K$. Then, model (1) becomes

$$\begin{cases} \mathrm{E}_\xi[y(\mathbf{x})] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u_1 z_1(\mathbf{x}) + \cdots + u_K z_K(\mathbf{x}), \mathbf{x} \in A, \\ \mathrm{V}_\xi[y(\mathbf{x})] = \sigma_\varepsilon^2, \mathbf{x} \in A. \end{cases} \tag{12}$$

If we assume that the response variables $y(\mathbf{x})$ were known for all $\mathbf{x}$ in A, we could fit the surface $y(\mathbf{x})$ estimating model (12) by means of the penalized least-square method, that is, minimizing the function

$$\int_A [y(\mathbf{x}) - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - u_1 z_1(\mathbf{x}) - \cdots - u_K z_K(\mathbf{x})]^2 d\mathbf{x} + \lambda \sum_{k=1}^{K} u_k^2. \tag{13}$$

Proceeding as in the case of a discrete population, the design-based estimator of the parameter vector $[\tilde{\boldsymbol{\beta}}', \tilde{\mathbf{u}}']'$ that minimizes function (13) is given by the following formula, formally equal to (7) (details are presented in Appendix III),

$$\begin{bmatrix} \tilde{\tilde{\boldsymbol{\beta}}} \\ \tilde{\tilde{\mathbf{u}}} \end{bmatrix} = \left[\begin{bmatrix} \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{X}_s & \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{Z}_s \\ \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{X}_s & \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{Z}_s \end{bmatrix} + \lambda \mathbf{D} \right]^{-1} \begin{bmatrix} \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{y}_s \\ \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{y}_s \end{bmatrix}. \tag{14}$$

Here, $\mathbf{X}_s$ is the $n \times 3$ matrix having as j-th row $[1, x_{j1}, x_{j2}]$, $j = 1, \ldots, n$; $\mathbf{Z}_s$ is an $n \times K$ matrix whose j-th row is given by $[z_1(\mathbf{x}_j), \ldots, z_K(\mathbf{x}_j)]$; $\boldsymbol{\Pi}_s = \mathrm{diag}(1/\pi(\mathbf{x}_1), \ldots, 1/\pi(\mathbf{x}_n))$ is the diagonal matrix having in its diagonal the inclusion probability densities of the sample locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Now, for each location $\mathbf{x} \in A$, we can predict the response variable $y(\mathbf{x})$ using the fitted model. We have $\hat{\tilde{y}}(\mathbf{x}) = \hat{\tilde{\beta}}_0 + \hat{\tilde{\beta}}_1 x_1 + \hat{\tilde{\beta}}_2 x_2 + \hat{\tilde{u}}_1 z_1(\mathbf{x}) + \cdots + \hat{\tilde{u}}_K z_K(\mathbf{x})$, $\mathbf{x} \in A$. Finally, we can write the spline model-assisted estimator of the mean of a continuous spatial population as follows:

$$\hat{\bar{Y}}_{spl} = \frac{1}{|A|} \int_A \hat{\tilde{y}}(\mathbf{x}) d\mathbf{x} + \frac{1}{|A|} \sum_{j=1}^{n} \frac{e(\mathbf{x}_j)}{\pi(\mathbf{x}_j)}, \tag{15}$$

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

Denoting by $\psi^{kk'}$ the generic entry of matrix $\boldsymbol{\Omega}^{-1/2}$, and putting $\hat{\tilde{\gamma}}_k = \hat{\tilde{u}}_1 \psi^{k1} + \cdots + \hat{\tilde{u}}_K \psi^{kK}$, $k = 1, \ldots, K$, we can express the first term on the right-hand side of equation (15) in the form $\int_A \hat{\tilde{y}}(\mathbf{x}) d\mathbf{x}/|A| = \hat{\tilde{\beta}}_0 + \hat{\tilde{\beta}}_1 \bar{x}_1 + \hat{\tilde{\beta}}_2 \bar{x}_2 + \hat{\tilde{\gamma}}_1 \bar{\breve{z}}_1 + \cdots + \hat{\tilde{\gamma}}_K \bar{\breve{z}}_K$, where $\bar{x}_1$ and $\bar{x}_2$ are the means in A of the geographic coordinates x_1 and x_2, while $\bar{\breve{z}}$ are the means in A of pseudocovariates $\breve{z}_k(\mathbf{x})$, $k = 1, \ldots, K$.

With similar arguments to those used to derive (11), it can be shown that an estimator of the variance of $\hat{\bar{Y}}_{spl}$ may be given by

$$\hat{\mathrm{V}}_p(\hat{\bar{Y}}_{spl}) = \frac{1}{|A|^2} \left[\sum_{j=1}^{n} \frac{e^2(\mathbf{x}_j)}{\pi^2(\mathbf{x}_j)} + \sum_{j=1}^{n} \sum_{j' \neq j}^{n} \frac{\pi(\mathbf{x}_j, \mathbf{x}_{j'}) - \pi(\mathbf{x}_j)\pi(\mathbf{x}_{j'})}{\pi(\mathbf{x}_j, \mathbf{x}_{j'})} \frac{e(\mathbf{x}_j)}{\pi(\mathbf{x}_j)} \frac{e(\mathbf{x}_{j'})}{\pi(\mathbf{x}_{j'})} \right].$$

A g-weighted version may also be considered.

Proceeding similarly as in the discrete case, we can prove that estimator (15) is $\sqrt{n}$- consistent. Assumption III reduces to

$$\begin{bmatrix} \hat{\tilde{\boldsymbol{\beta}}} \\ \hat{\tilde{\mathbf{u}}} \end{bmatrix} - \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = o_p(1), \quad \text{as } n \to \infty.$$

Note that the population is infinite and standard central limit theorem results can be applied.

4 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 $\hat{\bar{Y}}_{spl}$ with that of the Horvitz-Thompson estimator $\hat{\bar{Y}}_{HT} = N^{-1} \sum_{j=1}^{n} y(\mathbf{x}_j)/\pi(\mathbf{x}_j)$, under a stratified random sampling with a proportional allocation.

We considered some artificial populations defined on a grid of $60 \times 60 = 3{,}600$ spatial locations, $\mathbf{x} = (x_{l1}, x_{l'2})'$, with $l, l' = 1, \ldots, 60$, regularly located in the square $[0, 1] \times [0, 1]$, where $x_{lh} = (2l - 1)/120, l = 1, \ldots, 60, h = 1, 2$. The following finite populations were selected (they are sketched in Figure 2).

Population a. It was obtained by the Cholesky triangularization technique as a realization of the random variables $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_{3,600})$, whose covariance matrix has as typical element $\sigma^2 \rho^{d_{ii'}}$, with $\sigma^2 = 100$, $\rho = 0.5$, and $d_{ii'} = ||\mathbf{x}_i - \mathbf{x}_{i'}||$ (see Figure 2(a)).
Population b. It was obtained with the same procedure as for population *a*, but with $\rho = 0.8$ (see Figure 2(b)).
Population c. It was obtained from the function (see Figure 2(c))

$$f(x_{l1}, x_{l'2}) = 5[\sin(x_{l1})]^2 + 5[\cos(x_{l'2})]^2 + 5x_{l1}.$$

The 60×60 grid was divided into nine squares of 20×20 locations, to provide 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 $\mathbf{Z}$.

From each population, 2,000 stratified random samples with a proportional allocation were selected. 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 the value such that the trace of the matrix

$$\left[\begin{bmatrix} \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{X}_s & \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{Z}_s \\ \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{X}_s & \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{Z}_s \end{bmatrix} + \lambda \mathbf{D} \right]^{-1} \begin{bmatrix} \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{X}_s & \mathbf{X}_s' \boldsymbol{\Pi}_s \mathbf{Z}_s \\ \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{X}_s & \mathbf{Z}_s' \boldsymbol{\Pi}_s \mathbf{Z}_s \end{bmatrix},$$

which estimates the number of degrees of freedom of smoother $\tilde{\mathbf{S}}_\lambda$, is equal to a value r chosen in advance. Even though $\tilde{\mathbf{S}}_\lambda$ is actually known given λ (matrices $\mathbf{X}$ and $\mathbf{Z}$ are known for the complete population) in a practical context, the computational effort is lowered using the sample data. Then, the proposed estimator $\hat{\bar{Y}}_{spl}$ and the Horvitz-Thompson estimator were calculated. To estimate the variance of $\hat{\bar{Y}}_{spl}$, we chose, as an example, $\hat{\mathrm{V}}_p(\hat{\bar{Y}}_{spl})$ in equation (11). To compensate for the negative bias that might arise ignoring the variance component due to the estimation of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$, in particular when r is high, we heuristically used the (naïve) formula $\hat{\mathrm{V}}'_p(\hat{\bar{Y}}_{spl}) = \hat{\mathrm{V}}_p(\hat{\bar{Y}}_{spl})(n - 9)/(n - 9 - r)$, where $n-9$ is the conventional number of degrees of freedom associated to the estimated variance of the sample mean in a stratified design with nine equal-sized strata (see Thompson, 2002, p. 121).

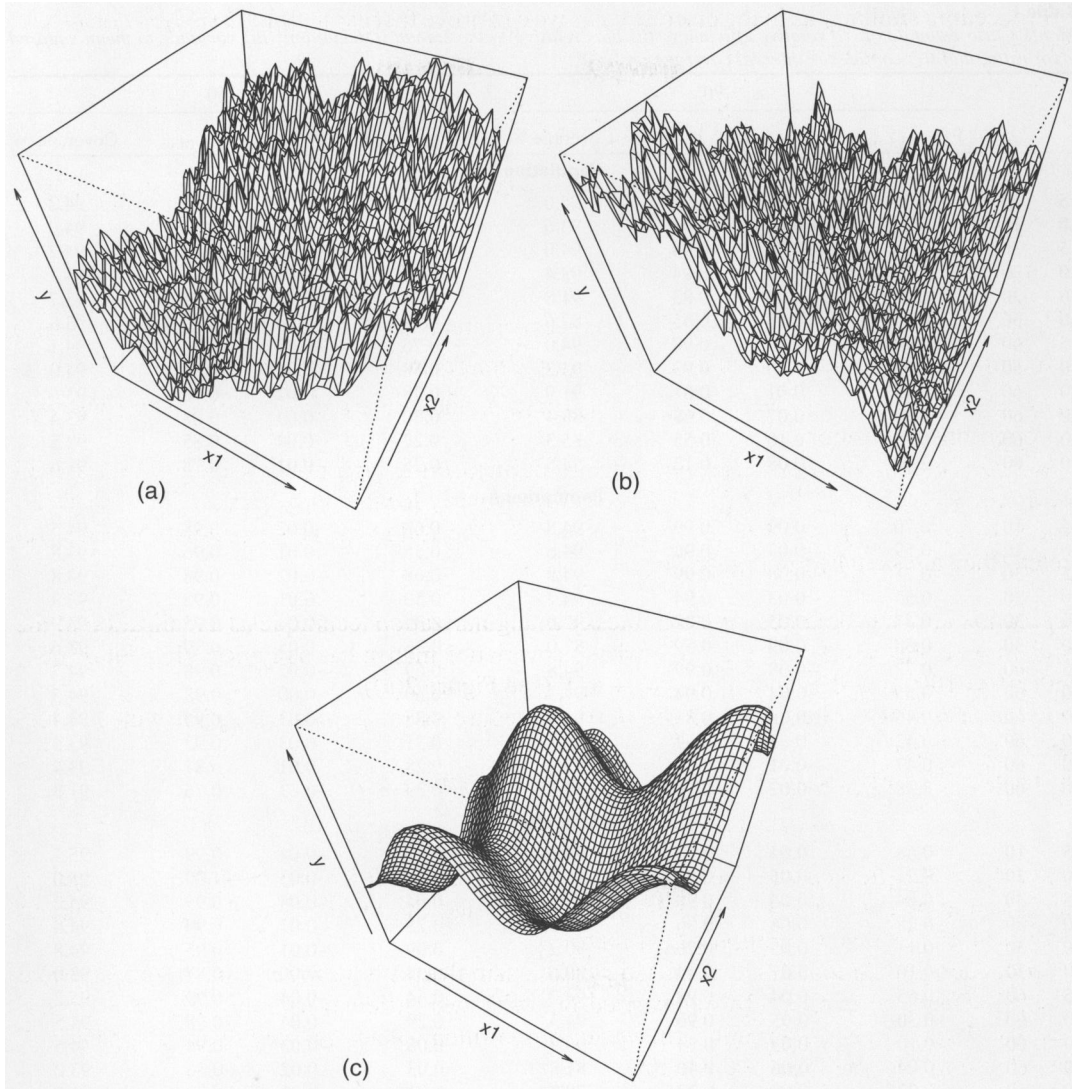

Figure 2. *Populations used in the simulation.*

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 $3 \leq r = \mathrm{tr}(\tilde{\mathbf{S}}_\lambda) \leq K + 3$ hold.

Denoting with E_{MC} and MSE_{MC}, the Monte Carlo empirical expected value and mean square error, respectively, the comparison of the two estimators was carried out by means of the following measures:

- Bias relative to the mean $R = \mathrm{E}_{MC}(\hat{\overline{Y}}_{spl} - \overline{Y})/\overline{Y}$.
- Bias relative to the standard deviation $R_{\mathrm{b/sd}} = [\mathrm{E}_{MC}(\hat{\overline{Y}}_{spl} - \overline{Y})]/[E_{MC}(\hat{V}'_p(\hat{\overline{Y}}_{spl}))]^{1/2}$.
- Relative efficiency $\mathrm{Eff}_{MC}(\hat{\overline{Y}}_{spl}) = \mathrm{MSE}_{MC}(\hat{\overline{Y}}_{spl})/\mathrm{MSE}_{MC}(\hat{\overline{Y}}_{HT})$ (with respect to the Horvitz-Thompson estimator)

Table 1
Monte Carlo estimate of: (i) relative efficiency; (ii) bias relative to the standard deviation; (iii) variance to mean squared error ratio; and (iv) confidence interval coverage.

		$n = 90$				$n = 360$			
r	K	$\text{Eff}_{MC}(\hat{\bar{Y}}_{spl})$	$R_{\text{b/sd}}$	$R_{\text{var/MSE}}$	Coverage %	$\text{Eff}_{MC}(\hat{\bar{Y}}_{spl})$	$R_{\text{b/sd}}$	$R_{\text{var/MSE}}$	Coverage %
					Population *a*				
5	10	0.74	−0.06	0.98	94.0	0.76	−0.01	0.97	94.2
10	10	0.53	−0.06	0.96	93.6	0.54	−0.01	0.94	94.0
5	30	0.74	−0.06	0.98	94.0	0.76	−0.02	0.97	94.2
10	30	0.56	−0.06	0.94	93.4	0.57	−0.01	0.94	94.1
20	30	0.42	0.04	0.83	91.8	0.38	0.04	0.91	93.7
30	30	0.47	0.05	0.65	87.6	0.32	0.04	0.90	93.8
5	60	0.74	−0.06	0.98	94.0	0.76	−0.02	0.97	94.1
10	60	0.58	−0.06	0.93	93.2	0.58	−0.01	0.94	94.0
20	60	0.42	0.01	0.81	91.9	0.38	0.02	0.90	93.6
30	60	0.37	0.07	0.68	88.4	0.30	0.04	0.88	93.4
40	60	0.40	0.10	0.55	85.3	0.27	0.04	0.85	92.5
60	60	2.00	0.08	0.12	54.5	0.26	−0.01	0.78	91.6
					Population *b*				
5	10	0.70	−0.09	0.99	94.8	0.68	−0.02	0.98	94.5
10	10	0.59	−0.03	0.96	94.6	0.55	0.01	0.96	93.8
5	30	0.71	−0.08	0.99	94.8	0.68	−0.02	0.98	94.8
10	30	0.55	−0.03	0.94	94.7	0.50	0.01	0.98	94.3
20	30	0.47	0.05	0.82	92.0	0.39	0.04	0.95	93.9
30	30	0.58	0.09	0.59	87.0	0.36	0.05	0.89	93.6
5	60	0.72	−0.08	0.99	94.8	0.69	−0.02	0.98	94.7
10	60	0.55	−0.04	0.94	94.3	0.50	0.00	0.98	94.5
20	60	0.46	0.02	0.81	91.5	0.38	0.03	0.95	94.4
30	60	0.43	0.03	0.67	88.5	0.31	0.03	0.92	93.5
40	60	0.47	−0.01	0.52	83.9	0.28	0.01	0.87	93.4
60	60	2.38	−0.02	0.11	52.6	0.29	−0.03	0.76	91.3
					Population *c*				
5	10	0.68	0.04	0.98	94.3	0.67	−0.04	0.99	95.1
10	10	0.21	0.05	0.98	94.7	0.20	−0.03	1.00	95.0
5	30	0.64	0.04	0.98	94.3	0.64	−0.04	0.99	95.2
10	30	0.28	0.06	0.96	94.6	0.28	−0.03	0.99	94.8
20	30	0.11	0.05	0.76	90.2	0.09	−0.01	0.95	94.8
30	30	0.10	−0.01	0.44	80.6	0.05	0.00	0.87	93.0
5	60	0.65	0.04	0.98	94.2	0.64	−0.04	0.99	95.2
10	60	0.30	0.05	0.96	94.3	0.29	−0.04	0.98	94.5
20	60	0.10	0.05	0.74	90.9	0.08	−0.03	0.94	94.5
30	60	0.04	0.06	0.46	81.8	0.03	−0.02	0.86	93.2
40	60	0.02	0.09	0.27	70.2	0.01	−0.01	0.78	91.6
60	60	0.03	0.03	0.06	42.0	0.00	0.10	0.69	88.9

- Variance to mean squared error ratio $R_{\text{var/mse}} = \text{E}_{MC}[\hat{\text{V}}'_p(\hat{\bar{Y}}_{spl})]/\text{MSE}_{MC}(\hat{\bar{Y}}_{spl})$.
- 95% confidence interval coverage (percentage).

The simulation results are presented in Table 1. The bias relative to the mean always takes negligible values and is omitted.

The new estimator appears to be far more efficient than the Horvitz-Thompson estimator. Its relative efficiency increases as r increases, with the following exceptions for $n = 90$: populations a and b, when $K = 30$ and $r > 20$ or $K = 60$ and $r > 30$; population c, when $K = 60$ and $r > 30$. The gain in efficiency is particularly high for population c generated by a model not including the noise component.

The proposed estimator behaves as approximately unbiased. The bias relative to the standard deviation is always small.

The ratio between the average value of $\hat{V}'_p(\hat{\bar{Y}}_{spl})$ and the empirical MSE$(\hat{\bar{Y}}_{spl})$ is always smaller than 1, which means that $\hat{V}'_p(\hat{\bar{Y}}_{spl})$ tends to underestimate the true design variance $V_p(\hat{\bar{Y}}_{spl})$. It can be observed that when the above ratio goes under 0.80, the confidence interval coverage deteriorates seriously, and it is different for more than three percentage points from the 95% nominal value. This occurs for $n = 90$ if $r > 10$, and for $n = 360$, if $r > 30$. Then, it seems that the underestimation of the variance of $\hat{\bar{Y}}_{spl}$ takes place if r is relatively high with respect to the sample size. In fact, high values of r allow the model to overfit the sample data and, as a result, to yield sample residuals smaller than those for nonsampled units. The findings offered by our simulation study suggest the following rule of thumb: given the sample size n, the dimension, r, of the spline regression model should satisfy the inequality $n/r > 10$. Since high values for r are preferable, less conservative rule of thumb might be achieved with alternative variance estimators, like the g-weighted version or some kind of resampling methods. However, this point is beyond the aim of the paper.

5 Concluding Remarks

We have proposed and studied an application of semiparametric methods in the model-assisted approach to the estimation of means or totals of spatial populations using the spatial coordinates as auxiliary variables. The idea is to assume a low-rank spline regression model as working model, and then to employ the resulting fitted values as predictors of the response values in a difference estimator. Our application allows approximately design unbiased and consistent estimators of the target parameters that capitalize on the spatial pattern in the data, captured by the fitted spline regression model.

The performance of the proposed estimator for finite-size samples has been investigated by means of a simulation study based on some artificial populations characterized by different levels of spatial structure. Substantial gains in efficiency with respect to the Horvitz-Thompson estimator, under a spatially stratified designs, are provided as long as the sample size is much larger than the number of degrees of freedom of the spline regression model. Under the same condition, the confidence interval coverage is quite near to the nominal level.

As regards the number of knots, first of all, we notice that the choice of this parameter is much less crucial than that of the smoothing parameter. The problem has been investigated mainly by means of simulation studies (see, for example, French *et al.*, 2001; Ruppert *et al.*, 2003). It is generally supported the idea that once there are enough knots to span domain A to fit features in the data, overfitting is controlled by the penalty term; hence, there is no risk for the performance of the spline regression model in fixing an high value for K, but this yields an increase of the computational burden.

A further problem is the knots location. A reasonable strategy "is to have knots mimic the distribution of the predictor space" (Ruppert *et al.*, 2003, p. 255). To do this, some algorithms have been proposed (see Johnson *et al.*, 1990; Nychka & Saltzman, 1998). Publicly available software is also provided by Nychka *et al.* (1998).

One important issue when applying our estimation method is the choice of the smoothing parameter λ (or, equivalently, the smoother number of degrees of freedom, r). Although our simulation demonstrates that significant gains in efficiency are obtained for a variety of choices of λ, it would be desirable to estimate this quantity on the basis of the sample observations. To do this, several techniques are available: the cross-validation criterion (see Ruppert *et al.*, 2003, pp. 114–116), the restricted maximum likelihood estimation, the generalized degree of freedom criterion (see Huang & Chen, 2007). Of course, these methods need to be suitably adjusted for

the effect of the sampling design. The issue deserves special attention and will be the object of further research.

At last, we note that the working model we have assumed to predict the response variable can also accommodate auxiliary variables other than the spatial coordinates, both quantitative and qualitative. The use of these covariates, often available, may increase the precision of the estimator and may capture some of the spatial dependence of the target variable. In such a case, the spline part of the model accounts for the spatial pattern of the residuals.

Acknowledgements

This work was 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" funded by the Italian government. We thank the anonymous referee for her/his valuable comments that improved the paper.

References

Barabesi, L. & Marcheselli, M. (2005). Monte Carlo integration strategies for design-based regression estimators of the spatial mean. *Environmetrics*, **16**, 803–817.

Bellhouse, D.R. (1977). Optimal designs for sampling in two dimensions. *Biometrika*, **64**, 605–611.

Breidt, F.J., Claeskens, G. & Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, **92**, 831–846.

Brus, D. (2000). Using regression models in design-based estimation of spatial means of soil properties. *European J. Soil Sci.*, **51**, 159–172.

Brus, D. & de Gruijter, J. (1997). Random sampling or geostatistical modeling? Choosing between design-based and model-based strategies for soil (with discussion). *Geoderma*, **80**, 1–59.

Brus, D. & Te Riele, W.J.M. (2001). Design-based regression estimators of spatial means of soil properties: the use of two-phase sampling when the means of the auxiliary variables are not known. *Geoderma*, **104**, 257–279.

Cordy, C.B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from continuous universe. *Statist. Probab. Lett.*, **18**, 353–362.

Cox, D.D., Cox, L.H. & Ensor, K.B. (1997). Spatial sampling and the environment: Some issues and directions. *Environ. Ecol. Stat.*, **4**, 219–233.

Cressie, N. (1993). *Statistics for Spatial Data*. New-York: Wiley.

de Gruijter, J., Brus, D., Bierkens, M. & Knotters, M. (2006). *Sampling for Natural Resource Monitoring*. Berlin: Springer-Verlag.

French, J.L, Kammann, E.E. & Wand, M.P. (2001). Comment on paper by Ke and Wang. *J. Amer. Statist. Assoc.*, **96**, 1285–1288.

Gregoire, T.G. & Valentine, H.T. (2008). *Sampling Strategies for Natural Resources and the Environment*. London: Chapman & Hall.

Huang, H.C. & Chen, C.S. (2007). Optimal geostatistical model selection. *J. Amer. Statist. Assoc.*, **102**, 1009–2024.

Hutchinson, M.F. & Gessler F.R. (1994). Splines: More than just a smooth interpolater. *Geoderma*, **62**, 45–67.

Iachan, R. (1985). Plane sampling. *Statist. Probab. Lett.*, **3**, 151–159.

Johnson, M.E., Moore, L.M. & Ylvisaker, D. (1990). Minimax and maximin distance designs. *J. Statist. Plann. Inference*, **26**, 131–148.

Laslett, G.M. (1994). Kriging and splines: An empirical comparison of their predictive performance in some applications. *J. Amer. Statist. Assoc.*, **89**, 391–405.

Nychka, D., Haaland, P., O'Connel, M. & Ellner, S. (1998). FUNFITS. Data analysis and statistical tools for estimating functions. In *Case Study in Environmental Statistics*, Eds. D. Nychka, W.W. Piegorsch & L.H. Cox, pp. 159–179. New York: Springer-Verlag.

Nychka, D. & Saltzman, N. (1998). Design of air quality monitoring networks. In *Case Study in Environmental Statistics*, Eds. D. Nychka, W.W. Piegorsch & L.H. Cox, pp. 51–76. New York: Springer-Verlag.

Overton, W.S. & Stehman, S.V. (1993). Properties of designs for sampling continuous spatial resources from a triangular grid. *Comm. Statist., Part A—Theory Methods*, **22**, 2641–2660.

Quenouille, M.H. (1949). Problems in plane sampling. *Ann. Math. Statist*, **20**, 356–375.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.
Särndal, C.E., Svensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
Stevens, D.L., Jr. (1994). Implementation of a national monitoring program. *J. Environ. Manag.*, **42**, 1–29.
Stevens, D.L., Jr. (1997). Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics*, **8**, 167–195.
Stevens, D.L., Jr. & Olsen A.R. (2004). Spatially balanced sampling of natural resources. *J. Amer. Statist. Assoc.*, **99**, 262–278.
Thompson, S.K. (2002). *Sampling*. New York: Wiley.

Résumé

Cet article traite de l'estimation de la moyenne d'une population spatiale. Dans le cadre d'une approche fondée sur un plan d'échantillonnage, un estimateur assisté par un modle de régression spline pénalisé est proposé et étudié. Nous montrons que cet estimateur est convergent (dans le cadre du plan) et établissons sa loi normale asymptotique. Une étude de simulation est menée afin d'étudier ses performances et l'estimation de sa variance, ainsi que les questions liées au biais relatif, à l'efficacité, et au taux de convergence des probabilités de couverture des intervalles de confiance correspondants. Ces simulations indiquent des gains d'efficacité considérables par rapport aux estimateurs découlant des méthodes d'échantillonnage classiques.

Appendix I—Proof of formula (7)

First note that in

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \left[\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix} + \lambda \mathbf{D} \right]^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

each element of the matrices on the right-hand side, apart from $\lambda\mathbf{D}$, is a population total of a cross product. Then, under *Assumption* I of Section 2.1, the element by element Horvitz-Thompson estimator is design-consistent for the corresponding element. This is accomplished in expression (7), which, therefore, is design-consistent for $[\tilde{\boldsymbol{\beta}}', \tilde{\mathbf{u}}']'$.

Appendix II—Proof of Theorem 1

Write estimator $\hat{\bar{Y}}_{spl}$ in the form

$$\begin{aligned} \hat{\bar{Y}}_{spl} &= \frac{1}{N}\sum_{i\in s}\frac{y(\mathbf{x}_i)}{\pi(\mathbf{x}_i)} + \frac{1}{N}[\mathbf{X}'\mathbf{1}_N - \mathbf{X}_s'\boldsymbol{\Pi}_s\mathbf{1}_n]'\hat{\tilde{\boldsymbol{\beta}}} + \frac{1}{N}[\mathbf{Z}'\mathbf{1}_N - \mathbf{Z}_s'\boldsymbol{\Pi}_s\mathbf{1}_n]'\hat{\tilde{\mathbf{u}}} \\ &= \frac{1}{N}\sum_{i\in s}\frac{y(\mathbf{x}_i)}{\pi(\mathbf{x}_i)} + \frac{1}{N}[\mathbf{X}'\mathbf{1}_N - \mathbf{X}_s'\boldsymbol{\Pi}_s\mathbf{1}_n]'\tilde{\boldsymbol{\beta}} + \frac{1}{N}[\mathbf{Z}'\mathbf{1}_N - \mathbf{Z}_s'\boldsymbol{\Pi}_s\mathbf{1}_n]'\tilde{\mathbf{u}} \\ &\quad + \frac{1}{N}[\mathbf{X}'\mathbf{1}_N - \mathbf{X}_s'\boldsymbol{\Pi}_s\mathbf{1}_n]'(\hat{\tilde{\boldsymbol{\beta}}} - \tilde{\boldsymbol{\beta}}) + \frac{1}{N}[\mathbf{Z}'\mathbf{1}_N - \mathbf{Z}_s'\boldsymbol{\Pi}_s\mathbf{1}_n]'(\hat{\tilde{\mathbf{u}}} - \tilde{\mathbf{u}}), \end{aligned}$$

where $\mathbf{1}_N$ and $\mathbf{1}_n$ are the vectors of ones of dimension N and n, respectively. By Assumption III, we have

$$\hat{\bar{Y}}_{spl} = \frac{1}{N}\sum_{i\in s}\frac{y(\mathbf{x}_i)}{\pi(\mathbf{x}_i)} + \frac{1}{N}[\mathbf{X}'\mathbf{1}_N - \mathbf{X}_s'\boldsymbol{\Pi}_s\mathbf{1}_n]'\tilde{\boldsymbol{\beta}} + \frac{1}{N}[\mathbf{Z}'\mathbf{1}_N - \mathbf{Z}_s'\boldsymbol{\Pi}_s\mathbf{1}_n]'\tilde{\mathbf{u}} + o_p(n^{-1/2}).$$

By Assumption I, it follows that

$$(\hat{\overline{Y}}_{spl} - \overline{Y})/\mathrm{V}_p(\hat{\overline{Y}}_{spl})^{1/2} \to \mathrm{N}(0, 1).$$

Furthermore, by Assumptions II and III, $\hat{\mathrm{V}}_p(\hat{\overline{Y}}_{spl}) = \mathrm{V}_p(\hat{\overline{Y}}_{spl}) + o_p(n^{-1})$ and the result follows.

Appendix III–Proof of formula (14)

The minimization of the objective function (13) gives rise to the following system:

$$\begin{cases}
\beta_0 \int_A d\mathbf{x} + \beta_1 \int_A x_1 d\mathbf{x} + \beta_2 \int_A x_2 d\mathbf{x} + u_1 \int_A z_1(\mathbf{x}) d\mathbf{x} + \cdots + u_K \int_A z_K(\mathbf{x}) d\mathbf{x} = \int_A y(\mathbf{x}) d\mathbf{x}, \\
\beta_0 \int_A x_1 d\mathbf{x} + \beta_1 \int_A x_1^2 d\mathbf{x} + \beta_2 \int_A x_1 x_2 d\mathbf{x} + u_1 \int_A x_1 z_1(\mathbf{x}) d\mathbf{x} + \cdots + u_K \int_A x_1 z_K(\mathbf{x}) d\mathbf{x} \\
\quad = \int_A x_1 y(\mathbf{x}) d\mathbf{x}, \\
\beta_0 \int_A x_2 d\mathbf{x} + \beta_1 \int_A x_1 x_2 d\mathbf{x} + \beta_2 \int_A x_2^2 d\mathbf{x} + u_1 \int_A x_2 z_1(\mathbf{x}) d\mathbf{x} + \cdots + u_K \int_A x_2 z_K(\mathbf{x}) d\mathbf{x} \\
\quad = \int_A x_2 y(\mathbf{x}) d\mathbf{x}, \\
\beta_0 \int_A z_k(\mathbf{x}) d\mathbf{x} + \beta_1 \int_A x_1 z_k(\mathbf{x}) d\mathbf{x} + \beta_2 \int_A x_2 z_k(\mathbf{x}) d\mathbf{x} + u_1 \int_A z_1(\mathbf{x}) z_k(\mathbf{x}) d\mathbf{x} \\
\qquad + \cdots + u_K \int_A z_k(\mathbf{x}) z_K(\mathbf{x}) d\mathbf{x} + \lambda u_k \int_A d\mathbf{x} = \int_A z_k(\mathbf{x}) y(\mathbf{x}) d\mathbf{x}, \quad k = 1, \ldots, K.
\end{cases}$$

A consistent design-based estimator of $[\tilde{\boldsymbol{\beta}}', \tilde{\mathbf{u}}']'$ can be obtained by substituting, in the above system, to each population total the corresponding Horvitz-Thompson estimator. The solution of the resulting system gives formula (14).

[*Received April 2011, accepted October 2011*]