METMA V International Workshop on Spatio-Temporal Modelling
Santiago de Compostela
30 Jun - 2 Jul 2010

Design-based and model-based inference in soil sciences: a way of reconciliation

G. Cicchitelli[1] and G.E. Montanari [2,*]

[1] *Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via A. Pascoli, 06100 Perugia, Italy; giuseppe.cicchitelli@stat.unipg.it*
[2] *Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via A. Pascoli, 06100 Perugia, Italy; giorgio@stat.unipg.it*
[*] *Corresponding author*

Abstract. *This paper deals with the estimation of the mean in soil surveys. After a brief recall of the two main different approaches based on the randomization distribution, on one side, and on the regionalised variable theory, on the other side, we propose an estimator assisted by a spline smoother of the population data under a design based perpective. The properties of the estimator are stated and some results from an ongoing simulation study carried out to investigate its performance in terms of relative bias and efficiency are quoted. The proposed estimator can represent a way of compromise between design-based and model-based paradigms taking advantages from both approaches.*

Keywords. *Spatial mean; Spline regression model; Horvitz-Thomspon estimator; Model-assisted estimator.*

1 Introduction

Estimating the mean or the total amount of a survey variable in a given area is a common problem in soil sciences. Traditionally, this problem was faced making use of sample survey methods. In the last decades, many soil scientists have switched to the regionalized variable theory and to its main tool, the kriging technique that finds the optimal linear predictor of soil properties.

For illustrative purposes, we will focus on the case of a spatially discrete population obtained by superimposing a very fine grid to a continuous study region. Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be the locations corresponding to the nodes of this grid, where $\mathbf{x}_i = [x_{i_1}, x_{i_2}]'$, $i = 1, \ldots, N$, is the vector of the geographical coordinates. The set of points defined in this way constitutes the finite population which will be denoted by

$U = \{1,\ldots,N\}$. Let $y(\mathbf{x}_1),\ldots,y(\mathbf{x}_N)$ be the values taken by survey variable $y(\mathbf{x})$ in U. Therefore, the population mean (block average) is expressed by

$$\bar{Y} = \frac{1}{N}\sum_{i\in U} y(\mathbf{x}_i).$$

Given the vector $\mathbf{y}_s = [y(\mathbf{x}_1),\ldots,y(\mathbf{x}_n)]'$ of n sample observations at locations $\mathbf{x}_1,\ldots,\mathbf{x}_n$, inference on $\bar{Y}$ can be conducted either according to the classic sampling theory or under the regionalized variable theory. In the sequel, we will use the expressions “design-based approach” and “model-based approach” to refer to the former and the latter paradigm, respectively.

In the design-based approach, $y(\mathbf{x}_1),\ldots,y(\mathbf{x}_N)$ are considered as fixed quantities and the randomness arises from the chance mechanism used for selecting the sample locations $\mathbf{x}_1,\ldots,\mathbf{x}_n$. The properties of the estimator of $\bar{Y}$ (bias and variance) are defined in terms of expectations over all possible samples which can be drawn with the sampling design. No assumptions on the population structure are needed for the validity of the inference: the precision of the estimates depends on the combination of a sampling design and an estimator, the so-called *strategy*. Besides, for a sufficiently large sample, the coverage of the confidence intervals equals approximately the desired confidence level, regardless of the structure of the population.

In the model-based approach, $y(\mathbf{x}_1),\ldots,y(\mathbf{x}_N)$ are assumed to be spatially dependent random variables whose joint distribution is described by a model, that is the stochastic mechanism which is assumed generating the data. The statistical properties of the related estimator of $\bar{Y}$ (bias and variance) are defined in terms of expectations over repeated realizations of this model. The main advantage of the model-based approach is its efficiency, if the population model is correctly specified. But, inappropriate modeling may cause biased estimates, loss of efficiency and problems in the confidence interval coverage.

For the debate on the advantages and disadvantages of the two approaches we refer to [3], [4] and to [8]. The second paper by Brus and de Gruijter presents also a large simulation study for the comparison, in repeated samples, of two strategies: on one side, the Horvitz-Thompson estimator is combined with stratified sampling, on the other side the kriging predictor is used in combination with the systematic sampling. The general conclusion is that the second strategy outperforms the first one, particularly for large samples and for local means. Similar results are obtained by [8], where the Horvitz-Thompson estimator and the kriging predictor are compared assuming simple random sampling for both strategies.

In the discussion of [4], Laslett stressed the fact that the design-based paradigm pays little heed to the information contained in sample labels. In soil surveys, the sample labels are the unit locations $\mathbf{x}_1,\ldots,\mathbf{x}_N$ and are generally employed, grossly, as basis for stratification. On the contrary, the model-based approach takes to an extreme such information through the modeling of the spatial dependency among $y(\mathbf{x}_1),\ldots,y(\mathbf{x}_N)$. One way to extract more information from the labels in the design-based framework is to use a superpopulation model according to the model-assisted approach proposed by [7]. In this perspective, [2] proposes to employ regression models assuming as independent variables the available covariates as well as the geographical coordinates of the locations.

In the present note, we go a step further making use of a penalized spline superpopulation model suitable to capture the correlation structure underlying the population data. This model is estimated on the basis of the sample observations and then the resulting fitted values are employed in a model-assisted estimator of the population mean.

2 The proposed Estimator

First of all we introduce a working model capable of describing the relationship between the survey variable $y(\mathbf{x})$ and the location $\mathbf{x}$. For this purpose, consider the quantities

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

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

where $\boldsymbol{\kappa}_1,\dots,\boldsymbol{\kappa}_K$ is a subset of the population locations $\mathbf{x}_1,\dots,\mathbf{x}_N$, called *knots*. Form the $N \times K$ matrix $\breve{\mathbf{Z}}$ and the $K \times K$ matrix $\boldsymbol{\Omega}$ having as typical elements c_{ik} and d_{kl} respectively. Define the $N \times K$ matrix $\mathbf{Z} = \breve{\mathbf{Z}}\boldsymbol{\Omega}_T^{-1/2}$ and consider the following spline regression superpopulation model having as spline basis the rows of $\mathbf{Z}$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}. \tag{1}$$

Here $\mathbf{y} = [y(\mathbf{x}_1),\dots,y(\mathbf{x}_N)]'$, $\mathbf{X}$ is an $N \times 3$ matrix having $[1, x_{i_1}, x_{i_2}]$ as i-th row, for $i = 1,\dots,N$, $\boldsymbol{\beta}$ and $\mathbf{u}$ are vectors of unknown constants, and $\boldsymbol{\epsilon} = [\varepsilon_1,\dots,\varepsilon_N]'$ is a random vector such that $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma_\varepsilon^2 \mathbf{I}_N$ ([6], p. 257) Following [1], a penalty criterion that restrict the variation of the spline coefficients to avoid data overfitting leads to the following minimum problem

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

We note that model (1) can be interpreted as a linear mixed model with

$$\mathrm{E}(\mathbf{u}) = \mathbf{0},\ \mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0},\ \mathrm{Var}(\mathbf{u}) = \sigma_u^2 \mathbf{I}_K,\ \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma_\varepsilon^2 \mathbf{I}_N.$$

and $\lambda = \sigma_\varepsilon^2/\sigma_u^2$. The solution of the minimum problem indicated above gives

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

where $\mathbf{D} = blockdiag(\mathbf{0}_{3\times 3}, \mathbf{I}_K)$. The resulting spline smoother of $\mathbf{y}$ is then given by

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

The fitted values above provide an approximation of the population values taking into account the spatial dependence.

The next step is to estimate $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ in a design based framework. Consider a random sample of locations $s \subset U$, of 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 $\mathbf{x}_i$, $i = 1,\dots,N$. A design-based consistent estimator of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ is given by

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

Here $\boldsymbol{\Pi}_s = \mathrm{diag}(1/\pi(\mathbf{x}_i))_{i\in s}$ is the sample submatrix of $\boldsymbol{\Pi} = \mathrm{diag}(1/\pi(\mathbf{x}_i))_{i\in U}$; similarly, $\mathbf{X}_s$ and $\breve{\mathbf{Z}}_s$ are the sub-matrices of $\mathbf{X}$ and $\breve{\mathbf{Z}}$ consisting of the rows for which $i \in s$. Hence, a design-based estimator of $\tilde{\mathbf{y}} = [\tilde{y}(\mathbf{x}_1),\dots,\tilde{y}(\mathbf{x}_N)]'$ is provided by

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

where $\hat{\tilde{\mathbf{y}}} = [\hat{\tilde{y}}(\mathbf{x}_1),\dots,\hat{\tilde{y}}(\mathbf{x}_N)]'$. It is clear that the number of knots and their placement, on one side, and the penalty parameter λ, on the other side, determine the performance of the fitted model. See [5] for more details on this.

Finally, using $\hat{\hat{\mathbf{y}}}$ as a predictor of $\mathbf{y}$, a model-assisted estimator of the population mean is given by

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

where $e(\mathbf{x}_i) = y(\mathbf{x}_i) - \hat{\hat{y}}(\mathbf{x}_i)$. It can be shown that $\widehat{\bar{Y}}_{bspl}$ is design consistent and has a normal limiting distribution. A design consistent estimator of the variance of $\widehat{\bar{Y}}_{bspl}$ is given by

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

where $\pi(\mathbf{x}_i, \mathbf{x}_j)$ is the joint inclusion probability of locations $\mathbf{x}_i, \mathbf{x}_j$ for $i, j \in s$ with $\pi(\mathbf{x}_i, \mathbf{x}_i) = \pi(\mathbf{x}_i)$ (the suffix p on the left-hand member of the previous equation indicates that here we are operating in the design-based framework, that is expectations are taken with respect to the sampling design).

The simulation studies carried out until now show that the proposed estimator is far more efficient than the Horvitz-Thompson estimator combined with the stratified sampling. Other empirical studies have been undertaken to compare the new estimator with the kriging predictor of the mean. A similar performance is conjectured under reasonable choices of the parameters that govern the smoothing process: in fact, there is a formal connection between the kriging methodology and splines for spatial prediction since both can be expressed as a mixed linear regression model by properly defining matrix **Z** in model (1).

The proposed model assisted approach can be of particular value for multipurpose soil surveys and when relevant and reliable prior information is not avalilable on the phenomenons to be surveyed.

References

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

[2] Brus D.J. (2000). Using regression models in design-based estimation of spatial means of soil properties. *European Journal of Soil Science* **51**: 159-172.

[3] Brus D.J., de Gruijter J.J. (1993). Design-based versus model-based estimation of spatial means-theory and application in environmental soil science. *Environmetrics* **4**: 123-152.

[4] Brus D.J., de Gruijter J.J. (1997). Random sampling or geostatistical modeling? Choosing between de-sign-based and sampling strategies for soil (with discussion). *Geoderma*, **80**, 1-44.

[5] Cressie, N.A.C. (1991). Statistics for spatial data. Wiley, New York.

[6] Ruppert D., Wand M.P., Carroll R.J. (2003). Semiparametric Regression. Cambridge University Press, Cambridge.

[7] Sărndal C.E., Swensson B., Wretman J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag: New York.

[8] Ver Hoef J. (2002). Sampling and geostatistics for spatial data. *Ecoscience*, **9**: 152-161.

[9] Wood S.N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society B*, **65**: 95-114.