# Model-assisted design-based estimation of a spatial mean

Cicchitelli, Giuseppe
*Università di Perugia, Dipartimento di Economia, Finanza e Statistica*
*Via A. Pascoli*
*06100 Perugia*
*E-mail: giuseppe.cicchitelli@stat.unipg.it*

Montanari, Giorgio Eduardo
*Università di Perugia, Dipartimento di Economia, Finanza e Statistica*
*Via A. Pascoli*
*06100 Perugia*
*E-mail: giorgioeduardo.montanari@unipg.it*

## 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 regions. We refer, for example, to mineral resources, vegetation cover, soil chemical composition, pollution concentration in soil, abundance of fish in a lake surface.

We assume that the response variable is described by an integrable function $y(\mathbf{x})$ defined for each location $\mathbf{x} = [x_1, x_2]'$, where $x_1$ and $x_2$ are the geographical coordinates, belonging to a bi-dimensional domain $A$. The population parameter we are interested in is the mean of the response variable, that is the quantity

$$\bar{Y} = \frac{1}{|A|}\int_A y(\mathbf{x})d\mathbf{x},$$

where $|A|$ denotes the area of domain $A$. Our aim is to estimate this parameter on the basis of a random sample $s = (\mathbf{x}_1,\dots,\mathbf{x}_n)$ of $n$ location points in $A$ drawn by a sampling design that assigns inclusion probability density $\pi(\mathbf{x})$ to location $\mathbf{x}$.

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 and de Gruijter, 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 consistency and an objective assessment of the uncertainty of the estimator, yielding confidence intervals with the correct coverage). On the other side a model-based approach is the best choice if we are interested in constructing a map, or in predicting the values of the response variable in small areas in the most efficient way. In this case, the efficiency is increased by postulating a model describing the spatial autocorrelation of data, which however 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 and Valentine, 2008). The model based-approach is the most used framework in geostatistics (mining, soil studies, air pollution monitoring), where modeling 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 (spatial 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. 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 and Stehman, 1993) and the generalized random tessellation stratified design (Stevens and 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 that adopt this perspective. Brus (2000) advocated the use of auxiliary variables in the form of a regression estimator within the model-assisted framework. Brus and Te Riele (2001) 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 and Marcheselli (2005) used the control-variate Monte Carlo integration method to increase the regression estimator accuracy.

In this paper, we present an application of semiparametric methods within the model-assisted approach to the estimation of means 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 or regression estimator.

## 2. The spline regression model assisted estimator

Following Ruppert *et al.* (2003, Chapter 13), we begin with the choice of $K$ knots, $\boldsymbol{\kappa}_1,\ldots,\boldsymbol{\kappa}_K$, in $A$, and with the definition of $K$ pseudo-covariate values as follows

$$[z_1(\mathbf{x}),\ldots,z_K(\mathbf{x})]=[\breve{z}_1(\mathbf{x}),\ldots,\breve{z}_K(\mathbf{x})]\,\mathbf{\Omega}^{-1/2},\quad \mathbf{x}\in A.$$

Here

$$\breve{z}_k(\mathbf{x})=(\|\mathbf{x}-\boldsymbol{\kappa}_k\|)^2\log(\|\mathbf{x}-\boldsymbol{\kappa}_k\|),\ \ k=1,\ldots,K,$$

and $\mathbf{\Omega}$ is a $K\times K$ matrix having as generic element the quantity

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

where $\|\cdot\|$ is the Euclidean norm.

Assume for $y(\mathbf{x})$ the following working model

$$\begin{cases}\mathrm{E}_\xi[y(\mathbf{x})]=\beta_0+\beta_1x_1+\beta_2x_2+u_1z_1(\mathbf{x})+\ldots+u_Kz_K(\mathbf{x}), & \mathbf{x}\in A\\ \mathrm{V}_\xi[y(\mathbf{x})]=\sigma^2,\quad \mathbf{x}\in A, & \end{cases} \tag{1}$$

where the regression coefficients are fixed, called spline regression model.

Fitting model (1) to the surface $y(\mathbf{x})$ by means of the penalized least-square method requires the minimization of the function

$$\int_A[y(\mathbf{x})-\beta_0-\beta_1x_1-\beta_2x_2-u_1z_1(\mathbf{x})-\ldots-u_Kz_K(\mathbf{x})]^2d\mathbf{x}+\lambda\sum_{k=1}^{K}u_k^2, \tag{2}$$

where $\lambda$ is the penalty factor. It can be shown that the design-based estimator of the parameter vector $[\tilde{\boldsymbol{\beta}}',\ \tilde{\mathbf{u}}']'$ which minimizes function (2) is given by the following formula (see Cicchitelli and Montanari, 2011)

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

where $\mathbf{X}_s$ is an $n \times 3$ matrix having as *i*-th row $[1, x_{i1}, x_{i2}]$, $i = 1, \dots, n$; $\mathbf{Z}_s$ is a $n \times K$ matrix whose *i*-th row is given by $[z_1(\mathbf{x}_i), \dots, z_K(\mathbf{x}_i)]$; $\mathbf{\Pi}_s = \text{diag}(1/\pi(\mathbf{x}_1), \dots, 1/\pi(\mathbf{x}_n))$ is the diagonal matrix having as elements the inclusion probability densities of sample locations $\mathbf{x}_1, \dots, \mathbf{x}_n$; $\mathbf{D} = \text{blockdiag } [\mathbf{0}_{3\times3}, \mathbf{I}_K]$. We notice that the trace of the projection matrix to achieve the fitted values gives the number of degrees of freedom, *r*, of the spline regression model, which in turn depends on the penalty factor $\lambda$.

Now, for each location $\mathbf{x} \in A$, we can predict the response values $y(\mathbf{x})$ by the fitted model

$$\hat{\hat{y}}(\mathbf{x}) = \hat{\hat{\beta}}_0 + \hat{\hat{\beta}}_1 x_1 + \hat{\hat{\beta}}_2 x_2 + \hat{\hat{u}}_1 z_1(\mathbf{x}) + \dots + \hat{\hat{u}}_K z_K(\mathbf{x}), \quad \mathbf{x} \in A.$$

Then, following Särndal (1992), a model-assisted design-based estimator of the population mean is given by

$$\hat{\hat{Y}}_{spl} = \frac{1}{|A|} \int_A \hat{\hat{y}}(\mathbf{x})\, d\mathbf{x} + \frac{1}{|A|} \sum_{i=1}^{n} \frac{e(\mathbf{x}_i)}{\pi(\mathbf{x}_i)}, \tag{3}$$

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

An estimator of its variance is given by

$$\hat{\text{V}}_p(\hat{\hat{Y}}_{spl}) = \frac{1}{|A|^2} \left[ \sum_{i=1}^{n} \sum_{j>i}^{n} \left( \frac{\pi(\mathbf{x}_i)\pi(\mathbf{x}_j)}{\pi(\mathbf{x}_i, \mathbf{x}_j)} - 1 \right) \left( \frac{e(\mathbf{x}_i)}{\pi(\mathbf{x}_i)} - \frac{e(\mathbf{x}_j)}{\pi(\mathbf{x}_j)} \right)^2 \right]$$

(the suffix *p* on the left-hand side of above equation indicates that we are operating in the design-based framework, i.e. that the expectation is taken with respect to the sampling design), where $\pi(\mathbf{x}_i, \mathbf{x}_j)$ is the second order inclusion density probability function.

It can be shown that the proposed estimator is approximately design unbiased and $\sqrt{n}$ - consistent.

## 3. Comparison with the kriging predictor

Now, it is of interest to compare our estimator to the kriging predictor under a design-based perspective. A natural objection against this exercise is that the optimality properties of the kriging and its variance hold within the model-based context. Nevertheless, studying the behavior of the predictor in repeated sampling from a fixed population may be useful to verify its suitability to be assumed as an estimator within the design-based framework. McArthur (1987) compared kriging and design-based methods on simulated spatial data, concluding that kriging is biased. Brus and de Gruijter (1997) presented a simulation study where a comparison was made between the Horvitz-Thompson estimator combined with the stratified random sampling and the kriging predictor combined with the systematic sampling. Their overall conclusion was that the kriging predictor is more efficient than Horvitz-Thompson estimator for large sample size, but it often presents poor confidence interval coverage rates due to the fact that kriging variance is not a good estimator of sampling variance of the kriging predictor. Ver Hoef (2002) compared the kriging predictor to the sample mean in repeated simple random samples drawn from an artificial population. He found that the kriging predictor is more efficient than the sample mean, and gives valid confidence intervals.

We now give a technical sketch of the kriging predictor of the population mean (better known as block kriging). First of all, we need to model the spatial autocorrelation of data. A common model is to assume that, under the model, $\text{E}_\xi[y(\mathbf{x})] = \mu$ and that the autocovariance between $y(\mathbf{x}_i)$ and $y(\mathbf{x}_j)$ is a function $C(\mathbf{h})$ which depends only on the distance $\mathbf{h}$ separating $y(\mathbf{x}_i)$ and $y(\mathbf{x}_j)$ (second-order stationarity). The autocovariance is generally expressed in a parametric form by means of parsimonious models, under the assumption of isotropy. An important class of isotropic covariance functions is the Matérn family, which involves a three parameter vector $\boldsymbol{\theta} = (\sigma^2, \rho, v)$, where $\sigma^2$ is the variance, $\rho$ is the range parameter (it controls how fast correlation decay with increasing distance) and $v$ is the smoothing parameter (it controls the smoothness of the resulting interpolating surface). The autocovariance is rarely known, so it needs to be estimated from the sample data.

The block kriging is the best linear unbiased predictor of the population mean and is given by (see Cicchitelli and Montanari, 1997, Ver Hoef, 2002)

$$\hat{\hat{Y}}_{kr} = \hat{\mu} + \mathbf{c}_s' \mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{1}_s \hat{\mu}), \tag{4}$$

where $\mathbf{1} = [1, 1, \ldots, 1]'$; $\hat{\mu} = (\mathbf{1}_s' \mathbf{V}_s^{-1} \mathbf{1}_s)^{-1} \mathbf{1}_s' \mathbf{V}_s^{-1} \mathbf{y}_s$ is the weighted least squares estimator of $\mu$; $\mathbf{c}_s$ is the $n$-dimensional vector whose generic entry, $c_i$, is given by

$$c_i = \frac{1}{|A|} \int_A C(\mathbf{x} - \mathbf{x}_i) d\mathbf{x};$$

$\mathbf{V}_s$ is the $n \times n$ dimensional matrix whose entries are the covariances between sample locations.

The prediction variance is given by

$$\mathrm{Var}_\xi(\hat{\hat{Y}}_{kr}) = \sigma_{A,A}^2 - \mathbf{c}_s' \mathbf{V}_s^{-1} \mathbf{c}_s + d^2 (\mathbf{1}_s' \mathbf{V}_s^{-1} \mathbf{1}_s)^{-1},$$

where

$$\sigma_{A,A}^2 = \frac{1}{|A|^2} \int_A \int_A C(\mathbf{x} - \mathbf{x}') d\mathbf{x} d\mathbf{x}' \quad \text{and} \quad d = 1 - \mathbf{1}' \mathbf{V}_s^{-1} \mathbf{c}_s.$$

## 4. Simulation study

Now we present the results of a simulation study aimed at comparing our estimator, given by equation (3), to the kriging predictor, expressed by equation (4). We considered the artificial population expressed by the following function (see Figure 1)

$$f(x_1, x_2) = 5[\sin(x_1)]^2 + 5[\cos(x_2)]^2 + 5x_1, \quad 0 < x_1 < 1, 0 < x_2 < 1.$$

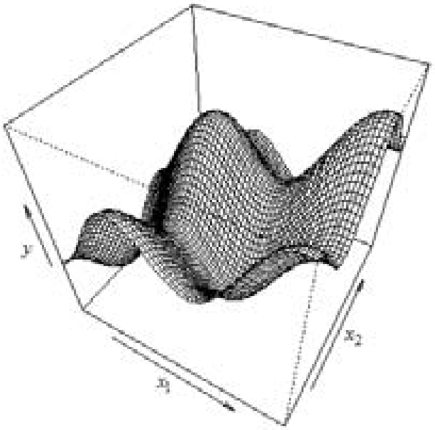

Figure 1: population surface $y(\mathbf{x})$

One thousand simple random samples were selected from the above population for sample sizes 100 and 250. We considered also the sample size $n = 500$, but, to reduce the bulk of computation time, the number of replications was reduced to 500. For each of two estimators (for sake of brevity we put $\hat{\bar{Y}}_1 = \hat{\bar{Y}}_{spl}$ and $\hat{\bar{Y}}_2 = \hat{\bar{Y}}_{kr}$), we obtained the Monte Carlo estimates of the following quantities:

- Relative bias to the mean $R = [\mathrm{E}_{MC}(\hat{\bar{Y}}_i) - \bar{Y}]/\bar{Y},\ \ i = 1, 2;$
- Relative efficiency (with respect to the sample mean, $\bar{y}$) $Eff_{MC}(\hat{\bar{Y}}_i) = \dfrac{\mathrm{mse}_{MC}(\hat{\bar{Y}}_i)}{\mathrm{mse}_{MC}(\bar{y})},\quad i = 1, 2;$
- Variance to mean squared error ratio $R_{\text{var/mse}} = \dfrac{\mathrm{E}_{MC}[\hat{\mathrm{V}}(\hat{\bar{Y}}_i)]}{\mathrm{mse}_{MC}(\hat{\bar{Y}}_i)},\ \ i = 1, 2;$
- Confidence interval coverage (percentage) for a 95% nominal level.

The knots appearing in model (1) were selected using the software due to Nychka *et al.* (1998). The penalty factor for $\hat{\bar{Y}}_{spl}$ has been chosen to achieve a predetermined number of degrees of freedom for the spline regressione model. The kriging predictor was computed assuming as autocovariance function the following isotropic exponential model

$$C(\mathbf{h}) = \begin{cases} \theta_1 + \theta_2, \mathbf{h} = \mathbf{0} \\ \theta_2 \exp(-\| \mathbf{h} \| / \theta_3),\ \ \mathbf{h} \neq \mathbf{0}, \end{cases}$$

whose parameters were estimated by means of the restricted maximum likelihood.

The main results of our simulation study are presented in Table 1 for different conbination of the number of knots, $K$, and the number of degrees of freedom, $r$.

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

| | | $\hat{\bar{Y}}_{spl}$ | | | | $\hat{\bar{Y}}_{kr}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $r$ | $K$ | $R$ | $Eff_{MC}$ | $R_{\text{var/mse}}$ | coverage % | $R$ | $Eff_{MC}$ | $R_{\text{var/mse}}$ | coverage % |
| | | | | | $n = 100$ | | | | |
| 3 | 100 | -0.0005 | 0.629 | 0.97 | 93.6 | -0.00021 | 0.042 | 1.74 | 98.5 |
| 9 | 100 | -0.0004 | 0.290 | 0.88 | 89.4 | | | | |
| 20 | 100 | -0.0003 | 0.088 | 0.64 | 87.6 | | | | |
| | | | | | $n = 250$ | | | | |
| 35 | 200 | 0.0002 | 0.025 | 0.64 | 89.1 | 0.00005 | 0.007 | 3.28 | 99.9 |
| 65 | 200 | 0.0002 | 0.006 | 0.32 | 73.8 | | | | |
| 90 | 200 | 0.0002 | 0.003 | 0.18 | 60.0 | | | | |
| | | | | | $n = 500$ | | | | |
| 35 | 200 | 0,0003 | 0.021 | 0.76 | 90.4 | 0.00002 | 0.002 | 5.00 | 100.0 |
| 65 | 200 | 0,0002 | 0.004 | 0.50 | 85.4 | | | | |
| 90 | 200 | 0,0002 | 0.002 | 0.33 | 79.0 | | | | |

The two estimators behave as approximately unbiased and are far more efficient than the sample mean. The efficiency of our estimator is close to or greater than that of the kriging when the number of degrees of freedom of the spline regression model is high. Both estimators suffer from unsatisfactory confidence interval coverage rates. For our estimator there is a general under-coverage due to under-estimation of the sampling variance, which becomes more and more serious as the number of degrees of freedom increases. In fact, when *r* is high the spline regression model tends to overfit the sample data and, as a result, to yield sample residuals smaller than those for non-sampled units. On the contrary, kriging confidence intervals suffer from over-coverage: the kriging variance overestimates the design-based true variance of the estimator.

Further research is needed, on one side, to find alternative variance estimators for our estimator which include the variance component due to the estimated regression coefficient, which is particularly important when the degrees of freedom are high with respect to the sample size; on the other side, to explore more in depth the characteristics of the kriging predictor in the design-based context.

## REFERENCES

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

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

Brus, D. and 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. and 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.

Cicchitelli, G. and Montanari, G.E. (1997) The kriging predictor for spatial finite population inference, *Metron*, **55**, 39-57.

Cicchitelli, G. and Montanari, G.E. (2011) Design-based estimation of a spatial mean, Submitted.

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

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

McArthur R.D. (1987) An evaluation of sample designs for estimating a locally concentrated pollutant, *Communications in Statistics – Simulation and Computation*, **16**, 735-759.

Nychka, D., Haaland, P., O'Connel, M. and Ellner, S. (1998) FUNFITS. Data Analysis and Statistical Tools for Estimating Functions. In D. Nychka. W.W. Piegorsch and L.H. Cox (Eds.), *Case Study in Environmental Statistics* (Lecture Notes in Statistics, **132**, 159-179), Springer-Verlag, New York.

Overton, W.S. and Stehman, S.V. 1993 Properties of designs for sampling continuous spatial resources from a triangular grid, *Communications in Statistics, Part A - Theory and Methods*, **22**, 2641-2660.

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

Särndal, C.E., Svensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Stevens, D.L., Jr. and Olsen A.R. (2004) Spatially balanced sampling of natural resources, *Journal of the American Statistical Association*, **99**, 262-278.

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