On the Design Based Inference for Continuous Spatial Populations

Cicchitelli Giuseppe - Montanari Giorgio Eduardo

Abstract This paper deals with the estimation of the mean of a continuous spatial population. Under a design-based approach to inference, an estimator assisted by a penalized spline regression model recently proposed is further investigated by means of a comparison with its counterpart in a model based approach, known as block kriging. A simulation study is carried out to compare the performance of the new estimator with that of the block kriging predictor within a design based framework to inference.

1 Introduction

Continuous 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 over a compact subset A of a planar bi-dimensional region. 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. The aim is to estimate this parameter within the design based paradigm for inference on the basis of a random sample $s = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of n location points $\mathbf{x} = (x_1, x_2)$ in A drawn by a given sampling

Cicchitelli Giuseppe, Università degli Studi di Perugia, email: *giuseppe.cicchitelli@stat.unipg.it*
Montanari Giorgio Eduardo, Università degli studi di Perugia, email: *gem@unipg.it*

design that assigns inclusion probability density $\pi(\mathbf{x})$ to location $\mathbf{x}$. A similar problem has been recently studied by Stevens and Olsen (2004), Barabesi and Franceschi (2011), and Barabesi *et al*. (2012). The above papers pursue the improvement of efficiency of the estimator of $\bar{Y}$ using at the design stage the auxiliary information provided by the spatial coordinates of location points in domain *A*, in order to select spatially balanced samples. In this paper, such auxiliary information is used at the estimation stage and the properties of a model-assisted estimator recently introduced by Cicchitelli and Montanari (2012) are further investigated. This estimator is based on a low-rank spline regression model capable of capturing the spatial correlation pattern in the data and it is proved to be design consistent and approximately unbiased. The performance of the estimator in now evaluated assuming as benchmark the kriging predictor, which is largely used in geostatistics in a model-based approach to inference. Here the latter predictor is assumed as an estimator of $\bar{Y}$ within the design-based approach framework, since many simulation studies appeared in the literature show that beside its high efficiency it is also approximately design unbiased (see, for example, McArthur, 1987, Brus and de Gruijter, 1997, Ver Hoef, 2002).

2 The model-assisted estimator

A common model for the spatial autocorrelation of data assumes that the covariance between $y(\mathbf{x})$ and $y(\mathbf{x}')$ is a function, $C(\mathbf{h})$, which depends only on the distance $\mathbf{h}$ separating $\mathbf{x}$ and $\mathbf{x}'$ (second-order stationarity). The covariance 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.

Let $\boldsymbol{\kappa}_1,\ldots,\boldsymbol{\kappa}_K$ be K knots conveniently chosen in *A*. Define K pseudo-covariates as follows

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

where $\breve{z}_k(\mathbf{x})=C(\|\mathbf{x}-\boldsymbol{\kappa}_k\|)$, $k=1,\ldots,K$, and $\boldsymbol{\Omega}$ is a $K\times K$ matrix whose generic entry is $C(\|\boldsymbol{\kappa}_k-\boldsymbol{\kappa}_l\|)$, $k, l = 1, 2, \ldots, K$. Assume for $y(\mathbf{x})$ the following working model

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

called spline regression model (the suffix ξ denotes expectation with respect to the model).

Fitting model (1) to the surface $y(\mathbf{x})$ by means of the penalized least-square method, we obtain the following design-based estimator of the parameter vector $[\tilde{\boldsymbol{\beta}}',\tilde{\mathbf{u}}']'$

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

where $\mathbf{X}_s$ is an $n\times 3$ matrix having as *i*-th row $[1, x_{i1}, x_{i2}]$, $i=1,\ldots,n$; $\mathbf{Z}_s$ is a $n\times K$ matrix whose *i*-th entry is $[z_1(\mathbf{x}_i),\ldots,z_K(\mathbf{x}_i)]$; $\boldsymbol{\Pi}_s=\mathrm{diag}(1/\pi(\mathbf{x}_1),\ldots,1/\pi(\mathbf{x}_n))$ is the

diagonal matrix having in its diagonal the inclusion probability densities of sample locations $\mathbf{x}_1,\ldots,\mathbf{x}_n$; λ is the penalization parameter.

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

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

Then, a model-assisted estimator of the population mean is given by

$$\hat{\bar{Y}}_{spl} = \frac{1}{|A|}\int_A \hat{\tilde{y}}(\mathbf{x})\, d\mathbf{x} + \frac{1}{|A|}\sum_{i=1}^{n} \frac{e(\mathbf{x}_i)}{\pi(\mathbf{x}_i)}, \tag{2}$$

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

An estimator of the variance of the proposed estimator is given by

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

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

3 The kriging predictor

We now give a technical sketch of the best linear unbiased predictor of the population mean, better known as block kriging. Assuming that $\mathrm{E}_\xi[y(\mathbf{x})] = \mu$ and that the covariance between $y(\mathbf{x})$ and $y(\mathbf{x}')$, $C(\mathbf{h})$, depends only on the distance $\mathbf{h}$ separating $\mathbf{x}$ and $\mathbf{x}'$, it can be shown that the block kriging predictor of the population mean is given by

$$\hat{\bar{Y}}_K = \hat{\mu} + \mathbf{c}_s' \mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{1}_s \hat{\mu}), \tag{3}$$

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 μ; $\mathbf{c}_s$ 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

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

where

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

4 Efficiency comparisons

Estimator (2) is compared to the kriging predictor (3) under a design-based approach by means of a simulation study using artificial populations and samples of different sizes drawn using the Uniform Random Sampling scheme. Since a natural objection against this exercise is that the kriging variance given above holds within the model-based context, the difficulty can be overcome using resampling techniques to insure the design-consistent estimation of the variance. The same techniques can be used also to estimate the variance of estimator (2). In fact, the variance estimator $\hat{V}_p(\bar{Y}_{spl})$ given above tends to underestimate the true variance, especially when the number of degrees of freedom of the spline regression model is high.

Monte Carlo estimation of the following quantities are provided: a) Bias relative to the population mean for both estimators; b) Relative efficiency of $\bar{Y}_{spl}$ with respect to the $\bar{Y}_K$; 95% confidence interval coverage for both estimators.

The main results show that the spline regression model assisted estimator, developed within the design based framework, can reach the efficiency of the kriging predictor, whose design based properties are intractable.

References

1 Barabesi L. and Franceschi S. (2011) Sampling properties of spatial total estimators under tessellation stratified designs, *Environmetrics*, **22**, 271-278.

2 Barabesi L., Franceschi S. and Marcheselli M. (2012) Properties of design-based estimation under stratified spatial sampling, *Annals of Applied Statistics*, to appear.

3 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.

4 Cicchitelli G. and Montanari G.E. (2011) Design-based estimation of a spatial mean, *International Statistical Review*, to appear.

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

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