Design-based mapping in environmental surveys

Mappe basate sul disegno nelle indagini ambientali

L. Fattorini, M. Marcheselli and C. Pisani

Abstract The estimation of the values of a survey variable throughout a continuum of points or in a finite population of spatial units is investigated when a sample of points or units is selected by a probabilistic sampling scheme. At each point or for each spatial unit, the value is estimated using an inverse distance weighting interpolator and conditions ensuring its design-based asymptotic unbiasedness and consistency are summarized.

Abstract *La stima dei valori di una variabile di interesse in un continuum di punti o in una popolazione finita di unità spaziali è considerata quando un campione di punti o unità è selezionato tramite uno schema di campionamento probabilistico. In ogni punto o per ogni unità spaziale, il valore è stimato usando un interpolatore spaziale e sono discusse le condizioni che ne assicurano la correttezza asintotica e la coerenza in un approccio basato sul disegno.*

Key words: Design consistency, Sampling, Spatial interpolation.

1 Introduction

The spatial pattern of natural resources is considerable in many environmental and ecological surveys and mapping is essential for a visual overview of the spatial

Lorenzo Fattorini
Department of Economics and Statistics, University of Siena, Piazza San Francesco 8, 53100 Siena, Italy e-mail: lorenzo.fattorini@unisi.it,

Marzia Marcheselli
Department of Economics and Statistics, University of Siena, Piazza San Francesco 8, 53100 Siena, Italy e-mail: marzia.marcheselli@unisi.it

Caterina Pisani
Department of Economics and Statistics, University of Siena, Piazza San Francesco 8, 53100 Siena, Italy e-mail: caterina.pisani@unisi.it

pattern of the variable of interest on the study region. For example, soil composition and mineral concentration are of interest in geology, soil and water pollution are crucial in ecology, species abundance and coverage are important in forestry.

Depending on the variable and the goals of the survey, the study region may be considered either a continuum of points or a finite population of spatial units, such as when the region is partitioned into a network of regular polygons or into a collection of irregular patches (e.g. Opsomer et al. 2007). Model-based estimation methods are adopted in most cases: uncertainty arises from the super-population probability model, which has been supposed to generate the surface or the population values, conditional on the sample of spatial units. By contrast, if any assumption about the super-population model generating the surface or the population values is avoided, uncertainty stems only from the sampling scheme and the surface or the population values are considered fixed.

Recently, in a design-based framework, Fattorini et al. (2018a) propose the use of the model-assisted inverse distance weighting interpolator (Bruno et al., 2013) for estimating a variable of interest when a finite population of spatial units is considered. Moreover Fattorini et al. (2018b) investigate the use of the inverse distance weighting interpolator to estimate the surface values in the continuous population setting. To render statistically sound a design-based map, conditions ensuring some sort of design-based asymptotic unbiasedness and consistency are obtained for both continuous and finite populations and asymptotically conservative estimators of the mean squared error are proposed (Fattorini et al., 2018a, 2018b).

2 Notation and setting

Consider a study region $\mathscr{A}$ of area A. Let $\mathscr{A}$ be a connected and compact set of $\mathbf{R}^2$. For $\mathbf{p},\mathbf{q}\in\mathscr{A}$, let $\|\mathbf{p}-\mathbf{q}\|$ be their distance, where $\|\cdot\|$ denotes a norm in $\mathbf{R}^2$.

If the value of the survey variable at $\mathbf{x}$, say $y(\mathbf{x})$, is defined for each point $\mathbf{x}\in\mathscr{A}$, the population is a continuum of points and the aim is the estimation of $y(\mathbf{x})$, at least ideally, for each $\mathbf{x}\in\mathscr{A}$ from a sample $(\mathbf{X}_1,\ldots,\mathbf{X}_n)$ of n points selected from $\mathscr{A}$ by means of a suitable sampling scheme (Cordy, 1993).

On the other hand, if $\mathscr{A}$ is partitioned into N spatial units $\mathbf{a}_1,\ldots,\mathbf{a}_N$ of area $a_1,...,a_N$, we are dealing with a finite population. In this case let U denote both the set of spatial units and the set of indexes $\{1,\ldots,N\}$. Moreover let y_j be the amount of the survey variable within $\mathbf{a}_j$ in such a way that $f_j=y_j/a_j$ is the density and the goal is estimating the value y_j for each $j\in U$ using a probabilistic sample S of n units selected from U. Because in most situations the a_js are known for all the population units, it is equivalent to interpolate the y_js or the f_js but density estimation is more suitable for working in an asymptotic scenario in which the a_js decrease and then the y_js may approach zero.

In both frameworks a model-assisted estimation may be performed by exploiting Tobler's law (Tobler, 1970) and adopting the inverse distance weighting interpolator (Bruno et al., 2013).

In the continuous population setting, following Fattorini et al. (2018b) the inverse distance weighting interpolator turns out to be

$$\hat{y}(\mathbf{x}) = I_{\{\cup_{i=1}^{n}(\mathbf{X}_i=\mathbf{x})\}} y(\mathbf{x}) + [1 - I_{\{\cup_{i=1}^{n}(\mathbf{X}_i=\mathbf{x})\}}] \frac{\sum_{i=1}^{n} y(\mathbf{X}_i)\phi(||\mathbf{x}-\mathbf{X}_i||)}{\sum_{i=1}^{n}\phi(||\mathbf{x}-\mathbf{X}_i||)}, \qquad (1)$$

where I_E is the indicator function of the event E and $\phi : [0,\infty) \to \mathbf{R}^+$ is a non-increasing continuous function on $(0,\infty)$, with $\phi(0) = 0$, $\lim_{d\to 0^+} \phi(d) = \infty$.

In the finite population setting, quoting Fattorini et al. (2018a) the inverse distance weighting interpolator is

$$\hat{f}_j = I_{\{j\in S\}} f_j + [1 - I_{\{j\in S\}}] \frac{\sum_{i\in S} f_i \phi(||\mathbf{c}_j - \mathbf{c}_i||)}{\sum_{l\in S}\phi(||\mathbf{c}_j - \mathbf{c}_l||)}, \qquad (2)$$

where $\mathbf{c}_j$ is the centroid of the spatial unit $\mathbf{a}_j$, $j = 1,\ldots,N$.

3 Asymptotic results

The design-based expectation and variance of (1) and (2) cannot be expressed in closed form, giving no insights about the bias and precision. Therefore, conditions providing asymptotic design-based unbiasedness and consistency are needed.

3.1 Continuous population

Suppose a sequence of fixed-size designs to select a sample of n_k points $\mathbf{X}_1^{(k)},\ldots,\mathbf{X}_{n_k}^{(k)}$, with $n_k \to \infty$ as k increases. Moreover, from (1), let $\hat{y}_k(\mathbf{x})$ be the inverse distance weighting interpolator for the k-th design. The asymptotic design-based properties of $\hat{y}_k(\mathbf{x})$ are derived as the sample size increases but the surface remains fixed.

To achieve design consistency suppose that y is a bounded measurable function on $\mathscr{A}$ and $\lim_{d\to 0^+} d^2\phi(d) = \infty$. Fattorini et al. (2018b) prove that, under mathematical conditions ensuring an asymptotic spatial balance of the sampling scheme as the sample size increases, if y is continuous at $\mathbf{x}$, $\hat{y}_k(\mathbf{x})$ is point wise design consistent at $\mathbf{x}$ while if y is continuous on $\mathscr{A}$ it is uniformly design consistent. Particularly, under uniform random sampling point wise design consistency holds while under systematic grid sampling and tessellation stratified sampling, if the study area is partitioned into a sequence of polygonal grids, $\hat{y}_k(\mathbf{x})$ is uniformly design consistent.

Moreover if y is a Lipschitz function in a neighbourhood of $\mathbf{x}$ and $\phi(d) = d^{-\alpha}$, with $\alpha > 2$, under systematic grid sampling and tessellation stratified sampling, design consistency is ensured with a $O(n_k^{(2-\alpha)/(\alpha+1)})$ convergence rate, and the use of very large α values in the distance function ϕ seems to be advisable. However, a trade-off choice of α between 3 and 5 seems suitable for moderate sample sizes, en-

suring a convergence rate of at least $O(n_k^{-1/4})$. In many real situations, the Lipschitz assumption within patches partitioning the study area is quite reasonable. Indeed, often there are parts of the study region in which the surface changes smoothly throughout space, well approaching the Lipschitz condition, while sudden variations only occur on borders, which may be realistically approximated by curves of measure 0. Therefore the surface shares the Lipschitz condition almost everywhere.

3.2 Finite population

Consider a sequence $\{U_k\}$ of partitions of $\mathscr{A}$. Each partition U_k is constituted of N_k spatial units $\mathbf{a}_1^{(k)},\ldots,\mathbf{a}_{N_k}^{(k)}$ of area $a_1^{(k)},...,a_{N_k}^{(k)}$ and centroids $\mathbf{c}_1^{(k)},\ldots,\mathbf{c}_{N_k}^{(k)}$. Denote by $y_j^{(k)}$ the amount of the survey variable within $\mathbf{a}_j^{(k)}$ and by $f_j^{(k)}$ its density. Suppose that $N_k \to \infty$, i.e. $\mathscr{A}$ is partitioned into an increasing number of spatial units: it is natural to suppose also that $\sup_{j\in U_k} \mathrm{diam}(\mathbf{a}_j^{(k)}) \to 0$. Finally suppose a sequence of designs to select a sample of size n_k from U_k. Since U_k is a sequence of finite populations of increasing sizes, each constituted by different spatial units, the sequence differs from those customarily adopted to achieve asymptotic results in finite population inference, which are constituted of nested populations with increasing totals.

Let $\hat{f}_j^{(k)}$ be the inverse distance weighting interpolator for the k-th design: the goal is to determine its asymptotic design-based behavior as the partition of the study area becomes thinner and thinner. It should be noticed that the unit j of the kth population is lost in the subsequent populations. This problem can be handled by introducing two sequences of functions from $\mathscr{A}$ onto R^+, say

$$f_k(\mathbf{x}) = \sum_{j\in U_k} f_j^{(k)} I_{\{\mathbf{x}\in\mathbf{a}_j^{(k)}\}}, \qquad \mathbf{x}\in\mathscr{A}$$

$$\hat{f}_k(\mathbf{x}) = \sum_{j\in U_k} \hat{f}_j^{(k)} I_{\{\mathbf{x}\in\mathbf{a}_j^{(k)}\}} \qquad \mathbf{x}\in\mathscr{A}.$$

Practically speaking, the N_k densities of the kth population are substituted by a piecewise constant function in $\mathscr{A}$, which equals $f_j^{(k)}$ onto $\mathbf{a}_j^{(k)}$.

Suppose that there exists a Riemann integrable function f from $\mathscr{A}$ onto $[0,L]$, which gives the density of the survey variable for any $\mathbf{x}\in\mathscr{A}$ and that the spatial units not only have diameters approaching to 0, but there is no excessively elongated unit. The assumption on the distance function already adopted in the continuous setting is also needed.

Fattorini et al. (2018a) prove that, under mathematical conditions ensuring an asymptotic spatial balance of the sampling scheme, if f is continuous at $\mathbf{x}$, $\hat{f}_k$ is point wise design consistent at $\mathbf{x}$ while, if f is continuous on $\mathscr{A}$, it turns out to be uniformly design consistent. Under simple random sampling without replacement and stratified sampling with proportional allocation point wise design consistency

holds while one per stratum stratified sampling and systematic sampling ensures uniform design consistency.

4 Mean squared error estimation

Any estimator of the mean squared error of the interpolators should not be computationally demanding. Therefore time-consuming resampling procedures, such as the bootstrap or jackknife, should be avoided.

Owing to Tobler's law, the value of the sampled point nearest to $\mathbf{x}$ is likely to be a good proxy for $y(\mathbf{x})$ and the density of the sampled unit nearest to unit j is likely to be a good proxy for f_j. Then, in the continuous setting a simple and asymptotically conservative estimator for the mean squared error of $\hat{y}(\mathbf{x})$ is

$$\hat{V}(\mathbf{x}) = \left\{\hat{y}(\mathbf{x}) - y(\mathbf{X}_{\text{near}(\mathbf{x})})\right\}^2, \tag{3}$$

where near$(\mathbf{x})$ is the index of the sample location that is nearest to $\mathbf{x}$. The finite population counterpart of (3)

$$\hat{V}_j = (\hat{f}_j - \hat{f}_{\text{near}(j)})^2,$$

where near(j) is the label of the sampled unit that is nearest to unit j, can be adopted to estimate the mean squared error of $\hat{f}_j$.

5 Simulation study

As to the continuous population setting, in Fattorini et al. (2018b) a simulation is performed on three artificial surfaces on the unit square, referred to as surf1, surf2 and surf3, and, respectively, defined at any point $\mathbf{x} = (x_1, x_2)$ as

$$y(\mathbf{x}) = C_1(\sin^2 x_1 + \cos^2 x_2 + x_1), \quad y(\mathbf{x}) = C_2(\sin 3x_1 \sin^2 3x_2)^2,$$

$$y(\mathbf{x}) = \begin{cases} C_3 x_1 x_2, & \min(x_1, x_2) < 0.5, \\ C_3(1 + x_1 x_2), & \text{otherwise} \end{cases}$$

where the constants C_1, C_2, and C_3 ensure a maximum value of 10. The first two surfaces are continuous while surf3 shows a discontinuity at the edge of the upper-right quadrant of the unit square.

For each surface, $R = 10,000$ samples of size $n = 25, 100, 225, 400$ are independently selected by uniform random sampling, tessellation stratified sampling and systematic grid sampling: the last two schemes are performed by partitioning the unit square into quadrats of equal size and then selecting one point per quadrat. Es-

timation is performed for each of the $N = 10,000$ centroids of the equally-spaced 100×100 grid by means of (1) using $\phi(d) = d^{-\alpha}$ with $\alpha = 2, 2.5, 3, 4$.

The simulation results confirm the theoretical findings: for both the continuous surfaces, a sharp decrease in the minima, maxima and averages of the absolute bias and mean squared error occurs for $\alpha > 2$ as the sample size increases while the decreases are less marked for $\alpha = 2$. For surf3, uniform design consistency is precluded and pointwise design consistency is ensured for $\alpha > 2$ for only the set of continuity points. As the sample size increases, sharp decreases occur for only the minima and averages of the absolute bias and mean squared error for $\alpha > 2$ while the maxima remain approximately constant. The decreases are less marked for $\alpha = 2$.

In the finite population setting, Fattorini et al. (2018a) adopt surf2 and surf3 to define two densities on the unit square. For any density, five spatial populations of sizes N=100, 400, 1600, 6400, and 25,600 are constructed by partitioning the unit square into grids of 10×10, 20×20, 40×40, 80×80, and 160×160 quadrats, respectively, and then taking the integrals of the density onto the quadrats as population values. For any population, $R = 10,000$ samples of size $n = N/10$ are independently selected by means of simple random sampling without replacement, one-per-stratum stratified sampling and systematic sampling. One-per-stratum stratified sampling and systematic sampling are performed by partitioning the grids into blocks of 2×5 contiguous quadrats and selecting one quadrat per block. Once the samples are selected, (2) is adopted to estimate densities for all the quadrats in the populations by using the same distance function adopted in the continuous framework. Simulation results support the theoretical results and are analogous to those obtained for continuous population.The continuity of surf2 for $\alpha > 2$ ensures point-wise design consistency under simple random sampling without replacement and uniform design consistency under sytematic and one-per-stratum stratified sampling. Uniform design consistency is precluded for surf3 and point-wise design consistency is ensured only away from the discontinuity lines for $\alpha > 2$ and for all the sampling schemes.

References

1. Bruno F., Cocchi, D., Vagheggini, A.: Finite population properties of individual predictors based on spatial pattern. Environ. Ecol. Stat. **20** 467–494 (2013)
2. Cordy, C.B.: An extension of the Horvitz–Thompson theorem to point sampling from a continuous universe. Stat. Probabil. Lett. **18**, 353–362 (1993).
3. Fattorini, L., Marcheselli, M., Pisani, C., Pratelli, L.: Design-based maps for finite populations of spatial units. J. Am. Stat. Assoc. (2018a) DOI:10.1080/ 01621459.2016.1278174
4. Fattorini, L., Marcheselli, M., Pisani, C., Pratelli, L.: Design-based maps for continuous spatial populations. Biometrika (2018b) DOI:10.1093/biomet/asy012
5. Opsomer, J. D., Breidt, F. G., Moisen, G. G., Kauermann, G.: Model-assisted estimation of forest resources with generalized additive models. J. Am. Stat. Assoc. **102**, 400–416 (2007)
6. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. Econ. Geogr. **46**, 234–240 (1970)