Individual spatial prediction under the design-based framework

F. Bruno[1], D. Cocchi[1] and A. Vagheggini[1,*]

[1] *Dipartimento di Scienze Statistiche, Università di Bologna, via delle Belle Arti 41, 40126 Bologna, Italy; alessandr.vagheggin2@unibo.it, francesca.bruno@unibo.it, daniela.cocchi@unibo.it*
**Corresponding author*

Abstract. *Spatial inference is usually carried out by means of model-based techniques, which estimate the underlying superpopulation model generating the data. However, at present, design-based methods for inference on spatial data are being rediscovered, even if the related techniques are mostly used for estimating synthetic population values, i.e. means and totals. The aim of this work is to develop a class of design-based individual spatial predictors able to exploit the spatial information available before sampling. Such predictors are able to replicate the observed values when the spatial location is sampled and otherwise predict unobserved values through weighted sums as is usual in spatial interpolation. The weights are constructed in order to assign higher influence to the observations close to the location to predict and fade away as the spatial lag increases. Moreover, as is customary, they are built in order to sum to one in the sample, needing a standardization that induces ratios of random variables. Therefore, their statistical properties can be assessed only in approximate way. Then, among all possible, an individual design-based predictor is compared with the kriging predictor through a Monte Carlo simulation showing that, especially at small sampling dimensions, its properties are quite similar to the kriging's.*

Keywords. *Spatial statistics; Design-based inference; Individual prediction; Monte Carlo simulation.*

1 Introduction

Spatial individual prediction is usually carried out under the model-based framework. Under this approach, the kriging is widely used due to its well known properties and is, therefore, considered as the benchmark. However, it is known that the kriging offers poor performances if the sample dimension is small; indeed, the estimates of the semivariogram model parameters may not be efficient when just few couples of locations are available at each lag for estimation. This may lead to the well known model misspecification problem. In contrast, one may want to adopt deterministic techniques; these prevent from the model misspecification problem, but they have not any uncertainty measure attached.

We aim at building a family of design-based individual predictors for spatial data. The starting point is to use a deterministic interpolator able to exploit the spatial relationship by means of functions of the Euclidean distances. Moreover, the known locations are seen as the realization of a probabilistic sampling design. In this work we propose a more generel version than the one presented by [1], being, at the same time, simpler in its formulation. The new theoretical framework allows to encompass different functions of the Euclidean distances between sites in the domain as well as any without replacement probabilistic sampling design.

2 Spatial interpolation in the design-based framework

A deterministic spatial interpolator assigns the observed values to the known locations belonging to the set L, while for the other sites in the domain computes a weighted sum of observed values, where the weights depend on some function of the Euclidean distances between locations

$$\widehat{z}(\mathbf{u}) = \begin{cases} \sum\limits_{i\in L} w_i z(\mathbf{u}_i), & \text{if } \mathbf{u} \notin L; \\ z(\mathbf{u}), & \text{otherwise.} \end{cases}$$

The class of interpolator we chose are a generalization of the one proposed by [4], where the weights are a standardized decreasing function of the Euclidean distances.

In order to fully contextualize this class of deterministic spatial interpolators in the framework of the design-based inference, we need to make some assumptions on the population. First, the population dimension needs to be considered as finite and let us indicate it with N. Second, the observed values are the outcome of a fixed, yet unknown function $z(\cdot)$ of the spatial coordinates $\mathbf{u} = (u_x, u_y)$. We can now consider the set L as the realization of a probabilistic sampling design, and we will indicate it with s as customary in the literature of finite populations. Then, the Bernoulli random variables $Q_i = I_{(i\in S)}$ uniquely manage inclusion in the sample, and their complements to 1 manage exclusion from the sample since the they are mutually exclusive events. The N Bernoulli random variables Q_i can be collected in the N-dimensional vector $\mathbf{Q}$, whose realization $\mathbf{q}$ has n unit values in correspondence of the sampled locations and null values elsewhere. Furthermore, let us define the $N \times N$ symmetric matrix $\mathbf{\Phi}$ having the decreasing function of the Euclidean distances in the off-diagonal elements and null diagonal (if we choose the inverse squared distance we obtain the interpolator proposed by [4]).

The individual design-based spatial predictor we propose is defined as

$$\hat{z}(\mathbf{u}_i) = (\mathbf{h}_i^\top \mathbf{1}_N)^{-1}\mathbf{h}_i^\top \mathbf{z}. \tag{1}$$

where the weighting vector $\mathbf{h}_i$ is defined as

$$\mathbf{h}_i = Q_i\mathbf{e}_i + (1 - Q_i)\mathbf{Q} \circ \phi_i,$$

where ϕ_i is the ith column of matrix $\mathbf{\Phi}$. The resulting predictor is able to replicate the observed value at the sampled locations, whereas compute a weighted sum for the unsampled ones. The resulting spatial individual design-based predictor is a ratio of random quantities.

3 Statistical properties

Given that predictor (1) is a ratio of random quantities, its statistical properties can be obtained only in approximate form.

The first-order Taylor expansion approximated expectation of predictor (1) is

$$\mathrm{E}[\hat{z}(\mathbf{u}_i)] = \frac{(\pi_i\mathbf{e}_i + (\boldsymbol{\pi} - \tilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{z}}{(\pi_i\mathbf{e}_i + (\boldsymbol{\pi} - \tilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{1}_N} + O_p(n^{-1}),$$

where we define vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^\top$ collecting the first-order inclusion probabilities and vector $\tilde{\boldsymbol{\pi}}_i = (\pi_{1i}, \dots, \pi_{(i-1)i}, \pi_i, \pi_{(i+1)i}, \dots, \pi_{Ni})^\top$ involving first- and second-order inclusion probabilities in the sample.

The first-order Taylor expansion approximated variance of predictor (1) is

$$\mathrm{V}[\hat{z}(\mathbf{u}_i)] = \mathbf{k}_i^\top \mathrm{E}[\mathbf{h}_i\mathbf{h}_i^\top]\mathbf{k}_i + O_p(n^{-2})$$

where we define vector

$$\mathbf{k}_i = \frac{(\pi_i\mathbf{e}_i + (\boldsymbol{\pi} - \tilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{1}_N\, \mathbf{z} - (\pi_i\mathbf{e}_i + (\boldsymbol{\pi} - \tilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{z}\, \mathbf{1}_N}{(\pi_i\mathbf{e}_i + (\boldsymbol{\pi} - \tilde{\boldsymbol{\pi}}_i) \circ \boldsymbol{\phi}_i)^\top \mathbf{1}_N)^2}$$

and the expectation of quantity $\mathbf{h}_i\mathbf{h}_j^\top$

$$\mathrm{E}[\mathbf{h}_i\mathbf{h}_j^\top] = \pi_i\mathbf{e}_i\mathbf{e}_i^\top + (\tilde{\boldsymbol{\Pi}} - \check{\boldsymbol{\Pi}}_i) \circ \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top$$

that involves matrix $\tilde{\boldsymbol{\Pi}}$ collecting column vectors $\tilde{\boldsymbol{\pi}}_i$ and matrix $\check{\boldsymbol{\Pi}}_i = \mathrm{E}[Q_i\mathbf{Q}\mathbf{Q}^\top]$ involving inclusion probabilities up to the third-order.

The IDW individual predictor is finite population consistent [3]

$$\lim_{n \to N} \hat{z}(\mathbf{u}_i) = z(\mathbf{u}_i).$$

4 Assessment of predictor performances

In order to evaluate the performances of the IDW individual predictor, a Monte Carlo experiment has been performed. We generate different spatial populations over a square domain with a superimposed regular grid of dimension 40 × 40 point units. These populations are the realizations of a Gaussian random field with expectation $\mu = 2$ and exponential semivariogram model with sill $\sigma^2 = 4$, nugget $\tau^2 = 0$ and different range parameters $\phi = 6, 15, 45$, respectively. Population A and C represents two extreme situations: the former present a very small range parameter and is very similar to a white noise spatial setting, whereas the latter assumes a very large range parameter leading to a spatial influence along the entire domain. In between, we generate Population B. For each dataset the Monte Carlo experiment consists in drawing 1000 samples using SRSWoR at four different sampling fractions ($f = 0.0125$, 0.025, 0.05 and 0.10) and computing predictor (1), the kriging predictor and the SRWoR estimator in predictive form [2].

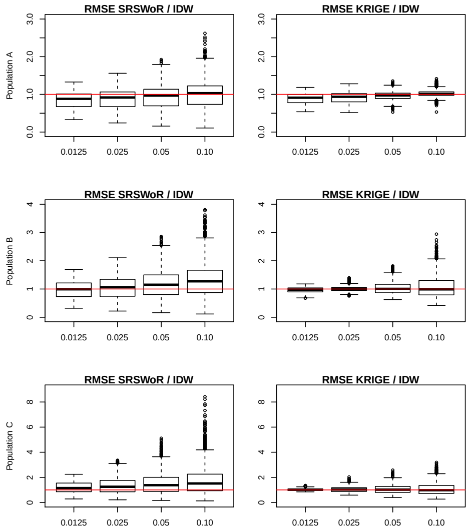

Figure 1: Boxplots of the ratio of the RMSE of the IDW predictor with SRSWoR estimator in predictive form(left-hand panels) and kriging (on the right hand panels).

Figure 1 summarizes the comparisons as ratios between the RMSE obtained via Monte Carlo experiment: for all panels the reference value is one, when the ratio is higher than one, predictor (1) outperforms the competitor. Starting from left-hand, as the spatial lag of the correlation increases, the performances of the IDW predictor increase highlighting the importance of using spatial information for estimation regardless of the sample dimension. As expected the IDW predictor's performances improve for increasing sample sizes. For Population A, the SRSWoR estimator in predictive form shows slightly better performances. The right-hand panels of Figure 1 show that the IDW and kriging predictors are almost equivalent in terms of RMSE. Indeed for almost all panels and all sample sizes these ratios are very close to 1; however, presenting many outliers at higher sampling fractions highlighting that in these cases the IDW performs slightly better than kriging.

Acknowledgments. The research work underlying this paper was funded by a FIRB 2012 grant (project no. RBFR12URQJ; title: Statistical modelling of environmental phenomena: pollution, meteorology, health and their interactions) for research projects by the Italian Ministry of Education, Universities and Research and by the agreement 7/2012 between the Department of Statistical Sciences and the Regional Agency for Prevention and Environment of Emilia-Romagna Region (ARPA ER).

References

[1] Bruno, F., Cocchi, D. and Vagheggini, A. (2013). Finite population properties of individual predictors based on spatial patterns. *Environmental and Ecological Statistics* **20**, 467–494.

[2] Bolfarine, H. and Zacks, S. (1992). *Prediction Theory for Finite Populations*. Springer-Verlag. New York.

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

[4] Shepard, D. (1968) A two-dimensional interpolation function for irregularly-spaced data. *Proceedings of the 1968 23rd Association for Computing Machinery national conference*, 517–523.