On the design-based properties of spatial interpolation[1]

Francesca Bruno, Daniela Cocchi, Alessandro Vagheggini
Dipartimento di Scienze Statistiche, Via Belle Arti 41 Bologna,
{francesca.bruno; daniela.cocchi; alessandr.vagheggin2}@unibo.it

Abstract: When spatial interpolation is carried out under a deterministic approach rather than according to the classical model-based approach known as kriging, the statistical properties of the predictor cannot be assessed. The aim of this work is to achieve these properties under a finite population design-based framework, that treats spatial locations as the outcome of a probabilistic sample.

Keywords: spatial sampling; ratio estimator, design based inference; spatial information in finite population inference.

1. Introduction

Given n locations $\mathbf{u}_1,\ldots,\mathbf{u}_n$ over a surface, let us consider a fixed but unknown deterministic function $z(\cdot)$ which generates the data $z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n)$. The inverse distance weighted interpolator (IDW, Shepard, 1968) for predicting the value in an unknown location (denoted by a Greek letter) is

$$\hat{\zeta}(\mathbf{u}_{\bar{\lambda}}) = \mathbf{z}'\mathbf{w}_{\bar{\lambda}} \quad , (1)$$

where the normalized inverse squared distances of the unknown location from all the sampled ones $w_i = \left\|\mathbf{u}_{\bar{\lambda}} - \mathbf{u}_i\right\|^{-2} \Big/ \sum_{j=1}^{n} \left\|\mathbf{u}_{\bar{\lambda}} - \mathbf{u}_j\right\|^{-2}$ are contained in the weighting vector $\mathbf{w}_{\bar{\lambda}} = (w_1,...,w_i,...,w_n)'$ and $\mathbf{z}$ is the n-dimensional vector of the observed values. The IDW properties are well known; the predictor conforms to the Tobler's law of geography. Here we propose to view this predictor under a design-based perspective.
Let us now consider the n locations as a probabilistic sample from a population of N (Barabesi, 2008): the unknown values at the unsampled locations are the object of the inference.

2. The Inverse Distance Weighted interpolator in the finite population framework

Under the design-based framework, the IDW interpolator can be seen as the result of a sampling procedure. Since each individual unobserved value depends only on its unique specific geographical relationship with the sampled locations, the simple random sampling without replacement is chosen.

[1] Work supported by the project PRIN 2008: New developments in sampling theory and practice, Project number 2008CEFF37, Sector: Economics and Statistics, awarded by the Italian Government.

The sampling design can be suitably taken into account through the use of random selection matrices (Bruno *et al*., 2011), that allow to pass from sample-based quantities to population-based ones. Expression (1) becomes

$$\hat{\zeta}(\mathbf{u}_{\bar{\lambda}}) = \mathbf{z}'\mathbf{w}_{\bar{\lambda}} = \frac{\boldsymbol{\zeta}'\mathbf{A}_{\bar{\lambda}}\boldsymbol{\Phi}\mathbf{b}_{\bar{\lambda}}}{\mathbf{1}'\mathbf{A}_{\bar{\lambda}}\boldsymbol{\Phi}\mathbf{b}_{\bar{\lambda}}} = \frac{\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}}{\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}} \quad , (2)$$

where $\boldsymbol{\Phi}$ is a $N \times N$ symmetric matrix containing the same function of the Euclidean distances $\left\|\mathbf{u}_{\bar{\lambda}} - \mathbf{u}_{\lambda}\right\|^{-2}$ of (1) before normalization with null diagonal, while $\boldsymbol{\varphi}_{\bar{\lambda}}$ is its $\bar{\lambda}$-th column vector ($\bar{\lambda} = 1,\ldots,N$). $\mathbf{A}_{\bar{\lambda}}$ is the diagonal matrix containing the random conditional indicator variables $I_{(\lambda \in s | \bar{\lambda} \notin s)}$ and $\mathbf{b}_{\bar{\lambda}}$ is the randomization of the $\bar{\lambda}$-th canonical basis vector $\mathbf{e}_{\bar{\lambda}}$ through the random indicator variable $I_{(\bar{\lambda} \notin s)}$. Using some matrix algebra and results for conditional random variables one can see that $\mathbf{K}_{\bar{\lambda}}$ is the diagonal matrix of the joint random indicator variables $I_{(\bar{\lambda} \notin s, \lambda \in s)}$. The IDW interpolator is written in (2) as a function of the *N*-dimensional vector of population values $\boldsymbol{\zeta}$ and of random indicator variables. Through the use of selection matrices, sampled and unsampled locations are associated in order to manage exclusion and conditional inclusion in the sample through random indicator variables. The resulting predictor turns out to be a design-based ratio-type estimator (Särndal *et al*., 1992).

3. Approximated first two moments of the IDW interpolator

Rewriting the IDW interpolator as in (2) allows the calculus of its statistical properties. Since it is a ratio of linear random combinations, its properties can be analytically computed only as approximations. For managing the involved random variables, we define the "association probabilities", linking a potentially unsampled location with all the others. These probabilities represent the starting point for the calculus of the statistical properties of the predictor: an uncertainty measure can, in this way, be associated to the deterministic IDW interpolator.

Theorem 1: The approximated expected value of (2) is

$$\mathrm{E}\left[\hat{\zeta}(\mathbf{u}_{\bar{\lambda}})\right] = \mathrm{E}\left[\frac{\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}}{\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}}\right] \cong \frac{\mathrm{E}\left[\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]}{\mathrm{E}\left[\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]} = \frac{\sum_{\lambda \neq \bar{\lambda}} \zeta_{\lambda}\varphi_{\lambda\bar{\lambda}}}{\sum_{\lambda \neq \bar{\lambda}} \varphi_{\lambda\bar{\lambda}}} = \frac{T_{1\bar{\lambda}}}{T_{2\bar{\lambda}}} \quad (3)$$

Proof.
It follows directly from the expected value of the random matrix $\mathbf{K}_{\bar{\lambda}}$ as

$$\mathrm{E}[\mathbf{K}_{\bar{\lambda}}] = \frac{N-n}{N}\frac{n}{N-1}\mathbf{D}_{\bar{\lambda}} \quad , (4)$$

where $\mathbf{D}_{\bar{\lambda}}$ is a diagonal matrix of unit values besides the null value at the $(\bar{\lambda},\bar{\lambda})$-th position. □

Let us define the difference between each $\bar{\lambda}$-th population value and its interpolation via the other $N-1$ values

$$\delta(\mathbf{u}_{\bar{\lambda}}) = \zeta(\mathbf{u}_{\bar{\lambda}}) - \left(\sum_{\lambda \neq \bar{\lambda}} \zeta_{\lambda} \varphi_{\lambda\bar{\lambda}} \Big/ \sum_{\lambda \neq \bar{\lambda}} \varphi_{\lambda\bar{\lambda}} \right), \quad (5)$$

as the "structural bias" associated to location $\mathbf{u}_{\bar{\lambda}}$. The bias of estimator (2), *i.e.* $\mathrm{E}\left[\hat{\zeta}(\mathbf{u}_{\bar{\lambda}})\right] - \zeta(\mathbf{u}_{\bar{\lambda}})$, is also not null. However, it can be seen that, as the sample size increases, $\hat{\zeta}(\mathbf{u}_{\bar{\lambda}})$ tends to its "true" value $T_{1\bar{\lambda}}/T_{2\bar{\lambda}}$ (3). Predictor (2) may exhibit a high "structural bias" due not to the sample size but to the nature of the interpolator.

Theorem 2. The approximated variance of (2) is

$$\mathrm{V}\left[\hat{\zeta}(\mathbf{u}_{\bar{\lambda}})\right] \approx \frac{\mathrm{V}\left[\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]}{\mathrm{E}\left[\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]^2} - 2\frac{\mathrm{Cov}\left(\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}, \mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right)}{\mathrm{E}\left[\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]^3} + \frac{\mathrm{E}\left[\boldsymbol{\zeta}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]^2 \mathrm{V}\left[\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]}{\mathrm{E}\left[\mathbf{1}'\mathbf{K}_{\bar{\lambda}}\boldsymbol{\varphi}_{\bar{\lambda}}\right]^4},$$

which, using a notation similar to (3), can be expressed as

$$\mathrm{V}\left[\hat{\zeta}(\mathbf{u}_{\bar{\lambda}})\right] \approx \frac{1}{cT_{2\bar{\lambda}}^4}\left[h\left(T_{3\bar{\lambda}}T_{6\bar{\lambda}} - 2T_{7\bar{\lambda}}T_{8\bar{\lambda}} + T_{5\bar{\lambda}}T_{1\bar{\lambda}}^2\right) + m\left(T_{4\bar{\lambda}}T_{6\bar{\lambda}} - 2T_{8\bar{\lambda}}^2 + T_{6\bar{\lambda}}T_{1\bar{\lambda}}^2\right)\right],$$

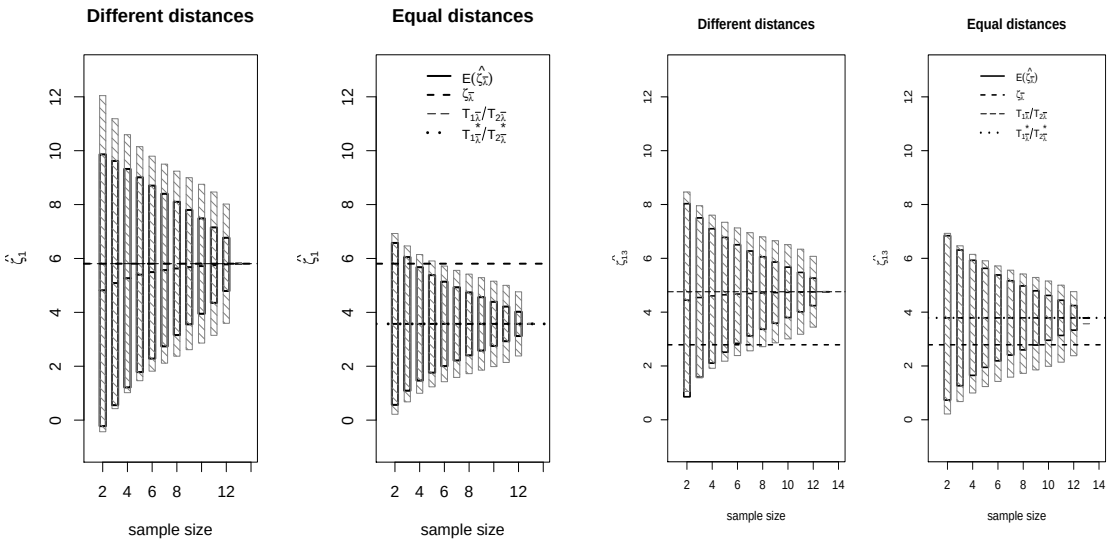
where c, h and m are population constants and quantities $T_{\square\bar{\lambda}}$ are similar to those in (3). For the proof, see Bruno *et al.* (2011). □

4. A simulation study

We assess the improvement in inference provided by the use of a weighting system based only on geographical distances. No model specification is required and the only assumption made is that data follow the Tobler's law. The weighting system we propose, suggested by the IDW interpolator (1), is the same for the whole population, but the weights change according to the location to predict. When geography is not important, it might be more useful to predict the unweighted mean of the *N*-1 population values, for the unknown location.

A simulation study has been carried out for evaluating the approximate properties of the IDW interpolator under the design-based framework. A population of fifteen sparse data points is considered. A map of the population under study, the table of the values of the variable and the "structural bias" associated to each point of the population are given in Bruno *et al.* (2011). We illustrate two opposite situations, in the four panels of Figure 1. For the first location, where $\zeta(\mathbf{u}_1) = 5.81$, the structural bias is null: the use of the distances, linked to the IDW predictor, leads to a better prediction (panel a) than the consideration of equal weights (panel b), as highlighted by the tendency of the expected value (3) to the real value. The other location, where $\zeta(\mathbf{u}_{13}) = 2.79$, presents a structural

bias $\delta(\mathbf{u}_{13}) = -1.97$ and (3) fails in properly predicting the true value (panel c). The unweighted version of the structural bias is on the contrary $\delta^*(\mathbf{u}_{13}) = -1.00$. For this point, the use of geography is misleading and a situation of unweighted estimation in simple random sampling would be preferable (panel d).



a) b) c) d)

Figure 1: Prediction with different and equal weights for two locations (as *n* increases). Location 1: panels a) and b); location 13: panels c) and d).

References

Barabesi L. (2008) Facoltà di Economia "R.M. Goodwin", Università degli Studi di Siena, mimeo.

Bruno, F., Cocchi, D. and Vagheggini, A. (2011) Spatial interpolation using a finite population approach, submitted.

Cressie N.A.C. (1993) *Statistics for spatial data*, Wiley, New York.

Shepard D. (1968) A two-dimensional interpolation function for irregularly-spaced data, *Proceedings of the 1968 23rd ACM national conference*, 517-524.

Särndal C.-E., Swensson B., Wretman J. (1992) *Model-assisted survey sampling*, Springer-Verlag, New York.

Stevens D.L. (2006) Spatial properties of design-based versus model-based approaches to environmental sampling, *American Statistical Association; Section on Statistics & the Environment Newsletter*, 10, 3-5.

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