

A Multivariate Geostatistical Technique: kriging on factors

Una Tecnica di analisi Geostatistica Multivariata: Il kriging sui fattori

Giovanna Jona Lasinio

Dip. Statistica, Probabilità e Statistiche Applicate

Università di Roma "La Sapienza"

Giovanna.jonalasinio@uniroma1.it

Silvia Loriga¹

ISTAT

silvia.loriga@istat.it

Abstract: In questo lavoro riportiamo brevemente alcuni risultati relativi a metodi statistici per l'analisi di dati spaziali e spazio-temporali multivariati. La base dell'approccio proposto è di tipo geostatistico, in quanto segue ciascuna osservazione viene rappresentata a mezzo di tre coordinate, due per lo spazio ed una per il tempo. La tecnica proposta procede sostanzialmente in due passi: il primo consiste nel proiettare il dataset multivariato in uno spazio appropriato ottenendo dei "fattori" tra loro "quasi" ortogonali (nello spazio di riferimento). In un secondo momento applichiamo uno stimatore di tipo kriging ai valori proiettati. Poiché tutte le trasformazioni coinvolte sono di tipo lineare è particolarmente semplice risalire dalle stime sui fattori alle stime riferite alle variabili originali. Si presentano molto brevemente alcuni risultati teorici, alcune simulazioni e applicazioni a dati reali.

Keywords: geostatistics, kriging, multivariate random fields, spatial data, space-time data.

1. Introduction

In this paper we briefly report some results obtained while working on multivariate statistical techniques aiming to describe and predict observations dependent in time and space. Our approach is substantially geostatistical (see for instance Cressie (1993), Wackernagel (1995) for detailed exposition of the topic, or Wackernagel *et al.* (2001) for an application of geostatistical techniques to space-time data). We analyze observations on several geographical locations repeated for several times. We deal with isotropic second order (or intrinsic) stationary multivariate random fields (MRF). In this setting we study a complex dependence structure involving the spatial arrangement of sites and the repetition of measures. Time is seen as a third dimension, then each observation is associated to a coordinate system in $\mathbb{R}^3(x, y, t)$. Following the approach to the analysis of multivariate spatial data proposed in Jona Lasinio (2001), we describe a non parametric kriging-type procedure for multivariate spatial and space-time data. The general idea is to find a representation of the multivariate (dependent) dataset such that correlation between variables is "almost" removed and then to perform standard kriging. This "de-correlation" is obtained using a Principal Component Analysis (PCA) with a specific metric (see next section for details) which allows us to keep the information on the space-time arrangement of observations. As all the transformations involved are linear, it's easy to go back to the original representation of the dataset. Our results have been compared, through simulations in

¹Il lavoro è stato svolto con finanziamento MURST - COFIN 2000 *Statistica nella valutazione del Rischio Ambientale*

the spatial setting, with the procedure proposed by Xie and Myers (1995a) (see Xie and Myers (1995b) too). An application to an agronomic dataset is reported in section 3.

2. Theoretical Results

Starting from the approach to the analysis of multivariate spatial data proposed in Jona Lasinio (2001), we describe a non parametric kriging-type procedure for multivariate spatial and space-time data. More precisely, let $\{\mathbf{Z}(s_j), s_j \in \mathcal{D}\}$ be a MRF with k components observed on the set $\mathcal{D} \subset \mathbb{R}^3$, h_{ij} be the value of the distance between sites $s_i, s_j \in \mathcal{D}$, $i, j = 1, \dots, n$ and let $\text{Var}(\mathbf{Z}(s_i) - \mathbf{Z}(s_j)) = 2\Gamma_Z(h_{ij}) = 2\Gamma_Z(h)$, $\forall s_i, s_j \in \mathcal{D}$ such that $h_{ij} = h$ be the crossvariogram matrix at lag h and let $\Gamma_Z = \sum_{h \in \mathcal{H}} \Gamma_Z(h)$ be a synthesis matrix where $\mathcal{H} = \{h_1, \dots, h_{max}\}$ is the set of lags. From the spectral decomposition of $\Gamma_Z = \Delta' \Psi \Delta$ we obtain the $n \times k$ factors' matrix $\mathbf{Y} = \mathbf{Z} \Delta'$ whose components are principal components in the Γ_Z metric of $\mathbf{Z}$. Notice that $\Gamma_Y(h) = \Delta \Gamma_Z(h) \Delta'$ then

$$\Gamma_Y = \sum_h \Gamma_Y(h) = \Delta \Gamma_Z \Delta' = \Psi$$

This implies that the k factors are on the average uncorrelated. By applying the ordinary kriging estimator to each factor and by projecting back the estimates to the original space, we are able to state the following proposition:

Proposition 1 For an isotropic second order or intrinsic stationary MRF the kriging on factors estimator is given by the following expression:

$$\hat{Z}_l(s_0) = \sum_{p=1}^k \sum_{j=1}^k \sum_{i=1}^n \lambda_i^p Z_j(s_i) \delta_{pj} \delta_{pl} \quad l = 1, \dots, k \quad s_0 \in \mathcal{D}; \quad (1)$$

where δ_l are eigenvectors of Γ_Z and λ^p are the kriging weights for each factor $p = 1, \dots, k$. The estimator in (1) is uniformly unbiased and its variance is:

$$\begin{aligned} \text{Var}(\hat{Z}_l(s_0)) &= \sum_{p=1}^k \text{Var}(\hat{Y}_p(s_0)) \delta_{pl}^2 + \\ &+ \sum_{p=1}^k \sum_{j \neq p=1}^k \delta_{pl} \delta_{jl} \sum_{i=1}^n \sum_{r=1}^n \lambda_i^p \lambda_r^j \delta_p' \cdot \mathbf{C}_Z(s_i, s_r) \delta_j. \end{aligned} \quad (2)$$

where $l = 1, \dots, k \quad s_0 \in \mathcal{D}$.

For proofs see Loriga (2001).

In Loriga (2001) further theoretical results are available, one of the most relevant being an upper bound for the estimation error in the case of second order stationary MRF.

3. Simulations and application

As a first step we analyze the performance of our technique via simulation in the pure spatial setting (no repeated observation in each site). Considering 16 different dependence structures, we simulate 500 MRF (5 components simulated on a 10×10 lattice) in each run

and for each run we compare our approach to the one proposed in Xie and Myers (1995a); there a simultaneous quasi-diagonalization (via least square) of all cross-variogram matrices is used. The simulation proceed along the lines proposed in Jona Lasinio (2001), the MRF are Gaussian then there is no need for MCMC type algorithms. The 16 dependence structures we consider combine several type of correlation structures between variables and different neighboring systems for the simulated locations.

Our method seems to give a better performance in all the considered situations except when the MRF components are uncorrelated; however in these cases both methods give "poor" performances. This is obvious when noticing that both techniques use the correlation between MRF components in order to obtain "better" predictions. Our algorithm is always computationally more efficient.

We apply our technique to a set of agrometeoreological and phenological data collected over 28 monitoring stations for one year (daily observations in 1998); the monitoring network is located in the "Fattoria di Poggio Casciano" in the Chianti region². The farm has an extension of 120 hectares and height above the sea level between 120 and 280 meters. In the farm territory 28 thermometric stations have been set in order to record the daily temperatures. Seventeen out of 28 stations lay in the grape yards. In this framework it is natural to choose as the main subject of study the growth and development of vine plants. In particular we estimate a bioclimatic index based on average daily temperature, which is a very good indicator of the phenological development of vine plants. Our study proceeds in two stages, the first one is a purely spatial analysis at each time during the blossoming period of the plants; the second one is a joint analysis (*i.e.* in space and time) of the data always during the same period of the year. The proposed technique is applied using empirical crossvariogram's matrices computed after binning the h_{ij} 's values. In the spatial case we estimate the phenological stage of vine plants in three moments of the blossoming season (May 25, May 30, June 4). We use topographical variables (height above sea level, distance from the bottom of the valley, incline of the slope) in order to estimate the bioclimatic index. When the index is greater then a given value (274 Celsius degrees) the blossoming is considered reached. In order to evaluate our technique's performances we estimate index's values on observed locations and then we compare estimated values with the observed ones. Results are satisfactory, in fact the only sites that have been poorly estimated (in the sense of square errors) are the ones (stations with labels 15, 19 and 20) which are located on the borders of the farm (they have less neighbouring sites to be used in the estimation procedure). Our technique can be successfully applied to interpolate values in order to build a continuous map from discrete observations. Results on this use of the method with topographic and bioclimatic variables we'll be shown.

In the space-time setting several approaches are possible. Our first application is to the treatment of missing data. Again we are interested in the estimation of a bioclimatic index and we use topographical variables and temperatures in the procedure. In this case we deal only with estimation inside $\mathcal{D}$. For instance, if a thermometric station is broken for a few days, we can try to fill the gap via kriging on factors. Results are encouraging, as obvious the reconstruction of missing measurements becomes worst as the number of missing days increases. When the missing data are one full day for all stations or it is of interest to interpolate data for all stations between two different times a smoothed version of kriging on factors gives encouraging results. If the aim of the study is to forecast values in space and time the kriging estimator has to be corrected (see Cressie (1993)) in order

²We thank Simone Orlandini and his group from DISAT, University of Florence, for giving us the data we used in this application

to take into account the position of to be estimated sites which are outside $\mathcal{D}$.

4. Conclusion

In the spatial framework the proposed technique gives very satisfactory results both on simulated and real data, being more efficient from both a computational and a statistical point of view than the one proposed by Xie and Myers. In the space-time setting the proposed technique can be used in several ways, from the estimation of missing data to the interpolation of space-time surfaces. In the applications we carried out best results are obtained when a smoothed version of kriging estimator is used, more precisely when estimates are averaged over neighbouring locations. Further theoretical developments may concern the asymptotic behaviour of the proposed estimator and the evaluation of finer upper bounds for the estimation error.

References

- Cressie N. (1993) *Statistics for Spatial Data*, Wiley, New York.
- Jona Lasinio G. (2001) Modeling and exploring multivariate spatial variation: A test procedure for isotropy of multivariate spatial data, *Journal of Multivariate Analysis*, 77, 295–317.
- Loriga S. (2001) *Problemi di stima per dati spaziali multivariati*, Tesi di dottorato in Statistica, Università di Bologna.
- Wackernagel H. (1995) *Multivariate Geostatistics*, Springer-Verlag, Berlin.
- Wackernagel H., Bertino L. and Lajaunie C. (2001) Geostatistical normalization- case studies, *Report No 3 of Contract IST-1999-11313, Technical Report N-31/01/G ENSMP - ARMINES, Centre de Géostatistique 35 rue Saint Honoré, F-77305 Fontainebleau, France* <http://cg.ensmp.fr>.
- Xie T. and Myers D. (1995a) Fitting matrix-valued variogram models by simultaneous diagonalization (part 1: Theory), *Mathematical Geology*, 27, 867–875.
- Xie T. and Myers D. (1995b) Fitting matrix-valued variogram models by simultaneous diagonalization (part 2: Application), *Mathematical Geology*, 27, 877–888.