SIS2016

Università degli Studi di Salerno

June 8th – June 10th, 2016

PROCEEDINGS

of the 48th scientific meeting of the
Italian Statistical Society

Editors: Monica Pratesi and Cira Pena

ISBN: 9788861970618

Object Oriented Geostatistical Simulation of Functional Compositions via Dimensionality Reduction in Bayes spaces

Simulazione geostatistica orientata agli oggetti per composizioni funzionali tramite riduzione dimensonale in Spazi di Bayes

Alessandra Menafoglio, Alberto Guadagnini and Piercesare Secchi

Abstract We address the problem of geostatistical simulation of spatial complex data, with emphasis on functional compositions (FCs). We pursue an object oriented geostatistical approach and interpret FCs as random points in a Bayes Hilbert space. This enables us to deal with data dimensionality and constraints by relying on a solid geometric basis, and to develop a simulation strategy consisting of: (i) optimal dimensionality reduction of the problem through a simplicial principal component analysis, and (ii) geostatistical simulation of random realizations of FCs via an approximate multivariate problem. We illustrate our methodology on a dataset of natural soil particle-size densities collected in an alluvial aquifer.

Abstract *Si considera il problema della simulazione geostatistica di dati complessi spazialmente distribuiti, con particolare riferimento a composizioni funzionali (FC). Si segue un approccio geostatistico orientato agli oggetti, interpretando le FC come punti aleatori in uno spazio di Hilbert Bayes. Questo consente di trattare la dimensionalità dei dati e i relativi vincoli poggiando su una solida base geometrica, e di sviluppare una strategia di simulazione in due passi: (i) riduzione dimensionale ottima attraverso un'analisi delle componenti principali funzionali simpliciali, e (ii) simulazione geostatistica di FC attraverso un problema multivariato approssimato. La metodologia proposta è illustrata attraverso la sua applicazione a un dataset di densità granulometriche osservate in un acquifero alluvionale.*

Key words: Object Oriented geostatistics, functional compositions, Bayes Hilbert spaces, uncertainty quantification

Alessandra Menafoglio; Piercesare Secchi
MOX, Department of Mathematics, Politecnico di Milano, Milano, Italy e-mail: alessandra.menafoglio@polimi.it; e-mail: piercesare.secchi@polimi.it

Alberto Guadagnini
Department of Civil and Environmental Engineering, Politecnico di Milano, Milano, Italy; Department of Hydrology and Water Resources, University of Arizona, 85721 Tucson, Arizona, USA
e-mail: alberto.guadagnini@polimi.it

1 Introduction

Environmental field studies are nowadays based on heterogeneous, complex and spatially dependent data, such as georeferenced functional data (e.g., curves or surfaces) or distributional data (e.g., probability density functions). The variety, dimensionality and complexity of these data pose new challenges for data-driven geoscience applications. Based on our recent work [5], in this communication we focus on the problem of uncertainty quantification via stochastic simulation, in the presence of complex spatial data such as functional compositions (FCs). FCs are infinite-dimensional data that provide only relative information, being constrained to be positive and integrate to a constant (e.g., probability density functions). FCs are found, e.g., in field studies relying upon particle-size densities (PSDs), as those shown in Fig. 1(a) and considered in [4, 5]. These data describe the local distribution of grain sizes for 60 soil samples collected along a borehole in a shallow aquifer near the city of Tübingen, Germany. PSDs are relevant to describe the textural properties of aquifer systems, as well as to estimate key parameters such as hydraulic conductivity. In this setting, stochastic simulation is geared at providing multiple realizations of the entire field of PSDs, consistent with available data. Following our proposal in [5], we pursue an object oriented approach (e.g., [3]) and interpret each datum as a point within the Bayes Hilbert space of [1, 6] whose elements are FCs. We here review the object oriented method for stochastic simulation we recently proposed in [5]. The method relies on (i) dimensionality reduction via simplicial functional principal component analysis (SFPCA, [2]), and (ii) geostatistical simulation of an approximate problem of reduced dimension. We demonstrate our methodology on the field data of PSDs depicted in Fig. 1(a).

2 Geostatistical simulation in Bayes spaces

Denote by $(\Omega, \mathfrak{F}, \mathbb{P})$ a probability space, by $D \subset \mathbb{R}^d$ a spatial domain, and let $\mathscr{X}_s$ be a random element in a Hilbert space $(H, +, \cdot, \langle \cdot, \cdot \rangle)$, referred to a location $s \in D$. For the dataset of PSDs in Fig. 1(a), $\mathscr{X}_s$ represents a random PSD at $s \in D$, and $D \subset \mathbb{R}$ denotes the target borehole (i.e., a 1D spatial domain). Even though H may denote a general Hilbert space, for the purpose of our application we here focus on the Bayes Hilbert space $A^2(\mathscr{T})$ (or A^2 for short) of [1], whose elements are FCs on $\mathscr{T} = [t_m, t_M] \subset \mathbb{R}$ with square-integrable logarithm. The space A^2, endowed with the perturbation $(+)$ and powering $(\cdot)$ operations (see [1, 6])

$$(f+g)(t) = \frac{f(t)g(t)}{\int_{\mathscr{T}} f(s)g(s)\,ds}, \quad (\alpha \cdot f)(t) = \frac{f(t)^\alpha}{\int_{\mathscr{T}} f(s)^\alpha\,ds},$$

and the inner product

$$\langle f,g\rangle = \frac{1}{2|\mathscr{T}|}\int_{\mathscr{T}}\int_{\mathscr{T}} \ln\frac{f(t)}{f(s)}\ln\frac{g(t)}{g(s)}\,dt\,ds, \quad f,g\in A^2(\mathscr{T}),$$

is a separable Hilbert space. We remark that $(A^2,+,\cdot,\langle\cdot,\cdot\rangle)$ is precisely designed to reflect the peculiar features of FCs, such as the properties of scale invariance and relative scale [6].

Given a set of locations $s_1,...,s_n$, and the observations $\mathscr{X}_{s_1},...,\mathscr{X}_{s_n}$ at these locations, we aim to provide random realizations of the element $\mathscr{X}_{s_0}$ at a target location $s_0\in D$, conditional to $\mathscr{X}_{s_1},...,\mathscr{X}_{s_n}$ (i.e., to perform *conditional simulation*). Note that the problem is particularly challenging, since we aim to sample from a distribution on an infinite-dimensional space of constrained objects. We assume that the $\mathscr{X}_{s_1},...,\mathscr{X}_{s_n}$ are a partial observation of a Gaussian stationary random field $\{\mathscr{X}_s, s\in D\}$ on H, with (constant) spatial mean $m=\mathbb{E}[\mathscr{X}_s]$, and cross-covariance operator C. The latter is defined, for $s_1,s_2\in D$, as

$$C(s_1-s_2)x = \mathbb{E}[\langle \mathscr{X}_{s_1}-m,x\rangle(\mathscr{X}_{s_2}-m)], \quad x\in H.$$

Given the Hilbert space structure of H and following [5], we consider for the field a truncated K-dimensional Karhunen-Loève expansion, which provides nested optimal approximations of the observations for any finite order K. Specifically, call (λ_k,e_k), $k\geq 1$, the eigenpairs of the zero-lag covariance operator $C(0)$, i.e., the covariance operator associated with each $\mathscr{X}_s$. In [5], we propose projecting the data over the first K eigenfunction $\{e_k, 1\leq k\leq K\}$, and accordingly model the data through the joint modeling of the coefficients $\{\xi_k(s), k=1,...,K, s\in D\}$, where $\xi_k(s)=\langle \mathscr{X}_s-m,e_k\rangle$ represents the projection of the centered observation at s along the eigenfunction e_k. We refer to [5] for further theoretical justification of the method and associated details.

In light of these observations, the following two step procedure can be considered for the stochastic simulation in A^2 (or, generally, in H) [5]: (i) compute the eigen-decomposition of the (empirical) zero-lag covariance, and the corresponding coefficients; (ii) provide conditional simulations of the multivariate random field of coefficients. The latter step can be performed through any of the widely employed geostatistical techniques for the simulation of multivariate random fields. We finally note that step (i) is based on a principal component analysis performed in the Bayes Hilbert space, i.e., on a simplicial principal component analysis (SFPCA, [2]). The latter identifies the directions in A^2 of maximum variability of the data. Given that A^2 is a Hilbert space, all the techniques which are useful to interpret principal components in multivariate or functional settings can be employed.

3 An application to particle-size densities at the Lauswiesen site

In this communication, we illustrate our method through its application to the field data depicted in Fig. 1(a), which represent a subset of the data considered in [5].

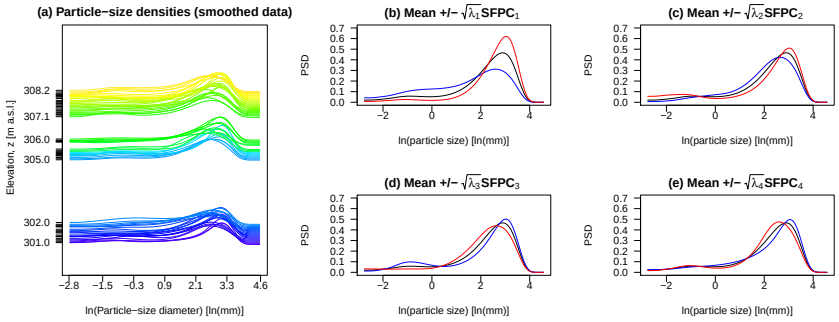

Fig. 1 Particle-size densities at the Lauswiesen site. Panel (a): Data considered in [4] (data were smoothed with a set of $m = 70$ Bernstein polynomials), represented according to the elevation of the corresponding soil sample. Panels (b) to (d): Plot of the mean (black line) $\pm$ the first SFPCs (a blue line indicates the sign $+$, a red line indicates the sign $-$), powered by the corresponding standard deviation, i.e., $\widehat{m} \pm \sqrt{\lambda}_i \cdot e_i$, $i = 1,...,4$

We estimate the zero-lag covariance operator through the empirical estimator $\widehat{S}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \langle x_{s_i} - \widehat{m}, \cdot \rangle \cdot (x_{s_i} - \widehat{m})$, with x_{s_i} the observation at $s_i \in D$ and $\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} x_{s_i}$ the sample mean. We numerically compute the eigen-decomposition of $\widehat{S}$ and retain $K = 4$ simplicial functional principal components (SFPCs), that together explain 97.4% of the data variability. In this regard, note that the proportion of the total variability explained by the k-th SFPC can be estimated through the ratio between the k-th eigenvalue and the sum of all the eigenvalues, i.e., $\lambda_k / \sum_{j=1}^{K} \lambda_j$ (as in multivariate principal component analysis). Interpretations of the SFPCs can be based on Fig. 1(b) to (e), that display the plots of the mean PSD perturbed by plus/minus the retained SFPCs scaled according to the corresponding standard deviations (i.e., $\widehat{m} \pm \sqrt{\lambda}_i \cdot e_i$, $i = 1,...,4$). In particular, the first two SFPCs provide information about the position of the mode and the modality of the PSD: high (low) scores along the first SFPC are associated with a larger (smaller) mode and mass concentration around it, whereas high (low) scores along the second SFPC are indicative of bimodal (unimodal) distributions.

Having estimated the multivariate cross-variogram structure of the scores along these SFPCs, we perform conditional Gaussian cosimulation of the fields of scores. Fig. 2(b) reports an example of conditional simulations over a fine grid along the vertical direction. This type of realizations is key in field applications to quantify the uncertainty associated with point-wise predictions, such as those provided by Kriging [4]. Kriging yields the best linear unbiased predictor of $\mathscr{X}_{s_0}$, that in the Gaussian case coincides with an estimate of the conditional expectation of $\mathscr{X}_{s_0}$ given the observations, i.e. $\mathbb{E}[\mathscr{X}_{s_0} | \mathscr{X}_{s_1}, ..., \mathscr{X}_{s_n}]$. Even though Kriging provides the best prediction (in the mean square sense), Kriging maps appear smoother than actual realizations of the field (compare Fig. 2(a) and Fig. 2(b)). In this sense, conditional simulation is relevant to reproduce the actual spatial variability of the phenomenon,

to be employed in a Monte Carlo setting for the characterization of the spatial distribution of aquifer properties (see Fig. 2(c) and (d)).

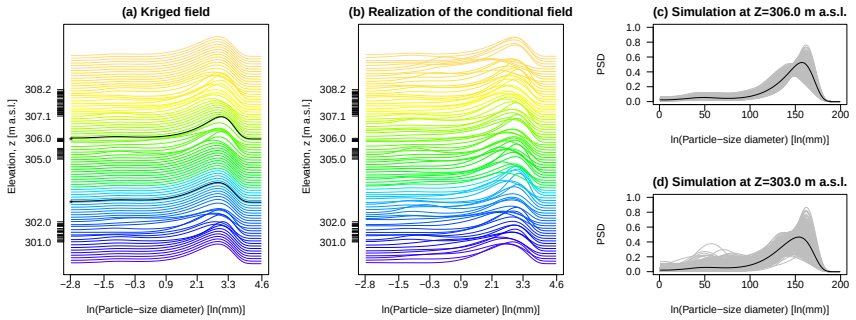

Fig. 2 Kriged field and conditional realizations. Panel (a): Kriging estimation over a grid along the vertical direction; black lines indicate predictions at elevations 303.0 and 306.0 m a.s.l.. Panel (b): a conditional realization on the same grid considered in panel (a). Panels (c) and (d): Kriging estimation at elevations 303.0 and 306.0 m a.s.l. (black line) and a sample of 1000 conditional simulations at the same sites (grey lines)

References

1. Egozcue, J. J., J. L. Díaz-Barrero, and V. Pawlowsky-Glahn (2006), Hilbert space of probability density functions based on Aitchison geometry, *Acta Mathematica Sinica, English Series*, *22*(4), 1175–1182.
2. Hron, K., A. Menafoglio, M. Templ, K. Hruzova, and P. Filzmoser (2016), Simplicial principal component analysis for density functions in Bayes spaces, *Computational Statistics & Data Analysis*, 94, 330–350.
3. Marron, J. S., A. M. Alonso, (2014), Overview of object oriented data analysis, *Biometrical Journal*, 56, 732–753.
4. Menafoglio, A., A. Guadagnini, and P. Secchi (2014), A Kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers, *Stochastic Environmental Research and Risk Assessment*, *28*(7), 1835–1851.
5. Menafoglio, A., P. Secchi, and A. Guadagnini (2015), Stochastic Simulation of Soil Particle-Size Curves in Heterogeneous Aquifer Systems through a Bayes space approach, *MOX-report* 59/2015.
6. van den Boogaart, K. G., J. Egozcue, and V. Pawlowsky-Glahn (2014), Bayes Hilbert spaces, *Australian & New Zealand Journal of Statistics*, *56*, 171–194.