

Kriging prediction for functional compositional data and application to particle-size curves

Previsione di Kriging per dati composizionali funzionali con applicazione a curve granulometriche

Alessandra Menafoglio, Alberto Guadagnini and Piercesare Secchi

Abstract We present a new geometric approach to kriging functional compositional data which embraces the viewpoints of both Functional and Compositional Data Analysis. Our theoretical framework enables one to characterize and predict random fields valued in the Hilbert space of functional compositions endowed with the Aitchison geometry. We show the application of the methodology to a field case scenario dealing with particle-size data collected within a heterogeneous aquifer near Tübingen, Germany. We consider particle-size densities, interpreted as functional compositional data, and perform kriging of these curves to obtain a complete characterization of the soil textural properties within the aquifer.

Abstract *Si presenta un nuovo approccio geometrico al kriging di dati composizionali funzionali che combina i punti di vista della Functional Data Analysis e della Compositional Data Analysis. Il contesto teorico proposto consente di caratterizzare e prevedere campi aleatori a valori nello spazio di Hilbert delle composizioni funzionali, dotato della geometria di Aitchison. Si mostra l'applicazione della metodologia a un caso studio relativo a dati granulometrici osservati in un acquifero eterogeneo vicino a Tübingen, in Germania. Si analizzano le densità granulometriche, interpretate come dati funzionali composizionali, e si applica la metodologia di kriging a tali curve al fine di ottenere una completa caratterizzazione delle proprietà tessiturali della matrice porosa nell'acquifero investigato.*

Key words: Compositional data, functional data, particle-size distributions

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

A. Guadagnini
Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Italy;
Department of Hydrology and Water Resources, The University of Arizona, USA
e-mail: alberto.guadagnini@polimi.it

1 Introduction

The increasing complexity of the type and quality of measurements collected in earth sciences motivates the development of new advanced statistical techniques to deal with spatially dependent high-dimensional and constrained data. This work focuses on the problem of the geostatistical prediction of functional compositional data (FCD). The latter are defined as infinite-dimensional objects constrained to be non-negative and to integrate to a constant [3] (e.g., probability density functions). We recall that a B -part composition is a B -dimensional vector whose components are proportions (or percent amounts) of a whole according to a certain partition of the domain (e.g., [1]). Thus, a functional compositional datum can be figured as a continuous composition obtained by refining the domain partition until having (infinite) infinitesimal parts [3].

To accommodate both the functional and the compositional nature of the data, we combine the viewpoints of Functional Data Analysis and Compositional Data Analysis. In this context we consider each datum as a point of an infinite-dimensional space endowed with a proper geometry to account for data constraint. We employ the Aitchison geometry ([1, 3]) within the Hilbert space method developed in [5], to perform kriging of a partially observed spatial field of functional compositions.

We illustrate the potential of the proposed method through its application to particle-size densities (PSDs), which are FCD commonly available in hydrological applications. PSDs describe the distribution of particle sizes within the soil sample they refer to, and are key to properly reconstruct the structural properties of the porous medium under study. We show that our approach leads to a precise characterization of the spatial field of PSDs by exploiting the entire information embedded in the PSDs through a global definition of spatial dependence.

2 Method

We consider the problem of predicting a spatial field of functional compositions from a sample of functional compositional data $\mathcal{Y}_{s_1}, \dots, \mathcal{Y}_{s_n}$ observed at a discrete set of locations $s_1, \dots, s_n$ in a spatial domain $D \subset \mathbb{R}^d$. Each datum $\mathcal{Y}_{s_i}$, $i = 1, \dots, n$, is a *functional composition* (FC): it is a non-negative real function on a compact domain $\mathcal{T} = [t_{min}, t_{max}]$ constrained to integrate to a constant (e.g., a continuous probability density function), i.e., $\mathcal{Y}_{s_i} : \mathcal{T} \rightarrow [0, +\infty)$ s.t. $\int_{\mathcal{T}} \mathcal{Y}_{s_i} = c$, for some $c \in \mathbb{R}$ (usually, $c = 1$). Note that the L^2 structure would not be appropriate to deal with FCs, because usual pointwise operations are not well defined in this context (e.g., the pointwise sum of two FCs is not a FC).

Following [3], we define two non-negative function f, g to be equivalent if there is an $\alpha > 0$ such that $f = \alpha g$; we then consider the space $A^2(\mathcal{T})$ of (equivalent classes of) non-negative real functions on the compact domain $\mathcal{T}$ with square integrable logarithm:

$$A^2(\mathcal{T}) = \{f : \mathcal{T} \rightarrow \mathbb{R} : f \geq 0 \text{ a.e.}, \log(f) \in L^2\}.$$

In the following, the representative of an equivalence class in A^2 will always be its element f integrating to 1. We then indicate with $\oplus, \odot$ the perturbation and powering operators in A^2 , respectively acting as:

$$f \oplus g = \frac{fg}{\int_{\mathcal{T}} fg} \quad f, g \in A^2; \quad \alpha \odot f = \frac{f^\alpha}{\int_{\mathcal{T}} f^\alpha}, \quad \alpha \in \mathbb{R}, f \in A^2.$$

We note that the difference operator $\ominus$ induced by perturbation $\oplus$ acts as $f \ominus g = f \oplus \frac{1/g}{\int_{\mathcal{T}} (1/g)}$, for $f, g \in A^2$. Moreover, we remark that, unlike pointwise L^2 operations, perturbation and powering of PDFs are well-defined. Finally, we denote with $\langle \cdot, \cdot \rangle$ the Aitchison inner product in A^2 :

$$\langle f, g \rangle = \int_{\mathcal{T}} [\log(f) \log(g)] - \frac{1}{|\mathcal{T}|} \int_{\mathcal{T}} \log(f) \int_{\mathcal{T}} \log(g), \quad f, g \in A^2.$$

In [3], $(A^2, \oplus, \odot, \langle \cdot, \cdot \rangle)$ is proven to be a separable Hilbert Space.

We assume the dataset $\mathcal{Y}_{s_1}, \dots, \mathcal{Y}_{s_n}$ to be the collection of the observations in n locations $s_1, \dots, s_n$ in D of a stochastic process $\{\mathcal{Y}_s, s \in D\}$ valued in A^2 and characterized by a globally second-order stationary model in the sense of [5], with trace-covariogram C and trace-semivariogram γ , i.e., for all $s, s_i, s_j \in D$,

$$\begin{aligned} \mathbb{E}[\mathcal{Y}_s] &= \underset{\mathcal{Y} \in A^2(\mathcal{T})}{\operatorname{arginf}} \mathbb{E}[\|\mathcal{Y}_s \ominus \mathcal{Y}\|_{A^2}^2] = m; \\ C(h) &= \operatorname{Cov}(\mathcal{Y}_{s_i}, \mathcal{Y}_{s_j}) = \mathbb{E}[\langle \mathcal{Y}_{s_i}, \mathcal{Y}_{s_j} \rangle], \quad h = \operatorname{dist}(s_i, s_j); \\ \gamma(h) &= \operatorname{Var}(\mathcal{Y}_{s_i}, \mathcal{Y}_{s_j}) = \mathbb{E}[\|\mathcal{Y}_{s_i} \ominus \mathcal{Y}_{s_j}\|], \quad h = \operatorname{dist}(s_i, s_j). \end{aligned}$$

We aim at predicting $\mathcal{Y}_{s_0}$ at an unsampled location $s_0 \in D$ through the Ordinary Kriging predictor. The latter is the best linear unbiased predictor $\mathcal{Y}_{s_0}^* = \bigoplus_{i=1}^n \lambda_i^* \odot \mathcal{Y}_{s_i}$, whose weights $\lambda_1^*, \dots, \lambda_n^* \in \mathbb{R}$ minimize the Aitchison variance of the prediction error under the unbiasedness constraint:

$$(\lambda_1^*, \dots, \lambda_n^*) = \underset{\substack{\lambda_1, \dots, \lambda_n \in \mathbb{R}: \\ \mathcal{Y}_{s_0}^\lambda = \bigoplus_{i=1}^n \lambda_i \odot \mathcal{Y}_{s_i}}}{\operatorname{argmin}} \operatorname{Var}_{A^2}(\mathcal{Y}_{s_0}^\lambda \ominus \mathcal{Y}_{s_0}) \quad \text{s.t.} \quad \mathbb{E}_{A^2}[\mathcal{Y}_{s_0}^\lambda] = m. \quad (1)$$

The well-posedness of problem (1) follows from the results of [5] for the general case of Hilbert-space valued random fields.

Proposition 1 ([5]). Assume that $\Sigma = (C(h_{i,j})) \in \mathbb{R}^{n \times n}$, $h_{i,j} = s_i - s_j$, $i, j = 1, \dots, n$, is a positive definite matrix. Then problem (1) admits a unique solution $(\lambda_1^*, \dots, \lambda_n^*) \in \mathbb{R}^n$, which is obtained by solving:

$$\left(\frac{C(h_{i,j})|1}{1|0} \right) \begin{pmatrix} \lambda_i \\ \zeta \end{pmatrix} = \begin{pmatrix} C(h_{0,i}) \\ 1 \end{pmatrix}, \quad (2)$$

ζ being the Lagrange multiplier associated with the unbiasedness constraint. The ordinary kriging variance of predictor $\mathcal{Y}_{s_0}^*$ is then

$$\sigma_*^2(s_0) = \text{Var}_{A^2}(\mathcal{Y}_{s_0}^*) = C(0) - \sum_{i=1}^n \lambda_i^* C(h_{i,0}) - \zeta^*. \quad (3)$$

An estimate of the trace-covariogram C , or, equivalently, of the trace-semivariogram γ is required to solve system (2). To this end, a method of moment (MoM) estimator $\hat{\gamma}$ can be employed:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j) \in N(h)} \|\mathcal{Y}_{s_i} \ominus \mathcal{Y}_{s_j}\|_{A^2}^2, \quad (4)$$

where $N(h)$ denotes the set of location pairs separated by h and $|N(h)|$ is its cardinality. A discretized version of $\hat{\gamma}$ is considered in typical applications and a valid variogram model is then fitted to observations [2]. To compute the norms appearing in (4) and the resulting kriging predictions we resort to a centered log-ratio transform, which defines an isometric isomorphism between the space A^2 and the space L^2 , where efficient routines are available. We refer to [4] for further details on the computational details.

3 A case study on particle-size distributions

We illustrate our approach through a field scale case study. The data we consider are part of the dataset collected within an alluvial aquifer with thickness of about 5 m near the city of Tübingen, Germany. We focus on the 60 particle-size curves (PSCs) collected at one of the monitoring wells at the site: in each location s_i , $i = 1, \dots, 60$, the PSC $\mathcal{X}_{s_i}$ expresses the relative amount $\mathcal{X}_{s_i}(t)$ of particles within a given soil sample having size smaller than or equal to t , where $t \in \mathcal{T} = [t_{\min}, t_{\max}]$ [7]. As such, PSCs describe the distribution of grain sizes within the soil sample. These data are relevant to applications related to groundwater hydrology, soil science and geochemistry, with particular emphasis on applications oriented towards modeling physical and chemical processes occurring in heterogeneous earth systems. We interpret PSCs as cumulative distribution functions and focus on their densities $\mathcal{Y}_{s_i}$, $i = 1, \dots, n$, which can be modeled as functional compositional data. We note that a data preprocessing is needed to compute density functions, because the available particle-size curves are reconstructed through grain sieve analysis performed on the basis of 12 sieve diameters. Here, we resort to the smoothing procedure based on Bernstein polynomials described in [4]. Smoothed particle-size densities are depicted in Fig. 1a.

The (smoothed) densities are then embedded in the Hilbert space A^2 of functional compositions and the methodology introduced in Section 2 is applied. Second-order stationarity is assumed for the field, this assumption being supported by prior knowl-

edge on the test site ([7] and references therein). Fig. 1b-1c depicts prediction results. These are seen to be satisfactory on the basis of cross-validation, evidencing a median and mean sum of squared error lower than 0.2% (Fig. 2a and c). Cross-validation predictions appear overall unbiased: this is evidenced by the cross-validation residuals depicted in Fig. 2b which are fairly spread across a uniform distribution (i.e., the neutral element of perturbation $\oplus$). Tests performed to assess the local quality of kriging predictions indicate that our methodology proves to be fairly accurate in predicting the local features of particle-size distributions.

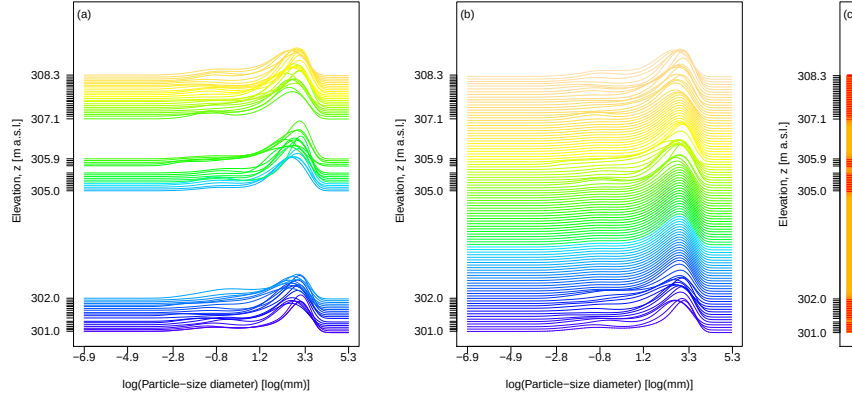


Fig. 1 (a) Particle-size densities (smoothed data); (b) kriged particle-size densities; (c) kriging variance, ranging between 0 (darkest shade) and 2.53 (lightest shade). Vertical coordinates correspond to data (sampled or predicted) locations.

We finally remark that a key advantage of our approach lies in the possibility of obtaining predictions of the entire particle-size curve at unsampled locations, as opposed to classical [2] or compositional kriging techniques [8] which allow only finite-dimensional predictions, based on a set of selected features (or synthetic indices) of the curve.

4 Conclusion

In this work, a novel methodology to krig functional compositions has been presented. Our approach enables one to predict the entire curve at unsampled locations on the basis of new definitions of spatial dependence which globally take into account the information content embedded in the curves. The method has been illustrated through its application to the spatial analysis of particle-size curves. Our kriging predictor for functional compositions grounds its strength on a very general method, which is appropriate for data belonging to any Hilbert space and is pre-

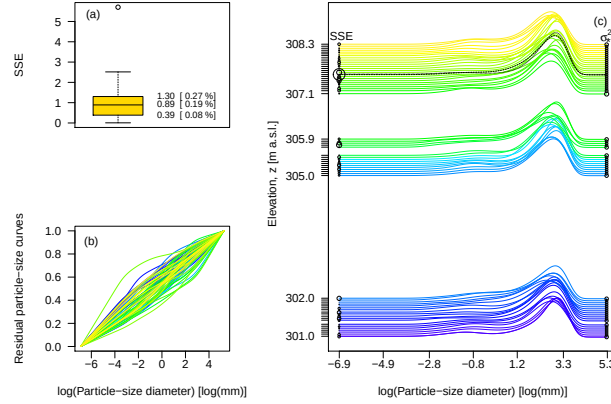


Fig. 2 Cross-validation results. (a) boxplot of the SSE, reporting the absolute and relative quartile values; (b) cross-validation residual particle-size curves; (c) cross-validation prediction of particle-size densities as a function of elevation (the size of the symbols is proportional to the associated cross-validation SSE/kriging variance); the curve highlighted is associated with the highest SSE.

sented in [5]. More generally, this approach can be exploited whenever a Hilbert space approximation can be used to describe data. For instance, it can be employed in the case of data belonging to a Riemannian manifold which can be locally approximated through a Hilbert tangent space where linear models and linear predictors can be defined. Such an idea, thoroughly explored in [6], opens new groundbreaking perspectives into the geostatistical analysis of shapes, M-reps or directional data.

References

1. J. Aitchison (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B*, 44(2), 139-177.
2. N. Cressie (1993). *Statistics for Spatial data*. John Wiley & Sons.
3. JJ. Egozcue, JL. Díaz-Barrero, V. Pawlowsky-Glahn (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Math. Sin. (English Ser.)*, 22(4), 1175-1182.
4. A. Menafoglio, A. Guadagnini, P. Secchi (2013). A Kriging Approach based on Aitchison Geometry for the Characterization of Particle-Size Curves in Heterogeneous Aquifers. *Stoch. Environ. Res. Risk Assess.*, in press.
5. A. Menafoglio, P. Secchi, M. Dalla Rosa (2013). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electron. J. Stat.*, 7, 2209-2240.
6. D. Pigoli, A. Menafoglio, P. Secchi (2013). Kriging prediction for manifold-valued random field. *MOX-rep.*, 63/2013. Politecnico di Milano.
7. M. Riva, L. Guadagnini, A. Guadagnini (2010). Effects of uncertainty of lithofacies, conductivity and porosity distributions on stochastic interpretations of a field scale tracer test. *Stoch. Environ. Res. Risk Assess.* 24, 955-970.
8. R. Tolosana-Delgado, K.G. van den Boogaart, V. Pawlowsky-Glahn (2011) *Geostatistics for Compositions*, Pawlowsky-Glahn & Buccianti edn, John Wiley & Sons, Ltd, 73-86