MODELLISTICA E CALCOLO SCIENTIFICO

MODELING AND SCIENTIFIC COMPUTING

MOX–Report No. 33/2013

A Kriging Approach based on Aitchison Geometry for the Characterization of Particle-Size Curves in Heterogeneous Aquifers

Menafoglio, A; Guadagnini, A; Secchi, P

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

A Kriging Approach based on Aitchison Geometry for the Characterization of Particle-Size Curves in Heterogeneous Aquifers

Alessandra Menafoglio[1], Alberto Guadagnini[2,3] and Piercesare Secchi[1]

[1]MOX-Department of Mathematics, Politecnico di Milano, Italy

[2]Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Italy

[3]Department of Hydrology and Water Resources, The University of Arizona, USA

alessandra.menafoglio@polimi.it

alberto.guadagnini@polimi.it

piercesare.secchi@polimi.it

Abstract

We consider the problem of predicting the spatial field of particle-size curves (PSCs) from a sample observed at a finite set of locations within an alluvial aquifer near the city of Tübingen, Germany. We interpret particle-size curves as cumulative distribution functions and their derivatives as probability density functions. We thus (a) embed the available data into an infinite-dimensional Hilbert Space of compositional functions endowed with the Aitchison geometry and (b) develop new geo-statistical methods for the analysis of spatially dependent functional compositional data. This approach enables one to provide predictions at unsampled locations for these types of data, which are commonly available in hydrogeological applications, together with a quantification of the associated uncertainty. The proposed functional compositional kriging (FCK) predictor is tested on a one-dimensional application relying on a set of 60 particle-size curves collected along a 5-m deep borehole at the test site. The quality of FCK predictions of PSCs is evaluated through leave-one-out cross-validation on the available data, smoothed by means of Bernstein Polynomials. A comparison of estimates of hydraulic conductivity obtained via our FCK approach against those rendered by classical kriging of effective particle diameters (i.e., quantiles of the PSCs) is provided. Unlike traditional approaches, our method fully exploits the functional form of particle-size curves and enables one to project the complete information content embedded in the PSC to unsampled locations in the system.

Keywords: Geostatistics; compositional data; functional data; particle-size curves; groundwater; hydrogeology

1 Introduction

The geostatistical characterization of the spatial distribution of particle-size curves (PSCs) is a key issue in earth sciences. These types of data are typically based on standard grain sieve analysis of soil samples, yielding a discrete representation of the curves by measuring selected particle diameters which, in turn, correspond to quantiles of the particle-size curve. The information can then be employed to classify soil types (e.g., Riva et al. (2006) and references therein), to infer hydraulic parameters such as porosity and hydraulic conductivity (e.g., amongst others, Lemke and Abriola (2003); Riva et al. (2006, 2008, 2010); Bianchi et al. (2011); Tong et al. (2010); Barahona-Palomo et al. (2011) and references therein), or, in the presence of inorganic compounds, to provide estimates of the porous medium sorption capacity (e.g., Hu et al. (2004) and references therein).

Classification of aquifer geomaterials and the estimation of their spatial arrangement is relevant to properly reconstruct the internal architecture of groundwater systems which can play a critical role in controlling contaminant spreading on different scales. Methodologies which are typically employed for the estimation of the location of internal boundaries between lithofacies take advantage of geological and/or hydraulic information and include, amongst other methods, sequential indicator approaches (Deutsch and Journel (1997); Guadagnini et al. (2004) and references therein), Nearest-neighbor classification (e.g., Tartakovsky et al. (2007)), or Support Vector Machines (Wohlberg et al., 2006).

Several techniques widely employed for the estimation of aquifer hydraulic parameters are based on particle-size information. They usually rely on spatially dependent particle-size data, measured from samples collected at a discrete set of points in a reservoir. In this context, the knowledge of the functional form of particle-size curves is not fully exploited in typical aquifer reconstruction practice. As an example of the way this information content is employed, we mention the work of Riva et al. (2006). These authors perform a geostatistical facies-based parametrization of the lithofacies occurring within a small scale alluvial aquifer system. They rely on sampled particle-size curves and apply a standard multivariate cluster analysis technique to classify these. They then perform indicator variography of the identified classes and provide estimates of the spatial distribution of lithotypes in the system. Hydraulic conductivity values are then assigned to the blocks of a numerical flow and transport model upon projecting only the 10^{th} and 60^{th} quantiles of the observed particle-size curves on the computational grid through kriging. A similar approach has been employed, amongst other authors, by Bianchi et al. (2011). In this sense, the information content embedded in the particle-size curve is only partially transferred to unsampled locations in the system, through few selected local features (in the example above, the 10^{th} and 60^{th} quantiles). Instead, a complete characterization of the spatial distribution of lithotypes in a reservoir attributes would require embedding the full particle-size curve into the geostatistical analysis.

In addition to this, having at our disposal the spatial arrangement of all the components of soil particle-size curves would allow improved predictions of soil hydraulic attributes through pedotransfer functions (e.g., Nemes et al. (2003); Pachepsky and Rawls (2004), and reference therein) as well as of soil geochemical parameters which are relevant in sorption/desorption and cation exchange processes.

These problems motivate the development of advanced geostatistical techniques which enable one to treat georeferenced particle-size curves. To this end, we model particle-size curves as cumulative distribution functions and analyze their derivatives, by coherently considering them as probability density functions. We use two viewpoints to interpret these types of data: (a) a Functional Data Analysis (FDA, Ramsay and Silverman (2005)) and (b) a Compositional Data Analysis (CoDa, Aitchison (1982, 1986); Pawlowsky-Glahn and Buccianti (2011)) approach. The key idea underlying FDA methods is to view each datum (i.e., each PSC), even though discretely observed, as a unique entity belonging to a suitable functional space. In this way, the curse of dimensionality is overcome allowing the statistical analysis of high-dimensional (virtually infinite-dimensional) data. On the other hand, CoDa deals with data which convey only relative information: a D-parts composition is a D-dimensional vector whose components are proportions (or percent amounts) of a whole according to a certain partition of the domain. Thus, a D-parts composition has D non-negative components, constrained to sum up to a constant (usually set to unity or 100) and belongs to a $(D-1)$-dimensional simplex. Probability density functions are functional and compositional data, i.e., they are infinite-dimensional objects which are constrained to be non-negative and to integrate to unity. They can be considered as compositional data obtained by refining the domain partition until (infinite) infinitesimal parts are obtained (Egozcue et al., 2006). In this framework, the geostatistical methodology we propose to treat spatially dependent functional compositional data takes advantage of the strengths of both the FDA and CoDa approaches.

An increasing body of literature on the geostatistical analysis of functional data is available, either in the stationary (e.g., Goulard and Voltz (1993); Nerini et al. (2010); Delicado et al. (2010) and references therein) or non-stationary setting (Menafoglio et al., 2012; Caballero et al., 2013). A relatively rich literature is also available in the field of spatially dependent compositional data (e.g., Tolosana-Delgado et al. (2011); Tolosana-Delgado et al. (2011); Pawlowsky-Glahn and Olea (2004); Leininger et al. (2013) and references therein). In this context, particle-size fractions have been treated as discrete compositional data (e.g., Odeh et al. (2003); Buchanan et al. (2012)) and compositional techniques have been employed to predict the soil composition at unsampled location. Albeit these techniques take properly into account the compositional constraints in PSCs, they are only suited for low-dimensional compositions and their application can be problematic if the dimensionality increases (i.e., curse of dimensionality). The data dimensionality is closely related to the resolution of the measurement technique which is employed: modern sieve-analysis techniques enable one to obtain high-resolution PSC, i.e., high-dimensional data, which need to be treated with advanced techniques. However, to the best of our knowledge, none of the available literature works addresses the problem of the geostatistical analysis of high-dimensional and functional compositional data.

Here, we focus specifically on the formulation of new geostatistical models and methods for functional compositional data. To do so, we move from the geostatistical methodology proposed in (Menafoglio et al., 2012) and the mathematical construction developed by Egozcue et al. (2006) and further investigated in (van den Boogaart et al., 2010). Our approach shares with FDA and CoDa the foundational role of geometry. Hilbert space theory allows FDA methods to cope with the infinite-dimensionality of the data (e.g., Ramsay and Dalzell (1991); Ferraty and Vieu (2006); Horváth and Kokoszka (2012) and references therein), while the log-ratio approach grounds the Aitchison geometry, which properly accounts for the compositional nature of the data (e.g., Pawlowsky-Glahn and Egozcue (2001, 2002)). Here, we employ Aitchison geometry within a Hilbert Space method to accommodate both the functional and compositional nature of the data.

Even though the developments illustrated in this work are motivated by the analysis of the particle-size data presented in Section 2, our methodology is indeed general and allows performing the geostatistical analysis of any kind of compactly supported functional compositional data, provided that these can be embedded in the Hilbert Space endowed with the Aitchison geometry described in Section 3. We thus introduce the model and illustrate the methodology within a stationary setting, in view of the considered application. For completeness, the theoretical developments associated with a non-stationary approach are reported in Appendix A.

Among the practical issues which need to be tackled when dealing with functional data, we consider the problem of their preprocessing when only discrete observations are available, as in our application: we propose the use of a smooth estimator based on Bernstein Polynomials and prove its consistency in Section 4. Section 5 illustrates applications of our functional compositional kriging technique to the target dataset.

2 Field data

The data we consider are part of the dataset collected at an experimental site located near the city of Tübingen, Germany. The aquifer is made up by alluvial material overlain by stiff silty clay and underlain by hard silty clay. The site characterization has been based on stratigraphic information collected at a set of monitoring and pumping wells (Martac and Ptak (2003) and references therein). The saturated thickness of the aquifer is about 5 m and all boreholes reach the bedrock which forms the impermeable aquifer base.

The extensive investigations performed at the site comprise field- and laboratory-scale data collection and analysis. Available data include particle-size curves, pumping and tracer tests as well as down-hole impeller flowmeter measurements. A complete description of the analyses performed at the site has been presented by Riva et al. (2006, 2008), to which we refer for additional details. The available data have been partially employed by Neuman et al. (2007, 2008) in the context of (a) the application of a stochastic interpretation of the results of a series of cross-hole pumping tests and (b) a geostatistically-based characterization of multiscale distribution of hydraulic conductivity at the site. Barahona-Palomo et al. (2011) compared hydraulic conductivity estimates obtained through particle-size curves and impeller flowmeter measurements. Riva et al. (2006, 2008, 2010) performed numerical Monte Carlo analyses of a tracer test and well-related capture zones at the site upon relying on the information provided by the available particle-size curves. The latter were measured on core samples associated with characteristic length ranging from 5 to 26.5 cm and indicating the occurrence of heterogeneous and highly conducive alluvial deposits. A total of 411 particle-size curves collected along 12 vertical boreholes are available within the site. Particle-size curves are reconstructed through grain sieve analysis performed with a set of 12 discrete sieve diameters. These data have been subject to cluster analysis to classify the spatial distribution of hydrofacies in the system through indicator-based variogaphy and Monte Carlo numerical simulations (Riva et al., 2006). Characteristic particle diameters estimated from the particle-size curves have been employed to provide estimates of porosity and hydraulic conductivity which have then formed the basis for three-dimensional simulations of the heterogeneous structure of the aquifer hydraulic attributes.

Here, we focus on the 60 particle-size curves which were collected at well B5 at the site. Figure 1 depicts the set of particle-size curves available at the well together with the vertical location of the sampling points. For ease of reference, each curve has been attributed to a vertical coordinate which coincides with the center of the sampling interval from which particle-size data have been extracted. The data are grouped within three main regions along the borehole and are mainly associated with (a) moderately sorted gravel with about 14% sand and very few fines, and (b) poorly sorted gravel with about 24% sand and few fines (Riva et al., 2006). This constitutes a rather unique data-set that enables us to explore extensively the key features and potential of the methodology we present which is conducive to the estimation of the complete particle-size distribution at unsampled locations.

3 A Kriging Approach for Particle-Size Distributions Characterization

3.1 A Stochastic Model for Particle-Size Distributions

Let $(\Omega, \mathfrak{F}, P)$ be a probability space and consider the random process $\{\chi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, $D \subset \mathbb{R}^d$, whose elements are particle-size curves. Each element $\chi_{\boldsymbol{s}}$, $\boldsymbol{s} \in D$, is a $[0,1]$ valued random function defined on $\mathcal{T} = [t_m, t_M]$, i.e., $\chi_{\boldsymbol{s}}$ is measurable and, for $\omega \in \Omega$, $\chi_{\boldsymbol{s}}(\omega, \cdot) : \mathcal{T} \to [0,1]$. Given a particle size $t \in \mathcal{T}$, $\chi_{\boldsymbol{s}}(\cdot, t)$ indicates the random fraction of grains with diameter smaller than or equal to t. Hence, each function $\chi_{\boldsymbol{s}}(\omega, \cdot)$ is a cumulative distribution function (CDF).

The usual vectorial structure for functional spaces, based on point-wise notions of sum and product by a real constant, is not appropriate when dealing with CDFs because the space of CDF is not closed with respect to such operations (for instance, the point-wise sum of two CDFs is not a CDF). Instead, a geometric approach based on Aitchison geometry (Aitchison, 1982, 1986) is more appropriate to treat distribution functions because it accounts for their compositional nature. In particular, Aitchison geometry is well suited for probability density functions (PDFs), which are (discrete or continuous) compositions, in the sense that they provide only relative information and are constrained to sum (or integrate) to a constant.

We thus consider the derivative process $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, defined on the probability space introduced above and such that, for $\boldsymbol{s} \in D$:

$$\mathcal{Y}_{\boldsymbol{s}}(\omega, \cdot) : \mathcal{T} \to [0, +\infty), \quad s.t. \quad \int_{\mathcal{T}} \mathcal{Y}_{\boldsymbol{s}}(\omega, t) dt = 1, \quad \omega \in \Omega.$$

We assume that, for $\omega \in \Omega$, $\mathcal{Y}_{\boldsymbol{s}}(\omega, \cdot) = d\chi_{\boldsymbol{s}}(\omega, \cdot)/dt$ is the density function of

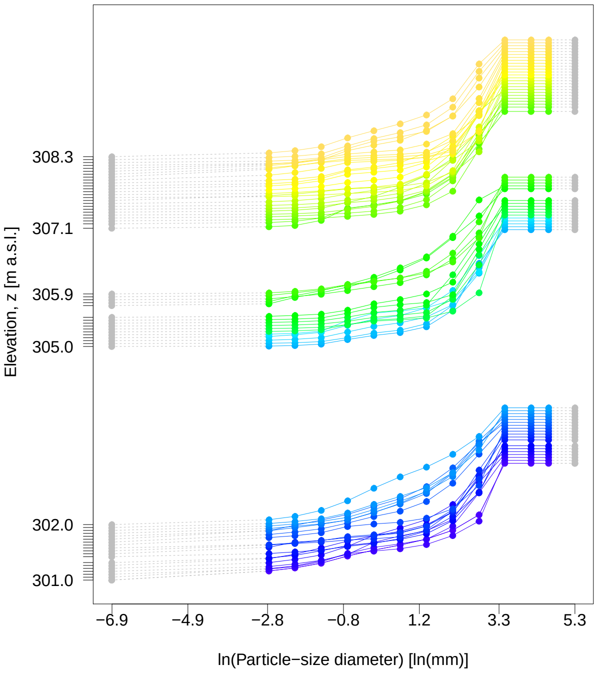

Figure 1: Available particle-size data. Data are represented on the vertical coordinate according to their sampled location and supported on the compact domain ($d_{min} = 0.001mm$; $d_{max} = 200mm$); d_{min} and d_{max} are the smallest and largest measured particle-size diameters, respectively; elevation is given in meters above sea level (m a.s.l.).

the random probability measure μ defined, for all $a \leq b$, by:

$$\mu_{\boldsymbol{s}}(\omega, (a, b]) = \chi_{\boldsymbol{s}}(\omega, b) - \chi_{\boldsymbol{s}}(\omega, a).$$

We call $\mathcal{Y}_{\boldsymbol{s}}$ the *particle-size density function* in $\boldsymbol{s} \in D$.

Let us denote with $A^2(\mathcal{T})$ the space of (equivalence classes of) non-negative real functions on $\mathcal{T}$ with square-integrable logarithm, i.e. (Egozcue et al., 2006):

$$A^2 = \{f : \mathcal{T} \to \mathbb{R}, \text{ such that } f \geq 0 \text{ a.e. and } \ln(f) \in L^2(\mathcal{T})\}.$$

In this work we assume that $\mathcal{Y}_{\boldsymbol{s}}(\omega, \cdot) \in A^2(\mathcal{T})$ for all $\boldsymbol{s} \in D$, $\omega \in \Omega$. In Subsection 3.3 we show that an isometric isomorphism exists between $A^2(\mathcal{T})$ and $L^2(\mathcal{T})$. Moreover, if we consider an orthonormal basis $\{\varphi_k\}_{k\geq 0}$ of $L^2(\mathcal{T})$, such that $\varphi_0 = 1/\sqrt{\eta}$ ($\eta = t_M - t_m$), and define the operator $T : A^2(\mathcal{T}) \to \ell^2$ as $Tf = \{\alpha_k\}_{k\geq 1}$, where α_k ($k \geq 1$) appears in the decomposition $\log(f) = \sum_{k\geq 0} \alpha_k \varphi_k$, then the following result holds.

Proposition 1 (Egozcue et al. (2006)). *$A^2(\mathcal{T})$ endowed with the Aitchison inner product*

$$\langle f, g\rangle_{A^2} = \langle Tf, Tg\rangle_{\ell^2}, \quad f, g \in A^2(\mathcal{T}), \tag{1}$$

and the induced norm is a separable Hilbert space.

Some of the basic definitions and properties of this functional space are recalled in the following Subsections. Additional properties and generalizations are reported in (Egozcue et al., 2006; van den Boogaart et al., 2010).

3.2 A Kriging predictor for Particle-Size Densities

We indicate with $\mathcal{C}[f]$ the closure of $f \in L^1(\mathcal{T})$, i.e.,

$$\mathcal{C}[f] = \frac{f}{\int_{\mathcal{T}} f(t)dt},$$

and denote with $\oplus, \odot$ the perturbation and powering operators in $A^2(\mathcal{T})$, respectively, acting as:

$$\begin{aligned} f \oplus g &= \mathcal{C}[fg], \quad f, g \in A^2(\mathcal{T}) \\ \alpha \odot f &= \mathcal{C}[f^\alpha], \quad \alpha \in \mathbb{R}, f \in A^2(\mathcal{T}), \end{aligned}$$

Note that the neutral elements of perturbation and powering are $e(t) = 1/\eta$ and 1, respectively, while in (Egozcue et al., 2006) it is proven that $(A^2(\mathcal{T}), \oplus, \odot)$ is a vector space. We denote with $f \ominus g$ the difference in the Aitchison geometry between f and g, namely the perturbation of f with the reciprocal of g, i.e., $f \ominus g = f \oplus \mathcal{C}[1/g]$, $f, g \in A^2(\mathcal{T})$.

For $s \in D$, we indicate with m_s the Fréchet mean of $\mathcal{Y}_s$ with respect to the Aitchison geometry on $A^2(\mathcal{T})$, namely:

$$m_{\boldsymbol{s}} = \mathbb{E}_{A^2}[\mathcal{Y}_{\boldsymbol{s}}] = \underset{\mathcal{Y}\in A^2(\mathcal{T})}{\operatorname{arginf}} \ \mathbb{E}[\|\mathcal{Y}_{\boldsymbol{s}} \ominus \mathcal{Y}\|^2_{A^2}] = \underset{\mathcal{Y}\in A^2(\mathcal{T})}{\operatorname{arginf}} \int_\Omega \|\mathcal{Y}_{\boldsymbol{s}}(\omega, \cdot) \ominus \mathcal{Y}(\cdot)\|^2_{A^2} P(d\omega).$$

Following (Menafoglio et al., 2012), for any given $\boldsymbol{s} \in D$, we represent the element $\mathcal{Y}_{\boldsymbol{s}}$ as a perturbation of the mean function $m_{\boldsymbol{s}}$ with a *neutral-mean* stochastic residual $\delta_{\boldsymbol{s}}$

$$\begin{aligned} \mathcal{Y}_{\boldsymbol{s}} &= m_{\boldsymbol{s}} \oplus \delta_{\boldsymbol{s}}, \\ \mathbb{E}_{A^2}[\delta_{\boldsymbol{s}}] &= 0_{\oplus} = 1/\eta. \end{aligned} \tag{2}$$

We assume that the process $\mathcal{Y}_{\boldsymbol{s}}$ can be represented by a global second-order stationary model. Hence, the process is characterized by a spatially constant mean function ($m_{\boldsymbol{s}} = m$, for all $\boldsymbol{s} \in D$), a trace-covariogram $C : \mathbb{R}^d \to \mathbb{R}$ and a trace-variogram $\gamma : \mathbb{R}^d \to \mathbb{R}$, which are respectively defined as:

$$C(\boldsymbol{s}_i - \boldsymbol{s}_j) = \mathrm{Cov}_{A^2}(\mathcal{Y}_{\boldsymbol{s}_i}, \mathcal{Y}_{\boldsymbol{s}_j}) = \mathbb{E}[\langle \mathcal{Y}_{\boldsymbol{s}_i} - m, \mathcal{Y}_{\boldsymbol{s}_j} - m\rangle_{A^2}], \quad \boldsymbol{s}_i, \boldsymbol{s}_j \in D (3)$$

$$2\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) = \mathrm{Var}_{A^2}(\mathcal{Y}_{\boldsymbol{s}_i} \ominus \mathcal{Y}_{\boldsymbol{s}_j}) = \mathbb{E}[\|\mathcal{Y}_{\boldsymbol{s}_i} \ominus \mathcal{Y}_{\boldsymbol{s}_j}\|^2_{A^2}], \quad \boldsymbol{s}_i, \boldsymbol{s}_j \in D. \tag{4}$$

Given a sample $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$ of $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, the Ordinary Kriging predictor of $\mathcal{Y}_{\boldsymbol{s}_0}$, at an unsampled location $\boldsymbol{s}_0 \in D$, is the best linear unbiased predictor (BLUP) in the Aitchison geometry:

$$\mathcal{Y}^*_{\boldsymbol{s}_0} = \bigoplus_{i=1}^{n} \lambda_i^* \odot \mathcal{Y}_{\boldsymbol{s}_i}. \tag{5}$$

Here, the weights $\lambda_1^*, ..., \lambda_n^* \in \mathbb{R}$ minimize the Aitchison variance of the prediction error under the unbiasedness constraint:

$$(\lambda_1^*, ..., \lambda_n^*) = \underset{\substack{\lambda_1, ..., \lambda_n \in \mathbb{R}: \\ \mathcal{Y}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} = \bigoplus_{i=1}^n \lambda_i \odot \mathcal{Y}_{\boldsymbol{s}_i}}}{\operatorname{argmin}} \mathrm{Var}_{A^2}(\mathcal{Y}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} \ominus \mathcal{Y}_{\boldsymbol{s}_0}) \quad \text{s.t.} \quad \mathbb{E}_{A^2}[\mathcal{Y}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0}] = m. \tag{6}$$

The problem of kriging of functional data has been tackled in (Menafoglio et al., 2012) within the general framework of (possibly non-stationary) functional processes valued in any separable Hilbert Space. Hence, problem (6) can be solved by exploiting this general approach which we recall here for a stationary setting.

Proposition 2 (Menafoglio et al. (2012))**.** *Assume that* $\Sigma = (C(\boldsymbol{h}_{i,j})) \in \mathbb{R}^{n \times n}$, $\boldsymbol{h}_{i,j} = \boldsymbol{s}_i - \boldsymbol{s}_j$, $i, j = 1, ..., n$, *is a positive definite matrix. Then problem* (6) *admits a unique solution* $(\lambda_1^*, ..., \lambda_n^*) \in \mathbb{R}^n$, *which is obtained by solving:*

$$\left(\begin{array}{c|c} C(\boldsymbol{h}_{i,j}) & 1 \\ \hline 1 & 0 \end{array}\right) \left(\begin{array}{c} \lambda_i \\ \hline \zeta \end{array}\right) = \left(\begin{array}{c} C(\boldsymbol{h}_{0,i}) \\ \hline 1 \end{array}\right), \tag{7}$$

ζ *being the Lagrange multiplier associated with the unbiasedness constraint. The ordinary kriging variance of predictor* (5) *is then*

$$\sigma^2_*(\boldsymbol{s}_0) = \mathrm{Var}_{A^2}(\mathcal{Y}^*_{\boldsymbol{s}_0}) = C(\boldsymbol{0}) - \sum_{i=1}^{n} \lambda_i^* C(\boldsymbol{h}_{i,0}) - \zeta^*. \tag{8}$$

In the light of expression (8), the following Čebyšëv inequality can be provided for the prediction errors:

$$P(\|\mathcal{Y}_{\boldsymbol{s}_0} \ominus \mathcal{Y}^*_{\boldsymbol{s}_0}\|_{A^2} > \kappa \cdot \sigma_*(\boldsymbol{s}_0)) < \frac{1}{\kappa^2}. \tag{9}$$

Note that this inequality can be used to elicit confidence bands on the norm of the prediction errors.

As in classical geostatistics, under stationarity conditions the only quantity which is required to be estimated is the trace-semivariogram γ, as $C(\boldsymbol{h}) = C(\boldsymbol{0}) - \gamma(\boldsymbol{h})$, $\boldsymbol{h} \in \mathbb{R}^d$ being a lag, or separation distance vector. To this end, a method of moments (MoM) estimator $\widehat{\gamma}$ can be employed:

$$\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{2|N(\boldsymbol{h})|} \sum_{(i,j)\in N(\boldsymbol{h})} \|\mathcal{Y}_{\boldsymbol{s}_i} \ominus \mathcal{Y}_{\boldsymbol{s}_j}\|^2_{A^2}, \tag{10}$$

where $N(\boldsymbol{h})$ denotes the set of location pairs separated by $\boldsymbol{h}$ and $|N(\boldsymbol{h})|$ its cardinality. A discretized version of $\widehat{\gamma}$ is considered in typical applications and a valid variogram model is fitted to observations.

The approach we present can also be employed in a non-stationary setting. For completeness, we report the details of this case in Appendix A.

3.3 Log-ratio transform

Here, we illustrate a representation of the process through a log-ratio transform. In addition to its theoretical value, this representation enables one to considerably simplify the computation of the quantities of interest (e.g., the trace-variogram). Our developments rely on the properties of the space $A^2(\mathcal{T})$ derived in (Egozcue et al., 2006).

Whenever $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ is a random field valued in A^2 which follows the dichotomy (2) and has finite variance, i.e., $E[\|\delta_{\boldsymbol{s}}\|^2_{A^2}] < +\infty$ for all $\boldsymbol{s} \in D$, there exist a (deterministic) sequence $\{\mu_k(\boldsymbol{s})\}_{k\geq 1}$, a zero-mean random sequence $\{\xi_k(\cdot, \boldsymbol{s})\}_{k\geq 1}$, both valued in ℓ^2, and an orthonormal basis $\{\Psi_k\}_{k\geq 1}$ of A^2, such that:

$$\mathcal{Y}_{\boldsymbol{s}}(\omega, \cdot) = \bigoplus_{k=1}^{\infty} (\mu_k(\boldsymbol{s}) + \xi_k(\omega, \boldsymbol{s})) \odot \Psi_k(\cdot), \quad \omega \in \Omega.$$

Here, $\mu_k(\boldsymbol{s}) = \langle m_{\boldsymbol{s}}, \Psi_k\rangle_{A^2}$, $\xi_k(\boldsymbol{s}) = \langle \delta_{\boldsymbol{s}}, \Psi_k\rangle_{A^2}$, $k \geq 1$, $\boldsymbol{s} \in D$. The sequences $\{\mu_k(\cdot, \boldsymbol{s})\}_{k\geq 1}$ and $\{\xi_k(\cdot, \boldsymbol{s})\}_{k\geq 1}$ satisfy the decomposition:

$$\log(\mathcal{Y}_{\boldsymbol{s}}(\omega, \cdot)) = \sum_{k=0}^{\infty} (\mu_k(\boldsymbol{s}) + \xi_k(\omega, \boldsymbol{s}))\varphi_k(\cdot), \quad \omega \in \Omega$$

provided that $\{\varphi_k\}_{k\geq 0}$ is an orthonormal basis of $L^2(\mathcal{T})$ such that $\varphi_0 = 1/\sqrt{\eta}$ ($\eta = t_M - t_m$ and $\Psi_k = \mathcal{C}[\exp\{\varphi_k\}]$.

The process $\{\mathcal{Z}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ defined on $(\Omega, \mathfrak{F}, \mathbb{P})$ as

$$\mathcal{Z}_{\boldsymbol{s}}(\omega, t) = \log(\mathcal{Y}_{\boldsymbol{s}}(\omega, t)) - \frac{1}{\eta}\int_{\mathcal{T}} \log(\mathcal{Y}_{\boldsymbol{s}}(\omega, z))dz, \quad \omega \in \Omega,\, t \in \mathcal{T},\, \boldsymbol{s} \in D, \tag{11}$$

satisfies

$$\mathcal{Z}_{\boldsymbol{s}}(\omega, \cdot) = \sum_{k=1}^{\infty} (\mu_k(\boldsymbol{s}) + \xi_k(\omega, \boldsymbol{s}))\varphi_k(\cdot), \tag{12}$$

since (see Egozcue et al. (2006))

$$\begin{aligned}\log(\mathcal{Y}_{\boldsymbol{s}}(\omega,\cdot)) &= \sum_{k=0}^{+\infty}(\mu_k(\boldsymbol{s})+\xi_k(\omega,\boldsymbol{s}))\varphi_k = \\ &= \sum_{k=1}^{+\infty}(\mu_k(\boldsymbol{s})+\xi_k(\omega,\boldsymbol{s}))\varphi_k + \frac{1}{\eta}\int_{\mathcal{T}}\log(\mathcal{Y}_{\boldsymbol{s}}(\omega,t))dt.\end{aligned}$$

Each element $\mathcal{Z}_{\boldsymbol{s}}$ of the process $\{\mathcal{Z}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ is a centered log-ratio (clr) transform of the corresponding element $\mathcal{Y}_{\boldsymbol{s}}$, in analogy with the finite- dimensional case (Pawlowsky-Glahn and Egozcue, 2001). Note that the A^2 inner product between two elements $f, g \in A^2(\mathcal{T})$ can be computed as an L^2 inner product between the clr transforms $clr(f)$, $clr(g) \in L^2(\mathcal{T})$:

$$\begin{aligned}\langle f,g\rangle_{A^2} &= \int_{\mathcal{T}}\log(f(t))\log(g(t))dt - \frac{1}{\eta}\int_{\mathcal{T}}\log(f(t))dt\int_{\mathcal{T}}\log(g(t))dt = \\ &= \int_{\mathcal{T}}\left(\log(f(t)) - \frac{1}{\eta}\int_{\mathcal{T}}\log(f(z))dz\right)\cdot \\ &\quad \left(\log(g(t)) - \frac{1}{\eta}\int_{\mathcal{T}}\log(g(z))dz\right)dt = \\ &= \langle clr(f), clr(g)\rangle_{L^2}\end{aligned}$$

the first equality above being proven by Egozcue et al. (2006).

The correspondence between the distributional features of the processes $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ in $A^2(\mathcal{T})$ and $\{\mathcal{Z}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ in $L^2(\mathcal{T})$ is apparent from identity (12), as the clr transform defines an isometric isomorphism between $A^2(\mathcal{T})$ and $L^2(\mathcal{T})$. In particular, the Fréchet mean of process $\{\mathcal{Y}_{\boldsymbol{s}}\}$ with respect to the Aitchison geometry on $A^2(\mathcal{T})$ coincides with the Fréchet mean of $\{\mathcal{Z}_{\boldsymbol{s}}\}$ with respect to $L^2(\mathcal{T})$. Moreover, stationarity and isotropy assumption for $\{\mathcal{Y}_{\boldsymbol{s}}\}$ in $A^2(\mathcal{T})$ can be stated in terms of the corresponding properties of $\{\mathcal{Z}_{\boldsymbol{s}}\}$ in $L^2(\mathcal{T})$. Notice that the definition (11) of process $\{\mathcal{Z}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ allows writing:

$$\begin{aligned}\mathrm{Cov}_{A^2}(\mathcal{Y}_{\boldsymbol{s}_i}, \mathcal{Y}_{\boldsymbol{s}_j}) &= \mathrm{Cov}_{L^2}\left(\mathcal{Z}_{\boldsymbol{s}_i}, \mathcal{Z}_{\boldsymbol{s}_j}\right); \\ \mathrm{Var}_{A^2}(\mathcal{Y}_{\boldsymbol{s}_i} \ominus \mathcal{Y}_{\boldsymbol{s}_j}) &= \mathrm{Var}_{L^2}\left(\mathcal{Z}_{\boldsymbol{s}_i} - \mathcal{Z}_{\boldsymbol{s}_j}\right).\end{aligned}$$

Therefore, the trace-variogram and the trace-covariogram of $\{\mathcal{Y}_{\boldsymbol{s}}\}$ in the Aitchison geometry coincide with the corresponding quantities associated with $\{\mathcal{Z}_{\boldsymbol{s}}\}$ with respect to the L^2 geometry.

The kriging prediction in $A^2(\mathcal{T})$ can then be performed by treating the transformed sample $Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n}$ in the $L^2(\mathcal{T})$ geometry, as:

$$\begin{aligned}\mathcal{Y}^*_{\boldsymbol{s}_0} &= \bigoplus_{i=1}^{n}\lambda_i^* \odot \mathcal{Y}_{\boldsymbol{s}_i} = \mathcal{C}\left[\prod_{i=1}^{n}\mathcal{Y}_{\boldsymbol{s}_i}^{\lambda_i^*}\right] = \mathcal{C}\left[\exp\left\{\sum_{i=1}^{n}\lambda_i^*\log(\mathcal{Y}_{\boldsymbol{s}_i})\right\}\right] = \\ &= \mathcal{C}\left[\exp\left\{\sum_{i=1}^{n}\lambda_i^*\mathcal{Z}_{\boldsymbol{s}_i}\right\}\right] = clr^{-1}(\mathcal{Z}^*_{\boldsymbol{s}_0}).\end{aligned}$$

The above isomorphism enables one to perform the required calculation by exploiting efficient routines which are designed for unconstrained data belonging to L^2, eventually back-transforming to A^2 the results. We adopt this strategy in our application, which is illustrated in Section 5.

4 Smoothing discrete particle-size data with Bernstein Polynomials

A key assumption underlying the spatial prediction methodology proposed here is that data are curves which can be evaluated at any point $t \in \mathcal{T}$. If particle-size curves were already observed in their functional form, the methodology illustrated in Section 3 could be directly applied, without any particular data preprocessing. A very close analogue to this kind of information could be obtained by employing modern sieve-analysis techniques which enable one to obtain the full PSC from a soil sample without being limited to a small number of discretely spaced sieve diameters. As detailed in Section 2, in the present case study an estimate of the PSC at a given spatial location $\boldsymbol{s}$ along the borehole is available only for a set of $N = 12$ sieve diameters, $t_1, ..., t_N$. In such a case, which is typically associated with several practical field situations, a preprocessing of the raw data is required to obtain smooth estimates of the PSCs and associated densities. Amongst different types of techniques, we propose (Subsection 4.1) and apply (Subsection 4.2) a smoothing procedure for particle-size distributions which is based on Bernstein Polynomials, following the approach of (Babu et al., 2002). For simplicity, we adopt here a general and immediate notation and omit the subscript $\boldsymbol{s}$ indicating spatial location, since we consider each particle-size curve separately.

4.1 A smooth estimator for cumulative distribution functions

Consider the problem of estimating a continuous and compactly supported CDF through a smooth estimator. We denote with F the (true) underlying CDF and assume it is supported on $[0, 1]$ (we invoke this assumption only for convenience of notation since this can be easily relaxed to consider a generic support $\mathcal{T} = [t_m, t_M]$ provided $x = (t - t_m)/(t_M - t_m) \in [0, 1]$ if $t \in \mathcal{T}$). If a sample $X_1, ..., X_\nu$ from F is available, the empirical cumulative distribution function (ECDF) F_ν, defined as:

$$F_\nu(x) = \frac{1}{\nu} \sum_{i=1}^{\nu} I_{[0,x]}(X_i),$$

I being the indicator function, is a (discontinuous) non-parametric estimator of F, which is strongly consistent because of the Glivenko-Cantelli Theorem. In our setting, the sample $X_1, ..., X_\nu$ would represent the set of (transformed) particle diameters measured within a soil sample extracted at a given location in the aquifer, ν being the number of diameters constituting the sample. When such a sample is available, the problem of smoothly estimating the particle-size curve, i.e., the underlying CDF, at a given location, could be solved by smoothing the ECDF by Bernstein Polynomials.

The use of Bernstein Polynomials to approximate a bounded continuous function, such as F, is supported by the following result.

Theorem 3 (Feller (1965), Theorem 1, Section VII.2)**.** *If $u(x)$ is a bounded and continuous function on the interval $[0, 1]$, then*

$$u_m^*(x) = \sum_{k=0}^{m} u(k/m) b_k(m, x) \to u(x)$$

as $m \to \infty$*, uniformly for* $x \in [0,1]$*. Here,*

$$b_k(m,x) = \binom{m}{k} x^k (1-x)^{m-k}, \quad k = 0, ..., m.$$

On these premisses, Babu et al. (2002) propose and explore the asymptotic properties of the smooth estimator $\widetilde{F}_{\nu,m} : [0,1] \to [0,1]$ defined as:

$$\widetilde{F}_{\nu,m}(x) = \sum_{k=0}^{m} F_\nu(k/m) b_k(m,x), \quad x \in [0,1], \tag{13}$$

whose density can be explicitly computed as:

$$\widetilde{f}_{\nu,m}(x) = m \sum_{k=0}^{m-1} (F_\nu((k+1)/m) - F_\nu(k/m)) b_k(m-1,x), \quad x \in [0,1]. \tag{14}$$

As opposed to kernel smoothing estimators (Rosenblatt, 1956; Parzen, 1962; Silverman, 1986), the estimator (14) is well suited for distributions with compact support, of the kind associated with the particle-size curves we analyze. In the cases of the kind we consider, where available particle-size data consist of a discrete set of observations of the ECDF, $\{F_\nu(x_1), ..., F_\nu(x_N)\}$, taken at prescribed (transformed) diameters $\{x_1, ..., x_N\}$, it is not possible to directly employ estimator (13), since the ECDF is not known for $x \in [0,1] \setminus \{x_1, ..., x_N\}$. Therefore, we propose to consider a modified smooth estimator $F^N_{\nu,m} : [0,1] \to [0,1]$ based on a linear interpolant of the ECDF samples $F_\nu(x_1), ..., F_\nu(x_N)$ and defined as:

$$F^N_{\nu,m}(x) = \sum_{k=0}^{m} F^{(1)}_\nu(k/m) b_k(m,t), \quad x \in [0,1], \tag{15}$$

$F^{(1)}_\nu$ being the linear interpolant of $F_\nu(x_1), ..., F_\nu(x_n)$, i.e.:

$$F^{(1)}_\nu(x) = \sum_{i=1}^{N+1} (F_\nu(x_{i-1}) + \frac{F_\nu(x_i) - F_\nu(x_{i-1})}{x_i - x_{i-1}}(x - x_{i-1})) I_{(x_{i-1},x_i]}(t), \quad x \in [0,1]$$

with $x_0 = 0$, $x_{N+1} = 1$ and $F_\nu(x_0) = 0$, $F_\nu(x_{N+1}) = 1$. Adopting (15) enables one to estimate the CDF F through an approximation $F^{(1)}_\nu$ of the ECDF F_ν combined with Bernstein Polynomials. Note that, while other approximations for F_ν could be employed, the linear approximation we consider (a) provides a balance between the precision of the approximation and the complexity of the function (and thus the computational cost), and (b) allows deriving an explicit expression of the corresponding PDF, say $f^N_{\nu,m}$, according to:

$$\widetilde{f}^N_{\nu,m}(x) = m \sum_{k=0}^{m-1} (F^{(1)}_\nu((k+1)/m) - F^{(1)}_\nu(k/m)) b_k(m-1,x), \quad x \in [0,1]. \tag{16}$$

Moreover, denoting with $\|\cdot\|_{\mathcal{C}^0}$ the uniform norm on the space of continuous functions, the following result holds (the proof is reported in Appendix B).

Theorem 4. *Let* F *be a continuous CDF on* $[0,1]$ *and assume* F *to be differentiable in* $(0,1)$ *with associated PDF,* f*. Suppose there exists* $\alpha^* > 0$ *such* *that* $\lim_{(N,x)\to(+\infty,x_0)} f(x)/N^{\alpha^*} = 0$ *for* $x_0 \in \{0,1\}$ *and there exists* $N^\star \geq 1$, $0 \leq \eta < \infty$ *such that* $\max_{i\in\{1,\ldots,N+1\}}(x_i - x_{i-1}) < \eta/N^{\alpha^*}$ *for any* $N \geq N^\star$. *Then*

$$\lim_{m,\nu,N\to+\infty} \|\widetilde{F}^N_{\nu,m} - F\|_{\mathcal{C}^0} = 0, \quad a.s. \tag{17}$$

Theorem 4 states that $F^N_{\nu,m}$ is a strongly consistent estimator for F, provided that the sampling design is compatible with the growth rate of the PDF f when approaching the boundary of the support.

Note that $\alpha^\star$ is allowed to attain any (positive) constant value as long as f is continuous on $[0,1]$ (and thus bounded by virtue of Weierstrass theorem) and the condition on the sampling design becomes very weak. On the contrary, the condition on the sampling design becomes stronger when f is discontinuous in $x = 0$ or $x = 1$. This is due to the observation that the requirement for information content increases with the growth rate of the PDF. For example, let us consider a Beta distribution with density

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x)$$

where α, β are positive parameters, I denotes the indicator function and B is the beta function. If α or β are lower than one, then one can note that f is still in $A^2(0,1)$, but is unbounded near the boundary of the support. In this case, the estimator (15) is strongly consistent provided that $\alpha* > \max\{1-\alpha, 1-\beta\}$, i.e., the number of samples which are needed to describe the curves is required to increase with a rate which is at least equal to $N^{\alpha^\star}$. However, it is remarked that the occurrence of an unbounded particle-size density is virtually impossible in practical hydrogeological applications, thus rendering boundedness a viable assumption.

Additionally, one can note that Theorem 4 implies the weak convergence of $\widetilde{\mu}^N_{\nu,m}$ to μ, as $\nu, m, N \to \infty$, $\widetilde{\mu}^N_{\nu,m}$ and μ being the probability measures associated with $\widetilde{F}^N_{\nu,m}$ and F, respectively.

Finally, we remark that Theorem 4 yields useful indications about the design of an experiment, in the sense that it is conducive to the identification of the most appropriate curve sampling strategy yielding an optimal smoothing. This is a feature which is not fully exploited in this work but constitutes a critical application-oriented element of our methodology, especially considering the high level of precision associated with modern techniques employed to record particle-size data.

4.2 Smoothing of particle-size data

The estimator (15) has been applied to each raw particle-size curve depicted in Figure 1. Particle diameters are log-transformed, as they are approximately uniformly distributed between log(0.063) and log(100) [log(mm)], when considered on a log-scale. The support of the particle-size curves has been assumed to be compact, upon setting the data support as $\mathcal{T} = [\log(0.001), \log(200)]$, consistent with the type of lithology at the site.

The number of Bernstein Polynomials employed for the smoothing procedure has been selected according to the median sum of squared error (SSE) between raw data and smoothed particle-size curves evaluated at the 12 observed particle

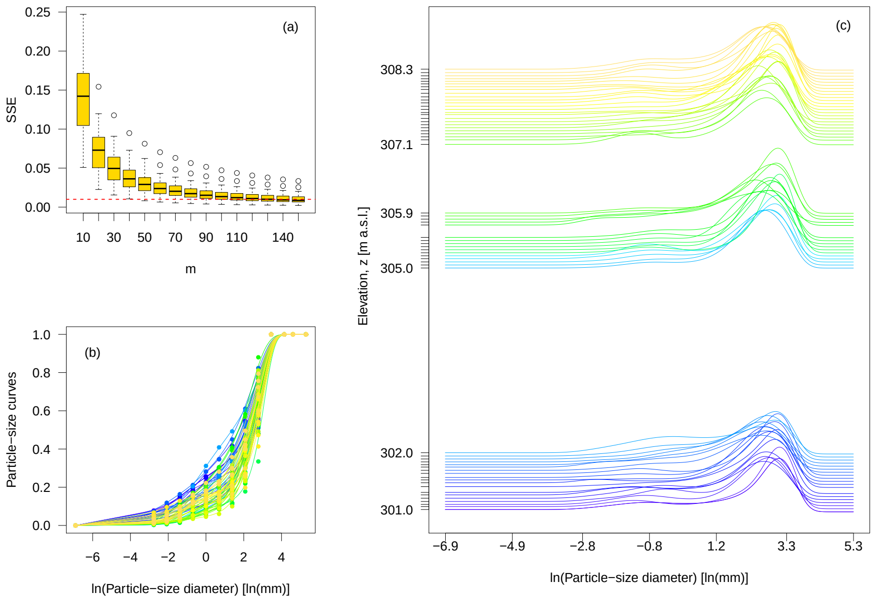

Figure 2: Smoothing procedure by Bernstein Polynomials. (a) boxplots of the SSE for $10 \leq m \leq 150$ (the threshold value of 0.01 is indicated by a dotted line); (b) raw particle-size curves (symbols) and particle-size curves smoothed by Bernstein Polynomials with $m = 140$ (solid lines); (c) vertical distribution of smoothed densities.

diameters. Figure 2a depicts boxplot of the SSE against the number of basis functions employed. No evident elbow in the median SSE appears in the figure. Therefore, the number of basis functions has been selected by setting a tolerance threshold of 0.01 (corresponding to $m = 140$) on the median SSE.

Figure 2b depicts the resulting smoothed curves (solid lines) juxtaposed to the available data (symbols). These results suggest that the overall features of the available dataset are well represented by the smoothing procedure. It can be noticed that the left tail of the distributions are associated with a quite uniform behavior, since the particle-size curves appear to display a linear dependence on the logarithm of the diameter. Note that direct observations are virtually absent at the left tails, as the smallest particle diameter recorded is equal to 0.063 mm. Hence, the observed uniform behavior of the smoothed curves can be considered as an artifact chiefly due to lack of a priori information on the left tail, leading to data censoring. This problem could eventually be circumvented upon adopting an improved experimental design, possibly based on the indications of Theorem 4.

Finally, Figure 2c depicts the vertical distribution of the particle-size densities computed according to (16) from the smoothed data reported in Figure 2b.

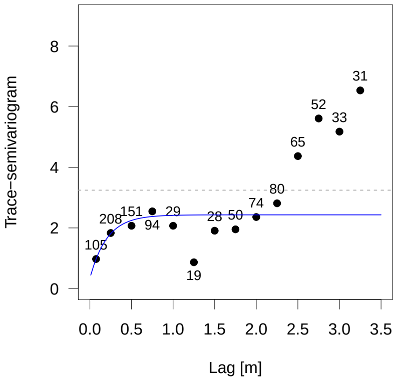

Figure 3: Estimated trace-semivariogram of the particle-size densities shown in Figure 2c: empirical trace-semivariogram (symbols), fitted model (solid line) and sample variance (dotted line). The number of pairs associated with each lag is reported

5 Results

5.1 Geostatistical analysis of the field data

Here, the notation introduced in Section 3 is employed as follows: quantities $\chi_{\boldsymbol{s}_1}, ..., \chi_{\boldsymbol{s}_n}$ denote the smoothed version of particle-size curves observed at locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ (solid lines in Figure 2b); $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$ indicate the smoothed particle-size densities depicted in Figure 2c and obtained as in (16). The functional dataset $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$ has been embedded into the space A^2 endowed with the Aitchison geometry and the methodology described in Section 3 has been coherently applied.

The stationarity assumption along the vertical direction is supported by prior knowledge of the field site (Riva et al., 2006, 2008, 2010); therefore, non-stationarity has not been considered in the present study. The structure of spatial dependence among the particle-size densities has been explored through the trace-semivariogram. The latter has been estimated from the data according to the discretized version of (10). Figure 3 depicts the empirical trace-semivariogram together with the selected fitted model. The empirical estimate displays a rapid growth up to a separation distance (lag) of about 0.6 m, where it stabilizes around a value of 2.4. The behavior displayed for the largest lags might be due to the decreasing number of data pairs available. On a cross-validation basis, an exponential structure (with calibrated partial sill of 2.09, practical range of 0.62 m, and nugget of 0.34) appeared to provide the most accurate results in terms of cross-validation SSE among different parametric semivariogram structures tested (spherical, hole and nested combinations).

Figure 4 depicts the results of the leave-one-out cross-validation procedure. Figure 4a shows the boxplot of the cross-validation SSE. The SSE for each sample $i = 1, ..., n$ has been computed as $\|\mathcal{Y}_{\boldsymbol{s}_i} \ominus \mathcal{Y}^{*(CV)}_{\boldsymbol{s}_i}\|^2_{A^2}$, $\mathcal{Y}^{*(CV)}_{\boldsymbol{s}_i}$ being the kriging prediction at $\boldsymbol{s}_i$ obtained upon removing the i-th datum (i.e., PSC) from

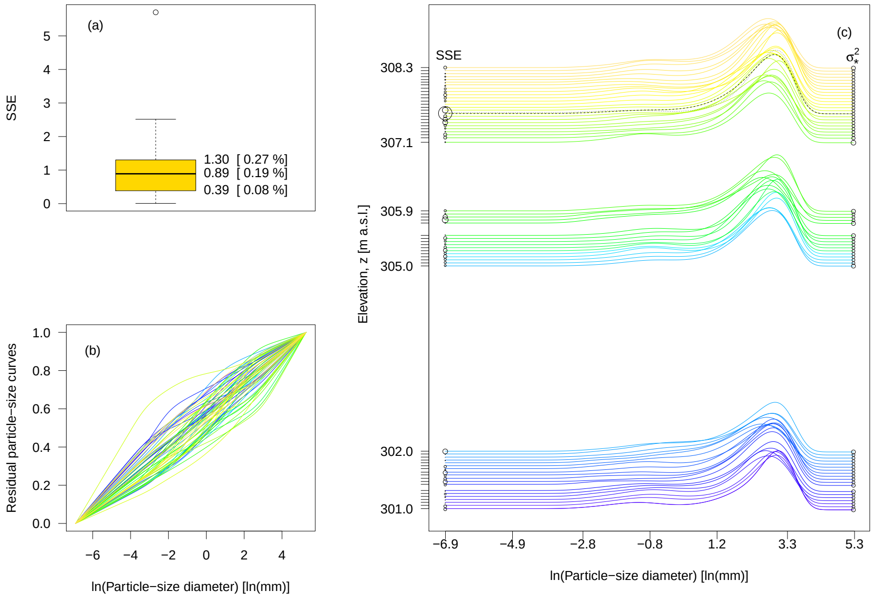

Figure 4: Cross-validation results. (a) boxplot of the SSE, reporting the absolute and relative quartile values; (b) cross-validation residual particle-size curves; (c) cross-validation prediction of particle-size densities as a function of elevation (the size of the symbols is proportional to the associated cross-validation SSE/kriging variance)

the dataset. The overall cross-validation error is very small when compared to the average squared norm of the data. One can note that both the median and the mean SSE are lower than 0.2% of the average squared norm of the data (median SSE: 0.986; mean SSE: 0.998). The spatial distribution of the SSE does not appear to be associated with a particular pattern, as evidenced by the seemingly random vertical distribution of the cross-validation SSE. Only one datum, corresponding to the vertical elevation $z = 307.53$ m and indicated with a dotted curve in Figure 4c is associated with a cross-validation SSE which is significantly larger than that of the remaining curves. This is due to a kriging prediction which is associated with a flattened peak of the particle-size density. With this exception, the key features of the data appear to be well reproduced by cross-validation predictions, with only a moderate smoothing effect.

All data but the PSC mentioned above are associated with a global prediction error which is lower than twice the kriging standard deviation. This result suggests that the 75% confidence bands constructed through the Čebyšëv inequality (9) tend to be quite conservative, being associated with an empirical level of 98.3%. The prediction provided by our proposed methodology appears to be overall unbiased, as shown by the cross-validation residual particle-size curves depicted in Figure 4b, which are fairly spread across a uniform cumulative distribution function (i.e., a straight line).

Prediction of the PSCs over a fine vertical grid with spacing of 1 mm has then been performed. Figure 5 depicts selected predicted particle-size curves

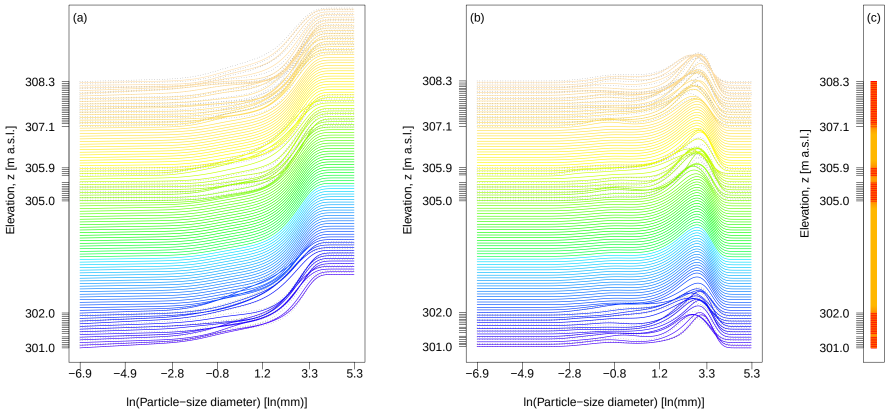

Figure 5: Vertical distribution of prediction results: (a) a sample of 100 of the 7190 predicted/kriged (solid curves) and observed (smoothed data; dotted curves) particle-size curves; (b) a sample of 100 of the 7190 predicted/kriged (solid curves) and observed (smoothed data; dotted curves) particle-size densities; (c) kriging variance. Kriging variance ranges between 0 (darkest color) and 2.53 (lightest color)

(Figure 5a), particle-size densities(Figure 5b) and the associated kriging variance (Figure 5c). The spatial prediction represents a smooth interpolation of the available data. Predictions follow the behavior of neighboring data for lags which are smaller than the calibrated trace-variogram range. Kriged curves tend to coincide with the estimated spatial mean (which is assumed to be constant) for greater lags. Hence kriged curves at unsampled locations which are far away from sampling points tend to be representative of a soil type which is associated with the mean particle-size curve.

5.2 Quantile assessment and hydraulic conductivity estimates

Knowing the estimates of the PSCs spatial distribution provides an exhaustive characterization of soil features which can be inferred from these curves. Our results enable one to provide estimates of desired particle-size quantiles to be employed, e.g., for facies identification, hydraulic conductivity assessment and/or geochemical parameters, at locations of interest. With reference to hydraulic conductivity estimates which can be inferred from particle-based formulations, here we compare the results which can be obtained through our functional compositional kriging approach against those associated with a classical kriging technique applied directly to quantiles of a PSC. These quantiles can be either directly measured or, as in (Riva et al., 2010), estimated through interpolation on the available measured particle sizes.

To this end, we remark that the proposed functional compositional kriging technique allows treating the complete set of information embedded in the available particle-size data within a framework based on global definitions of spatial dependence. On the other hand, classical approaches tend to characterize the spatial dependence of selected quantiles of the particle-size curve. In this sense, classical and functional approaches are markedly different from a methodological and application-oriented point of view. The functional approach allows modeling a global variogram for the functional process and the solution of the ensuing kriging system of equations is performed only once yielding the prediction (and associated prediction variance) of the complete particle-size curve at unsampled locations. On the other hand, typical geostatistical analyses (e.g., Riva et al. (2006, 2008, 2010); Bianchi et al. (2011) and references therein) treat each quantile separately (possibly introducing estimated cross-correlations in terms of cross-variograms) and project their predictions through kriging on a computational grid.

For the purpose of our application, we consider the log-transformed 10^{th} and 60^{th} quantiles of the particle-size distribution in $\boldsymbol{s}$, respectively indicated as $D_{10}(\boldsymbol{s})$ and $D_{60}(\boldsymbol{s})$, i.e.

$$D_{10}(\boldsymbol{s}) = \chi_{\boldsymbol{s}}^{-1}(0.10); \quad D_{60}(\boldsymbol{s}) = \chi_{\boldsymbol{s}}^{-1}(0.60); \quad \boldsymbol{s} \in D \tag{18}$$

or, equivalently,

$$\int_{t_m}^{D_{10}(\boldsymbol{s})} \mathcal{Y}_{\boldsymbol{s}}(t)dt = 0.10; \quad \int_{t_m}^{D_{60}(\boldsymbol{s})} \mathcal{Y}_{\boldsymbol{s}}(t)dt = 0.60; \quad \boldsymbol{s} \in D.$$

We remark that, for consistency, both classical and functional compositional geostatistical analyses are here performed on the quantities $D_{10}(\boldsymbol{s}_1), ..., D_{10}(\boldsymbol{s}_n)$ and $D_{60}(\boldsymbol{s}_1), ..., D_{60}(\boldsymbol{s}_n)$, i.e., the values associated with the $n = 60$ smoothed particle-size curves $\chi_{\boldsymbol{s}_1}, ..., \chi_{\boldsymbol{s}_n}$ obtained according to (18) (empty symbols in Figure 6a).

A classical study of the structure of spatial dependence of these log-quantiles is performed upon modeling the variograms of D_{10} and D_{60}. The cross-variogram has not been modeled because of the lack of cross-correlation between 10^{th} and 60^{th} log-quantiles at the site (Riva et al., 2010). Figure 6a and b depict the estimated empirical semivariograms (full symbols) together with the fitted valid models (solid curves). An exponential structure with nugget has been selected for both quantities. Variogram calibration results highlight that D_{10} shows a much higher variability than D_{60} (estimated sill is 0.58 and 0.04, with estimated nugget of 0.13 and 0.015, respectively for D_{10} and D_{60}). On the other hand, the range of the variogram of D_{10} appears to be about twice the one associated with D_{60} (estimated practical range is 0.62 and 0.28, respectively for D_{10} and D_{60}). These results are consistent with those obtained by Riva et al. (2010) who performed a geospatial analysis of D_{10} and D_{60} by considering all boreholes at the site, having clustered the data into two main soil types.

The fitted variogram structures reported in Figure 6a and b have been validated by means of a leave-one-out cross-validation analysis. Cross-validation predictions are reported in Figure 6e (crosses) together with the log-quantiles predictions obtained by the cross-validation predicted particle-size curves, computed according to (18) (solid circles). Kriging predictions obtained with the classical and functional compositional approaches appear to be very similar, displaying a moderate smoothing effect in both cases.

Table 1 lists the cross-validation SSEs associated with classical one-dimensional (1D K) and functional compositional kriging (FCK). For completeness

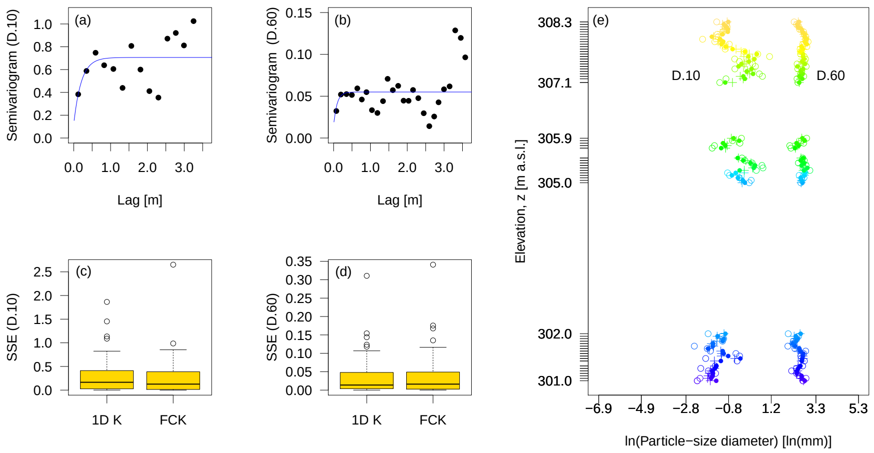

Figure 6: Comparison of cross-validation results of classical one-dimensional kriging (1D K) and functional compositional kriging (FCK): empirical variogram (symbols) and fitted models (solid curves) for (a) D_{10} and (b) D_{60}; boxplots of cross-validation SSEs for (c) D_{10} and (d) D_{60}; (e) D_{10} and D_{60} data (empty circles) together with cross-validation predictions with 1D K (crosses) and FCK (solid circles)

and ease of reference, these are also depicted in Figures 6c and d. Cross-validation results, as expressed by SSEs, appear to be comparable for the two approaches, as one can also notice by visual inspection of Figure 6e. The log-quantile D_{10} proves to be much more difficult to be predicted than D_{60}, due to its higher spatial variability. In this case, FCK yields slightly improved results in terms of SSEs. This might be due to the global nature of the approach embedded in FCK, which grounds its strength on the reliance on the entire curve for the prediction of local behaviors.

Finally, (log)hydraulic conductivities have been computed from cross- validation predictions. We recall that methods based on particle-size information to provide estimates of hydraulic conductivity, K, rely on formulations of the kind:

$$K = \frac{g}{v} C\, f(\phi)\, d_e^2 \tag{19}$$

where g is gravity, v is the fluid kinematic viscosity, $f(\phi)$ is a function of porosity, ϕ, d_e is an effective particle diameter, and C is defined as a sorting coefficient. The particular values of C and d_e, and the form of $f(\phi)$ depend on the formulation one employs. Empirical formulations which are usually adopted to obtain hydraulic conductivity from quantiles of particle-size curves of soil samples are collected by e.g., Vukovic and Soro (1992); Fetter (2001); Carrier (2003); Odong (2007).

Here, we consider two widely used formulations, corresponding to the Kozeny-Carman and Hazen equations. According to the Kozeny-Carman equation:

$$C = 8.3 \cdot 10^{-3}; \quad f(\phi) = \left[\frac{\phi^3}{(1-\phi)^2}\right]; \quad d_e = d_{10}. \tag{20}$$

	Method	Median SSE [%]	Mean SSE [%]
D_{10}	FCK	0.13 [9.5%]	0.28 [20.99%]
	1D K	0.17 [12.40%]	0.30 [22.15%]
D_{60}	FCK	$1.62 \cdot 10^{-2}$ [0.24%]	$3.76 \cdot 10^{-2}$ [0.56%]
	1D K	$1.38 \cdot 10^{-2}$ [0.21%]	$3.67 \cdot 10^{-2}$ [0.55%]
$\ln(K^{[KC]})$	FCK	0.50 [16.36%]	1.10 [35.98%]
	1D K	0.63 [20.76%]	1.16 [37.95%]
$\ln(K^{[H]})$	FCK	0.50 [13.59%]	1.11 [30.02%]
	1D K	0.65 [17.61%]	1.18 [32.70%]

Table 1: Comparison between cross-validation results related to quantile (D_{10} and D_{60}) and log-hydraulic conductivity assessment when considering Kozeny-Carman ($\ln(K^{[KC]})$) or Hazen ($\ln(K^{[H]})$) equations

Here, d_{10} is the particle diameter (in mm) associated with the 10% quantile of the particle-size curve and K is given in m/day. Estimates of ϕ can be obtained by (e.g., Vukovic and Soro (1992))

$$\phi = 0.255(1 + 0.83^U); \quad U = \left(\frac{d_{60}}{d_{10}}\right) \tag{21}$$

d_{60} being the 60% quantile of the particle-size curve. The Hazen equation is:

$$C = 6 \cdot 10^{-4}; \quad f(\phi) = 1 + 10(\phi - 0.26). \tag{22}$$

Note that log-hydraulic conductivity values at $\boldsymbol{s} \in D$ can be computed in both cases by a linear combination of the log-quantiles $D_{10}(\boldsymbol{s})$ and $D_{60}(\boldsymbol{s})$. Therefore, the BLU prediction of the log-hydraulic conductivities can be obtained from the BLU prediction of D_{10} and D_{60}, i.e., from the kriged log-quantiles.

The last rows of Table 1 reports the cross-validation median and mean SSE related to log-hydraulic conductivities computed by the Kozeny-Carman ($\ln(K^{[KC]})$) and Hazen ($\ln(K^{[H]})$) formulations. Functional compositional kriging provides improved results with respect to classical one-dimensional kriging in both cases. This might be due to the structure of the formulations considered and implies that the improvement in the D_{10} SSE is conducive to a corresponding improvement in log-hydraulic conductivities SSE, even though D_{60} appears to be slightly better predicted by classical kriging.

Finally, Figure 7 shows the predictions of the log-quantiles D_{10} and D_{60} (panel a) and the log-hydraulic conductivities $\ln(K^{[KC]})$ and $\ln(K^{[H]})$ (panels b and c, respectively), computed by traditional one-dimensional (dotted curves) and functional compositional (solid lines) kriging approaches. Predictions appear to be almost indistinguishable for quantiles and log-hydraulic conductivities. These results indicate that our proposed methodology (a) leads to the complete characterization of the soil textural properties and (b) proves to be fairly precise in predicting the local features of particle-size distributions, by means of a relatively simple procedure.

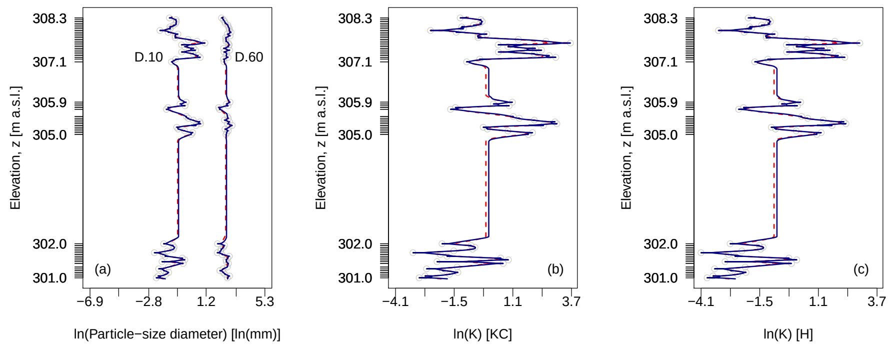

Figure 7: Comparison between kriging predictions obtained by 1D K (dotted curves) and FK (solid curves): (a) D_{10} and D_{60}; log-hydraulic conductivity based on (b) Kozeny-Carman [KC] and (c) Hazen [H] formulations. Data are indicated with symbols

6 Conclusions and further research

The main contributions of our work are both theoretical and application-oriented and our research leads to the following key conclusions.

1. Particle-size curves (PSCs), which constitute a typical information content employed in hydrogeology, soil science and geochemical applications, have been interpreted as functional compositional data. An original and general geostatistical methodology which enables one to treat spatially dependent functional compositional data has been proposed. Our approach rests on a kriging technique which is developed for variables belonging to general Hilbert spaces and that we have embedded in the space A^2 endowed with the Aitchison geometry. We investigate the relationship between the spaces A^2 and L^2 in view of bringing the theory to practical applications.

2. As PSCs are typically sampled at a discrete set of particle diameters, a smoothing method based on Bernstein Polynomials has been proposed (Section 4) and its consistency has been proven. In practical applications, different choices might be employed for data preprocessing or, in some instances, this preliminary data treatment might not be required. When the full PSC is available or is sampled through a fine resolution, the methodology we developed (Section 3) can be directly applied to the available data, without resorting to the procedure presented in Section 4.

3. Our methodology is demonstrated through an application relying on 60 PSCs sampled along a borehole within an alluvial aquifer near the city of Tübingen, Germany. On a cross-validation basis, the results obtained through our functional compositional kriging procedure proved to be conducive to satisfactory predictions (and associated uncertainty quantification) of PSCs at unsampled spatial locations.

4. Our approach enables one to provide estimates of desired quantiles of PSCs to be employed for hydraulic conductivity assessment at locations of interest. We compared the results which can be obtained with our functional compositional kriging approach against those associated with a classical kriging technique applied directly to quantiles which are either observed directly or, as in the current application, obtained through interpolation of the available particle-size data. We found the two methods to lead to consistent results, with a slightly improved performance of the functional compositional kriging on the basis of cross-validation results.

5. A key advantage of our functional approach to compositional data lies in the possibility of obtaining predictions of the entire particle-size curve at unsampled locations, as opposed to classical or compositional kriging techniques which allow capturing only selected local features of the curve. The information content provided by the full PSC is critical to proper modeling several physical and chemical processes occurring in heterogeneous earth systems and which are affected by the local composition of the host soil/rock matrix. In the light of the theoretical developments and results presented, further advancements include three-dimensional extensions to provide kriging predictions and stochastic simulation of PSCs associated with different soil types. In these scenarios, anisotropic and (possibly) non-stationary approaches are likely to be required to precisely characterize the heterogeneous (stochastic) nature of the particle-size curves within a given aquifer system.

Appendix A: the non-stationary case

In Subsection 3.1, second-order stationarity has been assumed in view of the particular application studied. However, the non-stationary case could be dealt with as well, by exploiting the estimators and the algorithms proposed in (Menafoglio et al., 2012).

In such a case, a linear model for the drift has to be formulated:

$$m_{\boldsymbol{s}} = \bigoplus_{l=0}^{L} f_l(\boldsymbol{s}) \odot a_l, \boldsymbol{s} \in D$$

and trace-covariogram and trace-variogram are to be defined in terms of the residual process $\{\delta_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, which is second-order stationary. The Universal Kriging predictor can be derived by solving the minimization problem:

$$(\lambda_1^*, ..., \lambda_n^*) = \underset{\substack{\lambda_1,...,\lambda_n \in \mathbb{R}: \\ \mathcal{Y}_{\boldsymbol{s}_0}^{\boldsymbol{\lambda}} = \bigoplus_{i=1}^{n} \lambda_i \odot \mathcal{Y}_{\boldsymbol{s}_i}}}{\operatorname{argmin}} \quad \mathrm{Var}_{A^2}(\mathcal{Y}_{\boldsymbol{s}_0}^{\boldsymbol{\lambda}} \ominus \mathcal{Y}_{\boldsymbol{s}_0}) \quad \text{s.t.} \quad \mathbb{E}_{A^2}[\mathcal{Y}_{\boldsymbol{s}_0}^{\boldsymbol{\lambda}}] = m_{\boldsymbol{s}_0},$$

which reduces to the linear system:

$$\left(\begin{array}{c|c} C(\boldsymbol{h}_{i,j}) & f_l(\boldsymbol{s}_i) \\ \hline f_l(\boldsymbol{s}_j) & 0 \end{array}\right) \left(\begin{array}{c} \lambda_i \\ \hline \zeta_l \end{array}\right) = \left(\begin{array}{c} C(\boldsymbol{h}_{0,i}) \\ \hline f_l(\boldsymbol{s}_0) \end{array}\right), \tag{23}$$

where $\zeta_0, ..., \zeta_L$ are $L+1$ Lagrange multipliers associated with the unbiasedness constraint. System (23) admits a unique solution provided $\Sigma = (C(\boldsymbol{h}_{ij}))$ is positive definite and $\mathbb{F} = (f_l(\boldsymbol{s}_i)) \in \mathbb{R}^{n \times (L+1)}$ is of full-rank.

The trace-semivariogram is required to be known or properly estimated to solve the Universal Kriging system. To this end, estimator (4) should not be used, because it can be severely biased if the mean function $m_{\boldsymbol{s}}$ is not spatially constant. Instead, a natural estimator for the trace-semivariogram is the (possibly discretized) MoM estimator from the residuals, i.e., following the notation introduced in Section 3,

$$\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{2|N(\boldsymbol{h})|} \sum_{(i,j)\in N(\boldsymbol{h})} \|\delta_{\boldsymbol{s}_i} \ominus \delta_{\boldsymbol{s}_j}\|^2_{A^2}.$$

Nevertheless, one is required to estimate the residuals $\delta_{\boldsymbol{s}_1}, ..., \delta_{\boldsymbol{s}_n}$. In the general context of data belonging to any Hilbert Space, Menafoglio et al. (2012) propose to estimate the residuals as a difference between observations and the generalized least squares (GLS) estimates of the drift at the sampled locations, i.e., in our setting, $\widehat{\delta}_{\boldsymbol{s}_i} = \mathcal{Y}_{\boldsymbol{s}_i} \ominus \widehat{m}^{GLS}_{\boldsymbol{s}_i}$, $i = 1, ..., n$. Furthermore, Menafoglio et al. (2012) derive the explicit expression of the GLS drift estimator and analyze its properties (in particular, it is proved that the GLS drift estimator is also the BLU estimator for the mean). Embedding the results of Menafoglio et al. (2012) within our framework yields to the following expression of the GLS drift estimator:

$$\widehat{\boldsymbol{m}}^{GLS}_{\boldsymbol{s}} = \mathbb{F}(\mathbb{F}^T\Sigma^{-1}\mathbb{F})^{-1}\mathbb{F}^T\Sigma^{-1} \odot \boldsymbol{\mathcal{Y}}_{\boldsymbol{s}}. \tag{24}$$

having adopted the vectorial notation: $(\mathbb{A} \odot f)_i = \bigoplus_{j=1}^n \mathbb{A}_{i,j} \odot f_j$, $\mathbb{A} = (\mathbb{A}_{ij}) \in \mathbb{R}^{n,n}$, $\boldsymbol{f} = (f_i)$, $f_i \in A^2$, $i = 1, 2, ..., n$. Note that the optimality of estimator (24) relies on a properly accounting for the structure of spatial dependence Σ, which is unknown. In order to cope with this problem, an iterative algorithm starting from an ordinary least squares estimate of the drift ($\widehat{\boldsymbol{m}}^{OLS}_{\boldsymbol{s}} = \mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}\mathbb{F}^T \odot \boldsymbol{\mathcal{Y}}_{\boldsymbol{s}}$) can be employed (Menafoglio et al. (2012), Section 4-5).

Therefore, even though our application relies on a stationary setting, the methodology we present could also be used to treat non-stationary settings upon applying the more general procedure which has been briefly illustrated in this Appendix.

Appendix B: proof of Theorem 4

Proof. First, write

$$\|\widetilde{F}^N_{\nu,m} - F\|_{\mathcal{C}^0} \leq \|\widetilde{F}^N_{\nu,m} - \widetilde{F}_{\nu,m}\|_{\mathcal{C}^0} + \|\widetilde{F}_{\nu,m} - F\|_{\mathcal{C}^0}. \tag{25}$$

The last term of (25) vanishes as $\nu, m \to +\infty$ (Babu et al. (2002), Theorem 2.1). Let us consider the second term and write its argument as:

$$\widetilde{F}^N_{\nu,m}(x) - \widetilde{F}_{\nu,m}(x) = \sum_{k=0}^{m} (F^{(1)}_\nu(k/m) - F_\nu(k/m)) b_k(m, x), \quad x \in [0, 1].$$

It is straightforward to see that, for $x \in [0, 1]$

$$F^{(1)}_\nu(x) - F_\nu(x) = \frac{1}{2}\sum_{i=1}^{N} [(F_\nu(x_i) - F_\nu(x)) - (F_\nu(x) - F_\nu(x_{i-1}))] I_{(x_{i-1},x_i]}(x).$$

Fix $\varepsilon > 0$, consider ν such that $\|F - F_\nu\| < \varepsilon/3$ and $N > N^\star$ such that $f(x)/N^\alpha < \frac{\varepsilon}{3\eta}$ for any $x \in [0,1]$. Then:

$$
\begin{aligned}
&\|\widetilde{F}^N_{\nu,m} - \widetilde{F}_{\nu,m}\|_{\mathcal{C}^0} = \max_{0\le k\le m} \left| F^{(1)}_\nu(k/m) - F_\nu(k/m)\right| \le \\
&\quad\le \max_{0\le k\le m} \left\{ \frac{1}{2} \sum_{i=1}^{N+1} \left|(F_\nu(x_i) - F_\nu(k/m)) - (F_\nu(k/m) - F_\nu(x_{i-1}))\right| I_{(x_{i-1},x_i]}(k/m) \right\} \le \\
&\quad\le \max_{0\le k\le m} \left\{ \frac{1}{2} \sum_{i=1}^{N+1} \left|4\varepsilon + (F(x_i) - F(k/m)) - (F(k/m) - F(x_{i-1}))\right| I_{(x_{i-1},x_i]}(k/m) \right\} \le \\
&\quad\le \max_{0\le k\le m} \left\{ \frac{1}{2} \left|4\varepsilon + (F(k/m + \eta/N^{\alpha^*}) - F(k/m)) - (F(k/m) - F(k/m - \eta/N^{\alpha^*}))\right| \right\} = \\
&\quad= 2\varepsilon + \max_{0\le k\le m} \left| \frac{\eta}{2N^{\alpha^*}} f(\widetilde{x}^{(k)}_1) - \frac{\eta}{2N^{\alpha^*}} f(\widetilde{x}^{(k)}_2)\right| < \varepsilon
\end{aligned}
$$

for some $\widetilde{x}^{(k)}_1 \in [k/m, k/m + \eta/N^{\alpha^*}]$, $\widetilde{x}^{(k)}_2 \in [k/m - \eta/N^{\alpha^*}, k/m]$. The thesis then follows from the arbitrariness of ε. □

Acknowledgements Funding from MIUR (Italian Ministry of Education, Universities and Research - PRIN2010-11; project: "Innovative methods for water resources under hydro-climatic uncertainty scenarios") is acknowledged.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological) 44*(2), 139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall.

Babu, G., A. Canty, and Y. Chaubey (2002). Application of Bernstein Polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference 105*, 377–392.

Barahona-Palomo, M., M. Riva, X. Sanchez-Vila, E. Vazquez-Sune, and A. Guadagnini (2011). Quantitative comparison of impeller flowmeter and particle-size distribution techniques for the characterization of hydraulic conductivity variability. *Hydrogeology Journal 19*(3), 603–61. doi:10.1007/s10040-011-0706-5.

Bianchi, M., C. Zheng, C. Wilson, G. Tick, G. Liu, and S. M. Gorelick (2011). Spatial connectivity in a highly heterogeneous aquifer: From cores to preferential flow paths. *Water Resour. Res. 47*, W05524.

Buchanan, S., J. Triantafilis, I. Odeh, and R. Subansinghe (2012). Digital soil mapping of compositional particle-size fractions using proximal and remotely sensed ancillary data. *Geophysics 77*(4), WB201–WB211. doi: 10.1190/GEO2012-0053.1.

Caballero, W., R. Giraldo, and J. Mateu (2013). A universal kriging approach for spatial functional data. *Stochastic Environmental Research and Risk Assessment*, 1–11.

Carrier, W. (2003). Goodbye, hazen; hello, kozeny-carman. *J Geotech Geoenviron Eng 129*(11), 1054–1056.

Delicado, P., R. Giraldo, C. Comas, and J. Mateu (2010). Statistics for spatial functional data. *Environmetrics 21*(3-4), 224–239.

Deutsch, C. and A. Journel (1997). *GSLIB: Geostatistical software library and user's guide* (second ed.). Oxford University Press, UK.

Egozcue, J., J. Díaz-Barrero, and V. Pawlowsky-Glahn (2006, Jul.). Hilbert space of probability density functions based on aitchison geometry. *Acta Mathematica Sinica, English Series 22*(4), 1175–1182.

Feller, W. (1965). *An introduction to probability theory and its applications*, Volume II. Wiley, New York.

Ferraty, F. and P. Vieu (2006). *Nonparametric functional data analysis : theory and practice.* Springer, New York.

Fetter, C. (2001). *Applied Hydrogeology.* Englewood Cliffs, New Jersey.

Goulard, M. and M. Voltz (1993). Geostatistical interpolation of curves: A case study in soil science. In A. Soares (Ed.), *Geostatistics Tróia '92*, Volume 2, pp. 805–816. Dordrecht: Kluwer Academic.

Guadagnini, L., A. Guadagnini, and D. Tartakovsky (2004). Probabilistic reconstruction of geologic facies. *J. Hydrol. 294*, 57–67.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications.* Springer Series in Statistics. Springer.

Hu, B., J. Wu, and D. Zhang (2004). A numerical method of moments for solute transport in physically and chemically nonstationary formations: linear equilibrium sorption with random k_d. *Stoch Environ Res Risk Assess 18*, 22–30. doi:10.1007/s00477-003-0161-5.

Leininger, T., A. Gelfand, J. Allen, and J. Silander (2013). Spatial regression modeling for compositional data with many zeros. *Journal of Agricultural, Biological, and Environmental Statistics*, 1–21.

Lemke, L. and L. Abriola (2003). Predicting dnapl entrapment and recovery: the influence of hydraulic property correlation. *Stoch Environ Res Risk Assess 17*, 408–418. doi:10.1007/s00477-003-0162-4.

Martac, E. and T. Ptak (2003). Data sets for transport model calibration/validation, parameter upscaling studies and testing of stochastic transport models/theory. Report D16 of Project "Stochastic Analysis of Well-Head Protection and Risk Assessment - W-SAHaRA", EU contract EVK1-CT-1999-00041, Milan, Italy.

Menafoglio, A., M. Dalla Rosa, and P. Secchi (2012). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. MOX-report 34/2012, Politecnico di Milano.

Nemes, A., M. Schaap, and J. WÃűsten (2003). Functional evaluation of pedotransfer functions derived from different scales of data collection. *Soil Science Society of America Journal 67*, 1093–1102.

Nerini, D., P. Monestiez, and C. Manté (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis 101*(2), 409–418.

Neuman, S., A. Blattstein, M. Riva, D. Tartakovsky, A. Guadagnini, and T. Ptak (2007). Type curve interpretation of late-time pumping test data in randomly heterogeneous aquifers. *Water Resour. Res. 43*(10), W10421.

Neuman, S., M. Riva, and A. Guadagnini (2008). On the geostatistical characterization of hierarchical media. *Water Resources Research 44*(2), W02403.

Odeh, I., A. Todd, and J. Triantafilis (2003). Spatial prediction of soil particle-size fractions as compositional data. *Soil Science 168*(7), 501–515. doi: 10.1097/00010694-200307000-00005.

Odong, J. (2007). Evaluation of empirical formulae for determination of hydraulic conductivity based on grain-size analysis. *J Am Sci 3*(3), 54–60.

Pachepsky, Y. and W. Rawls (2004). *Development of pedotransfer functions in soil hydrology*, Volume 30 of *Developments in Soil Science*. Elsevier, Amsterdam.

Parzen, E. (1962). On estimation of probability density and mode. *Ann. Math Statist. 33*, 1965–1070.

Pawlowsky-Glahn, V. and A. Buccianti (2011). *Compositional data analysis. Theory and applications*. Wiley.

Pawlowsky-Glahn, V. and J. Egozcue (2001). Geometric approach to statistical analysis in the symplex. *Stochastic Environmental Research and Risk Assessment 15*, 384–398.

Pawlowsky-Glahn, V. and J. Egozcue (2002). BLU Estimators and Compositional Data. *Mathematical geology 34*(3).

Pawlowsky-Glahn, V. and R. Olea (2004). *Geostatistical analysis of compositional data*. Oxford university press.

Ramsay, J. and C. J. Dalzell (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society 53*(3), 539–572.

Ramsay, J. and B. Silverman (2005). *Functional data analysis* (Second ed.). Springer, New York.

Riva, M., A. Guadagnini, D. Fernandez-Garcia, X. Sanchez-Vila, and T. Ptak (2008). Relative importance of geostatistical and transport models in describing heavily tailed breakthrough curves at the lauswiesen site. *J Contam Hydrol 101*, 1–13.

Riva, M., L. Guadagnini, and 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. doi:10.1007/s00477-010-0399-7.

Riva, M., L. Guadagnini, A. Guadagnini, T. Ptak, and E. Martac (2006). Probabilistic study of well capture zones distributions at the lauswiesen field site. *J Contam Hydrol 88*, 92–118.

Rosenblatt, M. (1956). Remarks on ome nonparametric estimated of density functions. *Ann. Math Statist. 27*, 832–837.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Tartakovsky, D., B. Wohlberg, and A. Guadagnini (2007). Nearest-neighbor classification for facies delineation. *Water Resour. Res. 43*, W07201. doi:10.1029/2007WR005968.

Tolosana-Delgado, R., J. Egozcue, A. Sánchez-Arcilla, and J. Gómez (2011). Classifying wave forecasts with model-based geostatistics and the aitchison distribution. *Stochastic Environmental Research and Risk Assessment 25*(8), 1091–1100.

Tolosana-Delgado, R., K. van den Boogaart, and V. Pawlowsky-Glahn (2011). *Geostatistics for Compositions*, pp. 73–86. John Wiley & Sons, Ltd.

Tong, J., B. . Hu, and J. Yang (2010). Using data assimilation method to calibrate a heterogeneous conductivity field conditioning on transient flow test data. *Stoch Environ Res Risk Assess 24*, 1211–1223. doi:10.1007/s00477-010-0392-1.

van den Boogaart, K., J. Egozcue, and V. Pawlowsky-Glahn (2010). Bayes linear spaces. *SORT 34*(2), 201–222.

Vukovic, M. and A. Soro (1992). Determination of hydraulic conductivity of porous media from grain-size composition. Littleton, Colorado.

Wohlberg, B., D. Tartakovsky, and A. Guadagnini (2006). Subsurface characterization with support vector machines. *IEEE Trans. on Geoscience and Remote Sensing 44*(1), 47–57. doi: 10.1109/TGRS.2005.859953.

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

32/2013 Taddei, T.; Perotto, S.; Quarteroni, A.
Reduced basis techniques for nonlinear conservation laws

31/2013 Dassi, F.; Ettinger, B.; Perotto, S.; Sangalli, L.M.
A mesh simplification strategy for a spatial regression analysis over the cortical surface of the brain

30/2013 Cagnoni, D.; Agostini, F.; Christen, T.; de Falco, C.; Parolini, N.; Stevanović, I.
Multiphysics simulation of corona discharge induced ionic wind

29/2013 Lassila, T.; Manzoni, A.; Quarteroni, A.; Rozza, G.
Model order reduction in fluid dynamics: challenges and perspectives

28/2013 Ekin, T.; Ieva, F.; Ruggeri, F.; Soyer, R.
Statistical Issues in Medical Fraud Assessment

27/2013 Tagliabue, A.; Dede', L.; Quarteroni, A.
Isogeometric Analysis and Error Estimates for High Order Partial Differential Equations in Fluid Dynamics

24/2013 Mazzieri, I.; Stupazzini, M.; Guidotti, R.; Smerzini, C.
SPEED-SPectral Elements in Elastodynamics with Discontinuous Galerkin: a non-conforming approach for 3D multi-scale problems

25/2013 Cattaneo, Laura; Zunino, Paolo
Computational models for coupling tissue perfusion and microcirculation

26/2013 Ieva, F.; Paganoni, A.M.
Detecting and visualizing outliers in provider profiling via funnel plots and mixed effect models

23/2013 Srensen, H.; Goldsmith, J.; Sangalli, L.M.
An introduction with medical applications to functional data analysis