ORIGINAL PAPER

A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers

Alessandra Menafoglio · Alberto Guadagnini · Piercesare Secchi

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 PSCs 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 geostatistical 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 PSCs 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 PSCs and enables one to project the complete information content embedded in the PSC to unsampled locations in the system.

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 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

A. Menafoglio (✉) · P. Secchi
MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy
e-mail: alessandra.menafoglio@polimi.it

A. Guadagnini
Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Milano, Italy

A. Guadagnini
Department of Hydrology and Water Resources, The University of Arizona, Tucson, AZ, USA

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 PSCs 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 PSCs 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 $10th$ and $60th$ quantiles of the observed PSCs 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 $10th$ and $60th$ 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 PSCs 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 PSCs. To this end, we model PSCs as cumulative distribution functions (CDFs) and analyze their derivatives, by coherently considering them as probability density functions (PDFs). 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); Aitchison (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 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. Thus, a B-part composition has B non-negative components, constrained to sum up to a constant (usually set to unity or 100) and belongs to a $(B - 1)$-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. 2013; 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 particle-size data.

Here, we focus specifically on the formulation of new geostatistical models and methods for functional compositional data. The problem of kriging PDFs has been considered also by Salazar Buelvas (2011), who exploits a

logarithmic transformation to deal with the data constraint. Our approach is however different and move from the geostatistical methodology proposed in (Menafoglio et al. 2013) 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); Pawlowsky-Glahn and Egozcue (2002)]. Here, we employ the Aitchison geometry within a Hilbert Space method to accommodate both the functional and the 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 Sect. 2, our methodology is general and allows performing a 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 Sect. 3. We 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 1.

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 Sect. 4 and Appendix 2, respectively. Section 5 illustrates applications of our functional compositional kriging (FCK) 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
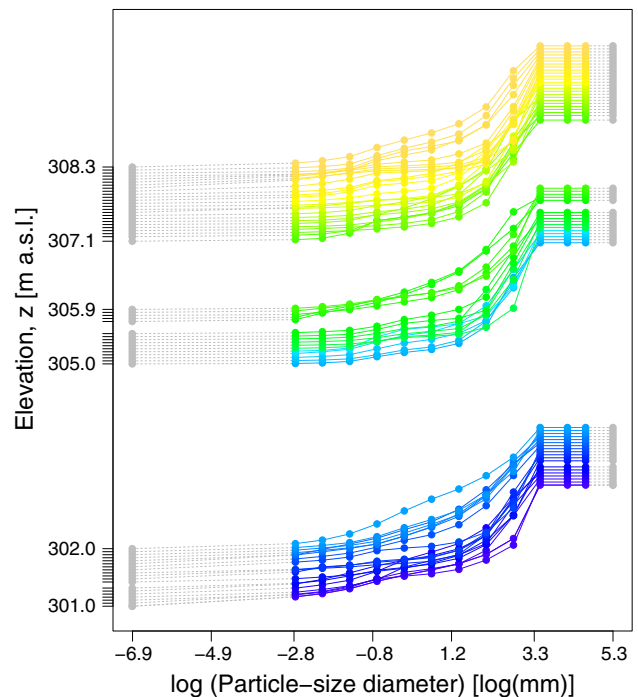

Fig. 1 Available particle-size data. Vertical coordinates correspond to the data sampling locations. Data are supported on the compact domain ($d_{min} = 0.001$ mm; $d_{max} = 200$ mm); d_{min} and d_{max} are the smallest and largest measured particle-size diameters, respectively; elevation is given in meters above sea level (m a.s.l.)

analysis. Available data include PSCs, 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 PSCs 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 PSCs. 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 PSCs collected along 12 vertical boreholes are available within the site. PSCs 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 PSCs 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 PSCs which were collected in the quaternary sand and gravel depositional unit at well B5 at the site. Figure 1 depicts the set of PSCs 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 $\{\chi_s, s \in D\}, D \subset \mathbb{R}^d$, whose elements are PSCs. Each element $\chi_s, s \in D$, is a [0,1] valued random function defined on a compact domain $\mathcal{T} = [t_m, t_M]$, i.e., χ_s is measurable and, for $\omega \in \Omega, \chi_s(\omega, \cdot) : \mathcal{T} \to [0,1]$. Given a particle size $t \in \mathcal{T}, \chi_s(\cdot, t)$ indicates the random fraction of grains with diameter smaller than or equal to t. Hence, each function $\chi_s(\omega, \cdot)$ is a 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 CDFs is not closed with respect to such operations (for instance, the point-wise sum of two CDFs is not a CDF). Instead, an approach based on the 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 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 $\{\mathcal{Y}_s, s \in D\}$, defined on the probability space introduced above and such that, for $s \in D, \omega \in \Omega$:

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

We assume that, for $\omega \in \Omega, \mathcal{Y}_s(\omega, \cdot) = d\chi_s(\omega, \cdot)/dt$ is the density function of the random probability measure μ defined, for all $a \leq b$, by:

$$\mu_s(\omega, (a, b]) = \chi_s(\omega, b) - \chi_s(\omega, a).$$

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

Let us denote with $L^2(\mathcal{T})$ the space of (equivalence classes of) real functions on $\mathcal{T}$ that are square-integrable with respect to the Lebesgue measure,

$$L^2(\mathcal{T}) = \left\{ f : \mathcal{T} \to \mathbb{R}, \text{ such that } \int_{\mathcal{T}} |f(t)|^2 dt < \infty \right\},$$

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

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

In this work we assume that $\mathcal{Y}_s(\omega, \cdot) \in A^2(\mathcal{T})$ for all $s \in D, \omega \in \Omega$.

Egozcue et al. (2006) provide A^2 with a Hilbert space structure on the basis of its relationship with the space ℓ^2 of square-summable real sequences, endowed with the inner product $\langle x, y \rangle = \sum_{i \geq 1} x_i y_i, x = \{x_i\}_{i \geq 1}$ and $y = \{y_i\}_{i \geq 1}$ being sequences in ℓ^2. Indeed, if $\{\varphi_k\}_{k \geq 0}$ is an orthonormal basis of $L^2(\mathcal{T})$ such that $\varphi_0 = 1/\sqrt{\eta}$ ($\eta = t_M - t_m$), each element $f \in A^2$ admits the representation $\log(f) = \sum_{k \geq 0} \alpha_k \varphi_k$. If we define the operator $T : A^2(\mathcal{T}) \to \ell^2$ as $Tf = \{\alpha_k\}_{k \geq 1}$, i.e., the operator associating to $f \in A^2$ the Fourier coefficients of $\log(f) \in L^2$ (except α_0), then the following result holds.

Proposition 1 [Egozcue et al. (2006)] $A^2(\mathcal{T})$ *endowed with perturbation $\oplus$ and powering $\odot$ operations,*

$$f \oplus g = T^{-1}(Tf + Tg); \quad c \odot f = T^{-1}(c \cdot T(f)),$$
$$f, g \in A^2, c \in \mathbb{R},$$

and the Aitchison inner product

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

is a separable Hilbert space.

Some of the basic definitions and properties of the functional space A^2 are recalled in the following Subsections. Furthermore, we show in Sect. 3.3 that an isometric isomorphism exists between $A^2(\mathcal{T})$ and $L^2(\mathcal{T})$. 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 an absolutely integrable function f on $\mathcal{T}$, 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:

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

Egozcue et al. (2006) prove that these two operators are the same as those defined in Proposition 1 and that $(A^2(\mathcal{T}), \oplus, \odot)$ is a vector space. Note that the neutral elements of perturbation and powering are $e(t) = 1/\eta$ and 1, respectively. 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 (Fréchet 1948):

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

Following (Menafoglio et al. 2013), for any given $s \in D$, we represent the element $\mathcal{Y}_s$ as a perturbation of the mean function m_s with a *neutral-mean* stochastic residual δ_s

$$\mathcal{Y}_s = m_s \oplus \delta_s,$$
$$\mathbb{E}_{A^2}[\delta_s] = 1/\eta. \tag{2}$$

We assume that the process $\mathcal{Y}_s$ can be represented by a global second-order stationary model. Hence, the process is characterized by a spatially constant mean function ($m_s = m$, for all $s \in D$), 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:

$$C(s_i - s_j) = \mathrm{Cov}_{A^2}(\mathcal{Y}_{s_i}, \mathcal{Y}_{s_j}) = \mathbb{E}[\langle \mathcal{Y}_{s_i} - m, \mathcal{Y}_{s_j} - m \rangle_{A^2}],$$
$$s_i, s_j \in D; \tag{3}$$

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

Given a sample $\mathcal{Y}_{s_1}, \ldots, \mathcal{Y}_{s_n}$ of $\{\mathcal{Y}_s, s \in D\}$, the Ordinary Kriging predictor of $\mathcal{Y}_{s_0}$, at an unsampled location $s_0 \in D$, is the best linear unbiased predictor (BLUP) in the Aitchison geometry:

$$\mathcal{Y}^*_{s_0} = \bigoplus_{i=1}^n \lambda_i^* \odot \mathcal{Y}_{s_i}. \tag{5}$$

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

$$(\lambda_1^*, \ldots, \lambda_n^*) = \underset{\substack{\lambda_1, \ldots, \lambda_n \in \mathbb{R}: \\ \mathcal{Y}^\lambda_{s_0} = \oplus_{i=1}^n \lambda_i \odot \mathcal{Y}_{s_i}}}{\argmin} \ \mathrm{Var}_{A^2}(\mathcal{Y}^\lambda_{s_0} \ominus \mathcal{Y}_{s_0}) \quad \text{s.t.}$$
$$\mathbb{E}_{A^2}[\mathcal{Y}^\lambda_{s_0}] = m. \tag{6}$$

The problem of kriging functional data has been tackled in (Menafoglio et al. 2013) 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. (2013)] *Assume that $\Sigma = (C(h_{i,j})) \in \mathbb{R}^{n \times n}, h_{i,j} = s_i - s_j, i, j = 1, \ldots, n,$ is a positive definite matrix. Then problem (6) admits a unique solution $(\lambda_1^*, \ldots, \lambda_n^*) \in \mathbb{R}^n$, which is obtained by solving:*

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

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

$$\sigma_*^2(s_0) = \mathrm{Var}_{A^2}(\mathcal{Y}^*_{s_0}) = C(\mathbf{0}) - \sum_{i=1}^n \lambda_i^* C(h_{i,0}) - \zeta^*. \tag{8}$$

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

$$P(\|\mathcal{Y}_{s_0} \ominus \mathcal{Y}^*_{s_0}\|_{A^2} > \kappa \cdot \sigma_*(s_0)) < \frac{1}{\kappa^2} \tag{9}$$

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

As in classical geostatistics (Cressie 1993), under stationarity conditions the only quantity which is required to be estimated is the trace-semivariogram γ, as $C(h) - C(\mathbf{0}) = \gamma(h)$, $h \in \mathbb{R}^d$ being a lag, or separation distance vector. To this end, a method of moments (MoM) estimator $\widehat{\gamma}$ can be employed:

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

where $N(\boldsymbol{h})$ denotes the set of location pairs separated by $\boldsymbol{h}$ and $|N(\boldsymbol{h})|$ its cardinality. A discretized version of $\hat{\gamma}$ is considered in typical applications and a valid variogram model is fitted to observations (Chilès and Delfiner 1999).

The approach we present can also be employed in a non-stationary setting. For completeness, we report the details of this case in Appendix 1.

3.3 Centered log-ratio transform

Here, we illustrate a representation of the process through a centered 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 model (2) and has finite variance, i.e., $\mathbb{E}[\|\delta_{\boldsymbol{s}}\|_{A^2}^2] < +\infty$ for all $\boldsymbol{s} \in D$, there exist a (deterministic) sequence $\{\mu_k(\boldsymbol{s})\}_{k \geq 0}$, a zero-mean random sequence $\{\xi_k(\cdot, \boldsymbol{s})\}_{k \geq 0}$, both valued in ℓ^2, and an orthonormal basis—also called Hilbert 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 $\{\mu_k(\cdot, \boldsymbol{s})\}_{k \geq 0}$ and $\{\xi_k(\cdot, \boldsymbol{s})\}_{k \geq 0}$ 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 $\Psi_k = \mathcal{C}[\exp\{\varphi_k\}], k \geq 1$ [see also (Egozcue et al. 2006)].

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

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

$$\omega \in \Omega, t \in \mathcal{T}, \boldsymbol{s} \in D, \qquad (11)$$

satisfies

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

since [see Egozcue et al. (2006)]

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

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

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 $clr(f), clr(g) \in L^2(\mathcal{T})$:

$$\langle f, g \rangle_{A^2} = \int_{\mathcal{T}} \log(f(t)) \log(g(t)) dt - \frac{1}{\eta} \int_{\mathcal{T}} \log(f(t)) dt$$

$$\times \int_{\mathcal{T}} \log(g(t)) dt = \int_{\mathcal{T}} \left\{ \left(\log(f(t)) - \frac{1}{\eta} \int_{\mathcal{T}} \log(f(z)) dz \right) \right.$$

$$\left. \times \left(\log(g(t)) - \frac{1}{\eta} \int_{\mathcal{T}} \log(g(z)) dz \right) \right\} dt$$

$$= \langle clr(f), clr(g) \rangle_{L^2}$$

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:

$$\text{Cov}_{A^2}(\mathcal{Y}_{\boldsymbol{s}_i}, \mathcal{Y}_{\boldsymbol{s}_j}) = \text{Cov}_{L^2}(\mathcal{Z}_{\boldsymbol{s}_i}, \mathcal{Z}_{\boldsymbol{s}_j});$$

$$\text{Var}_{A^2}(\mathcal{Y}_{\boldsymbol{s}_i} \ominus \mathcal{Y}_{\boldsymbol{s}_j}) = \text{Var}_{L^2}(\mathcal{Z}_{\boldsymbol{s}_i} - \mathcal{Z}_{\boldsymbol{s}_j}).$$

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 $Z_{\boldsymbol{s}_1}, \ldots, Z_{\boldsymbol{s}_n}$ in the $L^2(\mathcal{T})$ geometry, as:
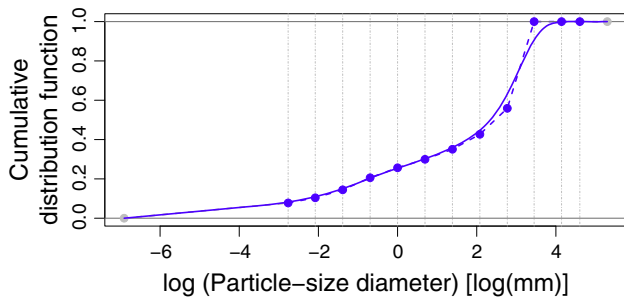
Fig. 2 Available particle-size raw datum at elevation $z = 301.03$ m a.s.l. (*symbols*), its interpolant (*dashed curve*) and PSC smooth estimate (*solid curve*). Vertical dotted lines mark the $N = 12$ log-transformed sieve diameters, $t_1, \ldots, t_N$

$$
\begin{aligned}
\mathcal{Y}^*_{s_0} &= \bigoplus_{i=1}^{n} \lambda^*_i \odot \mathcal{Y}_{s_i} = \mathcal{C}\left[\prod_{i=1}^{n} \mathcal{Y}^{\lambda^*_i}_{s_i}\right] \\
&= \mathcal{C}\left[\exp\left\{\sum_{i=1}^{n} \lambda^*_i \log(\mathcal{Y}_{s_i})\right\}\right] = \mathcal{C}\left[\exp\left\{\sum_{i=1}^{n} \lambda^*_i \mathcal{Z}_{s_i}\right\}\right] \\
&= clr^{-1}(\mathcal{Z}^*_{s_0}).
\end{aligned}
$$

The above isometric isomorphism enables one to perform the required calculation by exploiting efficient routines which are designed for data belonging to L^2, eventually back-transforming to A^2 the results. We note that the integrals appearing in (11) and (13) can be computed by matrix products when data are projected on typically employed bases (e.g., Fourier, B-splines) (Ramsay and Silverman 2005), or by quadrature schemes otherwise. We adopt the latter strategy in our application, which is illustrated in Sect. 5.

4 Smoothing discrete particle-size data with Bernstein Polynomials

A key assumption underlying the spatial prediction methodology here proposed is that data are curves which can be evaluated at any point $t \in \mathcal{T}$. If PSCs were already observed in their functional form, the methodology illustrated in Sect. 3 could be directly applied, without any particular data preprocessing. As detailed in Sect. 2, in the present case study an estimate of the particle-size curve at a given spatial location s is available only for a set of $N = 12$ sieve diameters.

As a way of example, Fig. 2a reports the raw datum available at the location s_1 (elevation = 301.03 m a.s.l.) for the log-transformed diameters, $t_1, \ldots, t_{12}$ (symbols). This corresponds to $N = 12$ observations $\{\chi_{s_1,v}(t_1), \ldots, \chi_{s_1,v}(t_{12})\}$ of the empirical cumulative distribution function (ECDF) $\chi_{s_1,v}$, v being the number of particle diameters

measured within the soil sample. The ECDF, in turn, is an estimate of the underlying CDF, i.e., the particle-size curve χ_{s_1} (see also Appendix 2). In this 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.

Here, we introduce a smoothing procedure for PSCs based on Bernstein Polynomials. The use of Bernstein Polynomials to provide a smooth estimate of a CDF is well documented in the literature [e.g., (Vitale 1975), (Petrone 1999), (Leblanc 2010) and references therein] and is particularly suited for CDFs with compact support (Babu et al. 2002). We follow the approach of Babu et al. (2002), who develop a smooth estimator for CDFs based on the ECDF. To cope with the partial information available from grain sieve analysis, we propose first to linearly interpolate the available measurements to obtain a continuous approximation of the ECDF and then to smooth this through Bernstein Polynomials. We remark that the estimator here defined provides also a smooth estimate of the particle-size density, unlike the linear interpolant of the available data. We refer to Appendix 2 for the mathematical construction of the estimator and the proof of the following properties: (a) the obtained estimates range in [0,1] and are monotonically non-decreasing (i.e., they fulfills the requirements of a CDF); (b) the corresponding densities can be explicitly computed and (c) the estimator is strongly consistent for the underlying CDF under infill domain sampling (Theorem 1).

The proposed estimator has been applied to each raw particle-size curve depicted in Fig. 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 PSCs 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 PSCs evaluated at the 12 observed particle diameters. Figure 3a depicts boxplots 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 3b 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 uniform behavior, since the PSCs appear
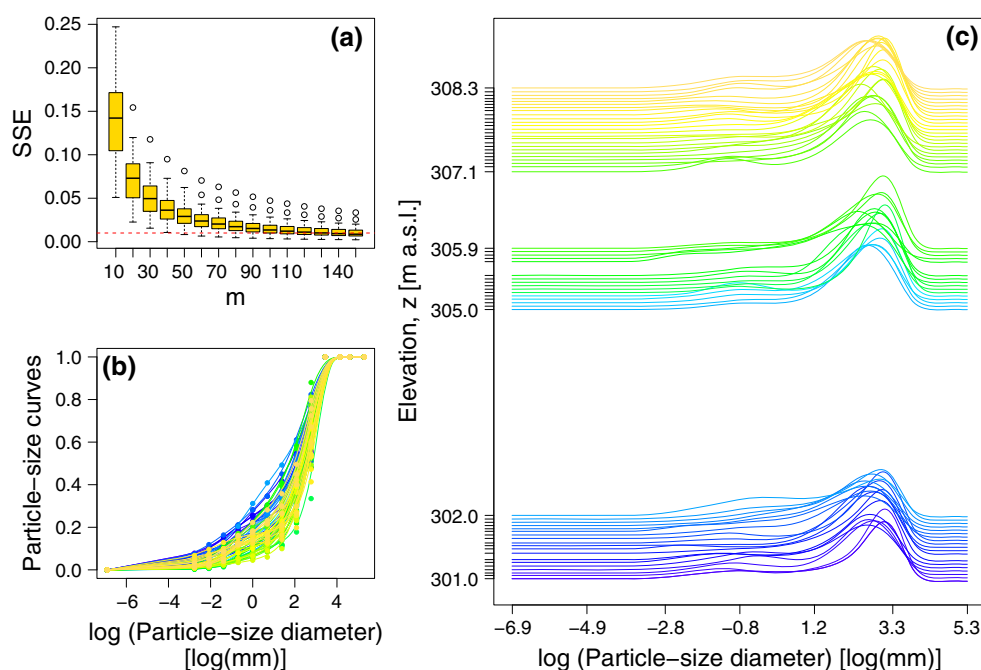
Fig. 3 Smoothing procedure by Bernstein Polynomials. **a** Boxplots of the SSE for $10 \leq m \leq 150$ (the threshold value of 0.01 is indicated by a *dotted line*); **b** raw PSCs (*symbols*) and particle-size curves smoothed by Bernstein Polynomials with $m = 140$ (*solid lines*); **c** vertical distribution of smoothed densities

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 1 in Appendix 2.

Finally, Fig. 3c depicts the vertical distribution of the particle-size densities computed from the smoothed data reported in Fig. 3b.

5 Results

5.1 Geostatistical analysis of the field data

Here, the notation introduced in Sect. 3 is employed as follows: quantities $\chi_{s_1}, \ldots, \chi_{s_n}$ denote the smoothed version of PSCs observed at locations $s_1, \ldots, s_n$ (solid lines in Fig. 3b); $\mathcal{Y}_{s_1}, \ldots, \mathcal{Y}_{s_n}$ indicate the smoothed particle-size densities depicted in Fig. 3c and obtained as in (26). The functional dataset $\mathcal{Y}_{s_1}, \ldots, \mathcal{Y}_{s_n}$ has been embedded into the space A^2 endowed with the Aitchison geometry and the methodology described in Sect. 3 has been coherently applied.
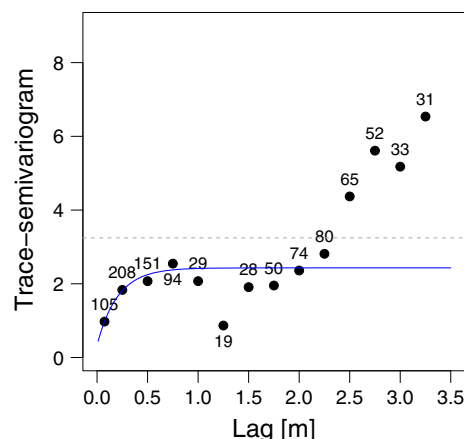

Fig. 4 Estimated trace-semivariogram of the particle-size densities shown in Fig. 3c: empirical trace-semivariogram (*symbols*), fitted model (*solid line*) and sample variance (*dotted line*). The number of pairs associated with each lag is reported

Visual inspection of the empirical trace-semivariogram depicted in Fig. 4 might indicate some degree of non-stationarity, even though it can be noted that the number of data pairs decreases with increasing lag. Since the stationarity assumption along the vertical direction is supported by prior knowledge of the field site (Riva et al. 2006, 2008, 2010), non-stationarity has not been considered in the present study. The structure of spatial dependence among the particle-size densities has been explored
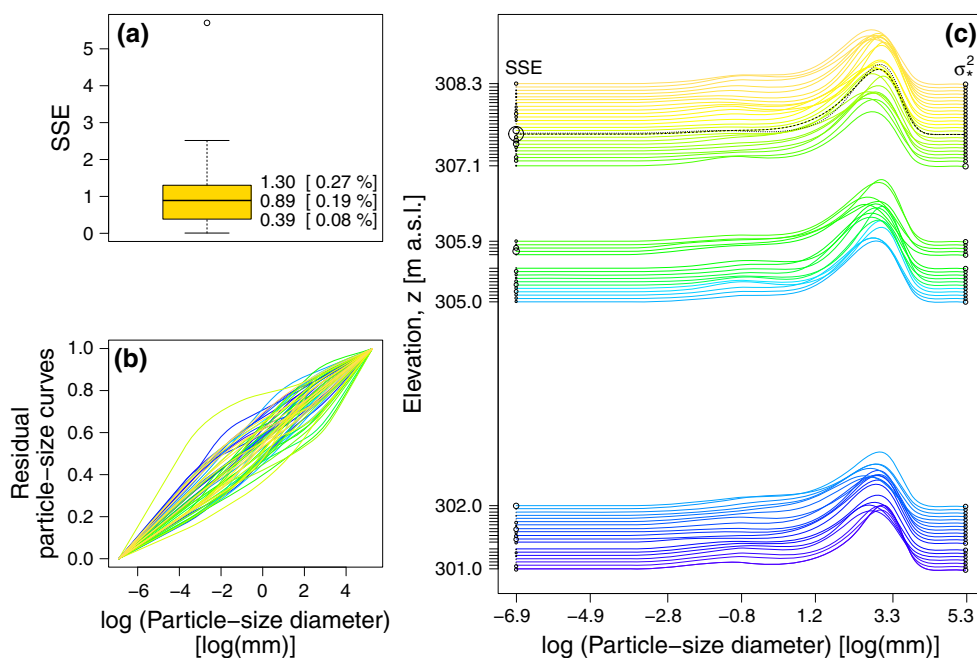
Fig. 5 Cross-validation results. **a** Boxplot of the SSE, reporting the absolute and relative quartile values; **b** cross-validation residual particle-size curves; **c** cross-validation prediction of particle-size densities as a function of elevation (the size of the *symbols* is proportional to the associated cross-validation SSE/kriging variance)

through the trace-semivariogram. The latter has been estimated from the data according to the discretized version of (10). Figure 4 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 5 depicts the results of the leave-one-out cross-validation procedure. Figure 5a shows the boxplot of the cross-validation SSE. The SSE for each sample $i = 1, \ldots, n$ has been computed as $\|\mathcal{Y}_{s_i} \ominus \mathcal{Y}_{s_i}^{*(CV)}\|_{A^2}^2$, $\mathcal{Y}_{s_i}^{*(CV)}$ being the kriging prediction at s_i obtained upon removing the i-th datum (i.e., PSC) from 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 Fig. 5c 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 Chebyshev 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 PSCs depicted in Fig. 5b, which are fairly spread across a uniform CDF (i.e., a straight line).

Prediction of the PSCs over a fine vertical grid with spacing of 1 mm has then been performed. Figure 6 depicts selected predicted PSCs (Fig. 6a), particle-size densities (Fig. 6b) and the associated kriging variance (Fig. 6c). Figure 7 depicts a detailed view of the interpolation results relative to the group of samples at elevations $307.1 \leq z \leq 308.3$ m a.s.l included in Fig. 6. The spatial prediction represents a smooth interpolation of the
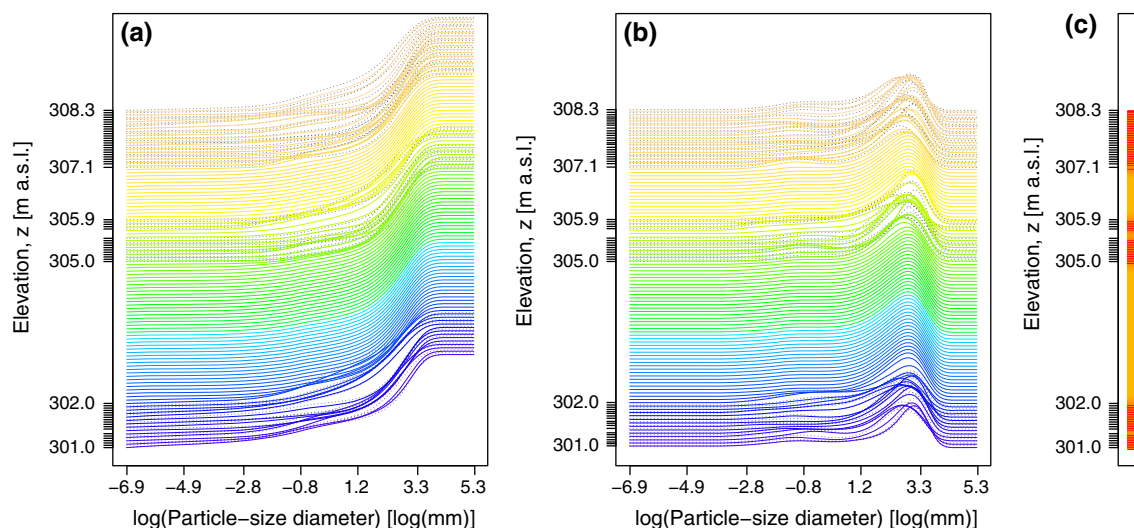
Fig. 6 Vertical distribution of prediction results: **a** a sample of 100 of the 7,190 predicted/kriged (*solid curves*) and observed (smoothed data; *dotted curves*) particle-size curves; **b** a sample of 100 of the 7,190 predicted/kriged (*solid curves*) and observed (smoothed data; *dotted curves*) particle-size densities; **c** kriging variance. Kriging variance ranges between 0 (*darkest shade*) and 2.53 (*lightest shade*)
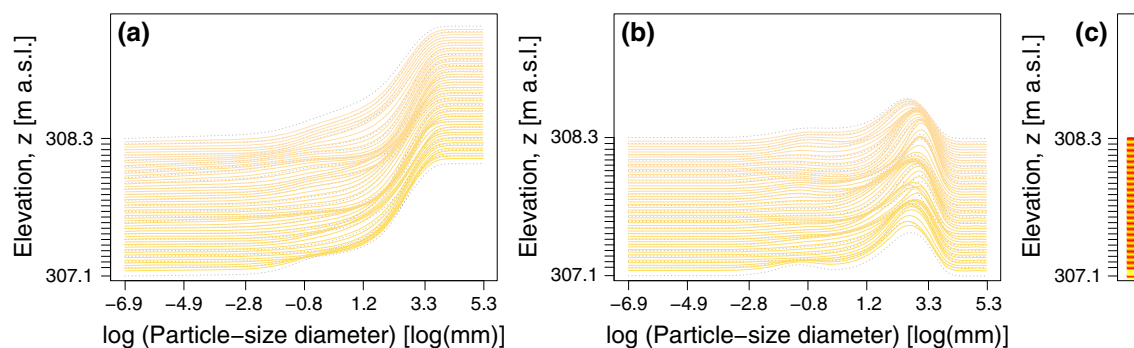

Fig. 7 Vertical distribution of prediction results for the group of samples at elevations $307.1 \leq z \leq 308.3$ m a.s.l.: **a** a sample of 50 predicted/kriged (*solid curves*) and observed (smoothed data; *dotted curves*) particle-size curves; **b** a sample of 50 predicted/kriged (*solid curves*) and observed (smoothed data; *dotted curves*) particle-size densities; **c** kriging variance. Kriging variance ranges between 0 (*darkest shade*) and 1.05 (*lightest shade*)

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, we compare the results which can be obtained through our FCK 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 FCK technique allows treating the complete set of information embedded in the available particle-size data within a framework based
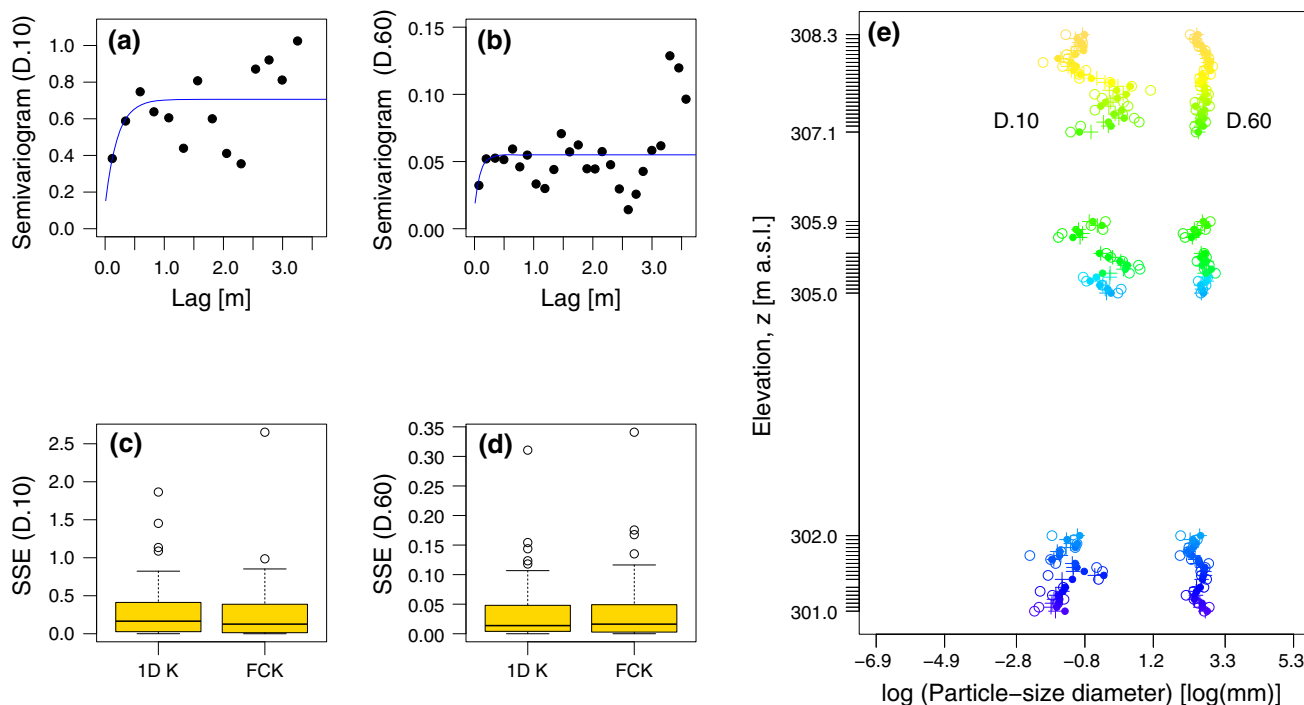
Fig. 8 Comparison of cross-validation results of classical one-dimensional kriging (1D K) and functional compositional kriging (FCK): empirical variogram (*symbols*) and fitted models (*solid curves*) for **a** D_{10} and **b** D_{60}; boxplots of cross-validation SSEs for **c** D_{10} and **d** D_{60}; **e** D_{10} and D_{60} data (*empty circles*) together with cross-validation predictions with 1D K (*crosses*) and FCK (*solid circles*)

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. Such an approach, besides being methodologically different from the one we propose, can produce inconsistent results (Tolosana-Delgado et al. 2008).

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

$$D_{10}(s) = \chi_s^{-1}(0.10); \quad D_{60}(s) = \chi_s^{-1}(0.60); \quad s \in D \quad (14)$$

or, equivalently,

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

We remark that, for consistency, both classical and functional compositional geostatistical analyses are here performed on the quantities $D_{10}(s_1), \ldots, D_{10}(s_n)$ and $D_{60}(s_1), \ldots, D_{60}(s_n)$, i.e., the values associated with the $n = 60$ smoothed PSCs $\chi_{s_1}, \ldots, \chi_{s_n}$ obtained according to (14) (empty symbols in Fig. 8a).

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 10th and 60th log-quantiles at the site (Riva et al. 2010). Figure 8a 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 partial sill of 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 of 0.62 and 0.28, respectively for D_{10} and D_{60}). These results are consistent

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

	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 %]
$\log(K^{[KC]})$	FCK	0.50 [16.36 %]	1.10 [35.98 %]
	1D K	0.63 [20.76 %]	1.16 [37.95 %]
$\log(K^{[H]})$	FCK	0.50 [13.59 %]	1.11 [30.02 %]
	1D K	0.65 [17.61 %]	1.18 [32.70 %]

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 Fig. 8a and b have been validated by means of a leave-one-out cross-validation analysis. Cross-validation predictions are reported in Fig. 8e (crosses) together with the log-quantiles predictions obtained by the cross-validation predicted PSCs, computed according to (14) (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 FCK. For completeness and ease of reference, these are also depicted in Fig. 8c, 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 Fig. 8e. 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, the natural logarithm of hydraulic conductivities has 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:

$$K = \frac{g}{\nu} C f(\phi) d_e^2 \tag{15}$$

where g is gravity, ν 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 PSCs 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:

$$C = 8.3 \cdot 10^{-3}; \quad f(\phi) = \left[\frac{\phi^3}{(1 - \phi)^2} \right]; \quad d_e = d_{10}. \tag{16}$$

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)]

$$\phi = 0.255(1 + 0.83^U); \quad U = \left(\frac{d_{60}}{d_{10}} \right) \tag{17}$$

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

$$C = 6 \cdot 10^{-4}; \quad f(\phi) = 1 + 10(\phi - 0.26). \tag{18}$$

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

The last rows of Table 1 reports the cross-validation median and mean SSE related to log-conductivities computed by the Kozeny-Carman ($\log(K^{[KC]})$) and Hazen ($\log(K^{[H]})$) 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-conductivities SSE, even though D_{60} appears to be slightly better predicted by classical kriging.

Finally, Fig. 9 shows the predictions of the log-quantiles D_{10} and D_{60} (panel a) and the log-conductivities $\log(K^{[KC]})$ and $\log(K^{[H]})$ (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-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.
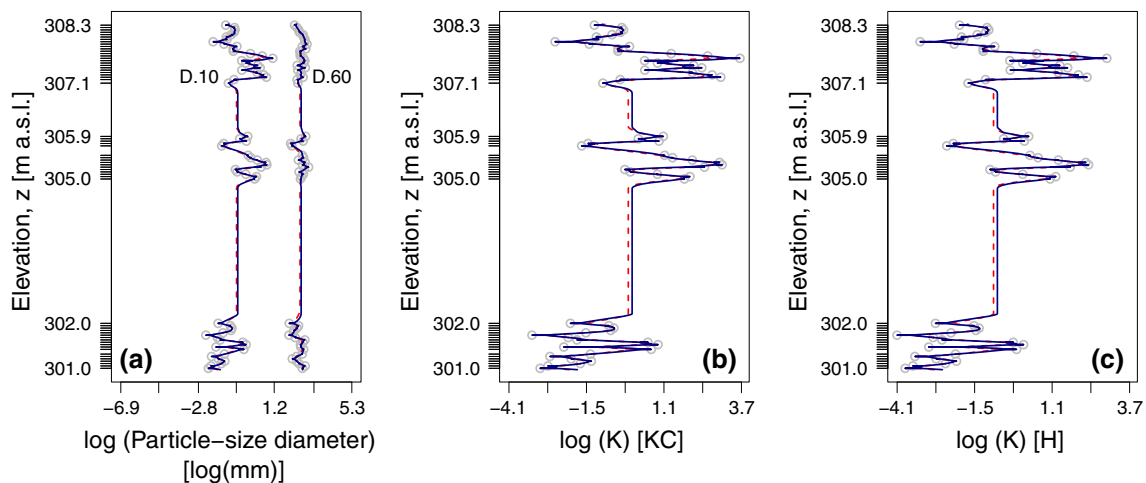
Fig. 9 Comparison between kriging predictions obtained by 1D K (*dotted curves*) and FK (*solid curves*): **a** D_{10} and D_{60}; hydraulic log-conductivity based on **b** Kozeny-Carman [KC] and **c** Hazen [H] formulations. Data are indicated with *symbols*

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] We interpret PSCs, which constitute a typical information content employed in hydrogeology, soil science and geochemical applications, 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 (Sect. 4) and its consistency has been proven (Appendix 2). 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 (Sect. 3) can be directly applied to the available data, without resorting to the procedure presented in Sect. 4 and Appendix 2.

[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 FCK 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 FCK 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 FCK 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 only finite-dimensional predictions, based on a set of selected features (or synthetic indices) of the curve. The information content provided by the full PSC is critical to the proper modeling of 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.

Appendices

Appendix 1: the non-stationary case

In Sect. 3.1, second-order stationarity has been assumed in view of the particular application studied. The non-stationary case could be dealt with as well, by exploiting the estimators and the algorithms proposed in (Menafoglio et al. 2013).

In such a case, for any given $s \in D$, we represent the element $\mathcal{Y}_s$ as:

$$\mathcal{Y}_s = m_s \oplus \delta_s, \tag{19}$$

$$m_s = \bigoplus_{l=0}^{L} f_l(s) \odot a_l, \tag{20}$$

$$\mathbb{E}_{A^2}[\delta_s] = 1/\eta,$$

where, for $l = 0, \ldots\ldots, L, a_l \in A^2$ and $f_l(s) \in \mathbb{R}$. In model (19), the drift m_s is assumed to follow a linear model with functional coefficients $\{a_l\}$ independent of the spatial location and scalar regressors $\{f_l(s), s \in D\}$ which are assumed to be known over the entire domain D. The residual process $\{\delta_s, s \in D\}$ is supposed to be neutral-mean and second-order stationary. This allows to define the trace-covariogram C and the trace-variogram γ as in (3) and (4), but in terms of the residual process.

Given a sample $\mathcal{Y}_{s_1}, \ldots, \mathcal{Y}_{s_n}$, the Universal Kriging predictor at a location $s_0 \in D$ is the BLUP, derived by solving the minimization problem:

$$(\lambda_1^*, \ldots, \lambda_n^*) = \underset{\substack{\lambda_1, \ldots, \lambda_n \in \mathbb{R}: \\ \mathcal{Y}_{s_0}^\lambda = \bigoplus_{i=1}^n \lambda_i \odot \mathcal{Y}_{s_i}}}{\operatorname{argmin}} \operatorname{Var}_{A^2}(\mathcal{Y}_{s_0}^\lambda \ominus \mathcal{Y}_{s_0}) \quad \text{s.t.}$$
$$\mathbb{E}_{A^2}[\mathcal{Y}_{s_0}^\lambda] = m_{s_0}. \tag{21}$$

We remark that the unbiasedness constraint in problem (21) is stronger that that of problem (6), since the drift in s_0, m_{s_0}, is governed by the linear model (20). Problem (21) reduces to the linear system (in block matrix form):

$$\begin{pmatrix} C(\boldsymbol{h}_{i,j}) & f_l(s_i) \\ f_l(s_j) & 0 \end{pmatrix} \begin{pmatrix} \lambda_i \\ \zeta_l \end{pmatrix} = \begin{pmatrix} C(\boldsymbol{h}_{0,i}) \\ f_l(s_0) \end{pmatrix}, \tag{22}$$

where $\zeta_0, \ldots, \zeta_L$ are $L + 1$ Lagrange multipliers associated with the unbiasedness constraint. System (22) admits a unique solution provided that $\Sigma = (C(\boldsymbol{h}_{i,j}))$ is positive definite and the design matrix $\mathbb{F} = (f_l(s_i)) \in \mathbb{R}^{n \times (L+1)}$ 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_s is not spatially constant. Indeed, the estimator $\widehat{\gamma}(\boldsymbol{h})$ defined by (4) provides an unbiased estimate of $\mathbb{E}[\|\chi_{s+\boldsymbol{h}} \ominus \chi_s\|^2]$, but the latter coincides with $\gamma(\boldsymbol{h})$ if and only if the mean is 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 Sect. 3,

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

In the general context of data belonging to Hilbert Spaces, Menafoglio et al. (2013) propose to estimate the residuals $\delta_{s_1}, \ldots, \delta_{s_n}$ as a difference between the observations and the generalized least squares (GLS) estimates of the drift at the sampled locations, i.e., in our setting, $\widehat{\delta}_{s_i} = \mathcal{Y}_{s_i} \ominus \widehat{m}_{s_i}^{GLS}, i = 1, \ldots, n$. Furthermore, Menafoglio et al. (2013) derive the explicit expression of the GLS drift estimator and analyze its properties. Embedding those results within our framework yields to the following expression of the GLS drift estimator:

$$\widehat{\boldsymbol{m}}_s^{GLS} = \mathbb{F}(\mathbb{F}^T \Sigma^{-1} \mathbb{F})^{-1} \mathbb{F}^T \Sigma^{-1} \odot \boldsymbol{\mathcal{Y}}_s, \tag{23}$$

where $\widehat{\boldsymbol{m}}_s^{GLS} = (\widehat{m}_{s_1}^{GLS}, \ldots, \widehat{m}_{s_n}^{GLS})^T, \boldsymbol{\mathcal{Y}}_s = (\mathcal{Y}_{s_1}, \ldots, \mathcal{Y}_{s_n})^T$ and having adopted the vectorial notation: $(\mathbb{A} \odot \boldsymbol{f})_i = \bigoplus_{j=1}^n \mathbb{A}_{i,j} \odot f_j, \mathbb{A} = (\mathbb{A}_{ij}) \in \mathbb{R} \quad n \times n, \boldsymbol{f} = (f_i), f_i \in A^2, i = 1, 2, \ldots, n$. Note that the optimality of estimator (23) relies on a proper account 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 can be employed [Menafoglio et al. (2013), Sects. 4–5].

Appendix 2: A smooth estimator for cumulative distribution functions

In this Appendix, we report the technical details leading to the construction of the smooth estimator for PSCs which has been illustrated and applied in Sect. 4. For ease of illustration, we adopt here a simplified notation by omitting the subscript s indicating spatial location, since we consider each particle-size curve separately.

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 $\mathcal{T} = [t_m, t_M]$ through the transformation $x = (t - t_m)/(t_M - t_m) \in [0,1]$ for $t \in \mathcal{T}$). If a sample $X_1, \ldots, X_v$ from F is available, the ECDF F_v is defined as $F_v(x) = \frac{1}{v}\sum_{i=1}^{v} I_{[0,x]}(X_i), I$ being the indicator function. F_v is a (discontinuous) non-parametric estimator of F, which is strongly consistent because of the Glivenko-Cantelli Theorem. In our setting, the sample $X_1, \ldots, X_v$ would represent the set of (transformed) particle diameters measured within a soil sample extracted at a given location in the aquifer. When such a sample is available, the problem of smoothly estimating the particle-size curve, i.e. the underlying CDF F, can be solved by smoothing the ECDF by means of Bernstein Polynomials.

The use of Bernstein Polynomials to approximate a bounded continuous function, such as F, is supported by the Bernstein-Weierstrass Approximation Theorem (Feller (1965), Theorem 1, Section VII.2). This states that any continuous function, $u(x)$, on the closed interval [0,1] can be uniformly approximated by:

$$u_m^*(x) = \sum_{k=0}^{m} u(k/m)b_{k,m}(x), \quad x \in [0,1],$$

where $b_{k,m}(x) = mkx^k(1-x)^{m-k}, k = 0, \ldots, m$. On these premises, Babu et al. (2002) propose and explore the asymptotic properties of the smooth estimator $F_{v,m}$: $[0,1] \to [0,1]$ defined as:

$$F_{v,m}(x) = \sum_{k=0}^{m} F_v(k/m)b_{k,m}(x), \quad x \in [0,1]. \quad (24)$$

As opposed to kernel smoothing estimators (Rosenblatt 1956; Parzen 1962; Silverman 1986), the estimator (24) is well suited for distributions with compact support, of the kind associated with the particle-size curves we analyze. We do not observe the data $X_1, \ldots, X_v$, but only the value of the ECDF F_v for given diameters $\{x_1, \ldots, x_N\}$. Therefore, we propose to consider a modified smooth estimator $F_{v,m}^N: [0,1] \to [0,1]$ based on a linear interpolant of the observed values of the ECDF and defined as:

$$F_{v,m}^N(x) = \sum_{k=0}^{m} F_v^{(1)}(k/m)b_{k,m}(x), \quad x \in [0,1], \quad (25)$$

$F_v^{(1)}$ being the linear interpolant of $F_v(x_1), \ldots, F_v(x_N)$, i.e., for $x \in [0,1]$:

$$F_v^{(1)}(x)$$
$$= \sum_{i=1}^{N+1}\left(F_v(x_{i-1}) + \frac{F_v(x_i) - F_v(x_{i-1})}{x_i - x_{i-1}}(x - x_{i-1})\right)I_{(x_{i-1},x_i]}(x),$$

with $x_0 = 0$, $x_{N+1} = 1$ and $F_v(x_0) = 0$, $F_v(x_{N+1}) = 1$. Adopting (25) enables one to estimate the CDF F through a continuous approximation $F_v^{(1)}$ of the ECDF F_v combined with Bernstein Polynomials. Note that, while other approximations for F_v 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 $f_{v,m}^N$:

$$f_{v,m}^N(x) = m\sum_{k=0}^{m-1}(F_v^{(1)}((k+1)/m) - F_v^{(1)}(k/m))b_{k,m-1}(x),$$
$$x \in [0,1].$$
$$(26)$$

We remark that the smooth estimate $F_{v,m}^N$ provided by (25) is a monotonically increasing function, since its derivative (26) is positive ($F_v^{(1)}$ is monotonically non-decreasing with $F_v^{(1)}(0) = 0$, $F_v^{(1)}(1) = 1$ and $b_{k,m-1}(x)$ is positive for all $m > 1, k = 0, \ldots, m, x \in (0,1)$). Moreover, denoting with $\|\cdot\|_{\mathcal{C}^0}$ the uniform norm on the space of continuous functions, the following result holds.

Theorem 1 *Let F be a continuous CDF on* [0,1]. *If*

$$\lim_{N\to\infty}\max_{i\in\{1,\ldots,N+1\}}(x_i - x_{i-1}) = 0,$$

then

$$\lim_{m,v,N\to+\infty}\|F_{v,m}^N - F\|_{\mathcal{C}^0} = 0, \quad a.s. \quad (27)$$

Proof First, note that

$$\|F_{v,m}^N - F\|_{\mathcal{C}^0} \le \|F_{v,m}^N - F_{v,m}\|_{\mathcal{C}^0} + \|F_{v,m} - F\|_{\mathcal{C}^0}. \quad (28)$$

The last term of (28) vanishes as $v, m \to +\infty$ [Babu et al. (2002), Theorem 2.1]. Consider the second term and write its argument as:

$$F_{v,m}^N(x) - F_{v,m}(x) = \sum_{k=0}^{m}(F_v^{(1)}(k/m) - F_v(k/m))b_{k,m}(x),$$
$$x \in [0,1].$$

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

$$F_v^{(1)}(x) - F_v(x) = \sum_{i=1}^{N+1}\left[(F_v(x_i) - F_v(x))\frac{x - x_{i-1}}{x_i - x_{i-1}}\right.$$
$$\left. - (F_v(x) - F_v(x_{i-1}))\frac{x_i - x}{x_i - x_{i-1}}\right]I_{(x_{i-1},x_i]}(x).$$
$$(29)$$

Hence:

$$\|F_{v,m}^N - F_{v,m}\|_{C^0} = \max_{x \in [0,1]} |F_{v,m}^N(x) - F_{v,m}(x)| \leq \max_{0 \leq k \leq m} |F_v^{(1)}(k/m)$$

$$-F_v(k/m)| \leq \max_{0 \leq k \leq m} \left\{ \sum_{i=1}^{N+1} \left| (F_v(x_i) - F_v(k/m)) \frac{k/m - x_{i-1}}{x_i - x_{i-1}} \right. \right.$$

$$\left. \left. -(F_v(k/m) - F_v(x_{i-1})) \frac{x_i - k/m}{x_i - x_{i-1}} \right| I_{(x_{i-1}, x_i]}(k/m) \right\}.$$

Fix $\varepsilon > 0$, consider v such that $\|F - F_v\|_{C^0} < \varepsilon/3$: such a v exists because of Glivenko-Cantelli theorem. Furthermore, since F is uniformly continuous by Heine-Cantor theorem, let δ be such that $|F(x) - F(y)| < \varepsilon/3$ for all $x, y \in [0,1]$ such that $|x - y| < \delta$. Finally, take N such that $\max_{i \in \{1, \dots, N+1\}}(x_i - x_{i-1}) < \delta$. Then:

$$\|F_{v,m}^N - F_{v,m}\|_{C^0}$$

$$\leq \max_{0 \leq k \leq m} \left\{ \sum_{i=1}^{N+1} \left| \frac{2\varepsilon}{3} + (F(x_i) - F(k/m)) \frac{k/m - x_{i-1}}{x_i - x_{i-1}} \right. \right.$$

$$\left. \left. +(F(k/m) - F(x_{i-1})) \frac{x_i - k/m}{x_i - x_{i-1}} \right| I_{(x_{i-1}, x_i]}(k/m) \right\} < \varepsilon.$$

The thesis then follows from the arbitrariness of ε.

Theorem 1 states that $F_{v,m}^N$ is a strongly consistent estimator for F. We remark that Theorem 1 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.

References

Aitchison J (1982) The statistical analysis of compositional data. J R Stat Soc Series B (Methodological) 44(2):139–177

Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London

Babu GJ, Canty AJ, Chaubey YP (2002) Application of Bernstein Polynomials for smooth estimation of a distribution and density function. J Stat Plan Inference 105:377–392

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

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

Buchanan S, Triantafilis J, Odeh IOA, Subansinghe R (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, Giraldo R, Mateu J (2013) A universal kriging approach for spatial functional data. Stoch Environ Res Risk Assess 27:1553–1563

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

Chilès JP, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York

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

Delicado P, Giraldo R, Comas C, Mateu J (2010) Statistics for spatial functional data: some recent contributions. Environmetrics 21:224–239

Deutsch CV, Journel AG (1997) GSLIB: Geostatistical software library and user's guide, 2nd edn. Oxford University Press, UK

Egozcue JJ, Díaz-Barrero JL, Pawlowsky-Glahn V (2006) Hilbert space of probability density functions based on aitchison geometry. Acta Math Sinica, Engl Series 22(4):1175–1182

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

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

Fetter CW (2001) Applied hydrogeology. Englewood Cliffs, New Jersey

Fréchet M (1948) Les éléments Aléatoires de Nature Quelconque dans une Espace Distancié. Annales de L'Institut Henri Poincaré 10(4):215–308

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

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

Horváth L, Kokoszka P (2012) Inference for functional data with applications. Springer series in statistics, Springer, New York

Hu BX, Wu J, Zhang D (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

Leblanc A (2010) A bias-reduced approach to density estimation using bernstein polynomials. J Nonparametric Stat 22:459–475

Leininger TJ, Gelfand AE, Allen JM, Silander JAJ (2013) Spatial regression modeling for compositional data with many zeros. J Agric Biol Environ Stat 18(3):314–334

Lemke LD, Abriola LM (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, Ptak T (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, Secchi P, Dalla Rosa M (2013) A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. Electron J Stat 7:2209–2240. doi:10.1214/13-EJS843

Nemes A, Schaap MG, Wösten JHM (2003) Functional evaluation of pedotransfer functions derived from different scales of data collection. Soil Sci Soc Am J 67:1093–1102

Nerini D, Monestiez P, Manté C (2010) Cokriging for spatial functional data. J Multivar Anal 101(2):409–418

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

Neuman SP, Riva M, Guadagnini A (2008) On the geostatistical characterization of hierarchical media. Water Resour Res 44(2):W02403

Odeh IOA, Todd AJ, Triantafilis J (2003) Spatial prediction of soil particle-size fractions as compositional data. Soil Sci 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, Rawls WJ (2004) Development of pedotransfer functions in soil hydrology, developments in soil science, vol 30. Elsevier, Amsterdam

Parzen E (1962) On estimation of probability density and mode. Ann Math Stat 33:1965–1070

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

Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis in the simplex. Stoch Environ Res Risk Assess 15:384–398

Pawlowsky-Glahn V, Egozcue JJ (2002) BLU estimators and compositional data. Math geol 34(3):259–274

Pawlowsky-Glahn V, Olea RA (2004) Geostatistical analysis of compositional data. Oxford university press, New York

Petrone S (1999) Bayesian density estimation using bernstein polynomial posteriors. Can J Stat 27:105–126

Ramsay J, Silverman B (2005) Functional data analysis, 2nd edn. Springer, New York

Ramsay JO, Dalzell CJ (1991) Some tools for functional data analysis. J Roy Stat Soc B 53(3):539–572

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

Riva M, Guadagnini A, Fernández-Garcáa D, Sánchez-Vila X, Ptak T (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, Guadagnini L, Guadagnini A (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

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

Salazar Buelvas EJ (2011) Predicción Espacial de Funciones de Densidad de Probabilidad. Master's thesis, (in Spanish), Universidad Nacional de Colombia, Facultad de Ciencias, Departamento de Estadística, Cartagena de Indias D.T. y C., Colombia

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

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

Tolosana-Delgado R, Pawlowsky-Glahn V, Egozcue JJ (2008) Indicator kriging without order relation violations. Math Geosci 40(3):327–347

Tolosana-Delgado R, Egozcue JJ, Sánchez-Arcilla A, Gómez J (2011a) Classifying wave forecasts with model-based geostatistics and the Aitchison distribution. Stoch Environ Res Risk Assess 25(8):1091–1100

Tolosana-Delgado R, van den Boogaart KG, Pawlowsky-Glahn V (2011b) Geostatistics for Compositions, Pawlowsky-Glahn & Buccianti edn, Wiley, New York, pp 73–86

Tong J, Hu B, Yang J (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, Egozcue JJ, Pawlowsky-Glahn V (2010) Bayes linear spaces. SORT 34(2):201–222

Vitale R (1975) A bernstein polynomial approach to density function estimation. Stat Inference Relat Top 2:87–99

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

Wohlberg B, Tartakovsky DM, Guadagnini A (2006) Subsurface characterization with support vector machines. IEEE Trans Geosci Remote Sens 44(1):47–57. doi:10.1109/TGRS.2005.859953