Noname manuscript No.
(will be inserted by the editor)

A Class-Kriging predictor for Functional Compositions with Application to Particle-Size Curves in Heterogeneous Aquifers

Alessandra Menafoglio · Piercesare Secchi · Alberto Guadagnini

Received: date / Accepted: date

Abstract This work addresses the problem of characterizing the spatial field of soil particle-size distributions within a heterogeneous aquifer system. The medium is conceptualized as a composite system, characterized by spatially varying soil textural properties associated with diverse geomaterials. The heterogeneity of the system is modeled through an original hierarchical model for particle-size distributions that are here interpreted as points in the Bayes space of functional compositions. This theoretical framework allows performing spatial prediction of functional compositions through a functional compositional Class-Kriging predictor. To tackle the problem of lack of information arising when the spatial arrangement of soil types is unobserved, a novel clustering method is proposed, allowing to infer a grouping structure from sampled particle-size distributions. The proposed methodology enables one to project the complete information content embedded in the set of heteroge-

A. Menafoglio
MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy
E-mail: alessandra.menafoglio@polimi.it

P. Secchi
MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy

A. Guadagnini
Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Italy
Department of Hydrology and Water Resources, The University of Arizona, USA

neous particle-size distributions to unsampled locations in the system. These developments are tested on a field application relying on a set of particle-size data observed within an alluvial aquifer in the Neckar river valley, in Germany.

Keywords Geostatistics · Functional Compositions · Clustering · Particle-size curves · Groundwater · Hydrogeology

1 Introduction

The quality of groundwater flow and transport predictions in natural aquifer systems is markedly dependent on the way one can provide a proper representation of the heterogeneous spatial distribution of geomaterials and their associated hydraulic/transport parameters at a given model grid scale. Amongst the set of available analysis techniques, particle-size curves (PSCs) are widely employed to provide relatively inexpensive estimates of the types of geo-materials forming the internal architecture of an aquifer, and the associated values of hydraulic conductivity (Riva et al., 2006, 2010, 2014, and references therein). Particle-size data are routinely inferred from laboratory analyses of soil samples, typically upon relying on the successive use of sieves of variable grid size, according to defined standards. These data allow identifying a set of representative (or effective) grain diameters defined as the representative particle-size diameter which corresponds to a given quantile of a particle-size curve.

Typical hydrogeological studies rely only on a discrete set of quantiles (i.e., the effective diameters), which can be related through a set of empirical formulations to parameters such as hydraulic conductivity or porosity (Vukovic and Soro, 1992). These diameters are then subject to geostatistical analysis and projected onto a computational grid by Kriging. In this sense, the complete set of information embedded in a PSC is not fully exploited in typical hydrogeological analyses. As a key element of innovation in the geostatistical char-

acterization of PSCs, Menafoglio et al. (2014) propose to analyze particle-size distributions through their densities, interpreted as functional compositions (FCs). The latter are functions constrained to be non-negative and to integrate to a constant and are the infinite-dimensional counterparts of compositional data, that is, multivariate observations whose components are proportion or relative amounts of a whole according to a given domain partition. FCs can be considered as compositions whose domain partition has been refined until obtaining (infinite) infinitesimal parts (Egozcue et al., 2006).

The statistical analysis of FCs with compact support has been the subject of an increasing body of literature, starting from the pioneering work of Egozcue et al. (2006). These authors establish a Hilbert space structure for FCs based on the log-ratio approach, upon which the Aitchison geometry is grounded (Pawlowsky-Glahn and Egozcue, 2001; Pawlowsky-Glahn and Buccianti, 2011, and references therein). Additional developments are then proposed by Egozcue et al. (2013); van den Boogaart et al. (2010), who introduce and explore the theory of Bayes linear spaces for FCs and assign an algebraic interpretation to several basic notions of mathematical statistics (e.g., the Bayes theorem). van den Boogaart et al. (2014) extend the theory of Bayes spaces to FCs which are not necessarily compactly supported. Applications of the theory of Bayes spaces for compactly supported FCs are found within the framework of classification (Nerini and Ghattas, 2007), dimensionality reduction (Delicado, 2011; Hron et al., 2015) and spatial prediction (Menafoglio et al., 2014).

In this context, each PSD is here interpreted as a unique entity, that is, an *object datum* (Marron and Alonso, 2014; Sangalli et al., 2014), which is embedded into the Hilbert space of FCs endowed with the generalized Aitchison geometry. This geometric perspective bases its strength on the concepts of functional (FDA, Horváth and Kokoszka, 2012; Ramsay and Silverman,

2005, and references therein) and compositional data analysis (CoDa, Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn et al., 2015, and references therein). A similar approach is considered by Menafoglio et al. (2014), who propose a functional compositional Kriging (FCK) methodology relying upon the Universal Kriging theory for Hilbert data proposed by Menafoglio et al. (2013). Unlike traditional methods in hydrogeology, the functional-compositional viewpoint is a powerful approach conducive to obtaining predictions of the complete information content embedded in PSCs.

Particle-size distributions are also closely related to soil textural properties. For instance, Martìn et al. (2005) employ discrete characterizations of PSCs to propose a soil texture classification based on self-similar fractal features of the observed PSCs. Riva et al. (2006) rely on multivariate techniques to classify a set of discrete PSCs and, on this basis, to provide estimates and multiple Monte Carlo realizations of the spatial distribution of sedimentological facies, to be then employed in a stochastic model of flow and transport at an experimental site.

Here, the focus is on the heterogeneity of the system which can be ascribed to the existence of a grouping structure within observed PSDs, associated with diverse soil textural properties. In this setting, a novel hierarchical geostatistical model for PSDs is introduced, and a functional compositional Class-Kriging (FCCK) predictor is developed. The latter combines the information content associated with the spatial arrangement of the soil types with that provided by the model of spatial variability of the PSDs.

These methodological developments are illustrated in Sect. 4, upon relying on the Hilbert space structure introduced in Sect. 3. Emphasis is given to the practical issues arising from the lack of information which typically plagues our knowledge of environmental systems. Here, the scenario where the system composition is only partially observable is considered, in view of analyzing

the field data detailed in Sect. 2. In this context, an original unsupervised classification method, consistent with the proposed FCCK methodology, is introduced to address the problem of identifying soil types characterizing the aquifer when only a sample of unclassified PSDs is available. The application of these developments to the target dataset is illustrated in Sect. 5.

2 Field Data

For the purpose of application, an extensive dataset of PSCs, sampled at an experimental site located near the city of Tübingen (Germany), is considered.

The investigated aquifer body is essentially formed by alluvial material overlaid by stiff silty clay and underlaid by hard silty clay. A dense borehole network provides the elements for a high level hydrogeological characterization of the site. The saturated thickness of the aquifer is of about 5 m. All boreholes are fully penetrating until the bedrock which constitutes a practically impervious aquifer base. A recounting of the hydrogeological, hydraulic, sedimentological and geophysical analyses conducted at the site is offered by Riva et al. (2006, 2008), to which the reader is referred for additional details. Amongst the available data, the focus is on 406 PSCs sampled along twelve vertical boreholes. These data were adopted by Riva et al. (2006, 2008, 2010) in numerical Monte Carlo analysis and interpretation of a tracer test, and to provide a probabilistic delineation of well-related capture zones. Riva et al. (2014) rely on these data to support their analytical developments leading to a set of relationships between the spatial covariance of the (natural) logarithm of hydraulic conductivity and that of representative soil particle sizes and porosity. A subset of these data was employed by Menafoglio et al. (2014) to test their FCK methodology, in a stationary setting.

The available PSCs were measured on soil samples of characteristic length ranging from 5 to 26.5 cm. A set of twelve discrete sieve diameters (i.e., 0.063, 0.125, 0.25, 0.50, 1.0, 2.0, 4.0, 8.0, 16.0, 31.5, 63.0 and 100.0 mm) were employed to reconstruct these curves by way of grain sieve analysis. Application of commonly used empirical relationships between characteristic PSCs diameters and medium permeability supports the picture according to which the site is formed by highly conducive and heterogeneous alluvial deposits. Figure 1 depicts a sketch of the sampling network at the site.

A classification of the spatial distribution of sedimentological facies at the site is provided by Riva et al. (2006) who grouped the sampled PSCs into three main clusters upon relying on a multivariate K-mean cluster analysis technique (Mc Queen, 1967). The clusters identified by these authors correspond to the following sedimentological facies: (i) about 53% of the samples can be described as moderately sorted gravel with approximately 14% sand and very few fines; (ii) about 44% of the samples consist of poorly sorted gravel with about 24% sand and few fines; and (iii) about 3% of the samples are represented by well sorted sand with very few fines and about 23% gravel. Riva et al. (2006, 2008, 2010) base their estimates of hydraulic conductivities for each of these facies on characteristic particle diameters. Here, the application of the theoretical developments presented in Sect. 4 relies on these PSC data.

3 The Space A^2 of Functional Compositions

Most multivariate methods for compositional data are grounded on the log-ratio approach via the Aitchison geometry (Aitchison, 1982; Pawlowsky-Glahn et al., 2015, and references therein). This has been recently generalized to functional compositions, that are infinite-dimensional objects which convey only relative information, as they are constrained to be positive and to integrate to

a constant (Egozcue et al., 2006; van den Boogaart et al., 2010, 2014). This section briefly recalls the basic notions of the Aitchison geometry for functional compositions; the reader is referred to Egozcue et al. (2006, 2013); van den Boogaart et al. (2010, 2014) for further details.

Two functional compositions f, g are considered equivalent if there exists $\alpha > 0$ such that $f = \alpha g$. This kind of equivalence is known in the multivariate setting as the *scale invariance* property (Egozcue, 2009). It reflects the observation that the information content embedded into the parts of a composition does not depend on the constant describing the measure of the whole (e.g., unity or 100), such a constant representing a convention rather than an informative quantity. In the following, $A^2(\mathcal{T})$ will denote the space of (equivalence classes of) non-negative real functions on a compact domain $\mathcal{T}$ with square integrable logarithm

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

Hereafter, the representative of an equivalence class will be its element integrating to 1, since this always exists in the considered field case study. Following Egozcue et al. (2006), one can otherwise consider the element whose logarithm integrates to 0.

Egozcue et al. (2006) define on A^2 the perturbation ($\oplus$) and powering operations ($\odot$) as

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

Note that the difference operator $\ominus$ induced by the perturbation $\oplus$ acts as $f \ominus g = f \oplus \frac{1/g}{\int_{\mathcal{T}}(1/g(t))dt}$, for $f, g \in A^2$, while the neutral element of perturbation is $0_\oplus = 1/\eta$, η being the measure of the compact set $\mathcal{T}$ (e.g., $\eta = t_M - t_m$ if $\mathcal{T}$ is the closed interval $\mathcal{T} = [t_m, t_M]$). Finally, Egozcue et al. (2006) introduce

the Aitchison inner product $\langle \cdot, \cdot \rangle_{A^2}$ as

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

and prove that $(A^2, \oplus, \odot, \langle \cdot, \cdot \rangle_{A^2})$ is a separable Hilbert space.

This work focuses on functional compositions with a compact support. The theory has been recently extended to deal with compositions with infinite supports (van den Boogaart et al., 2014). As noted by Delicado (2011); Menafoglio et al. (2014); Hron et al. (2015), inferior and superior extremes for the support can be identified without a substantial loss of generality in most real-life case studies and this contributes to a substantial simplification of the technicalities involved in the data analysis. As an alternative, conditional distributions may be considered, upon focusing on the conditional densities within the range $[t_m, t_M]$ of values which are actually observed. The latter approach is adopted in the field application which is illustrated in Sect. 5.

4 A Class-Kriging Predictor for Particle-Size Densities

4.1 A Hierarchical Functional Model for Particle-Size Densities

Let $\{\mathcal{X}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ be the random field of particle-size curves over the three-dimensional aquifer $D \subset \mathbb{R}^3$, defined on a probability space $(\Omega, \mathfrak{F}, P)$. Each element $\mathcal{X}_{\boldsymbol{s}}$ is a random particle-size curve on $(\Omega, \mathfrak{F}, P)$: $\mathcal{X}_{\boldsymbol{s}}$ maps each particle size $t \in \mathcal{T} = [t_m, t_M]$ into the random relative amount $\mathcal{X}_{\boldsymbol{s}}(\cdot, t)$ of particles having size smaller than or equal to t. As such, $\mathcal{X}_{\boldsymbol{s}} : \Omega \times \mathcal{T} \to [0, 1]$ is a random cumulative distribution function.

Following Menafoglio et al. (2014), consider the derivative random field $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ defined on $(\Omega, \mathfrak{F}, P)$, where $\mathcal{Y}_{\boldsymbol{s}}$, $\boldsymbol{s} \in D$, is a probability density

function, referred to as particle-size density (PSD) of the particle-size curve $\mathcal{X}_{\boldsymbol{s}}$. For any given $\boldsymbol{s} \in D$, the PSD $\mathcal{Y}_{\boldsymbol{s}}$ is here treated as an element of A^2.

The heterogeneous structure of the aquifer D is here modeled through K soil types $\tau^{(k)}$, $k = 1, ..., K$, whose spatial arrangement determines the drift of the field $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$. Specifically, let $\{\mathit{\Pi}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ be the random field, defined over $(\Omega, \mathfrak{F}, P)$, whose generic element $\mathit{\Pi}_{\boldsymbol{s}} = \left(\mathit{\Pi}_{\boldsymbol{s}}^{(1)}, \mathit{\Pi}_{\boldsymbol{s}}^{(2)}, ..., \mathit{\Pi}_{\boldsymbol{s}}^{(K)}\right)$ is a random probability vector, that determines the probability of occurrence of the soil type $\tau^{(k)}$ in location $\boldsymbol{s} \in D$. As such, $\mathit{\Pi}_{\boldsymbol{s}}$ is a random element in the $(K-1)$-dimensional simplex $\mathit{\Delta}^{(K-1)}$. Hereafter, $\boldsymbol{\pi}_{\boldsymbol{s}}$ denotes a realization of $\mathit{\Pi}_{\boldsymbol{s}}$. Note that, similar to continuous density functions, the probability vectors $\boldsymbol{\pi}_{\boldsymbol{s}}$ can be interpreted as K-part compositions (i.e., positive vectors summing up to a constant). The $\mathit{\Pi}_{\boldsymbol{s}}$ are thus modeled through the Aitchison geometry in the simplex and the random field $\{\mathit{\Pi}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ is assumed to be second order stationary in the space $\mathit{\Delta}^{(K-1)}$ endowed with the Aitchison geometry (Tolosana-Delgado et al., 2011). Conditionally to the field $\{\mathit{\Pi}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, the spatial field of soil types $\{T_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ is modeled as a collection of independent discrete random variables, each $T_{\boldsymbol{s}}$ being valued in $\{1, ..., K\}$ with probability mass function equal to $\mathit{\Pi}_{\boldsymbol{s}}$.

Given the spatial arrangement of the soil types, for any $\boldsymbol{s} \in D$, denote by $m_{\boldsymbol{s}}$ the Fréchet mean, also called drift, of the PSD $\mathcal{Y}_{\boldsymbol{s}}$, which is defined for $\mathcal{Y} \in A^2$ as (Fréchet, 1948)

$$\mathbb{E}[\mathcal{Y}] = \underset{y \in A^2(\mathcal{T})}{\operatorname{arginf}} \ \mathbb{E}[\|\mathcal{Y} \ominus y\|_{A^2}^2].$$

Call $\delta_{\boldsymbol{s}} = \mathcal{Y}_{\boldsymbol{s}} \ominus m_{\boldsymbol{s}}$ the residual at $\boldsymbol{s}$ in D. By construction, $\delta_{\boldsymbol{s}}$ is a centred random element of A^2: the Fréchet mean of $\delta_{\boldsymbol{s}}$ is the neutral element of perturbation, that is, $\mathbb{E}[\delta_{\boldsymbol{s}}] = 1/\eta$. The PSD $\mathcal{Y}_{\boldsymbol{s}}$ is thus represented as a perturbation of its drift – which is determined by the soil type – by the neutral-mean

stochastic residual $\delta_{\boldsymbol{s}}$. Specifically, for any $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$, the following model is assumed hereafter

$$\mathcal{Y}_{\boldsymbol{s}} | \{\Pi_{\boldsymbol{s}} = \boldsymbol{\pi}_{\boldsymbol{s}}, T_{\boldsymbol{s}} = \tau^{(k)}\} = m^{(k)} \oplus \delta_{\boldsymbol{s}}. \tag{1}$$

The residual process $\{\delta_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ is assumed to be independent of the fields $\{T_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ and $\{\Pi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ (i.e., its distribution does not depend on the soil type), and to follow a second-order stationary model in the sense of Menafoglio et al. (2013), with trace-covariogram C and trace-variogram 2γ. That is, for any $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$

$$C(\boldsymbol{s}_i - \boldsymbol{s}_j) = \mathrm{Cov}_{A^2}(\delta_{\boldsymbol{s}_i}, \delta_{\boldsymbol{s}_j}) = \mathbb{E}[\langle \delta_{\boldsymbol{s}_i}, \delta_{\boldsymbol{s}_j} \rangle_{A^2}]; \tag{2}$$

$$2\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) = \mathrm{Var}_{A^2}(\delta_{\boldsymbol{s}_i} \ominus \delta_{\boldsymbol{s}_j}) = \mathbb{E}[\|\delta_{\boldsymbol{s}_i} \ominus \delta_{\boldsymbol{s}_j}\|^2_{A^2}]. \tag{3}$$

The trace-covariogram and trace-variogram are a generalization to the functional setting of the usual notions of covariogram and variogram. In this context, they assume the same role as their finite-dimensional counterparts and allow the description of the spatial dependence of the field. Their analysis leads to the same kind of interpretations in terms of stationarity/isotropy of the field.

As in classical geostatistics, expressing the drift of the field via a linear model considerably simplifies the problem of linear prediction, as it allows employing a Universal Kriging strategy. Thus, denote by $\{\psi_k(\boldsymbol{s}), k = 1, ..., K-1\}$ a set of binary variable, which represent indicators associated with the sediment type: for $k = 1, ..., K-1$, $\psi_k(\boldsymbol{s}) = 1$ if $T_{\boldsymbol{s}} = \tau^{(k)}$, and $\psi_k(\boldsymbol{s}) = 0$ otherwise; if $T_{\boldsymbol{s}} = \tau^{(K)}$ then $\psi_k(\boldsymbol{s}) = 0$ for every $k = 1, ..., K-1$. Using these indicators as regressors and in light of model (1), the drift in $\boldsymbol{s} \in D$ can be

then described through the following linear model in A^2

$$\mathbb{E}[\mathcal{Y}_{\boldsymbol{s}}|\, \Pi_{\boldsymbol{s}} = \boldsymbol{\pi}_{\boldsymbol{s}}, T_{\boldsymbol{s}} = \tau^{(k)}] = a_0 \oplus \bigoplus_{l=1}^{K-1} \psi_l(\boldsymbol{s}) \odot a_l, \tag{4}$$

where $a_0, ..., a_{K-1}$ are (possibly unknown) deterministic coefficients in A^2. According to model (4), one has

$$\begin{cases} m^{(k)} = a_0 \oplus a_k, & k = 1, ..., K-1, \\ m^{(k)} = a_0, & k = K. \end{cases} \tag{5}$$

Hence, coefficients $a_0, ..., a_{K-1}$ represent the discrepancy of the drift in the k-th group from that of a reference soil type. Here, without loss of generality the K-th soil type is set as the reference one. Finally, let us introduce a matrix expression for the drift at a set of measurement location $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ in the domain D. Let Ψ be the design matrix of model (4): the first column of Ψ is made of ones while, for $2 \leq j \leq K$, its (i, j)-th element is $\psi_{j-1}(\boldsymbol{s}_i)$. Denote by $\boldsymbol{a} = (a_0, ..., a_{K-1})^T$ the vector of coefficients (i.e., a vector of elements in A^2) and call $\boldsymbol{m} = (m_{\boldsymbol{s}_1}, ..., m_{\boldsymbol{s}_n})^T$ the vector of drifts at the sample locations. Denote by $\boxdot$ the natural extension consistent with perturbation of the rows-by-columns matrix multiplication: $(\mathbb{A} \boxdot f)_i = \bigoplus_{j=1}^{n} \mathbb{A}_{ij} \odot f_j$, $\mathbb{A} = (\mathbb{A}_{ij}) \in \mathbb{R}^{n \times n}$, $\boldsymbol{f} = (f_i)$, $f_i \in A^2$, $i = 1, 2, ..., n$. The model for the drift vector can be then express as

$$\boldsymbol{m} = \Psi \boxdot \boldsymbol{a}. \tag{6}$$

4.2 The Class-Kriging Predictor

Given a set of locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$, and the observations of the PSD process at these locations, $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$, the aim is to predict the element $\mathcal{Y}_{\boldsymbol{s}_0}$ at the

unobserved location $\boldsymbol{s}_0$ through the best linear unbiased predictor (BLUP) $\mathcal{Y}^*_{\boldsymbol{s}_0}$ conditional to the spatial arrangement of the soil types.

The BLUP has the form $\mathcal{Y}^*_{\boldsymbol{s}_0} = \bigoplus_{i=1}^n \lambda_i^* \odot \mathcal{Y}_{\boldsymbol{s}_i}$, that is the best linear combination in the geometry of the space A^2. This guarantees that the resulting prediction is a density function. Note that this would not be the case, in general, if one relied on other widely employed geometries, such as the L^2 geometry. The aim is thus to find the Class-Kriging predictor $\mathcal{Y}^*_{\boldsymbol{s}_0}$, whose weights minimize the (conditional) variance of prediction error under the unbiasedness constraint, that is, solve

$$\begin{aligned} &\min_{\substack{\lambda_1,...,\lambda_n \in \mathbb{R}: \\ \mathcal{Y}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} = \bigoplus_{i=1}^n \lambda_i \odot \mathcal{Y}_{\boldsymbol{s}_i}}} \mathrm{Var}_{A^2}\left(\mathcal{Y}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} \ominus \mathcal{Y}_{\boldsymbol{s}_0} \middle| T_{\boldsymbol{s}_0} = \tau^{(k_0)}, T_{\boldsymbol{s}_i} \in \tau^{(k_i)}, i = 1, ..., n \right) \\ &\text{subject to} \quad \mathbb{E}_{A^2}\left[\mathcal{Y}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} \middle| T_{\boldsymbol{s}_0} = \tau^{(k_0)}, T_{\boldsymbol{s}_i} \in \tau^{(k_i)}, i = 1, ..., n \right] = m^{(k_0)}. \end{aligned} \tag{7}$$

Having observed the spatial arrangement of soil types, problem (7) can be solved by relying upon the Universal Kriging theory for Hilbert space valued random fields developed in (Menafoglio et al., 2013). The following proposition states that if the spatial arrangement of the soil types is known at the measurement and target locations $\boldsymbol{s}_0, \boldsymbol{s}_1, ..., \boldsymbol{s}_n$ and under suitable assumption on the sampling design, problem (7) admits a unique solution.

Proposition 1 (Menafoglio et al. (2013)) *Assume that the covariance matrix of the observation $\Sigma = (C(\boldsymbol{h}_{i,j})) \in \mathbb{R}^{n \times n}$, $\boldsymbol{h}_{i,j} = \boldsymbol{s}_i - \boldsymbol{s}_j$, $i, j = 1, ..., n$, is a positive definite matrix. Assume further that the design matrix $\Psi \in \mathbb{R}^{n \times K}$ is of full rank. Then problem* (7) *admits a unique solution $\boldsymbol{\lambda}^* = (\lambda_1^*, ..., \lambda_n^*)^T \in \mathbb{R}^n$, which is obtained by solving the system of $n + K$ linear equations*

$$\begin{pmatrix} \Sigma & \Psi \\ \Psi^T & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\zeta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\sigma}_0 \\ \boldsymbol{\psi_0} \end{pmatrix}, \tag{8}$$

where $\boldsymbol{\zeta} = (\zeta_0, ..., \zeta_{K-1})^T$ *are* K *Lagrange multipliers associated with the unbiasedness constraint, whereas* $\boldsymbol{\sigma}_0 = (C(\boldsymbol{h}_{i,0})) \in \mathbb{R}^n$*, and* $\boldsymbol{\psi_0} = (\psi_k(\boldsymbol{s}_0)) \in \mathbb{R}^K$*. Conditionally on* $T_{\boldsymbol{s}_0}, T_{\boldsymbol{s}_1}, ..., T_{\boldsymbol{s}_n}$*, and denoting by* $(\boldsymbol{\lambda}^{*T}, \boldsymbol{\zeta}^{*T})^T$ *the solution of Eq.* (8)*, the Universal Kriging variance of predictor* $\mathcal{Y}^*_{\boldsymbol{s}_0}$ *is then*

$$\begin{aligned}\sigma^2_*(\boldsymbol{s}_0) &= \mathrm{Var}_{A^2}\left(\mathcal{Y}^*_{\boldsymbol{s}_0} \middle| T_{\boldsymbol{s}_0} = \tau^{(k_0)}, T_{\boldsymbol{s}_i} \in \tau^{(k_i)}, i = 1, ..., n\right) = \\ &= C(\mathbf{0}) - \sum_{i=1}^{n} \lambda^*_i C(\boldsymbol{h}_{i,0}) - \sum_{k=0}^{K-1} \zeta^*_k \psi_k(\boldsymbol{s}_0). \end{aligned} \tag{9}$$

Note that Eq. (8) is a linear system of $n+K$ equations, which has the very same form as its counterpart employed in classical geostatistics. Further, from expression (9), the following Chebyshev inequality for the prediction errors can be derived

$$P\left(\|\mathcal{Y}^*_{\boldsymbol{s}_0} \ominus \mathcal{Y}_{\boldsymbol{s}_0}\|_{A^2} > \kappa\sigma_*(\boldsymbol{s}_0) \middle| T_{\boldsymbol{s}_0} = \tau^{(k_0)}, T_{\boldsymbol{s}_i} \in \tau^{(k_i)}, i = 1, ..., n\right) < \frac{1}{\kappa^2}, \kappa > 0. \tag{10}$$

4.3 Estimating the Structure of Spatial Dependence and the Drift

As in the classical geostatistical framework, the solution of system (8) requires the structure of spatial dependence to be estimated if it is not a priori known. For this purpose, the estimate of the variogram can be employed (Cressie, 1993). A method of moment estimator $\widehat{\gamma}(\boldsymbol{h})$ from the residuals is here adopted

$$\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{2|N(\boldsymbol{h})|} \sum_{(i,j) \in N(\boldsymbol{h})} \|\delta_{\boldsymbol{s}_i} \ominus \delta_{\boldsymbol{s}_j}\|^2_{A^2}, \tag{11}$$

followed by a fitting of a valid model via weighted least squares. The trace-semivariogram γ defined in (3) is a real valued function fulfilling the same set of properties as its classical counterpart (e.g., conditional negative definiteness). Hence, usual parametric structures (e.g., exponential, spherical,

Matérn) can be employed as valid models. Note that estimator (11) depends on the residuals, which are usually unobserved. Menafoglio et al. (2013), in the general context of Hilbert data, propose to estimate the residuals by difference from the generalized least squares (GLS) estimate of the drift at the measurement locations. Specifically, let $\boldsymbol{\mathcal{Y}}_{\boldsymbol{s}} = (\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n})^T$ be the vector of observations, call $\widehat{\boldsymbol{m}}_{\boldsymbol{s}}^{GLS} = (\widehat{m}_{\boldsymbol{s}_1}^{GLS}, ..., \widehat{m}_{\boldsymbol{s}_n}^{GLS})^T$ the GLS estimator of the drift and $\widehat{\boldsymbol{a}}^{GLS} = (a_0^{GLS}, ..., a_{K-1}^{GLS})^T$ the GLS estimator of the coefficients vector $\boldsymbol{a} = (a_0, ..., a_{K-1})^T$. It follows from Eq. (6) that $\widehat{\boldsymbol{m}}_{\boldsymbol{s}}^{GLS} = \Psi \boxdot \widehat{\boldsymbol{a}}_{\boldsymbol{s}}^{GLS}$. Estimator $\widehat{\boldsymbol{a}}^{GLS}$ can be explicitly determined as (Menafoglio et al., 2013)

$$\widehat{\boldsymbol{a}}^{GLS} = (\Psi^T \Sigma^{-1} \Psi)^{-1} \Psi^T \Sigma^{-1} \boxdot \boldsymbol{\mathcal{Y}}_{\boldsymbol{s}}. \tag{12}$$

Note that Eq. (12) depends not only on the design matrix Ψ, but also on Σ, the covariance matrix of the residuals.

To jointly estimate the residuals and their covariance matrix, an iterative algorithm, initialized to an ordinary least squares estimate of the drift, is here employed. The algorithm closely follows the iterative algorithm which is widely employed in classical geostatistics to estimate the drift in a Universal Kriging context (Cressie, 1993, p.23). For completeness, the algorithm is presented in Appendix A; the reader is referred to Menafoglio et al. (2013, Sect. 4) for further details.

4.4 Assessing the Spatial Arrangement of Soil Types

In some field studies the spatial arrangement of the soil types is only observed at a few sampled locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$. In this case, to predict the unobserved PSD $\mathcal{Y}_{\boldsymbol{s}_0}$ in $\boldsymbol{s}_0$ one needs to first reconstruct the indicators $\psi_1(\boldsymbol{s}_0), ..., \psi_{K-1}(\boldsymbol{s}_0)$ of the soil types at $\boldsymbol{s}_0$, possibly employing the information related to the

probability of occurrence of the soil types, which is described by the field $\{\Pi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$.

If a complete observation of the field of probabilities $\{\Pi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ were available, one could assign to location $\boldsymbol{s}_0$ the most likely soil type, that is, the one associated with the highest probability in $\boldsymbol{\pi}_{\boldsymbol{s}_0}$. Instead, in the presence of a partial observation $\boldsymbol{\pi}_{\boldsymbol{s}_1}, ..., \boldsymbol{\pi}_{\boldsymbol{s}_n}$ of the random field $\{\Pi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, one could first predict $\boldsymbol{\pi}_{\boldsymbol{s}_0}$, and consequently assign the soil type in $\boldsymbol{s}_0$. Recall that each element of the field $\{\Pi_{\boldsymbol{s}}\}$ is a random probability vector. Hence an appropriate Kriging method should be applied. Application of a standard Cokriging to probability vectors does not guarantee that the prediction fulfils the properties of a probability vector (e.g., positivity, unit sum constraint). Since the $\boldsymbol{\pi}_{\boldsymbol{s}}$'s can be interpreted as K-part compositions, and consistent with the Bayes space approach, Kriging of the $\boldsymbol{\pi}_{\boldsymbol{s}}$'s is here performed with a log-ratio technique, namely the Simplicial Kriging (SK) of Tolosana-Delgado et al. (2008a, 2011). The SK consists of Cokriging the probability vectors $\boldsymbol{\pi}_{\boldsymbol{s}_1}, ..., \boldsymbol{\pi}_{\boldsymbol{s}_n}$ within the $(K-1)$-dimensional simplex $\Delta^{(K-1)}$ endowed with the Aitchison geometry. Tolosana-Delgado et al. (2008a, 2011) prove that this is equivalent to employ a standard Cokriging procedure based on the n vectors $\boldsymbol{l}_{\boldsymbol{s}_i} = \left(l_{\boldsymbol{s}_i}^{(1)}, ..., l_{\boldsymbol{s}_i}^{(K-1)}\right)^T$, $i = 1, ..., n$, where $\boldsymbol{l}_{\boldsymbol{s}_i}$ stands for the isometric log-ratio transform (ilr, Pawlowsky-Glahn and Buccianti, 2011) of $\boldsymbol{\pi}_{\boldsymbol{s}_i}$ in $\boldsymbol{s}_i$. This transform maps each compositional vector $\boldsymbol{\pi}_{\boldsymbol{s}_i}$ onto a $(K-1)$-dimensional vector of coordinates with respect to a given orthonormal basis of the simplex. The vector of coordinates can then be treated according to the usual Euclidean geometry.

The random field $\{\Pi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ is in general latent. In this case, one may approximate the probability vector in $\boldsymbol{s}_0$ through a generalized indicator, as suggested by Tolosana-Delgado et al. (2008a). The generalized indicator at a location $\boldsymbol{s}$ in D assigns a high probability (e.g., 0.95, 0.99) to the soil type which

is actually observed, and uniformly assigns the probability of the remaining soil types. That is, for $\boldsymbol{s} \in D$ and $k = 1, ..., K$, the generalized indicator $p_{\boldsymbol{s}}^{(k)}$ is defined as

$$p_{\boldsymbol{s}}^{(k)} = \begin{cases} 1-b, & T_{\boldsymbol{s}} = \tau^{(k)} \\ \frac{b}{K-1}, & T_{\boldsymbol{s}} \neq \tau^{(k)}, \end{cases} \tag{13}$$

where b is a (small) parameter usually set to $b = 0.05$, or $b = 0.1$ (Tolosana-Delgado et al., 2008a). The vectors of generalized indicators $\boldsymbol{p}_{\boldsymbol{s}_1}, ..., \boldsymbol{p}_{\boldsymbol{s}_n}$, with $\boldsymbol{p}_{\boldsymbol{s}_i} = \left(p_{\boldsymbol{s}_i}^{(1)}, ..., p_{\boldsymbol{s}_i}^{(K)}\right)^T$, are then used in place of the latent probability vectors $\boldsymbol{\pi}_{\boldsymbol{s}_1}, ..., \boldsymbol{\pi}_{\boldsymbol{s}_n}$ for SK prediction purposes. The SK method returns the BLU prediction – in the sense of the Aitchison geometry on $\Delta^{(K-1)}$ (Tolosana-Delgado et al., 2008a) – at $\boldsymbol{s}_0$, $(p_{\boldsymbol{s}_0}^{(1)*}, ..., p_{\boldsymbol{s}_0}^{(K)*})^T$, which can be then employed to assign at location $\boldsymbol{s}_0$ the most likely soil type $p_{\boldsymbol{s}_0}^{(k)*}$ and then solve system (8).

4.5 SFC K-Mean: a K-Mean Method for Spatially Dependent Functional Compositions

The spatial arrangement of the soil types often is not observed anywhere in the system. In these cases, the information content embedded in the particle-size distributions can be employed to identify the soil types through a cluster analysis. This section introduces an original unsupervised classification method, the SFC K-mean, which is coherent with the model introduced above and allows to cluster objects that are spatially dependent, functional and compositional. The proposed method is inspired by the K-mean clustering method (Mc Queen, 1967), but properly tailored to the framework of this work: dissimilarities between data are here computed according to the Aitchison geometry and spatial dependence is taken into account by computing the cluster centroids $\mathcal{C}_1, ..., \mathcal{C}_K$ through the GLS estimators $\{\widehat{m}_1^{GLS}, ..., \widehat{m}_K^{GLS}\}$, obtained accord-

ing to (12). Note that, if the K clusters $C_1, ..., C_K$ correctly represent the soil types, $\widehat{m}_1^{GLS}, ..., \widehat{m}_K^{GLS}$ provide the BLU estimates of the Frechét means $\{m^{(1)}, ..., m^{(K)}\}$, which in turn minimize of the global variance within the clusters; that is, for $k = 1, ..., K$

$$m^{(k)} = \underset{\xi \in A^2(\mathcal{T})}{\operatorname{arginf}} \ \mathbb{E}\left[\|\mathcal{Y}_{\boldsymbol{s}} \ominus \xi\|_{A^2}^2 \big| \, T_{\boldsymbol{s}} = \tau^{(k)} \right].$$

Assignment of PSDs to clusters is performed by minimizing the empirical total variance – in the Aitchison sense – within the clusters, that is, $\sum_{k=1}^{K} \sum_{\mathcal{Y}_{\boldsymbol{s}_i} \in C_k} \|\mathcal{Y}_{\boldsymbol{s}_i} \ominus \mathcal{C}_k\|_{A^2}^2$. The proposed method is sketched in Algorithm 1.

Algorithm 1 Given the realizations $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$ of the field $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ in locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$, and a number K of target clusters:

0. Initialization:

Fix K initial centroids $\mathcal{C}_1, ..., \mathcal{C}_K \in A^2(\mathcal{T})$ (e.g., by randomly sampling K of the n data $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$);

1. Assignment:

For each $i = 1, ..., n$, assign the data-point $\mathcal{Y}_{\boldsymbol{s}_i}$ to the k-th cluster, $k \in \{1, ..., K\}$, if its centroid $\mathcal{C}_k$ is the nearest one

$$k = \operatorname{argmin}\{\|\mathcal{Y}_{\boldsymbol{s}_i} - \mathcal{C}_j\|_{A^2}, \ j = 1, ..., K\};$$

2. Representation:

For each $k = 1, ..., K$, update the centroid $\mathcal{C}_k$ to the generalized least squares estimate of the within-cluster mean

$$\begin{cases} \mathcal{C}_k = \widehat{a}_0^{GLS} \oplus \widehat{a}_k^{GLS}, & k = 1, ..., K-1, \\ \mathcal{C}_k = \widehat{a}_0^{GLS}, & k = K, \end{cases}$$

where $(\widehat{a}_0^{GLS}, ..., \widehat{a}_{K-1}^{GLS})^T$ is given by Eq. (12), with $\Psi_{i,1} = 1$, for all $i = 1, ..., n$, $\Psi_{i,k} = 1$ if $\mathcal{Y}_{\boldsymbol{s}_i}$ belongs to cluster k, $\Psi_{i,k} = 0$ otherwise;

3. *Iteration:*

   Repeat 1. and 2. until no change in assignment occurs or a given maximum number of iterations is reached.

Whenever the structure of spatial dependence is unknown, Step 2 of Algorithm 1 requires the use of the iterative algorithm described in Subsect. 4.1 (and fully reported in Appendix A) to jointly estimate the covariance structure and the clusters centroids. The resulting nested iterative algorithm yields: the identified clusters, the estimates of the within-cluster drifts (i.e., the centroids) and the estimated structure of spatial dependence. From a computational viewpoint, the algorithm may become computationally demanding as the number n of data increases. However, Menafoglio et al. (2013) show via simulation that the iterative algorithm which may be involved in Step 2 typically converges within five iterations. As in standard K-mean methods, Algorithm 1 may suffer from the occurrence of local minima, which can be encountered depending on the initialization of the algorithm. This issue may become particularly evident in the presence of highly unbalanced cluster sizes, for which the starting condition may become quite influential. This issue can then be circumvented by running the algorithm for multiple initializations.

Finally note that, similar to the K-mean method, the SFC K-mean requires the number of clusters K to be known or chosen prior to applying Algorithm 1. A proper selection of K can be performed by any of the available standard techniques, for instance, one can minimize the total dissimilarity within clusters over a feasible range of K, as illustrated in Sect. 5.

5 Geostatistical Analysis of Field Data

In this section, the procedure detailed in Sect. 4 is applied to the field data described in Sect. 2. First note that the left tails of the observed particle-size distributions are censored, due to the sieve-measurement procedure. For illustrative purposes, the focus is here posed on the particle-size density conditional to the domain of observation $\mathcal{T} = [t_1, t_{12}]$, that is proportional – thus equivalent in the sense of the Bayes space – to the PSD restricted to the domain $\mathcal{T}$. For a possible way to account for the information related to the mass in the left tail the reader is referred to Appendix B.

5.1 Analysis of Conditional Particle-Size Densities

A smooth version of (conditional) PSD can be obtained from raw data by following the approach proposed by Menafoglio et al. (2014). First, particle-size data conditional to the domain $\mathcal{T} = [t_1, t_{12}]$ are obtained from raw data as

$$\widetilde{\mathcal{X}}^{(c)}_{\boldsymbol{s}_i}(t_j) = \frac{\widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t_j) - \widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t_1)}{\widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t_{12}) - \widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t_1)},$$

for $j = 1, ..., 12$, $i = 1, ..., 406$. For each $i = 1, ..., 406$, a set of $m = 70$ Bernstein polynomials is then employed to smooth the linear interpolation of $\widetilde{\mathcal{X}}^{(c)}_{\boldsymbol{s}_i}(t_j)$, $j = 1, ..., 12$, upon considering log-transformed particle diameters (hereafter t denotes log-transformed diameters; Fig. 2). The selected number of Bernstein polynomials guarantees a tolerance of 0.01 in the median sum of squared errors (SSE) between the raw observations and the values attained by the smoothed curves for the adopted grain sieve sizes. Notice that different smoothing techniques might be employed for data preprocessing; in some cases, data smoothing might not be required.

The notation of Sect. 4 is here employed as follows. The functional dataset of smoothed conditional PSCs at location $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$, $n = 406$, is denoted by $\mathcal{X}_{\boldsymbol{s}_1}, ..., \mathcal{X}_{\boldsymbol{s}_n}$ (Fig. 2(a)). Their densities $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$ (Fig. 2(b)), determined in closed form from smoothed PSCs (Menafoglio et al., 2014), are embedded into the space $A^2(\mathcal{T})$ over the compact domain $\mathcal{T} = [\log(0.063), \log(100)]$ and analyzed according to the methodology described in Sect. 4.

5.1.1 Clustering of the Data via SFC K-mean

As recalled in Sect. 2, previous analyses at the site employed standard multivariate techniques to classify the raw PSCs and infer the spatial arrangement of three main lithotypes within the system. Consistent with the compositional approach adopted in this work, the lithotypes within the system are here identified by applying the SFC K-mean method devised in Algorithm 1.

Based on a preliminary analysis on directional variograms performed for the complete set of available PSCs, a geometric anisotropy with anisotropy ratio of $R = 0.04$ between the horizontal and vertical directions is assumed for the field. Hereafter, the estimates are referred to a rescaled isotropic spatial domain obtained by dilating vertical coordinates z by a factor $1/R = 25$. Similar to Menafoglio et al. (2014), an exponential variogram model with nugget is selected, and its parameters are estimated via weighted least squares.

The number K of clusters is identified upon evaluating the residual dissimilarity between the PSDs and the cluster centroids. The lower the residual dissimilarity, the higher the fidelity with which the cluster centroids characterize the group components. The presented results are based on K ranging in $\{1, ..., 10\}$. Figure 3 depicts on a log-scale the boxplots of the dissimilarities between each data-point $\mathcal{Y}_{\boldsymbol{s}_i}$ and the center of the cluster to which it is assigned, that is, $d_i = \|\mathcal{Y}_{\boldsymbol{s}_i} - m^{(k)}\|^2_{A^2}$, if $T_{\boldsymbol{s}_i} = \tau^{(k)}$. An elbow in the median dissimilarity is clearly visible for $K = 2$, even though the corresponding box-

plot evidences the presence of five outliers. These are collected into a separate cluster when $K = 3$, leading to an elbow in the mean dissimilarity. These observations motivates the choice of $K = 3$, which is also consistent with the findings of Riva et al. (2006). The clustering results associated with $K = 3$ are depicted in Fig. 4.

The cluster centroids correspond to the estimates of the drift in Eq. (1) and are displayed in Fig. 4(b). The centroids of the first two clusters are interpreted as a characterization of two different behaviors within the right tail of the particle-size distribution, the first cluster featuring a lighter tail than the other one. The differentiation between the two clusters is consistent with the results of Riva et al. (2006), who highlight that a key difference between the two main sedimentological facies identified is ascribed to the proportions of gravel and sand material, which are associated with the largest particle diameters. The third cluster represents 1% of the sample and is associated with a centroid displaying its main peak at a grain size of about 0.4 mm (Fig. 4(b)). This is consistent with the main sedimentological composition of the corresponding cluster identified by Riva et al. (2006).

Inspection of Fig. 4(b) and Table 1 evidences that the first cluster appears to be mainly associated with the northern boreholes B1-B5, and the second cluster with the boreholes F0-F6. This is partly consistent with the observation that the former group of boreholes is located in an area where the Neckar river displays a bend, thus favoring the accumulation of the finer sediments in this area. Further, note that the particle-size densities at borehole B5, which are considered in the analysis of Menafoglio et al. (2014), are all assigned to the first cluster, coherent with the stationarity assumption considered by these authors.

5.1.2 Kriging Predictions

The generalized indicators are computed as in Eq. (13) with $b = 0.01$, and their spatial dependence is analyzed via simplicial variography (Tolosana-Delgado et al., 2008b, 2011). First note that a possible dependence of the cluster assignment on the horizontal x and y coordinates may be recognized by inspection of Fig. 4(c). This would support the introduction of a nonstationary model for the indicators. However, a cross-validation analysis (not reported here) does not support the adoption of a non-stationary model. Therefore, the results illustrated hereinafter are obtained under a stationarity assumption.

A geometric anisotropy with anisotropy ratio $R = 0.04$ between the horizontal and vertical directions is here considered, consistent with the preliminary results discussed in Subsect. 5.1.1. The empirical estimate of indicator variograms (referred to the rescaled spatial domain) is depicted in Fig. 5(a), together with the fitted exponential models. The results of the SK interpolation are depicted in Fig. 5(b), colors being associated with the predicted probability of occurrence of the soil types identified via SFC K-mean. The latter reflects the association of the first cluster of soil material with the Northern part of the aquifer (borehole B1-B5) which has been noted from the inspection of Fig. 4(c).

Finally, Fig. 6 depicts the results of the Kriging interpolation based on the residual semivariogram estimated through Algorithm 1 (Fig. 4(a)), and the SK prediction (Fig. 5(b)), in terms of both point-wise predictions and the associated Kriging variance. The kriged field provides a smooth interpolation of the available data. As such, outlying observations (e.g., the blue curve at $z = 305.5$ at borehole F6 in Fig. 6(b)) locally influence prediction results. Instead, for distances higher than the estimated range, the kriged field is representative of mean particle-size distribution. With reference to this point, note that a

sharp assignment has been here considered for Class-Kriging prediction, that is, the information content embedded into the kriged indicators $\boldsymbol{\pi}_{\boldsymbol{s}}^*$ has been used to derive the binary information associated with the soil type assignment. Nevertheless, the information provided by the kriged indicators $\boldsymbol{\pi}_{\boldsymbol{s}}^*$ may be employed as a further indication of the uncertainty associated with the drift estimate and Kriging prediction.

5.2 Cross-Validation Results

In this section a leave-one-out cross-validation analysis is performed to assess the quality of the SK predictions of generalized indicators, and the FCCK predictions of conditional PSDs.

The confusion matrix whose entries are listed in Table 2 is considered to evaluate the quality of the SK predictions. Even though the first cluster appears to be well-predicted (error: 14%), the prediction within the second cluster is affected by some errors, which are mostly registered at boreholes B1-B5. Overall, the discrepancy between the prediction via SK and the SFC K-mean assignment is 25.37%. Improved results are expected under conditions of stronger spatial dependence between generalized indicators or in the presence of a partial observation (or prior knowledge) of the random field $\{\Pi_{\boldsymbol{s}}\}$. The latter is here completely unobserved, due to the lack of prior knowledge on the spatial distribution of soil types disjoint from the information content of the PSCs.

Assessment of the impact of the SK prediction error on the FCCK prediction is performed by measuring the cross-validation SSE when cross-validation analysis is carried out (i) jointly on the indicators and on the curves, and (ii) only on the PSDs. Here, the SSE is computed as $\|\mathcal{Y}_{\boldsymbol{s}_i} - \mathcal{Y}_{\boldsymbol{s}_i}^{(CV)}\|_{A^2}$, $\mathcal{Y}_{\boldsymbol{s}_i}^{(CV)}$ denoting the Kriging prediction at $\boldsymbol{s}_i$ obtained upon removing the i-th data-

point $\mathcal{Y}_{\boldsymbol{s}_i}$ from the dataset. The SSE results associated with case (i) appear fairly satisfactory if compared to the mean norm of the data, with a 8.51% relative median SSE. However, a relative mean SSE of 20.36% is observed, due to several outliers in the SSE. These results are comparable with those obtained via the stationary FCK of Menafoglio et al. (2014) (median SSE: 5.23%, mean SSE: 20.23%). However, the overall quality of Kriging predictions is significantly improved when the indicators are not cross-validated (case (ii)), with a 0.96% and 3.50% median and mean SSE, respectively. The latter is actually the error which is ascribed to the FCCK predictor, the uncertainty on the cluster assignment being responsible for the remaining portion of the prediction error. This result can be expected, as Hron et al. (2015) note that the Bayes space geometry is very sensitive to the information content within the tails of the distribution, due to the *relative scale* property of compositions. The information content related to small values is extremely relevant when analyzing the data through log-ratios, and this reflects the observation that a small variation on a small probability value (e.g., from 0.05 to 0.1, that is 2 multiple) is more influential than the same variation over a high probability (from 0.5 to 0.55, i.e., 1.1 multiple). Therefore, an accurate description of the right tails of the PSDs in terms of cluster-varying drift turns into a significant gain in terms of prediction error.

Finally the cross-validation results of case (ii) are employed to evaluate the empirical coverage of Chebyshev inequality (10). A total of 95.57% of the PSCs are associated with a global prediction error which is comprised within the Chebyshev band built upon setting $\kappa = 2$, against a theoretical level of 75%. The conservative nature of Chebyshev bands is also supported by the results obtained for $\kappa = 3, 4$ (empirical vs theoretical coverage: 97.54% vs 88.89% ($\kappa = 3$) and 98.78% vs 93.75% ($\kappa = 4$)). This result is also consistent with Menafoglio et al. (2014), who found the Chebyshev bands to be quite

conservative when applied to a one-dimensional field dataset. Improvement of the uncertainty assessment may be obtained, for instance, upon resorting to alternative approaches, such as semiparametric bootstrap (Pigoli et al., 2013; Franco-Villoria and Ignaccolo, 2014, in the context of object oriented data analysis and FDA, respectively). A detailed analysis of this aspect in the context of the experimental dataset here analyzed is outside the scope of this work.

6 Conclusions and Further Research

In this work, an original theoretical framework for the geostatistical characterization of a set of heterogeneous particle-size densities (PSDs) has been established. PSDs are directly associated with particle-size curves (PSCs), which are routinely measured in hydrogelogical, hydrogeophysical and soil science applications. PSDs have been interpreted as FCs, and analyzed through the Aitchison geometry. The FCCK methodology relies on a novel hierarchical model for FCs and constitutes a generalization of the FCK methodology introduced by Menafoglio et al. (2014). These developments allows treating PSDs which are featured by a grouping structure driven by the mean soil textural properties of the system.

Application-oriented challenges associated with the lack of information about the spatial arrangement of the soil types have been addressed by proposing a novel clustering method for spatially dependent FCs. This method is consistent with the FCCK model and enables one to infer a grouping structure from a set of observed PSDs associated with spatially varying soil textural properties.

The quality of the predictions has been assessed via cross-validation and appears satisfactory, even as it proved to be strongly dependent on a proper

assessment of the spatial arrangement of the soil types. Indeed, a precise description of the right tails of PSDs, which are closely related to the cluster assignment, proved to be key to enhance the FCCK prediction performances with respect to the stationary FCK approach of Menafoglio et al. (2014). Additional research along these lines includes the improved assessment of prediction uncertainty, possibly upon resorting to computer intensive methods, such as semiparametric bootstrap.

Acknowledgements Financial support of MIUR (Project "Innovative methods for water resources under hydro-climatic uncertainty scenarios", PRIN 2010/2011) is gratefully acknowledged. Support from the European Union's Horizon 2020 Research and Innovation programme (Project "Furthering the knowledge Base for Reducing the Environmental Footprint of Shale Gas Development" FRACRISK - Grant Agreement No. 640979) is also acknowledged.

Appendix A: An Iterative Algorithm to Jointly Estimate the Drift and the Trace-Variogram

This appendix reports the iterative algorithm which can be employed to jointly estimate the drift and the covariance matrix of the residuals. The reader is referred to Menafoglio et al. (2013) for further details.

Algorithm 2 Given $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$, observations of the field $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ in the sites $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$, and the design matrix Ψ:

0. Initialization:

Estimate the drift coefficient vector $\boldsymbol{a}$ via the ordinary least squares estimator $\widehat{\boldsymbol{a}}^{OLS} = (\Psi^T\Psi)^{-1}\Psi^T \boxdot \boldsymbol{\mathcal{Y}}$, Ψ denoting the design matrix appearing in Eq. (6). Set $\widehat{\boldsymbol{a}} := \widehat{\boldsymbol{a}}^{OLS}$;

1. Trace-variogram estimate:

Estimate the trace-semivariogram $\gamma(\cdot)$ from the estimated residuals $\widehat{\boldsymbol{\delta}} =$

$\boldsymbol{\mathcal{Y}} \ominus (\Psi \boxdot \widehat{\boldsymbol{a}})$ via estimator (11), and fit a valid model. Derive from this the estimate $\widehat{\Sigma}$ of Σ;

2. *Drift estimate:*

   Estimate $\boldsymbol{a}$ with $\widehat{\boldsymbol{a}}^{GLS}$ according to (12) with $\widehat{\Sigma}$ in place of Σ, and set $\widehat{\boldsymbol{a}} := \widehat{\boldsymbol{a}}^{GLS}$;

3. *Iteration:*

   Repeat 1.–2. until convergence;

Appendix B: Dealing with Data Censoring

As a result of the sieve measurement procedure employed to collect the field data of Sect. 2, the left tails of the observed particle-size distributions are censored. Nevertheless, the proportions of particles within the censored left tails are known, as they coincide with the observations of the PSCs $\widetilde{\mathcal{X}}_{\boldsymbol{s}_i}$, $i = 1, ..., n$, at the first sieve t_1, namely $\widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t_1)$, $i = 1, ..., n$. This Appendix illustrates a method to account for such information, that is neglected when analyzing conditional PSDs.

Although one could select a priori a distribution to represent the left tail $\{\widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t), t < t_1\}$, $i = 1, ..., n$, (e.g., uniform, Menafoglio et al., 2014), this assumption might be highly influential on the analysis, especially in the presence of large variability in the $\widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t_1)$ for $i = 1, ..., n$. Instead, for each observed location $\boldsymbol{s}_i$, $i = 1, ..., n$, one can decouple the available information into the particle-size density conditional to the domain of observation $\mathcal{T} = [t_1, t_{12}]$ and a two-dimensional vector $\boldsymbol{\zeta}_{\boldsymbol{s}_i} = (\zeta_{\boldsymbol{s}_i}, 1 - \zeta_{\boldsymbol{s}_i})$, respectively collecting the mass within the censored left tail (i.e., for $t < t_1$) and the observed domain $[t_1, t_{12}]$ (Fig. 7). Note that $\boldsymbol{\zeta}_{\boldsymbol{s}_i}$ is a probability vector, that can be thus interpreted as a two-parts composition.

To treat the entire information available at the sample sites one can then proceed as follows: first, consider the conditional PSDs as functional compositions and proceed as in Sect. 5; second, separately treat the above mentioned two-part compositions via appropriate geostatistical methods, and finally combine the results to provide a complete description of predicted PSDs at unsampled locations.

Note that to predict the two-part compositions $\boldsymbol{\zeta}_{\boldsymbol{s}_i}$, $i = 1, ..., n$, one can employ the SK methodology of Tolosana-Delgado et al. (2008a, 2011), which has been recalled in Subsect. 4.4.

The results obtained on field data are now illustrated. Second-order stationarity is assumed, since no evident pattern can be recognized in the spatial arrangement of $\boldsymbol{\zeta}_{\boldsymbol{s}_i}$. The simplicial variogram of the compositions is modeled via a spherical model with nugget, fitted to the empirical estimate via weighted least squares (Fig. 8(a)). The corresponding SK predictions are depicted in Fig. 8(b). Cross-validation results show that the quality of SK predictions is quite satisfactory. The kriged field of PSDs, obtained by combining the results of FCCK and SK, is depicted in Fig. 9. These results are obtained by multiplying each predicted conditional PSD $\mathcal{Y}^*_{\boldsymbol{s}_0}$, $\boldsymbol{s}_0 \in D$, by the kriged mass $1 - \boldsymbol{\zeta}^*_{\boldsymbol{s}_0}$.

The availability of the complete information content related to predicted PSDs may be useful, for instance, for classification of sediments. For instance, these results can be employed to represent predictions over the soil textural triangle, that is widely employed in field investigations to classify geomaterials according to their textural properties.

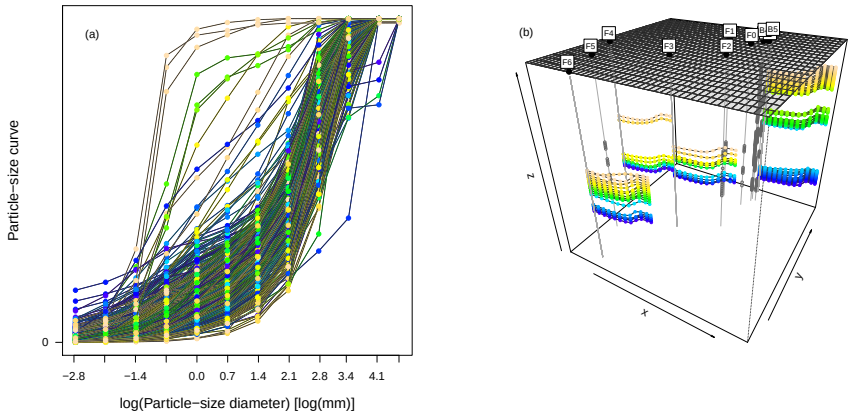
(a)
Particle−size curve
0
−2.8
−1.4
0.0
0.7
1.4
2.1
2.8
3.4
4.1
log(Particle−size diameter) [log(mm)]
(b)
F1
F2
F3
F4
F5
z
x
y

Fig. 1

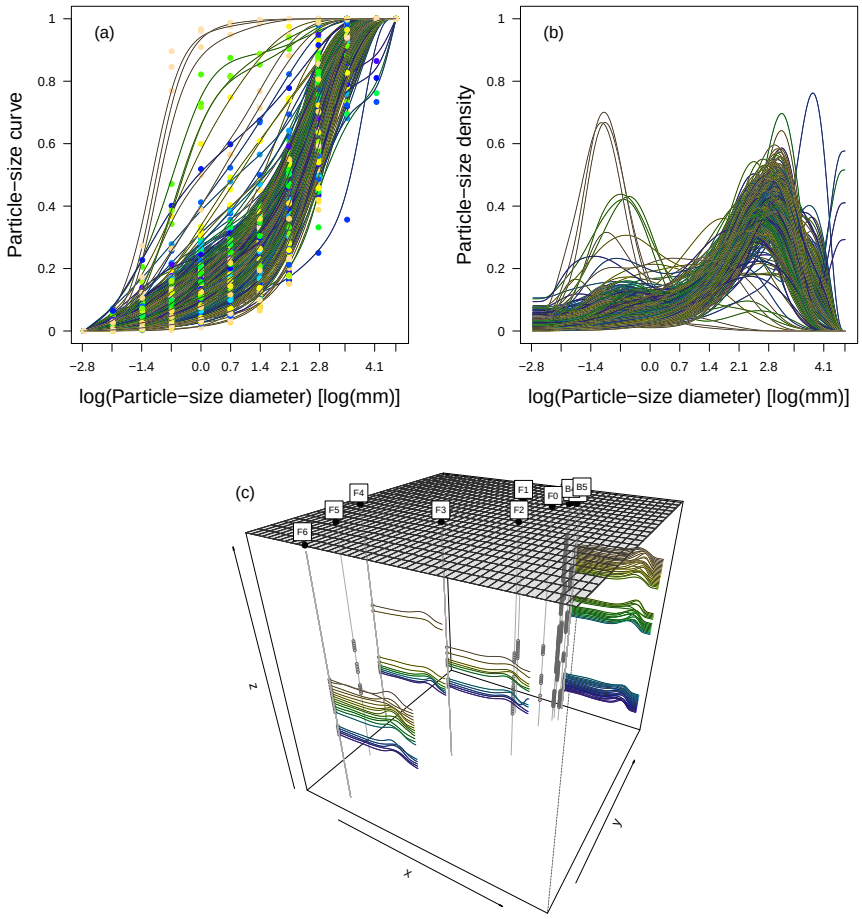
(a)
Particle−size curve
log(Particle−size diameter) [log(mm)]
(b)
Particle−size density
log(Particle−size diameter) [log(mm)]
(c)
F0
F1
F2
F3
F4
F5
F6
B
B5
z
x
y

Fig. 2

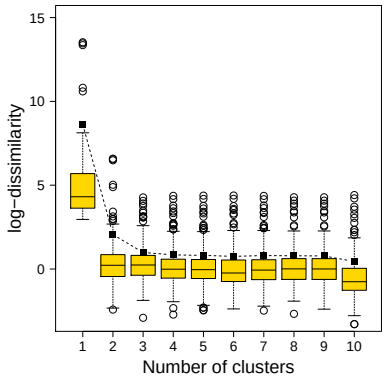

Fig. 3

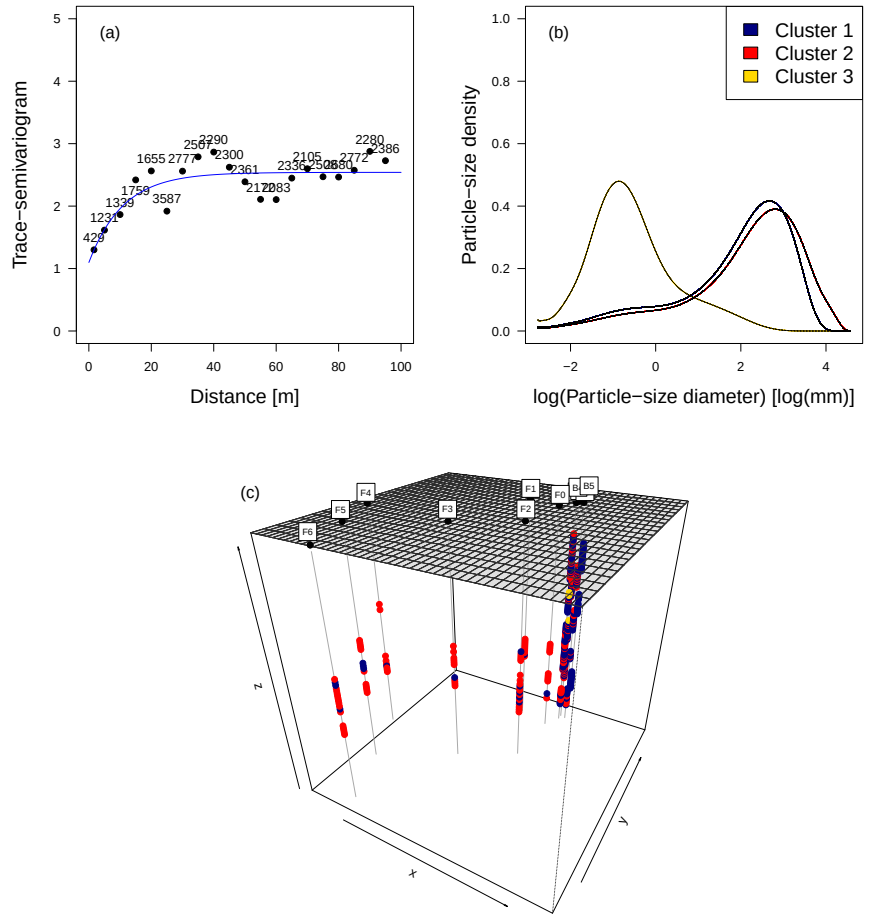
(a)
Trace−semivariogram
Distance [m]
(b)
Particle−size density
log(Particle−size diameter) [log(mm)]
Cluster 1
Cluster 2
Cluster 3
(c)
F0
F1
F2
F3
F4
F5
F6
B5
z
x
y

Fig. 4

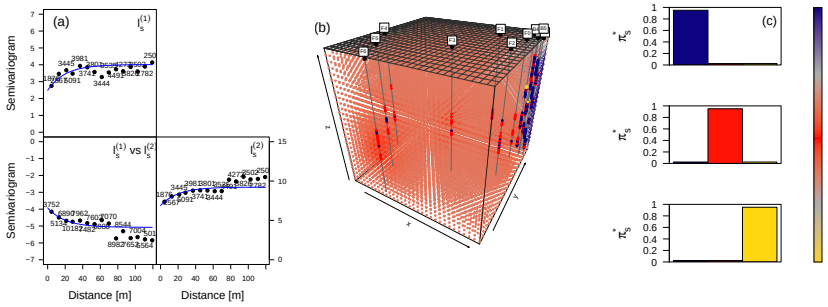

Fig. 5

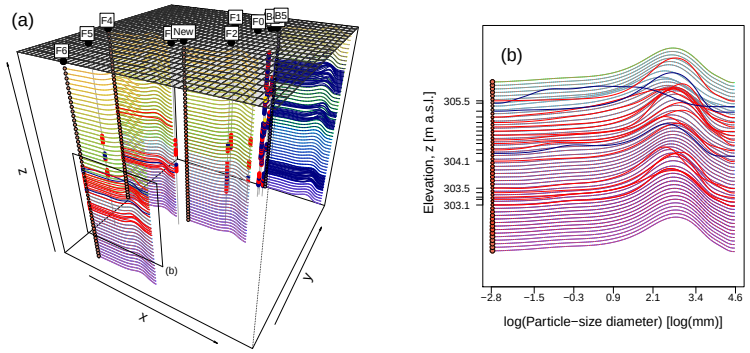
(a)
F6
F5
F4
New
F2
F1
F0
B5
z
x
y
(b)
(b)
Elevation, z [m a.s.l.]
305.5
304.1
303.5
303.1
−2.8
−1.5
−0.3
0.9
2.1
3.4
4.6
log(Particle−size diameter) [log(mm)]

Fig. 6

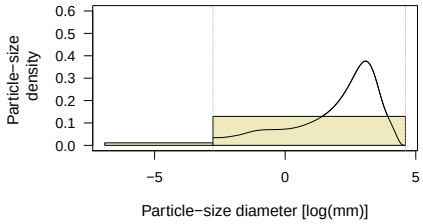
Particle−size
density
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−5
0
5
Particle−size diameter [log(mm)]

Fig. 7

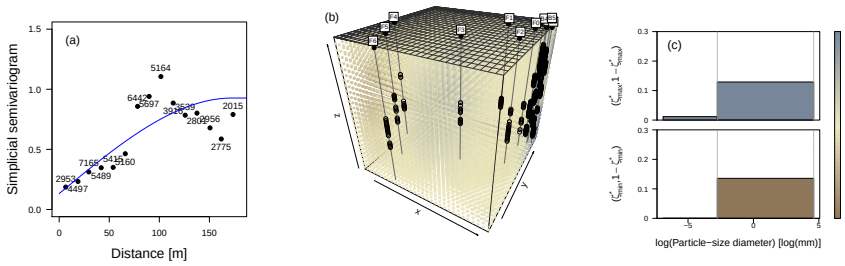
(a)
Simplicial semivariogram
Distance [m]
(b)
(c)
log(Particle−size diameter) [log(mm)]

Fig. 8

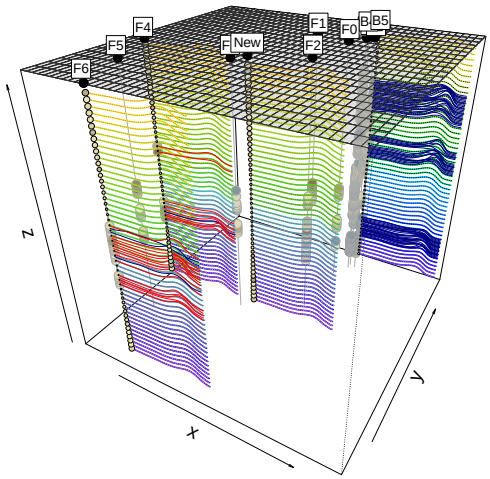

Fig. 9

Fig. 1 Raw data: (a) complete set of the raw PSCs (b) raw PSCs along boreholes B5, F3, F4 and F6. Colors indicate the depth of the sampling locations

Fig. 2 From field data to functional compositions: (a) raw (symbols) and smoothed (solid lines) PSCs; (b) smoothed PSDs; (c) smoothed PSDs along boreholes B5, F3, F4 and F6. Colors indicate the depth of the sampling locations

Fig. 3 Selection of the number of clusters K. On a log-scale: boxplots of dissimilarities d_i and the mean dissimilarity (symbols) for K ranging within $\{1, ..., 10\}$

Fig. 4 SFC K-mean results for $K = 3$: (a) empirical trace-semivariogram (symbols) and fitted model (solid line), along with the number of pairs associated with each lag; (b) estimated cluster centroids; (c) three-dimensional representation of the data assignment to the three identified clusters

Fig. 5 SK of generalized indicators: (a) Empirical variograms and cross-variogram (symbols), and fitted models (solid line) of the ilr transforms; (b) kriged field of generalized indicators; (c) representation of the generalized indicators corresponding to the extreme and central points of the color scale

Fig. 6 Conditional PSDs predicted via FCCK: (a) results at boreholes B5, F4 and F6 and at an undrilled location ("*New*") with coordinates (3508600,5377670). (b) Vertical distribution of predicted PSDs, for the group of samples at elevations $301 \leq z \leq 306$ m above sea level (a.s.l.), at borehole F6. In both panels: colors of the solid curves indicate depth; colors of the symbols indicate the cluster assignment; the size of the symbols is proportional to the Kriging variance; smoothed data are represented with solid curves colored according to the cluster assignment

Fig. 7 An example of the way the information content of a PSD is decoupled into the conditional PSD within the domain of observation (solid line) and the two-part composition of mass within the two subdomains $[t_m, t_1]$ and $[t_1, t_{12}]$ (represented via a histogram)

Fig. 8 SK of two-part compositions. (a) Empirical variogram (symbols) and fitted model (solid line). (b) Kriged field: predictions range in $[3.62 \cdot 10^{-4}, 5.25 \cdot 10^{-2}]$; colors are given on a log-scale. (c) Representation of the compositions corresponding to the extrema of the color scale

Fig. 9 Predicted PSDs at boreholes B5, F4 and F6 and at an undrilled location ("*New*") with coordinates (3508600,5377670). Colors of the solid curves indicate the depth. Colors of the symbols along the boreholes indicate the value of $\zeta^*_{\boldsymbol{s}}$, their size being proportional to the Kriging variance of SK. Smoothed data are represented with solid curves colored according to the cluster assignment

	Cluster=1	Cluster=2	Cluster=3
B1	39	20	0
B2	38	16	1
B3	43	24	0
B4	39	26	4
B5	62	0	0
F0	1	11	0
F1	1	9	0
F2	5	15	0
F3	1	11	0
F4	1	8	0
F5	3	11	0
F6	3	14	0

Table 1

		SK CV		
		Cluster=1	Cluster=2	Cluster=3
SFC K-mean	Cluster=1	205	30	1
	Cluster=2	71	94	0
	Cluster=3	1	0	4

Table 2

Table 1 Cluster assignment along the drilled boreholes

Table 2 Cross-validation results for SK based prediction of generalized indicators: clustering results (SFC K-mean) vs cross-validation (SK CV)

References

Aitchison, J. (1982). The statistical analysis of compositional data. *J Roy Stat Soc B 44*(2), 139–177.

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

Cressie, N. (1993). *Statistics for Spatial data*. John Wiley & Sons, New York.

Delicado, P. (2011). Dimensionality reduction when data are density functions. *Comput Stat Data An 55*(1), 401 – 420.

Egozcue, J. J. (2009). Reply to "On the Harker Variation Diagrams; ..." by J.A. Cortés. *Math Geosci 41*(7), 829–834.

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

Egozcue, J. J., V. Pawlowsky-Glahn, R. Tolosana-Delgado, M. Ortego, and K. van den Boogaart (2013). Bayes spaces: use of improper distributions and exponential families. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas 107*(2), 475–486.

Franco-Villoria, M. and R. Ignaccolo (2014). *Uncertainty evaluation in functional kriging with external drift*, pp. 113–118. In: Contributions in infinite-dimensional statistics and related topics.

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

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

Hron, K., A. Menafoglio, M. Templ, K. Hruzova, and P. Filzmoser (2015). Simplicial principal component analysis for density functions in Bayes spaces. *Comput Stat Data An*. DOI: 10.1016/j.csda.2015.07.007.

Marron, J. S. and A. M. Alonso (2014). Overview of object oriented data analysis. *Biometrical J 56*(5), 732–753.

Martìn, M. A., J. M. Rey, and F. J. Taguas (2005). An entropy-based heterogeneity index for mass-size distributions in earth science. *Ecol Model 182*, 221–228.

Mc Queen, J. (1967). Some methods for classification and analysis of multivariate observations. *5th Berkeley Symposium on Mathematics, Statistics and Probability 1*, 281–298.

Menafoglio, A., A. Guadagnini, and P. Secchi (2014). A Kriging Approach based on Aitchison Geometry for the Characterization of Particle-Size Curves in Heterogeneous Aquifers. *Stoch Env Res and Risk Assess 28*(7), 1835–1851.

Menafoglio, A., P. Secchi, and M. Dalla Rosa (2013). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electron J Stat 7*, 2209–2240.

Nerini, D. and B. Ghattas (2007). Classifying densities using functional regression trees: Applications in oceanology. *Comput Stat Data An 51*(10), 4984 – 4993.

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

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis in the symplex. *Stoch Env Res and Risk Assess 15*, 384–398.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). *Modeling and Analysis of Compositional Data*. Statistics in Practice. Wiley.

Pigoli, D., A. Menafoglio, and P. Secchi (2013). Kriging prediction for manifold-valued random field. CRiSM Paper No. 13-18, University of Warwick.

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

Riva, M., A. Guadagnini, D. Fernández-García, X. Sánchez-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.

Riva, M., X. Sánchez-Vila, and A. Guadagnini (2014). Estimation of spatial covariance of log-conductivity from particle-size data. *Water Resour Res.* in press.

Sangalli, L. M., P. Secchi, and S. Vantini (2014). Object oriented data analysis: A few methodological challenges. *Biometrical J 56*(5), 774–777.

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

Tolosana-Delgado, R., V. Pawlowsky-Glahn, and J. J. Egozcue (2008b). Simplicial indicator kriging. *J China Univ Geosci 19*(1), 65 – 71.

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

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

van den Boogaart, K. G., J. J. Egozcue, and V. Pawlowsky-Glahn (2014). Bayes hilbert spaces. *Aust & NZ J Stat 56*, 171–194.

Vukovic, M. and A. Soro (1992). *Determination of Hydraulic Conductivity of Porous Media from Grain-Size Composition*. Water Resources Publications, Littleton, Colorado.