MOX–Report No. 58/2014

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

Menafoglio, A.; Secchi, P.; Guadagnini, A.

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

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

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

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

[1]MOX-Department of Mathematics, Politecnico di Milano, Italy
[2]Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Italy
[3]Department of Hydrology and Water Resources, The University of Arizona, USA
alessandra.menafoglio@polimi.it
piercesare.secchi@polimi.it
alberto.guadagnini@polimi.it

Abstract

We address the problem of characterizing the spatial field of soil particle-size curves (PSCs) within a heterogeneous aquifer system. We conceptualize the medium as a composite system, associated with an uncertain spatial arrangement of geomaterials. We tie the identification of the latter to the spatially varying arrangement of soil textural properties, which is in turn estimated by an available set of observed PSCs. We analyze these PSCs through their particle-size densities (PSDs), which are interpreted as points in the infinite-dimensional Hilbert space of functional compositions (FCs). To model the heterogeneity of the system, we introduce an original hierarchical model for FCs, conducive to a Functional Compositional Class-Kriging (FCCK) predictor. To tackle the problem of lack of information when the spatial arrangement of soil types is unobserved, we propose a novel clustering method for spatially dependent FCs. The latter allows inferring a grouping structure from sampled PSDs, consistent with our theoretical framework. This enables one to project the complete information content embedded in the set of heterogenous PSDs to unsampled locations in the system, thus providing predictions of the spatial arrangement of (a) regions associated with each identified textural class, and (b) the PSDs within each region. Our methodological developments are tested on a field application relying on a set of particle-size curves 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 (a) the types of geomaterials forming the internal architecture of an aquifer, and (b) the associated values of hydraulic conductivity (see e.g., 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 are related through a set of empirical formulations to parameters such as hydraulic conductivity (e.g., 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 characterization 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 unity and are the infinite-dimensional counterparts of compositional data, i.e., 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 (e.g., Aitchison, 1982; Pawlowsky-Glahn and Buccianti, 2011, and references therein). Additional developments are then proposed by van den Boogaart et al. (2010); Egozcue et al. (2013), 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., 2014) and spatial prediction (Menafoglio et al., 2014).

In this context, we interpret each PSD as a unique entity, i.e., an *object datum* (Marron and Alonso, 2014; Sangalli et al., 2014), which is embedded into the Hilbert space of FCs endowed with the Aitchison geometry. This geometric perspective bases its strength on the concepts of functional (FDA, e.g., Ramsay and Silverman, 2005; Horváth and Kokoszka, 2012, and references therein)

and compositional data analysis (CoDa, e.g., Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011, 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, we focus 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, we introduce a novel hierarchical geostatistical model for PSDs which is conducive to the development of a Functional Compositional Class-Kriging (FCCK) predictor. 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. In this work we consider the scenario where the system composition is only partially observable, as is the case for our field data, detailed in Sect. 2. In this context, we address the problem of identifying soil types characterizing the aquifer when only a sample of unclassified PSDs is available. We do so by introducing an original unsupervised classification method, which is consistent with the proposed FCCK methodology. The application of our developments to the target dataset is illustrated in Sect. 5.

2 Field data

For the purpose of our application we consider an extensive dataset of PSCs sampled at an experimental site located near the city of Tübingen, Germany.

The investigated aquifer body is essentially formed by alluvial material overlain by stiff silty clay and underlain 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 we refer for additional details. Amongst the available data, we focus on 406 PSCs sampled along 12 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

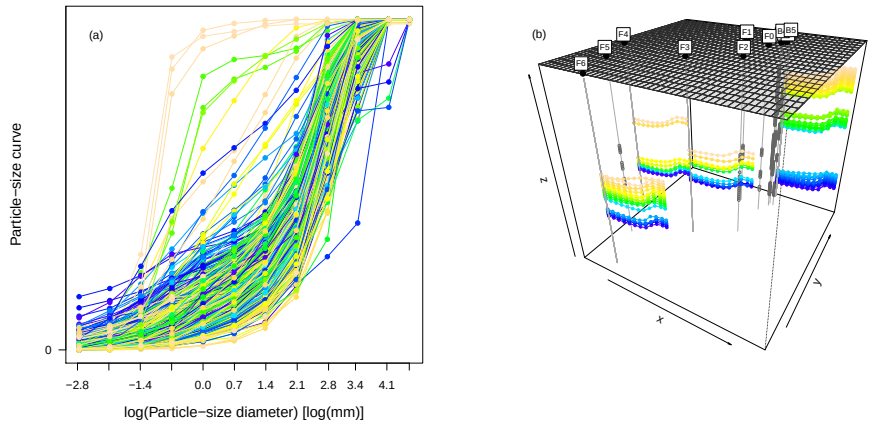


Figure 1: Raw data: (a) raw PSCs (b) raw PSCs along boreholes B5, F3, F4 and F6. Colors indicate the depth of the sampling locations

developments leading to a set of relationships between the spatial covariance of the (natural) logarithm of hydraulic conductivity (K) 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 12 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 group 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, we rely on these PSC data to provide an application of the theoretical developments presented in Sect. 4.

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; Buccianti et al., 2006, and references therein). This has been recently generalized to functional

compositions, i.e., infinite-dimensional objects which convey only relative information, as they are constrained to be non-negative 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; we refer the reader to (Egozcue et al., 2006, 2013; van den Boogaart et al., 2010, 2014) for further details.

We consider two functional compositions f, g to be 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). We call $A^2(\mathcal{T})$ 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 field case study we consider. 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

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

We 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 $0_\oplus = 1/\eta$, with $\eta = |\mathcal{T}|$. 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 compact support. The theory has been recently extended to deal with compositions with infinite supports (van den Boogaart et al., 2014). However, as noted by Delicado (2011); Menafoglio et al. (2014); Hron et al. (2014), 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. We adopt the latter approach 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 $(\Omega, \mathfrak{F}, P)$: $\mathcal{X}_{\boldsymbol{s}}$ associates with each particle size t in a compact domain $\mathcal{T} = [t_m, t_M]$ 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), we 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 $\boldsymbol{s} \in D$, we treat the PSD $\mathcal{Y}_{\boldsymbol{s}}$ as an element of A^2.

We model the heterogeneous structure of the aquifer D 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, we consider the random field $\{\Pi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, defined over $(\Omega, \mathfrak{F}, P)$, whose generic element $\boldsymbol{\pi}_{\boldsymbol{s}} = \left(\pi_{\boldsymbol{s}}^{(1)}, \pi_{\boldsymbol{s}}^{(2)}, ..., \pi_{\boldsymbol{s}}^{(K)}\right)$ is a random probability vector. The latter determines the probability of occurrence of the soil type $\tau^{(k)}$ in location $\boldsymbol{s} \in D$ and is a random element in the $(K-1)$-dimensional simplex $\Delta^{(K-1)}$. The random field $\{\Pi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ is assumed to be second order stationary in the space $\Delta^{(K-1)}$ endowed with the Aitchison geometry (Tolosana-Delgado et al., 2011). Conditionally to the field $\{\Pi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, we model the spatial field of soil types $\{T_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ as a collection of independent discrete random variables, each $T_{\boldsymbol{s}}$ being valued in $\{1, ..., K\}$ with probability mass function equal to $\boldsymbol{\pi}_{\boldsymbol{s}}$.

Given the spatial arrangement of the soil types, for any $\boldsymbol{s} \in D$, we represent the PSD $\mathcal{Y}_{\boldsymbol{s}}$ as a perturbation of its (deterministic) Fréchet mean – which is called drift and is determined by the soil type – with a neutral-mean stochastic residual $\delta_{\boldsymbol{s}}$. Specifically, for any $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$, we assume the following model

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

where the residual process $\{\delta_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ is independent of both $\{T_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ and $\{\Pi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, and follows 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 $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$

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

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

We note that, in light of model (1), we can describe the drift in $\boldsymbol{s} \in D$ through a linear model with $K-1$ binary regressors $\{\psi_l(\boldsymbol{s}), l = 1, ..., K-1\}$. Indeed,

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

where, for $k = 1, ..., K-1$, $\psi_l(\boldsymbol{s}) = 1$ if $l = k$, and $\psi_l(\boldsymbol{s}) = 0$ otherwise; if $T_{\boldsymbol{s}} = \tau^{(K)}$ then $\psi_l(\boldsymbol{s}) = 0$ for every $l = 1, ..., K-1$. Therefore, one has

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

4.2 The Class-Kriging predictor

Given a set of locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$, and the observations of the PSD process in these locations, $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$, we aim at predicting 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. We thus look for the Class-Kriging predictor $\mathcal{Y}^*_{\boldsymbol{s}_0} = \bigoplus_{i=1}^n \lambda_i^* \odot \mathcal{Y}_{\boldsymbol{s}_i}$, whose weights are given by the solution of the minimization problem

$$\min_{\substack{\lambda_1,...,\lambda_n \in \mathbb{R}: \\ \mathcal{Y}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} = \bigoplus_{i=1}^n \lambda_i \odot \mathcal{Y}_{\boldsymbol{s}_i}}} \mathrm{Var}_{A^2}\left(\mathcal{Y}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} \ominus \mathcal{Y}_{\boldsymbol{s}_0} \middle| T_{\boldsymbol{s}_0} = \tau^{(k_0)}, T_{\boldsymbol{s}_i} \in \tau^{(k_i)}, i = 1, ..., n \right)$$

$$\text{subject to} \quad \mathbb{E}_{A^2}\left[\mathcal{Y}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} \middle| T_{\boldsymbol{s}_0} = \tau^{(k_0)}, T_{\boldsymbol{s}_i} \in \tau^{(k_i)}, i = 1, ..., n \right] = m^{(k_0)}. \tag{5}$$

Having observed the spatial arrangement of soil types, problem (5) can be solved through the Universal Kriging predictor for Hilbert space valued random fields developed in (Menafoglio et al., 2013). Note that if each regressor ψ_k is known over the entire domain D and under suitable assumption on the sampling design, problem (5) is well posed.

Proposition 1 (Menafoglio et al. (2013))**.** *Assume that* $\Sigma = (C(\boldsymbol{h}_{i,j})) \in \mathbb{R}^{n \times n}$, $\boldsymbol{h}_{i,j} = \boldsymbol{s}_i - \boldsymbol{s}_j$, $i, j = 1, ..., n$, *is a positive definite matrix. Assume further that the design matrix* $\Psi = (1, \psi_k(\boldsymbol{s}_i)) \in \mathbb{R}^{n \times K}$ *is of full rank. Then problem* (5) *admits a unique solution* $(\lambda_1^*, ..., \lambda_n^*) \in \mathbb{R}^n$, *which is obtained by solving*

$$\left(\begin{array}{c|cc} C(\boldsymbol{h}_{i,j}) & 1 & \psi_k(\boldsymbol{s}_i) \\ \hline 1 & 0 & 0 \\ \psi_k(\boldsymbol{s}_j) & 0 & 0 \end{array} \right) \left(\begin{array}{c} \lambda_i \\ \hline \zeta_0 \\ \zeta_k \end{array} \right) = \left(\begin{array}{c} C(\boldsymbol{h}_{0,i}) \\ \hline 1 \\ \psi_k(\boldsymbol{s}_0) \end{array} \right), \tag{6}$$

where $\zeta_0, ..., \zeta_{K-1}$ *are* K *Lagrange multipliers associated with the unbiasedness constraint. Conditionally on* $T_{\boldsymbol{s}_0}, T_{\boldsymbol{s}_1}, ..., T_{\boldsymbol{s}_n}$, *the Universal Kriging variance of predictor* $\mathcal{Y}^*_{\boldsymbol{s}_0}$ *is then*

$$\begin{aligned} \sigma^2_*(\boldsymbol{s}_0) &= \mathrm{Var}_{A^2}\left(\mathcal{Y}^*_{\boldsymbol{s}_0} \middle| T_{\boldsymbol{s}_0} = \tau^{(k_0)}, T_{\boldsymbol{s}_i} \in \tau^{(k_i)}, i = 1, ..., n \right) = \\ &= C(\boldsymbol{0}) - \sum_{i=1}^n \lambda_i^* C(\boldsymbol{h}_{i,0}) - \sum_{k=0}^{K-1} \zeta_k^* \psi_k(\boldsymbol{s}_0). \end{aligned} \tag{7}$$

Note that from expression (7), the following Chebyshev inequality for the prediction errors can be derived

$$P\left(\|\mathcal{Y}^*_{\boldsymbol{s}_0} \ominus \mathcal{Y}_{\boldsymbol{s}_0}\|_{A^2} > \kappa \sigma_*(\boldsymbol{s}_0) \middle| T_{\boldsymbol{s}_0} = \tau^{(k_0)}, T_{\boldsymbol{s}_i} \in \tau^{(k_i)}, i = 1, ..., n \right) < \frac{1}{\kappa^2}, \kappa > 0. \tag{8}$$

As in the classical geostatistical framework, the solution of system (6) 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 (e.g., Cressie, 1993). We adopt a method of moment estimator $\widehat{\gamma}(\boldsymbol{h})$ from the residuals

$$\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{2|N(\boldsymbol{h})|} \sum_{(i,j) \in N(\boldsymbol{h})} \|\delta_{\boldsymbol{s}_i} \ominus \delta_{\boldsymbol{s}_j}\|^2_{A^2}, \tag{9}$$

followed by a fitting of a valid model via weighted least squares. We remark that 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. We note that estimator (9) 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 estimator of the drift $\widehat{\boldsymbol{m}}_{\boldsymbol{s}}^{GLS} = (\widehat{m}_{\boldsymbol{s}_1}^{GLS}, ..., \widehat{m}_{\boldsymbol{s}_n}^{GLS})^T$. The latter is obtained as $\widehat{\boldsymbol{m}}_{\boldsymbol{s}}^{GLS} = \Psi \odot \widehat{\boldsymbol{a}}_{\boldsymbol{s}}^{GLS}$ where Ψ indicates the design matrix (i.e., for $i = 1, ..., n$, $\Psi_{i,1} = 1$, $\Psi_{i,k+1} = \psi_k(\boldsymbol{s}_i)$, $k = 1, ..., K-1$), and $\widehat{\boldsymbol{a}}^{GLS}$ denotes the coefficient vector estimator

$$\widehat{\boldsymbol{a}}^{GLS} = (\Psi^T \Sigma^{-1} \Psi)^{-1} \Psi^T \Sigma^{-1} \odot \boldsymbol{\mathcal{Y}}_{\boldsymbol{s}} \tag{10}$$

with $\widehat{\boldsymbol{a}}^{GLS} = (\widehat{a}_0^{GLS}, ..., \widehat{a}_L^{GLS})^T$, $\boldsymbol{\mathcal{Y}}_{\boldsymbol{s}} = (\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n})^T$; here we have adopted the vectorial notation: $(\mathbb{A} \odot f)_i = \bigoplus_{j=1}^n \mathbb{A}_{ij} \odot f_j$, $\mathbb{A} = (\mathbb{A}_{ij}) \in \mathbb{R}^{n \times n}$, $\boldsymbol{f} = (f_i)$, $f_i \in A^2$, $i = 1, 2, ..., n$. To jointly estimate the residuals and the residual variogram, we resort to an iterative algorithm, which is initialized to an ordinary least square estimate of the drift, that is, $\widehat{\boldsymbol{m}}_{\boldsymbol{s}}^{OLS} = \widehat{\boldsymbol{m}}_{\boldsymbol{s}}^{GLS}\Big|_{\Sigma=\sigma^2\mathbb{I}}$, with $\sigma > 0$, $\mathbb{I}$ being the identity matrix (Menafoglio et al., 2013, Sect. 4).

4.3 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 predict the regressors $\psi_1(\boldsymbol{s}_0), ..., \psi_{K-1}(\boldsymbol{s}_0)$ at $\boldsymbol{s}_0$.

If a complete observation of the field $\{\Pi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ were available, one could assign the most likely soil type, i.e., the one associated with the highest probability in $\boldsymbol{\pi}_{\boldsymbol{s}_0}$, to location $\boldsymbol{s}_0$. 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 resort to the Simplicial Kriging (SK) (Tolosana-Delgado et al., 2008a, 2011) to predict $\boldsymbol{\pi}_{\boldsymbol{s}_0}$, and consequently assign the soil type in $\boldsymbol{s}_0$. 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, e.g., Pawlowsky-Glahn and Buccianti, 2011) of $\boldsymbol{\pi}_{\boldsymbol{s}_i}$ in $\boldsymbol{s}_i$.

However, the random field $\{\Pi_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ is in general latent. In this case, we approximate the probability vector in $\boldsymbol{s}_0$ through a generalized indicator, following the approach of Tolosana-Delgado et al. (2008a). For $\boldsymbol{s} \in D$ and $k = 1, ..., K$, we define a generalized indicator $p_{\boldsymbol{s}}^{(k)}$ as

$$p_{\boldsymbol{s}}^{(k)} = \begin{cases} 1-b, & T_{\boldsymbol{s}} = \tau^{(k)} \\ \frac{b}{K-1}, & T_{\boldsymbol{s}} \neq \tau^{(k)}, \end{cases} \tag{11}$$

where b is a (small) parameter usually set to $b = 0.05$, or $b = 0.1$ (Tolosana-Delgado et al., 2008a). The generalized indicator vectors $\boldsymbol{p}_{\boldsymbol{s}_1}, ..., \boldsymbol{p}_{\boldsymbol{s}_n}$, with $\boldsymbol{p}_{\boldsymbol{s}_i} = (p_{\boldsymbol{s}_i}^{(1)}, ..., p_{\boldsymbol{s}_i}^{(K)})^T$, are then used in place of the unobserved 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 soil type associated with the highest $p_{\boldsymbol{s}_0}^{(k)*}$ and then solve system (6).

4.4 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 infer the associated soil types through a cluster analysis. In this Section, we introduce an original unsupervised classification method, the SFC K-mean, which is coherent with the model introduced above and allows for the clustering of objects that are Spatially dependent, Functional and Compositional. The method we propose is inspired by the K-mean clustering method (Mc Queen, 1967), but properly tailored to our framework: 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 according to (10). We 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 are the minimizers of the global variance within the clusters; that is, for $k = 1, ..., K$,

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

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

We first assume the structure of spatial dependence to be known. In this case, the method we propose is sketched in Algorithm 2.

Algorithm 2. *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 out of the n data $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$);

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

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

2. Representation:
   For each $k = 1, ..., K$, update the centroid $\mathcal{C}_k$ to the generalized least square 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* (10)*, 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.*

Algorithm 2 returns: (a) the clusters compositions and (b) the estimates of the within-cluster drifts (i.e., the centroids).

The structure of spatial dependence is often unknown in field studies. In this case, Step 2 of Algorithm 2 can be replaced by the iterative algorithm described in Subsect. 4.1, which can be employed to jointly estimate the mean and the structure of spatial dependence of the field. For the sake of clarity, Algorithm 3 details the estimation procedure which will be employed for the case study illustrated in Sect. 6.

Algorithm 3. *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 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 *out of the* n *data* $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$*);*

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

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

2. Representation:

   I. Estimate the drift coefficient vector $\boldsymbol{a}$ *via* $\widehat{\boldsymbol{a}}^{OLS} = (\Psi^T\Psi)^{-1}\Psi^T\boldsymbol{\mathcal{Y}}$*, with* $\Psi = (\Psi_{ik})$*, with* $\Psi_{i1} = 1$*, for all* $i = 1, ..., n$*,* $\Psi_{i,k} = 1$ *if* $\mathcal{Y}_{\boldsymbol{s}_i}$ *belongs to cluster* k*,* $\Psi_{i,k} = 0$ *otherwise. Set* $\widehat{\boldsymbol{a}} := \widehat{\boldsymbol{a}}^{OLS}$*;*

   II. Estimate the trace-semivariogram $\gamma(\cdot)$ *from the estimated residuals* $\widehat{\boldsymbol{\delta}} = \boldsymbol{\mathcal{Y}} - \Psi \odot \widehat{\boldsymbol{a}}$ *via estimator* (9)*, and fit a valid model. Derive from this the estimate* $\widehat{\Sigma}$ *of* Σ*;*

   III. Estimate $\boldsymbol{a}$ *with* $\widehat{\boldsymbol{a}}^{GLS}$ *according to* (10) *with* $\widehat{\Sigma}$ *in place of* Σ*, and set* $\widehat{\boldsymbol{a}} := \widehat{\boldsymbol{a}}^{GLS}$*;*

   IV. Repeat II.–III. until convergence;

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

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

3. Iteration:
   Repeat 1. *and* 2. *until no change in assignment occurs or a prescribed maximum number of iterations is reached.*

Algorithm 3 return: (a) the clusters compositions, (b) the estimates of the within-cluster drifts (i.e., the centroids) and (c) the estimated structure of spatial dependence. From a computational viewpoint, Algorithm 3 is a nested iterative algorithm, as Step 2 requires the computation of the GLS estimator $\widehat{\boldsymbol{a}}^{GLS}$, which in turn needs to be performed iteratively. Hence, Algorithm 3 may become computationally demanding as the number n of data increases. However, Menafoglio et al. (2013) show via simulation that the convergence of the iterative algorithm employed in Step 2. proves to be very fast (typically within five iterations).

We 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 3. A proper selection of K can be performed by any of the available standard techniques, e.g., 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, we apply the procedure detailed in Sect. 4 to the field data described in Sect. 2. We first note that the left tails of the observed particle-size distributions are censored, due to the sieve-measurement procedure. Nevertheless, the proportions of particles within the censored left tails are known, as these coincide with the observations of the PSCs $\widetilde{\mathcal{X}}_{\boldsymbol{s}_i}$, $i = 1, ..., n$, at the first sieve t_1, namely $\widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t_1)$, $i = 1, ..., n$. Although one could select a priori a given 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 would be highly influential on the analysis here presented, due to the high variability of $\widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t_1)$ for $i = 1, ..., n$. Thus, for each observed location $\boldsymbol{s}_i$, $i = 1, ..., n$, we decouple the available information into (a) the particle-size density conditional to the domain of observation $\mathcal{T} = [t_1, t_{12}]$ and (b) a 2-part composition $\boldsymbol{\zeta}_{\boldsymbol{s}_i} = (\zeta_{\boldsymbol{s}_i}, 1 - \zeta_{\boldsymbol{s}_i})$, respectively collecting the mass within the censored left tail (i.e., for $t < t_1$) and the observed domain $[t_1, t_{12}]$ (Fig. 2). We then treat the available data as follows: (a) we consider the conditional PSDs as functional compositions and apply to these the methodology illustrated in Sect. 4; (b) we separately treat the above mentioned 2-part compositions via simplicial geostatistical methods (Tolosana-Delgado et al., 2008a, 2011), and (c) combine the results to provide a complete description of PSDs predictions at unsampled locations. We remark that performing a separate analysis of the 2-part compositions $\boldsymbol{\zeta}_{\boldsymbol{s}_1}, ..., \boldsymbol{\zeta}_{\boldsymbol{s}_n}$ enables one to avoid setting a priori the parameter that controls the lowest particle diameter which can be experimentally observed. This is a remarkable advantage, because, as stated above, this choice was verified to be influential on the analysis of the considered dataset (not shown).

5.1 Analysis of Conditional Particle-Size Densities

We follow the approach proposed in (Menafoglio et al., 2014) to obtain the conditional particle-size densities from the raw data. We smooth the raw conditional particle-size curves, that is, $\widetilde{\mathcal{X}}^{(c)}_{\boldsymbol{s}_i}(t_j) = \frac{\widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t_j) - \widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t_1)}{\widetilde{\mathcal{X}}_{\boldsymbol{s}_i}(t_{12}) - \widetilde{\mathcal{X}}_{\boldsymbol{s}}(t_1)}$, $j = 1, ..., 12$, $i = 1, ..., 406$, by means of $m = 70$ Bernstein polynomials (Fig. 3), upon con-

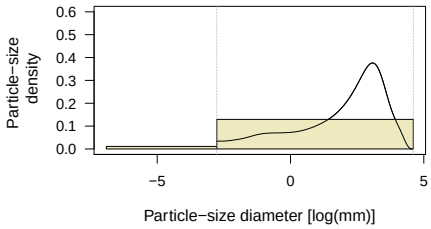


Figure 2: An example of the way the information content of a PSD is decoupled into (a) the conditional PSD within the domain of observation (solid line) and (b) the 2-part composition of mass within the two subdomains $[t_m, t_1]$ and $[t_1, t_{12}]$ (represented via a histogram)

sidering log-transformed particle diameters (hereafter t denotes log-transformed diameters). 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.

The notation of Sect. 4 is here employed as follows. We consider the functional dataset of smoothed conditional PSCs $\mathcal{X}_{\boldsymbol{s}_1}, ..., \mathcal{X}_{\boldsymbol{s}_n}$ at location $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$, $n = 406$ (Fig. 3a) and analyze their densities $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$ (Fig. 3b) as functional compositions. These are embedded into the space $A^2(\mathcal{T})$ over the compact domain $\mathcal{T} = [\log(0.063), \log(100)]$ and modeled 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 our compositional approach, we identify the lithotypes within the system by applying the SFC K-mean method devised in Algorithm 2.

Based on a preliminary analysis on directional variograms performed for the complete set of available PSCs, we assume the field to be geometrically anisotropic with anisotropy ratio of $R = 0.04$ between the horizontal and vertical directions. Hereafter, we refer the estimates to a rescaled isotropic spatial domain obtained by dilating vertical coordinates z by a factor $1/R = 25$. Similar to Menafoglio et al. (2014), we select for our interpretations an exponential variogram model with nugget, and estimate its parameters via weighted least square.

We identify the number K of clusters 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. Our results are based on K ranging in $\{1, ..., 10\}$. Figure 4 depicts on a log-scale the boxplots of the dissimilarities between each datum $\mathcal{Y}_{\boldsymbol{s}_i}$ 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 $K = 2$,

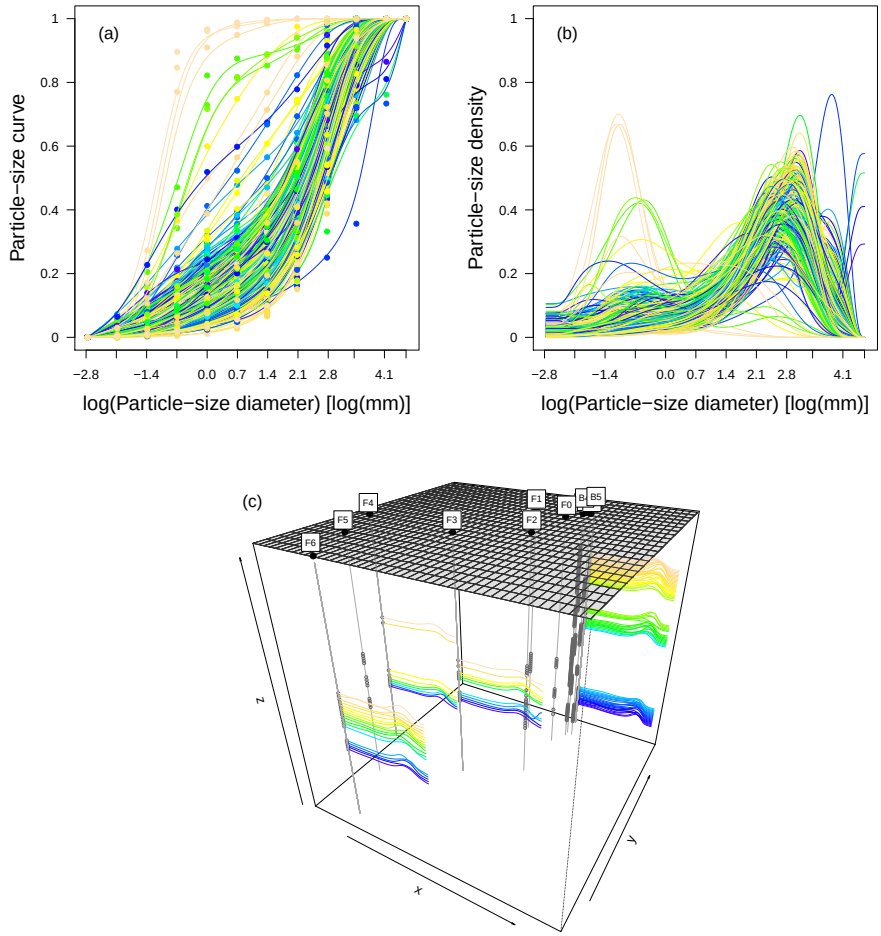


Figure 3: From field data to functional compositions: (a) raw (symbols) and smoothed (solid lines) PSCs; (b) smoothed PSDs; (c) smoothed PSDs along boreholes B5, F3, F4 and F6. Colors indicate the depth of the sampling locations

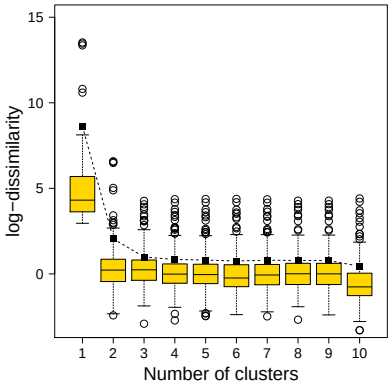


Figure 4: Selection of the number of clusters K. On a log-scale: boxplots of dissimilarities d_i and the mean dissimilarity (symbols) for K ranging within $\{1, ..., 10\}$

	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: Cluster assignment along the drilled boreholes.

even though the corresponding boxplot evidences the presence of five outliers. These are collected into a separate cluster when $K = 3$, leading to an elbow in the mean dissimilarity. This observation motivates our 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. 5.

The cluster centroids correspond to the estimates of the drift in (1) and are displayed in Fig. 5b. 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 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. 5b). This is consistent with

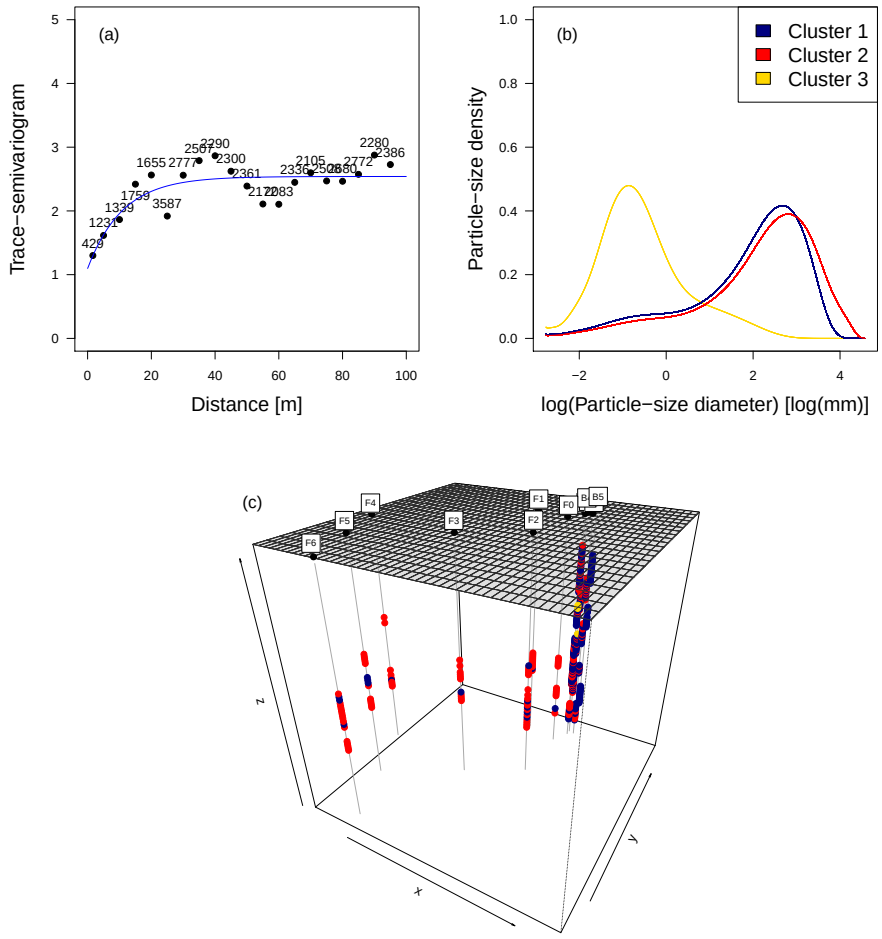


Figure 5: SFC K-mean results for $K = 3$: (a) Empirical trace-semivariogram (symbols) and fitted model (solid line); the number of pairs associated with each lag is also reported; (b) Estimated cluster centroids; (c) Three-dimensional representation of the data assignment to the two identified clusters

		MVT K-mean			
		Cluster 1	Cluster 2	Cluster 3	n_i
SFC K-mean	Cluster 1	132	98	6	236
	Cluster 2	49	116	0	165
	Cluster 3	0	0	5	5
	n_i	181	214	11	$n = 406$

Table 2: Comparison between the clustering results obtained via our SFC K-mean and via multivariate K-mean on raw data (MVT K-mean - results of Riva et al., 2006). The Table lists the number of elements within each cluster identified by SFC K-mean (rows), which are assigned to each of the three clusters obtained by Riva et al. (2006) (columns).

the main sedimentological composition of the corresponding cluster identified by Riva et al. (2006).

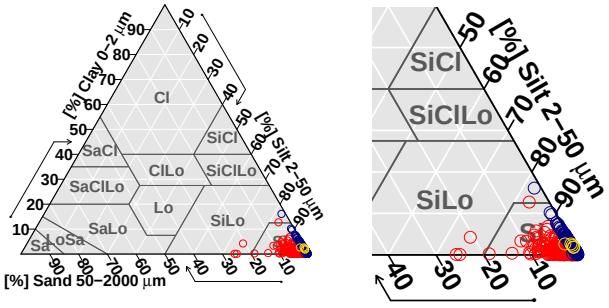
Inspection of Fig. 5b and Tab. 1 evidences that the first cluster appears to be mainly associated with the northern boreholes B1-B5, and the second cluster with the boreholes F0-F6. This is partly consistent with the observation that the former group of boreholes is located in an area where the Neckar river displays a bend, thus favoring the accumulation of the finer sediments in this area. We 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.

Table 2 lists the elements of the confusion matrix between our results and those of Riva et al. (2006). Each row lists the number of elements within the corresponding cluster identified by the SFC K-mean (k=1,2,3) which is assigned to each of the clusters of Riva et al. (2006). Inspection of these results reveals that the two classification strategies agree on about 62% of the data. Some discrepancy between these results is to be expected, because they are based on different measures of dissimilarity, namely, a functional Aitchison geometry and the Euclidean geometry on the 12-dimensional vectors of raw PSCs.

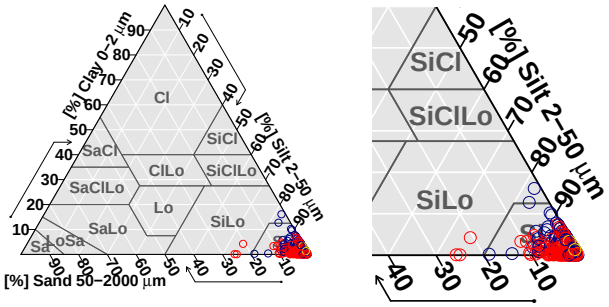
The use of a dissimilarity measure driven by the Aitchison geometry renders the classification compatible with a physical interpretation in terms of soil classes associated with a soil textural triangle. Figure 6 depicts such representation based on the results of our analysis as well as those from Riva et al. (2006). We remark that the classification via SFC K-mean induces an evident partition on the soil textural triangle, in the sense that soil samples associated to different clusters tend to separate on the soil textural triangle.

5.1.2 Kriging predictions

We compute the generalized indicators as in (11) and analyze their spatial dependence via simplicial variography (Tolosana-Delgado et al., 2008b, 2011). We first note that a possible dependence of the cluster assignment on the horizontal x and y coordinates may be recognized by inspection of Fig. 5c. 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, hereinafter we illustrate the results obtained under a stationarity assumption.



(a) SFC K-mean



(b) Multivariate K-mean

Figure 6: Representation of the clustering results on the USDA soil textural triangle (results are computed via `soiltexture` R Package, Moeys and Shangguan, 2014)

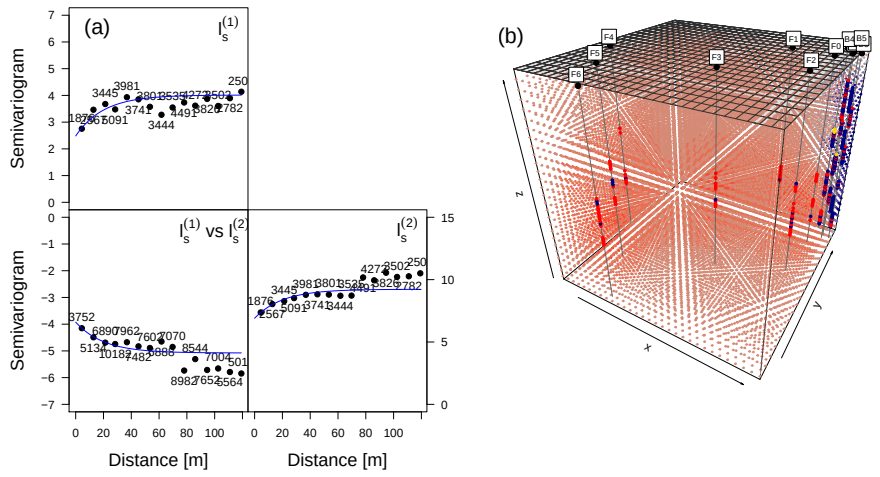


Figure 7: Simplicial Indicator Kriging of generalized indicators: (a) Empirical variograms and cross-variogram (symbols), and fitted models (solid line) of the ilr-transforms $l_{\boldsymbol{s}_i}^{(k)}$, $i = 1, ..., n$, $k = 1, 2$; (b) kriged field

We consider a geometric anisotropy with anisotropy ratio $R = 0.04$ between the horizontal and vertical directions, 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. 7a, together with the fitted exponential models. The results of the SK interpolation are depicted in Fig. 7b. 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. 5c.

Finally, Fig. 8 depicts the results of the Kriging interpolation based on (a) the residual semivariogram estimated through Algorithm 3 (Fig. 5a), and (b) the SK prediction (Fig. 7b) in terms of both point-wise predictions and the associated Kriging variance. The kriged field provide a smooth interpolation of the available data, which is representative of mean particle-size distribution for distances higher than the estimated range. With reference to this point, we note that a sharp assignment has been here considered for Class-Kriging prediction, i.e., 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. We note that 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 Compositional analysis of left tails and prediction of the PSDs

To provide the prediction of the PSDs, we analyze the 2-part compositions $\boldsymbol{\zeta}_{\boldsymbol{s}_i}$, $i = 1, ..., n$ obtained as described in Sect. 5. To this end, we employ the

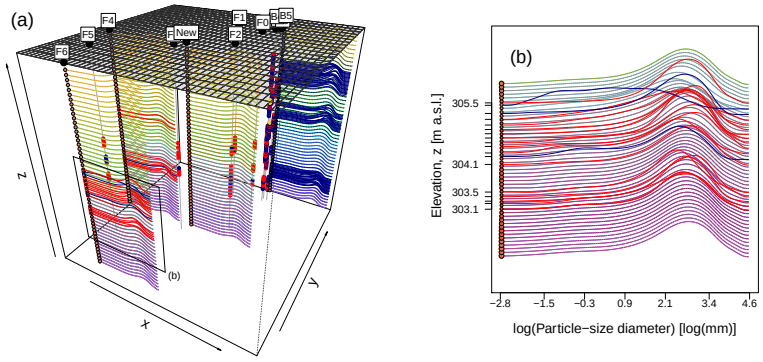


Figure 8: Conditional PSDs predicted via FCCK: (a) three-dimensional representation of results at boreholes B5, F4 and F6 and at an undrilled location ("*New*") with coordinates (3508600,5377670). (b) Vertical distribution of predicted PSDs, for the group of samples at elevations $301 \leq z \leq 306$ m a.s.l., at borehole F6. In both panels: colors of the solid curves indicate depth; colors of the superimposed dashed curves indicate the cluster assignment; the size of the symbols is proportional the Kriging variance

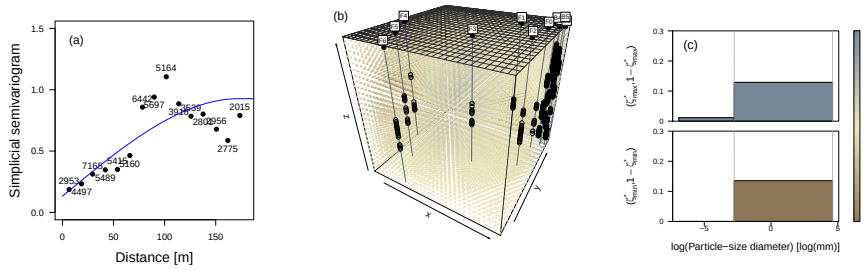


Figure 9: Simplicial Kriging of 2-part compositions $\boldsymbol{\zeta}_{\boldsymbol{s}_i}$, $i = 1, ..., n$. (a) Simplicial empirical variogram (symbols) and fitted model (solid line). (b) Kriged field $\{\zeta^*_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$: predictions range in $[3.62 \cdot 10^{-4}, 5.25 \cdot 10^{-2}]$; colors are given on a log-scale. (c) Representation of the compositions $\zeta^*_{\boldsymbol{s}}$ corresponding to the extrema of the color scale.

simplicial geostatistical methodology of Tolosana-Delgado et al. (2008a, 2011), which has been recalled in Subsect. 4.3.

We assume that second-order stationarity holds, since no evident pattern can be recognized in the spatial arrangement of $\boldsymbol{\zeta}_{\boldsymbol{s}_i}$. We model the simplicial variogram of the compositions via a spherical model with nugget, fitted to the empirical estimate via weighted least squares (Fig. 9a). The corresponding SK predictions are depicted in Fig. 9b.

The kriged field of PSDs, obtained by combining the results of FCCK and SK, is depicted in Fig. 10. These results are obtained by multiplying each predicted conditional PSD $\mathcal{Y}^*_{\boldsymbol{s}_0}$, $\boldsymbol{s}_0 \in D$, by the kriged mass $1 - \boldsymbol{\zeta}^*_{\boldsymbol{s}_0}$.

5.3 Cross-validation results

In this Section we assess the quality of (a) the SK predictions of generalized indicators, (b) the FCCK predictions of conditional PSDs and (c) the SK predictions of the 2-part compositions $\boldsymbol{\zeta}_{\boldsymbol{s}_i}$, $i = 1, ..., n$, through a leave-one-out cross-validation analysis.

We consider the confusion matrix whose entries are listed in Table 3 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 error rate of the classification is 25.37%. We remark that 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. We compute the SSE as $\|\mathcal{Y}_{\boldsymbol{s}_i} - \mathcal{Y}^{(CV)}_{\boldsymbol{s}_i}\|_{A^2}$, $\mathcal{Y}^{(CV)}_{\boldsymbol{s}_i}$ denoting the Kriging

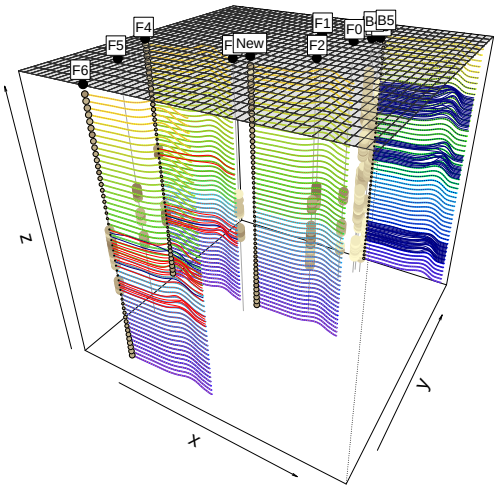


Figure 10: Predicted PSDs at boreholes B5, F4 and F6 and at an undrilled location ("*New*") with coordinates (3508600,5377670). Results are computed as combination of (a) the FCCK predictions of conditional PSDs (Fig. 8) and (b) the Simplicial Kriging predictions of $\boldsymbol{\zeta}_{\boldsymbol{s}_i}$, $i = 1, ..., n$ (Fig. 9). Colors of the solid curves indicate the depth, colors of the superimposed dashed curves indicate the cluster assignment. Colors of the symbols along the boreholes indicate the value of $\zeta^*_{\boldsymbol{s}}$, their size being proportional to the Kriging variance of Simplicial Kriging

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

prediction at $\boldsymbol{s}_i$ obtained upon removing the i-th datum $\mathcal{Y}_{\boldsymbol{s}_i}$ 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. (2014) note that the Aitchison geometry is very sensitive to the information content within the tails of the distribution, due to the *relative scale* property of compositions. 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.

The quality of the SK predictions of the compositions $\boldsymbol{\zeta}_{\boldsymbol{s}_i}$, $i = 1, ..., n$, is assessed in terms of multivariate SSE computed according to the Aitchison geometry for 2-part compositions, namely $\text{SSE}_{\boldsymbol{\zeta}} = \|\boldsymbol{\zeta}_{\boldsymbol{s}_i} - \boldsymbol{\zeta}_{\boldsymbol{s}_i}^{(CV)}\|_A^2$. Cross-validation results are in this case remarkably satisfactory when compared with the mean norm of the data, with a relative $\text{SSE}_{\boldsymbol{\zeta}}$ of 0.47% in median and 1.66% in mean.

We finally employ the cross-validation results of case (ii) to evaluate the empirical coverage of Chebyshev inequality (8). 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 (see e.g., in the framework of object oriented data analysis, Pigoli et al., 2013). 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

The theoretical and application-oriented contributions of our research lead to the following key conclusions.

1. We established an original theoretical framework for the geostatistical characterization of a set of heterogeneous Particle-Size Densities (PSDs). These are directly associated with Particle-Size Curves (PSCs), which are routinely measured in hydrogelogical, hydrogeophysical and soil science applications. PSDs have been interpreted as Functional Compositions (FCs), and analyzed through the Aitchison geometry. Our Functional and Compositional Class-Kriging (FCCK) methodology relies on a novel hierarchical model for FCs and constitutes a generalization of the FCK

methodology introduced by Menafoglio et al. (2014). Our developments allows treating PSDs which are featured by a grouping structure driven by the mean soil textural properties of the system.

2. 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. The latter enables one to infer a grouping structure from a set of observed PSDs associated with spatially varying soil textural properties. This method is consistent with the FCCK model. Results associated with the analyzed field data can be interpreted in view of the classical soil textural triangle.

3. The problem of data censoring due to the sieve measuring procedure has been tackled by decoupling each PSD into (a) a conditional PSD within the compact domain of observation determined by grain sieve analysis and (b) a 2-part composition collecting the mass within the left tail and the domain of observation. These have been separately analyzed via FCCK and multivariate techniques, respectively. This approach enables one to avoid setting a priori the parameter controlling the minimum size of soil particles, which may be influential for the analysis. Improved characterization of the full particle-size distributions might be obtained through a joint modeling of the conditional PSDs and the 2-part compositions. To this end, a possible strategy may consist in developing a Hilbert structure for mixed distributions (i.e., continuous and discrete) and coherently apply a FCCK model. We note however that the censoring problem is closely related to the measurement procedure, and can be overcome by modern methods for PSCs collection, e.g., sedigraph or laser diffraction.

4. The quality of our 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.

References

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

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

Buccianti, A., G. Mateu-Figueras, and V. P. Glahn (2006). *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological society, special pubblication 264.

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

Delicado, P. (2011). Dimensionality reduction when data are density functions. *Computational Statistics & Data Analysis 55*(1), 401 – 420.

Egozcue, J. J. (2009). Reply to "On the Harker Variation Diagrams; ..." by J.A. Cortés. *Mathematical Geosciences 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 Mathematica Sinica, English Series 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.

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 (2014). Simplicial principal component analysis for density functions in Bayes spaces. MOX-report 25/2014, Politecnico di Milano.

Marron, J. S. and A. M. Alonso (2014). Overview of object oriented data analysis. *Biometrical Journal 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. *Ecological Modelling 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. *Stochastic Environmental Research and Risk Assessment 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. *Electronic Journal of Statistics 7*, 2209–2240.

Moeys, J. and W. Shangguan (2014). *soiltexture: Functions for soil texture plot, classification and transformation*. R package version 1.2.13.

Nerini, D. and B. Ghattas (2007). Classifying densities using functional regression trees: Applications in oceanology. *Computational Statistics & Data Analysis 51*(10), 4984 – 4993.

Pawlowsky-Glahn, V. and A. Buccianti (2011). *Compositional data analysis. Theory and applications*. 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-Garcáa, 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. Sanchez-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 Journal 56*(5), 774–777.

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

Tolosana-Delgado, R., V. Pawlowsky-Glahn, and J. J. Egozcue (2008b). Simplicial indicator kriging. *Journal of China University of Geosciences 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. *Australian & New Zealand Journal of Statistics 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.

MOX Technical Reports, last issues

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

58/2014 Menafoglio, A.; Secchi, P.; Guadagnini, A.
A Class-Kriging predictor for Functional Compositions with Application to Particle-Size Curves in Heterogeneous Aquifers

57/2014 Giverso, C.; Verani, M.; Ciarletta P.;
Branching instability in expanding bacterial colonies

55/2014 Antonietti, P. F.; Houston P.; Sarti, M.; Verani, M.
Multigrid algorithms for hp-version Interior Penalty Discontinuous Galerkin methods on polygonal and polyhedral meshes

56/2014 Antonietti, P. F.; Sarti, M.; Verani, M.; Zikatanov, L. T.
A uniform additive Schwarz preconditioner for the hp-version of Discontinuous Galerkin approximations of elliptic problems

54/2014 Ferrario, E.; Pini, A.
Uncertainties in renewable energy generation systems: functional data analysis, monte carlo simulation, and fuzzy interval analysis

53/2014 Ieva, F.; Paganoni, A.M., Pietrabissa, T.
Dynamic clustering of hazard functions: an application to disease progression in chronic heart failure

52/2014 Dede , L..; Quarteroni, A.; S. Zhu, S.
Isogeometric analysis and proper orthogonal decomposition for parabolic problems

51/2014 Dassi, F.; Perotto, S.; Formaggia, L.
A priori anisotropic mesh adaptation on implicitly defined surfaces

50/2014 Bartezzaghi, A.; Cremonesi, M.; Parolini, N.; Perego, U.
An explicit dynamics GPU structural solver for thin shell finite elements

49/2014 D. Bonomi, C. Vergara, E. Faggiano, M. Stevanella, C. Conti, A. Redaelli, G. Puppini et al
Influence of the aortic valve leaflets on the fluid-dynamics in aorta in presence of a normally functioning bicuspid valve