MOX-Report No. 59/2015

Stochastic Simulation of Soil Particle-Size Curves in Heterogeneous Aquifer Systems through a Bayes space approach

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

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Stochastic Simulation of Soil Particle-Size Curves in Heterogeneous Aquifer Systems through a Bayes space approach

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

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

Abstract

We address the problem of stochastic simulation of soil particle-size curves (PSCs) in heterogeneous aquifer systems. Unlike traditional approaches that focus solely on a few selected features of PSCs (e.g., selected quantiles), our approach is conducive to stochastic realizations of the spatial distribution of the entire particle-size distribution which can optionally be conditioned on available measured data. We model PSCs as cumulative distribution functions, and their densities as functional compositions in a Bayes Hilbert space. This enables us to employ an appropriate geometry to deal with the data dimensionality and constraints, and to develop a simulation method for particle-size densities (PSDs) based upon a suitable and well defined projection procedure. The new theoretical framework enables us to represent and reproduce the complete information content embedded in PSC data. As a first field application, we test the quality of unconditional and conditional simulations obtained with our methodology by considering as a test bed a set of particle-size curved collected within a shallow alluvial aquifer in the Neckar river valley, Germany.

Keywords: Geostatistics; Functional Compositions; Particle-size distribution; Groundwater; Hydrogeology

1 Introduction

Characterization of natural heterogeneity of aquifer bodies relies on diverse sets of observations. These include, for example, direct measurements/estimates of hydraulic parameters such as hydraulic conductivity and porosity and data which enables to infer a classification of soil types. Merging all available information within a unique theoretical and operational framework would form the basis for a robust system characterization. A stochastic approach is nowadays recognized as a viable tool to quantify the way uncertainty propagates from incomplete knowledge of the properties of the host porous medium (in terms of

spatial distribution of geomaterials and associated parameters) to state variables of interest (including, e.g, groundwater fluxes and chemical concentrations).

Here, we focus on the way the information content embedded in particle-size curves (PSCs) can be effectively employed to assist the stochastic characterization of a natural aquifer. These types of data are routinely available in field studies performed in diverse settings. They are usually obtained through relatively simple and inexpensive methods, such as those relying on traditional grain sieve analysis or other techniques, based on, e.g., sedigraph or laser diffraction methods. Information content which can be extracted by PSCs include a set of representative particle diameters that are defined as average soil particle sizes corresponding to given quantiles of the PSC. Representative diameters can then be employed within existing empirical formulations relating them to aquifer parameters such as porosity and/or saturated hydraulic conductivity [e.g., Rosas et al., 2014, Vienken and Dietrich, 2011, Vukovic and Soro, 1992]. In a few other cases [e.g., Rogiers et al., 2012] a site-specific model is proposed to assess the possibility of estimating saturated hydraulic conductivity from the complete dataset characterizing the PSCs. These can also be employed for the purpose of soil textural classification, according to a variety of approaches [e.g., Riva et al., 2006, Martìn et al., 2005, and references therein]. In this sense, texture data consisting in percentage values of sand, silt and clay (which can be inferred from PSCs) can be employed together with other quantities, including e.g., bulk density of soil, as input to pedotransfer functions to estimate soil hydraulic properties [e.g., Rawls et al., 1982, Pachepsky et al., 2006, Schaap et al., 2001, Schaap, 2013, and references therein]. An alternative approach is grounded on concepts of similar media scaling [e.g., Miller and Miller, 1956, Vogel et al., 1991] to exploit the dependence of hydraulic properties on pore size and key geometrical descriptors of the pore space. The latter approach enables one to scale hydraulic properties of multiple soils to unique reference water retention curves and partially saturated relative hydraulic conductivity functions [e.g.,amongst others, Tuli et al., 2001, Das et al., 2005, Nasta et al., 2013]

In this broad framework, hydrogeological investigations commonly employ a number of discrete quantiles of an available PSC which are then subject to geostatistical analysis and then (a) projected onto a grid through kriging or (b) employed in a numerical Monte Carlo setting to generate multiple realizations of the spatial distribution of aquifer properties and/or textural composition [e.g., Riva et al., 2006, 2008, 2010, Hu et al., 2009, Bianchi et al., 2011]. As recently pointed out by Menafoglio et al. [2014, 2015], these standard approaches suffer from two major drawbacks: (a) they require the joint geostatistical analysis of multiple characteristic particle diameters with an ordering constraint, thus entailing, e.g., calibration of multiple variogram and cross-variogram models, and (b) they are not conducive to exploiting fully the richness of the information content associated with available PSCs.

Having at our disposal advanced techniques for the geostatistical simulation of an entire particle-size distribution instead of, e.g., selected quantiles, would dramatically improve our ability to represent and reproduce the complete information content embedded in PSC data. The development and establishment of the theoretical basis underpinning these concepts and their translation into operational simulation algorithms and tools has therefore the clear potential to provide a remarkably improved characterization of the variability of the system to be embedded within Monte Carlo based stochastic simulation procedures of

groundwater flow and chemical transport. To the best of our knowledge, the challenging problem of performing geostatistical simulations of soil particle-size distributions has not yet been explored in the literature.

Here, we focus on this problem and present theoretical developments and associated computational algorithms and operational procedures to generate multiple replicates of spatial distributions of PSCs. These can optionally be conditional on available observed PSCs at a set of discrete locations in the system. In the latter case, the simulations interpolate the data observed at the sampled locations. We demonstrate our approach by way of a field-scale analysis grounded on observed PSCs and obtain (conditional and unconditional) realizations of PSCs which are amenable to be included in a Monte Carlo based approach aimed at the statistical characterization of groundwater flow and transport in randomly heterogeneous aquifer systems.

We ground our approach on a non-parametric framework, which combines the point of view of geostatistics [Chilès and Delfiner, 1999], Functional Data Analysis [FDA, Ramsay and Silverman, 2005] and Compositional Data Analysis [CoDa, Pawlowsky-Glahn et al., 2015]. We do so by following the concepts first introduced by Menafoglio et al. [2014, 2015]: we model PSCs as cumulative distribution functions and interpret their densities, termed particle-size densities (PSDs), as functional compositional data.

We employ an appropriate geometry to deal with the compositional nature of the data, and develop a simulation method for PSDs based upon a suitable and well defined projection strategy. The latter enables us to (a) reduce the dimensionality of the problem by guaranteeing a high degree of precision, and (b) geostatistically characterize and simulate PSDs via an approximated multivariate problem.

The work is organized as follows. Section 2 describes the field data that are employed as a test bed to illustrate our methodology, while Section 3 introduces the basic notions on Bayes space theory that are here employed. Section 4 illustrates our simulation strategy in the unconditional and conditional settings. Section 5 presents our simulation results obtained at the target field site. Section 6 concludes the work.

2 Experimental site and available data

We consider here a dataset obtained at the Lauswiesen site, located in the Neckar river valley near the city of Tübingen, Germany. The subsurface system in the area has been characterized through extensive information obtained at a number of boreholes, which are employed to perform sedimentological as well as hydraulic analyses. A relatively regular upper clay layer with a thickness of 1 - 2 m overlies a conductive Quaternary sand and gravel deposit. The latter rests on a layer of Keuper marl which is considered to define an impervious bedrock boundary of the aquifer hosted in the Quaternary sand and gravel system. The saturated thickness of the aquifer we are considering is approximately 5 m. All boreholes penetrate the aquifer down to bedrock. Details of site hydrogeology are given by Riva et al. [2006] and references therein. Available pumping test data have been employed by Neuman et al. [2007] for the stochastic analysis of late-time drawdowns and by Panzeri et al. [2015] for the application of data assimilation techniques based on the concept of Moment Equation Ensemble

Kalman Filter.

Of specific relevance to our study are the available 406 PSCs sampled along 12 fully penetrating vertical boreholes. The dataset was employed by Riva et al. [2006, 2008, 2010], Barahona-Palomo et al. [2011] and Riva et al. [2014] in the context of stochastic modeling studies aimed at (a) providing a probabilistic analysis of solute residence times within well capture zones, (b) interpreting an observed tracer test in a numerical Monte Carlo framework, (c) assessing the link between the spatial covariance functions of the (natural) logarithm of hydraulic conductivity (K) and of soil particle representative diameters, and (d) characterizing the correlation between hydraulic conductivity values estimated through impeller flowmeter downhole measurements and by way of empirical formulations based on PSC representative diameters.

The available PSCs were measured on soil samples of characteristic length ranging from 5 to 26.5 cm. A number of 12 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 in the sieve analysis procedure. Figure 1c depicts a sketch of the borehole network and sampling locations at the site. Applying traditional empirical relationships between characteristic soil diameters and permeability indicates that the site is mainly constituted by heterogeneous and conducive deposits of alluvial origin.

Particle size curves associated with one of the available boreholes (borehole B5 in Figure 1) have been employed by Menafoglio et al. [2014] to perform a geostatistical analysis of PSCs through the corresponding densities, interpreted as Functional Compositions (FCs). These authors embed this latter concept within the geostatistical framework of Menafoglio et al. [2013] through which they project (Kriging) estimates of the full PSC on a computational grid, together with the associated Kriging variance. The geostatistical setting of Menafoglio et al. [2014] has been extended by Menafoglio et al. [2015] to characterize the complete set of PSCs at the site and to properly account for the information content related to the local occurrence of diverse soil types (or textural classes). The key result of the authors is the formulation of an original theoretical framework according to which one can take full advantage of the complete set of information embedded in measured PSCs to (a) classify PSCs into clusters which represent the occurrence at a site of diverse soil types (b) characterize the spatial distribution of each identified textural class, and (c) predict the heterogeneous distribution of PSCs within each region which contributes to form the internal architecture of the geological system.

Menafoglio et al. [2014] and Menafoglio et al. [2015] analyze available PSDs by resorting to a smoothing procedure for PSCs based on Bernstein polynomials. This enabled them to obtain the smooth estimates of PSDs from raw data (Figure 1a and b), and to embed these in their geostatistical analyses. These data refer to the particle-size distribution within the domain of available observation, i.e., associated to the grain dimensions between the minimum and the maximum sieve diameters. For the purpose of illustration, we here consider a subset of the smoothed data of Menafoglio et al. [2015], as detailed in Section 5; the reader is referred to Menafoglio et al. [2015] for further details on data preprocessing.

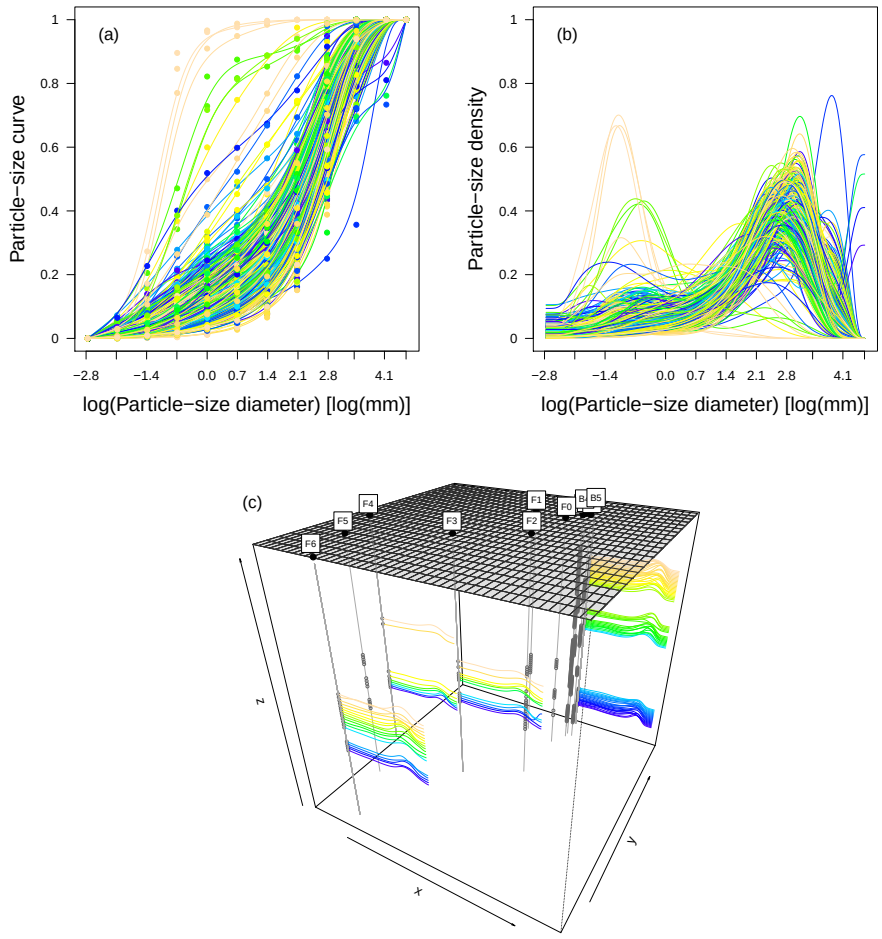

Figure 1: From field data to a sample of PSDs: (a) raw (symbols) and smoothed (solid lines) conditional PSCs; (b) smoothed conditional PSDs; (c) smoothed conditional PSDs along boreholes B5, F3, F4 and F6. Colors indicate the depth of the sampling locations

3 Density functions as elements of a Bayes space

A proper (geo)statistical analysis and simulation of PSDs should account for the peculiar nature of this kind of constrained (compositional) data. The log-ratio approach for the statistical analysis of multivariate compositions was pioneered by Aitchison [1986], Pawlowsky-Glahn and Egozcue [2001] and is well established in the statistical literature. It is based on the key observation that constant-sum objects convey only relative information. Indeed, one can readily see that a component (or part) of a compositional vector does not provide information *per se*, but relative to the measure of the whole – i.e., the constant they sum up to – and to the remaining parts of the composition. Note that the measure of the whole (e.g., unity, 100) is in general a convention rather than an informative element for the analysis. The Aitchison geometry then yields a proper setting to perform the statistical analysis, by accounting for the data constraints via the log-ratio approach.

In this setting, density functions, such as PSDs, can be viewed as Functional Compositions (FCs), i.e., compositional vectors with infinitely-many parts, that are constrained to be positive and to integrate to a constant. As such, they inherit the key properties of multivariate compositions. Recent works of Egozcue et al. [2006, 2013], van den Boogaart et al. [2010] and van den Boogaart et al. [2014] extend the Aitchison geometry to the infinite-dimensional setting through the theory of Bayes spaces, with the aim of providing the space of FCs with a geometrical structure consistent with the key properties of compositions and allowing for their statistical analysis. As in Menafoglio et al. [2014, 2015], we here focus on continuous FCs with compact support $\mathcal{T} = [t_{min}, t_{max}]$. We say that two FCs f, g are equivalent if they are proportional, i.e., $f = c \cdot g$, for $c > 0$. Note that this equivalence relation reflects the so-called *scale invariance* property of FCs upon which the log-ratio approach is grounded. Indeed, proportional FCs convey the same set of *relative* information; in other words, the measure of the whole is of no interest in a compositional analysis. Hereinafter, we will always consider as representative of an equivalence class of FCs its element integrating to 1.

We term $A^2(\mathcal{T})$ (or A^2 for short) the space of (equivalence classes of) FCs on $\mathcal{T}$, whose logarithms are squared integrable, i.e.,

$$A^2 = \left\{ f : \mathcal{T} \to (0, +\infty),\ \int_{\mathcal{T}} \log^2(f) < +\infty \right\}. \tag{1}$$

Following Egozcue et al. [2006] and van den Boogaart et al. [2014], we define on A^2 the operation of perturbation $\oplus$ and powering $\odot$

$$f \oplus g = \mathcal{C}(fg); \quad \alpha \odot f = \mathcal{C}(f^\alpha), \quad f, g \in A^2,\ \alpha \in \mathbb{R}, \tag{2}$$

where $\mathcal{C}(f) = \int_{\mathcal{T}} f$ is the closure operation, which maps a FC in the representative of its equivalence class that integrates to 1. Note that the neutral elements of perturbation and powering are $0_\oplus \equiv 1/|\mathcal{T}|$, with $|\mathcal{T}|$ the length of $\mathcal{T}$, and 1, respectively. Egozcue et al. [2006] prove that $(A^2, \oplus, \odot)$ is a vector space, perturbation and powering playing the role of sum and product by a constant, respectively. In this setting, we denote by $f \ominus g$ the difference, in the geometry of A^2, between f and g, namely the perturbation of f with the reciprocal of g, i.e., $f \ominus g = \mathcal{C}[f \oplus 1/g]$, for f, g in A^2.

To endow A^2 with a Hilbert space structure, Egozcue et al. [2006] equip the vector space $(A^2, \oplus, \odot)$ with the inner product

$$\langle f, g \rangle_{A^2} = \int_{\mathcal{T}} [\log(f) \log(g)] - \frac{1}{|\mathcal{T}|} \int_{\mathcal{T}} \log(f) \int_{\mathcal{T}} \log(g), \quad f, g \in A^2. \tag{3}$$

Egozcue et al. [2006] prove that $(A^2, \oplus, \odot, \langle \cdot, \cdot \rangle_{A^2})$ is a Hilbert space.

An isometric isomorphism exists between the space $A^2(\mathcal{T})$ and the space $L^2(\mathcal{T})$ of (equivalence classes of) square-integrable functions on $\mathcal{T}$. An example of such an isometric isomorphism is the centred log-ratio (clr) transformation $\mathrm{clr} : A^2 \to L^2$, that acts on a FC $f \in A^2$ as

$$\mathrm{clr}(f) = \log(f) - \frac{1}{|\mathcal{T}|} \int_{\mathcal{T}} \log(f). \tag{4}$$

From the computational viewpoint, the use of clr-transforms is convenient, as it allows mapping the problem in L^2, where most methods of FDA can be applied. Since clr is an isometric isomorphism, one has

$$\mathrm{clr}(f \oplus g) = \mathrm{clr}(f) + \mathrm{clr}(g), \quad \mathrm{clr}(\alpha \odot f) = \alpha \cdot \mathrm{clr}(f), \quad \langle f, g \rangle_{A^2} = \langle \mathrm{clr}(f), \mathrm{clr}(g) \rangle_{L^2}, \tag{5}$$

with $\langle x, y \rangle_{L^2} = \int_{\mathcal{T}} xy$.

We make extensive use of the Hilbert space geometry of the space $(A^2, \oplus, \odot, \langle \cdot, \cdot \rangle_{A^2})$ in the following Sections and show how resorting to a clr-transform simplifies the computations.

4 A projection strategy for the simulation of Particle-Size Densities

We consider the probability space $(\Omega, \mathfrak{F}, \mathbb{P})$, Ω being a non-empty set (i.e., a space of events), $\mathfrak{F}$ a $\sigma-$algebra of subsets of Ω, and $\mathbb{P}$ a probability measure defined on $\mathfrak{F}$. We denote by $D \subset \mathbb{R}^3$ a three-dimensional domain, and by $\{\mathcal{X}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ the random field, defined on $(\Omega, \mathfrak{F}, \mathbb{P})$ whose generic element $\mathcal{X}_{\boldsymbol{s}} : \mathcal{T} \to [0, 1]$, indexed by the location $\boldsymbol{s}$ in D, is a random particle-size curve defined on the (same) compact domain $\mathcal{T}$. For any $t \in \mathcal{T}$ and $\boldsymbol{s} \in D$, $\mathcal{X}_{\boldsymbol{s}}(t)$ denotes the random proportion of particles having size smaller than or equal to t. Following [Menafoglio et al., 2014], we interpret $\mathcal{X}_{\boldsymbol{s}}$ as an absolutely continuous cumulative distribution function and focus on the derivative field $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, whose elements are probability density functions. We call $\mathcal{Y}_{\boldsymbol{s}}$ the *particle-size density* (PSD) of the particle-size curve $\mathcal{X}_{\boldsymbol{s}}$. Hereafter, we consider each $\mathcal{Y}_{\boldsymbol{s}}$ as an element of the Hilbert space of functional compositions, A^2, introduced in Section 3.

For any $\boldsymbol{s}$ in D, we denote by $m_{\boldsymbol{s}}$ the Fréchet mean of $\mathcal{Y}_{\boldsymbol{s}}$, i.e. [Fréchet, 1948]

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

We indicate with C the covariance function of the field $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$. Function C maps any pair of locations $\boldsymbol{s}_1, \boldsymbol{s}_2$ in D into the cross-covariance operator $C(\boldsymbol{s}_1, \boldsymbol{s}_2)$ between the elements of the field at such locations, i.e.,

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2)x = \mathbb{E}[\langle \mathcal{Y}_{\boldsymbol{s}_1} \ominus m_{\boldsymbol{s}_1}, x \rangle_{A^2} \odot (\mathcal{Y}_{\boldsymbol{s}_2} \ominus m_{\boldsymbol{s}_2})], \quad x \in A^2. \tag{6}$$

In the following, we assume $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ to be a stationary Gaussian random field in A^2 [Bogachev, 1998, Bosq, 2000]. This implies that the mean function $m_{\boldsymbol{s}} = m$ is spatially constant, and there exists a function $\widetilde{C}$ such that $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \widetilde{C}(\boldsymbol{s}_1 - \boldsymbol{s}_2)$ for any $\boldsymbol{s}_1, \boldsymbol{s}_2$ in D. For ease of notation, hereafter we denote $\widetilde{C}$ by C.

Let $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ be sampling/measurement locations in D. Given the observation of $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$ at these locations, our goal is to provide a simulation (or realization) of the PSD $\mathcal{Y}_{\boldsymbol{s}_0}$ at a given location $\boldsymbol{s}_0$ in D. The target simulations may be either *unconditional* or *conditional*. The former are realizations from the (estimated) distribution of the field $\{\mathcal{Y}_{\boldsymbol{s}}\}$, whereas the latter are realizations from the (estimated) conditional distribution of $\{\mathcal{Y}_{\boldsymbol{s}}\}$ given the observations $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$. As such, conditional simulations reproduce the actual data at the measurement locations.

We consider for each $\boldsymbol{s} \in D$ the expansion

$$\mathcal{Y}_{\boldsymbol{s}}(\omega) = m \oplus \bigoplus_{k=1}^{+\infty} \xi_k(\boldsymbol{s}; \omega) \odot u_k, \quad \omega \in \Omega, \tag{7}$$

where $\{u_k, k \geq 1\}$ is a given orthonormal basis of A^2, and $\xi_k(\boldsymbol{s}; \omega) = \langle \mathcal{Y}_{\boldsymbol{s}}(\omega) \ominus m, u_k \rangle_{A^2}$. The basis $\{u_k, k \geq 1\}$ and the expansion (7) are well defined by virtue of the Hilbert space structure of the space A^2.

If we could jointly simulate random realizations of all the real (random) coefficients $\{\xi_k(\boldsymbol{s}_0)\}_{k \geq 1}$, we would obtain a random realization of $\mathcal{Y}_{\boldsymbol{s}_0}$ through (7). However, it should be noted that this is practically unaffordable because the effort required to simulate a multivariate random field increases with its dimensionality. We circumvent this issue by considering, for $\boldsymbol{s}$ in D and ω in Ω, the sequence of truncated expansions

$$\mathcal{Y}_{\boldsymbol{s}}^K(\omega) = m \oplus \bigoplus_{k=1}^{K} \xi_k(\boldsymbol{s}; \omega) \odot u_k, \quad K \geq 1. \tag{8}$$

The element $\mathcal{Y}_{\boldsymbol{s}}^K$ associated with a truncation order K yields an approximation of $\mathcal{Y}_{\boldsymbol{s}}$ such that

$$\mathbb{E}[\|\mathcal{Y}_{\boldsymbol{s}}^K \ominus \mathcal{Y}_{\boldsymbol{s}}\|_{A^2}^2] = \sum_{k=K+1}^{+\infty} \mathbb{E}[|\xi_k(\boldsymbol{s})|^2] = \sum_{k=K+1}^{+\infty} \langle C(\boldsymbol{0}) u_k, u_k \rangle, \tag{9}$$

which approaches 0 as K increases to infinity. Note that the term at the right hand side of (9) does not depend on the spatial index $\boldsymbol{s}$ in D. Thus, for any given tolerance, one can determine a truncation order K such that $\mathcal{Y}_{\boldsymbol{s}}^K$ approximates $\mathcal{Y}_{\boldsymbol{s}}$ – in the mean square sense – with a desired precision, uniformly in D.

Given a truncation order K, a random field $\{\mathcal{Y}_{\boldsymbol{s}}^K, \boldsymbol{s} \in D\}$ whose elements are given by (8) can be defined on $(\Omega, \mathfrak{F}, \mathbb{P})$ in A^2. The distributional properties of such a field are determined by m and by those of the zero-mean multivariate random field $\{\boldsymbol{\xi}(\boldsymbol{s}), \boldsymbol{s} \in D\}$, $\boldsymbol{\xi}(\boldsymbol{s})$ indicating the K-dimensional coefficient vector of the basis expansion (8) in $\boldsymbol{s}$, i.e., $\boldsymbol{\xi}(\boldsymbol{s}) = (\xi_1(\boldsymbol{s}), ..., \xi_K(\boldsymbol{s}))^T$. Note that both $\mathcal{Y}_{\boldsymbol{s}}^K$ and $\boldsymbol{\xi}(\boldsymbol{s})$ are Gaussian random fields (in A^2 and $\mathbb{R}^K$, respectively) by virtue of the Gaussian assumption on the field $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$. Additionally, we observe that the element $\mathcal{Y}_{\boldsymbol{s}}^K$ has mean $m_{\boldsymbol{s}}^K = m$ by virtue of (8), and the following

matrix representation of the covariance function C^K of the field $\{\mathcal{Y}_{\boldsymbol{s}}^K\}$ holds

$$C^K(\boldsymbol{h})x = \bigoplus_{j=1}^{K} \bigoplus_{k=1}^{K} (\mathbb{C}_{jk} x_j) \odot u_k, \tag{10}$$

where $x_j = \langle x, u_j \rangle_{A^2}$ and $\mathbb{C}_{jk} = \langle C(\boldsymbol{h}) u_j, u_k \rangle_{A^2} = \mathbb{E}[\xi_j(\boldsymbol{s}) \xi_k(\boldsymbol{s})]$.

In light of these observations, a natural strategy to obtain either conditional or unconditional simulations of the field $\{\mathcal{Y}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ is to resort to approximation (8) for an appropriate order K and then perform simulations of the multivariate random field $\{\boldsymbol{\xi}(\boldsymbol{s}), \boldsymbol{s} \in D\}$.

In principle, a large value for parameter K would be preferable, to obtain improved approximations of $\mathcal{Y}_{\boldsymbol{s}}$ through $\mathcal{Y}_{\boldsymbol{s}}^K$. Nevertheless, the value of K has a dramatic effect on the computational cost which is required for the simulation because it controls the dimensionality of the field $\{\boldsymbol{\xi}(\boldsymbol{s}), \boldsymbol{s} \in D\}$. Thus, one needs to consider a balance between limited computational power and accuracy.

We also note that the quality of a K-th order approximation of the kind (8) varies according to the basis $\{u_k, k \geq 1\}$ which is employed. Given $K \geq 1$, the mean square error of approximating $\mathcal{Y}_{\boldsymbol{s}}$ through the projection (8) over the first K elements of the basis $\{u_k, k \geq 1\}$ is bounded below by [see, e.g., Horváth and Kokoszka, 2012, Theorem 3.2]

$$\mathbb{E}[\|\mathcal{Y}_{\boldsymbol{s}}^K \ominus \mathcal{Y}_{\boldsymbol{s}}\|_{A^2}^2] \geq \sum_{k=K+1}^{+\infty} \lambda_k, \tag{11}$$

where (λ_k, e_k), $k \geq 1$, represent the eigenpairs of $C(\mathbf{0})$, with eigenvalues ordered in decreasing order $\lambda_1 \geq \lambda_2 \geq ...$. Given K, a sensible choice of the basis should then attain a mean square error of approximation as close as possible to the lower bound (11). It can be proved [e.g., Horváth and Kokoszka, 2012, Theorem 3.2] that the bound in Eq. (11) is reached when considering $u_1, ..., u_K$ to be precisely the set of the first K eigenvectors of $C(\mathbf{0})$, $e_1, ..., e_K$. The eigenvalue λ_k $(k = 1, 2, ...)$ then represents the proportion of the total variability which is captured by projecting the data along direction e_k. One can then set the truncation order K as the minimum order that allows explaining a given amount of the total variability (e.g., 90% or 95%) or, depending on the case analyzed, K can be identified as the minimum order at which an elbow starts to be appear in the scree plot, where the proportion of variability explained by the eigenvectors is plotted as a function of K.

In most studies, the zero-lag covariance operator is not known a priori. In this case, one can apply the so called Simplicial Functional Principal Component Analysis [SFPCA Hron et al., 2015] to (a) estimate from available data the zero-lag covariance operator $C(\mathbf{0})$ through the empirical estimator

$$Sx = \frac{1}{n} \sum_{i=1}^{n} \langle \mathcal{Y}_{\boldsymbol{s}_i} \ominus \widehat{m}, x \rangle_{A^2} (\mathcal{Y}_{\boldsymbol{s}_i} \ominus \widehat{m}), \quad x \in A^2, \tag{12}$$

$\widehat{m} = \frac{1}{n} \bigoplus_{i=1}^{n} \mathcal{Y}_{\boldsymbol{s}_i}$ denoting the sample mean, (b) compute the eigen-pairs $(\widehat{\lambda}_k, \widehat{e}_k)$, $k = 1, ..., n-1$, of this estimate, and (c) project the observations on the first K eigenvectors (or simplicial functional principal components, SFPCs) of S to

obtain the representation

$$\mathcal{Y}_{\boldsymbol{s}_i} \approx \widehat{m} \oplus \bigoplus_{k=1}^{K} \widehat{\xi}_k(\boldsymbol{s}_i) \odot \widehat{e}_k. \tag{13}$$

Here $\widehat{\xi}_k(\boldsymbol{s}_i) = \langle \mathcal{Y}_{\boldsymbol{s}_i} \ominus \widehat{m}, \widehat{e}_k \rangle_{A^2}$ is called *score* and is the projection of $(\mathcal{Y}_{\boldsymbol{s}_i} \ominus \widehat{m})$ along the k-th SFPC $\widehat{e}_k$. Note that SFPCA is the infinite-dimensional counterpart of principal components analysis, which is widely employed in the multivariate framework to perform optimal dimensionality reduction of a multivariate dataset. In general, most of the techniques that are commonly employed in the multivariate framework to identify and interpret principal components can be extended to the Bayes Hilbert space setting, as shown by Hron et al. [2015].

In this work, we ground the computation of the SFPCs and expansion (13) on the centered log-ratio (clr) transformation, that maps the space A^2 of functional compositions into the space L^2 of square-integrable functions, as recalled in Section 3. For $\omega \in \Omega$, we denote by $\mathcal{Z}_{\boldsymbol{s}}(\omega)$ the clr transform of $\mathcal{Y}_{\boldsymbol{s}}(\omega)$, i.e., $\mathcal{Z}_{\boldsymbol{s}}(\omega) = \mathrm{clr}(\mathcal{Y}_{\boldsymbol{s}}(\omega))$; we then define $\widehat{m}_c = \mathrm{clr}(\widehat{m})$. Hron et al. [2015] prove that, for any given K, the problem of finding the first K SFPCs of $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$ can be solved by back-transforming via clr^{-1} the first K eigenfunctions $\{\widehat{w}_k\}_{k=1,...,K}$ of the sample covariance operator S_{clr} of $\mathcal{Z}_{\boldsymbol{s}_1}, ..., \mathcal{Z}_{\boldsymbol{s}_n}$, that is defined as

$$S_{clr}x = \frac{1}{n}\sum_{i=1}^{n} \langle \mathcal{Z}_i - \widehat{m}_c, x \rangle_{L^2} (\mathcal{Z}_i - \widehat{m}_c), \quad x \in L^2.$$

Therefore, for $k = 1, ..., K$, one can compute the k-th SFPC as $\widehat{e}_k = \mathrm{clr}^{-1}(\widehat{w}_k)$, and use (5) to obtain the expansion

$$\mathcal{Z}_{\boldsymbol{s}}^{K}(\omega) = \widehat{m} + \sum_{k=1}^{K} \widehat{\xi}_k(\boldsymbol{s};\omega) \cdot \widehat{w}_k, \quad \omega \in \Omega, \tag{14}$$

where $\widehat{\xi}_k(\boldsymbol{s};\omega) = \langle \mathcal{Z}_{\boldsymbol{s}}^{K}(\omega) - \widehat{m}_c, \widehat{w}_k \rangle_{L^2}$ is the projection of $\mathcal{Z}_{\boldsymbol{s}}^{K}(\omega)$ along the k-th principal direction in the clr-space, identified by $\widehat{w}_k$. Note that the basis coefficients $\widehat{\xi}_k(\boldsymbol{s};\omega)$ appearing in (14) coincide with those in (13), as $\langle \mathcal{Y}_{\boldsymbol{s}}^{K}(\omega) \ominus \widehat{m}, \widehat{e}_k \rangle_{A^2} = \langle \mathcal{Z}_{\boldsymbol{s}}^{K}(\omega) - \widehat{m}_c, \widehat{w}_k \rangle_{L^2}$.

Given the optimal expansion (13), one can then employ multivariate techniques [e.g., Chilès and Delfiner, 1999, Mariethoz and Caers, 2015] to perform unconditional or conditional geostatistical simulations of the K-dimensional vectors of scores $\widehat{\boldsymbol{\xi}}(\boldsymbol{s}_i) = \left(\widehat{\xi}_1(\boldsymbol{s}_i), ..., \widehat{\xi}_K(\boldsymbol{s}_i)\right)^T$. Here, we illustrate the field application of our approach by employing the multivariate Gaussian simulator available in the package `gstat` [Pebesma, 2004] of software R [R Core Team, 2013]. Conditional simulations of Section 5.4 are based on the sequential Gaussian method of Abrahamsen and Benth [2001]. It is remarked that any multivariate simulation method could be employed as well, without substantial modifications to the overall strategy here proposed.

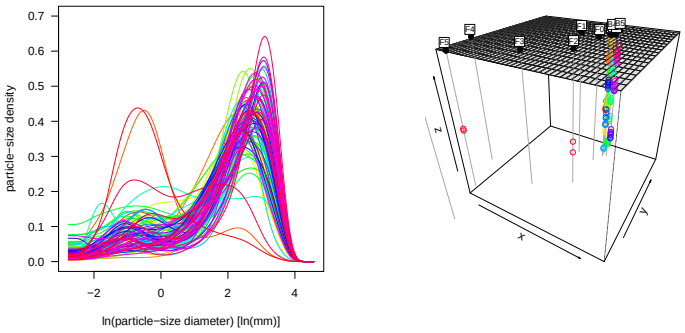

Figure 2: Sub-sample of conditional PSDs at the Lauswiesen field site.

5 Example of Application: Simulation of Particle-Size Densities at the Lauswiesen test site

We illustrate here our methodology for the simulation of PSDs on the basis of field data presented in Section 2. As a test bed, we consider the subset of the complete dataset depicted in Figure 2, formed by 100 PSDs randomly sampled from the set of data belonging to the second cluster singled out by Menafoglio et al. [2015]. As a first step, we apply SFPCA to this data set in Subsection 5.1 and obtain the best empirical basis for the representation of the data. In the following Subsections we illustrate the results of unconditional and conditional simulation at the site.

5.1 Simplicial Functional Principal Component Analysis of PSDs at the field site

Following the approach based on clr transform described in Section 4, we perform SFPCA of the dataset depicted in Figure 2. For the sake of simplicity, we estimate the mean m via the sample estimator $\widehat{m} = \frac{1}{n} \bigoplus_{i=1}^{n} \mathcal{Y}_{\boldsymbol{s}_i}$; more refined estimate may be employed [e.g., via generalized least squares Menafoglio et al., 2013, 2014]. Figure 3 depicts the key results of the analysis. Based on the scree plot in Figure 3a and on the scores boxplots in Figure 3b, we set the truncation order to $K = 4$. This choice enables us to explain 97% of the total variability of the dataset. The first $K = 4$ SFPCs $\{\widehat{e}_1, ..., \widehat{e}_4\}$ and their clr-transform $\{\widehat{w}_1, .., \widehat{w}_4\}$ are depicted in Figure 3c and d, respectively.

Figures 3e to h depict the mean function perturbed by plus/minus the eigenfunctions powered by twice the standard deviation along the corresponding direction, i.e., $\widehat{m} \oplus \left(\pm 2\sqrt{\widehat{\lambda}_k} \right) \odot \widehat{e}_k$, $k = 1, ..., 4$. The curves in Figure 3e-h are representative of the patterns characterizing the observations presenting high/low scores along the corresponding SFPCs. In this sense, the first SFPC captures the variability in the position of the mode and in the mass concentration around it. High scores along SFPC $\widehat{e}_1$ are represented by the blue curve in Figure 3e, which depicts a PSD with larger mode and higher mass concentration than $\widehat{m}$, the opposite behavior being depicted as a red curve in Figure 3e. The second

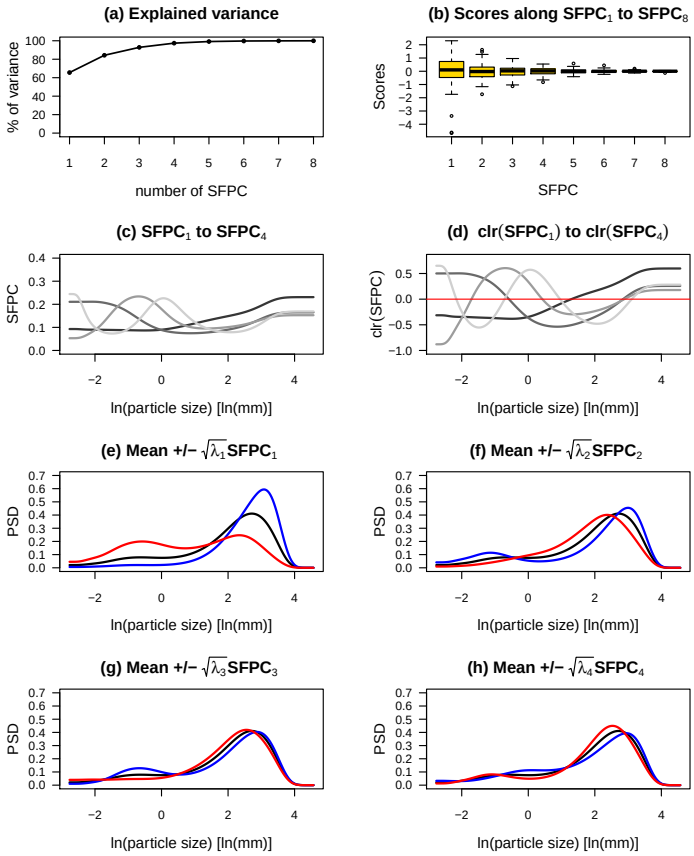

Figure 3: Results of SFPCA on the dataset of PSDs. Panels (d) to (g): the solid black curve indicates the mean function, the red curve indicates the mean $\ominus$ the SFPCs, the blue curve indicates the mean $\oplus$ the SFPC.

SFPC is interpreted in terms of the modality of the distribution (Figure 3f): high scores along the SFPC $\widehat{e}_2$ are registered for bimodal densities (blue curve), whereas low scores are associated with unimodal distributions. A correspondingly strong interpretation for the remaining SFPCs is not emerging as clearly as for the first two SFPCs.

Figures 4a and b compare the original data and their approximation based on the truncated expansion (13) with $K = 4$. Inspection of Figure 4 allows recognizing that the approximated curves provide a viable reproduction of all the main features of the original densities.

5.2 Geostatistical modeling of the scores

Once the approximation (13) has been obtained, simulation of a PSD $\mathcal{Y}_{\boldsymbol{s}_0}$ at a target location $\boldsymbol{s}_0$ in D requires the geostatistical characterization of the vectors of scores $\widehat{\boldsymbol{\xi}}(\boldsymbol{s}_1), ..., \widehat{\boldsymbol{\xi}}(\boldsymbol{s}_n)$. Consistent with the assumption of Section 4, we consider $\widehat{\boldsymbol{\xi}}(\boldsymbol{s}_1), ..., \widehat{\boldsymbol{\xi}}(\boldsymbol{s}_n)$ to be a partial observation of a K-dimensional stationary Gaussian random field $\{\widehat{\boldsymbol{\xi}}(\boldsymbol{s}), \boldsymbol{s} \in D\}$. Following Menafoglio et al. [2015], we

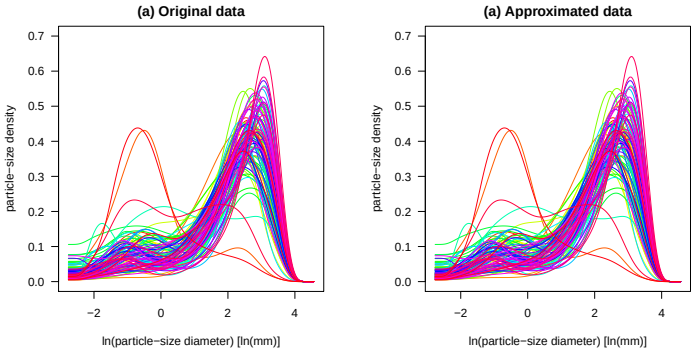

Figure 4: Original smoothed dataset and approximated PSDs obtained via (13).

consider a geometric anisotropy at the site, characterized by anisotropy ratio of $R = 0.04$ between the horizontal and vertical directions. Thus, hereafter we refer all our estimates and simulated quantities to an isotropic spatial domain obtained by dilation of the actual vertical coordinate by a factor $1/R = 25$. Figure 5 depicts the variograms and cross-variograms estimated from the scores $\widehat{\boldsymbol{\xi}}(\boldsymbol{s}_1), ..., \widehat{\boldsymbol{\xi}}(\boldsymbol{s}_n)$. We fit a valid model to these estimates by employing a Linear Model of Coregionalization [LMC, e.g., Chilès and Delfiner, 1999] based on an exponential model with nugget. We note that speed up of computations could be achieved upon employing simplifying assumptions on the vector of scores, e.g., by modeling the fields $\{\widehat{\boldsymbol{\xi}}_k(\boldsymbol{s}), \boldsymbol{s} \in D\}$, $k = 1, ..., K$, as uncorrelated. This simplifying assumption might be considered as a viable approximation at the site on the basis of the results depicted in Figure 5. For the sake of completeness, in our application described in the following Subsections we prefer to consider the complete LMC estimated as in Figure 5.

5.3 Unconditional simulation of PSDs

We illustrate an example of unconditional simulation of PSDs by considering a two-dimensional computational grid $D_0 \subset D$ which comprises 625 points, at a fixed elevation of 300 m a.s.l. Based on the LMC estimated in Subsection 5.2, we perform unconditional Gaussian cosimulation of the K-dimensional vectors $\widehat{\boldsymbol{\xi}}(\boldsymbol{s}_0)$, $\boldsymbol{s}_0 \in D_0$. Figure 6 depicts a selected realization simulated on the grid D_0 according to the proposed methodology.

We test the quality of the simulation by generating $NMC = 1000$ Monte Carlo replicates of the field on D_0. The CPU time required for the computations based on the R package `gstat`, within R version 3.0.2 was approximately 70'55" (CPU time refers to an Intel® Core™ i7-3517U CPU @ 1.90 GHz). We then compute the empirical variogram associated with each realization as well as directional sample variograms based on the collection of the NMC generated fields. Figure 7 depicts the generating variogram models together with the NMC variograms associated with (i) the generated fields and (ii) the sample variogram calculated along two mutual normal directions for a reference point located at the center of the simulation domain. Visual inspection of the results suggests that the generating variogram models are always fairly reproduced in

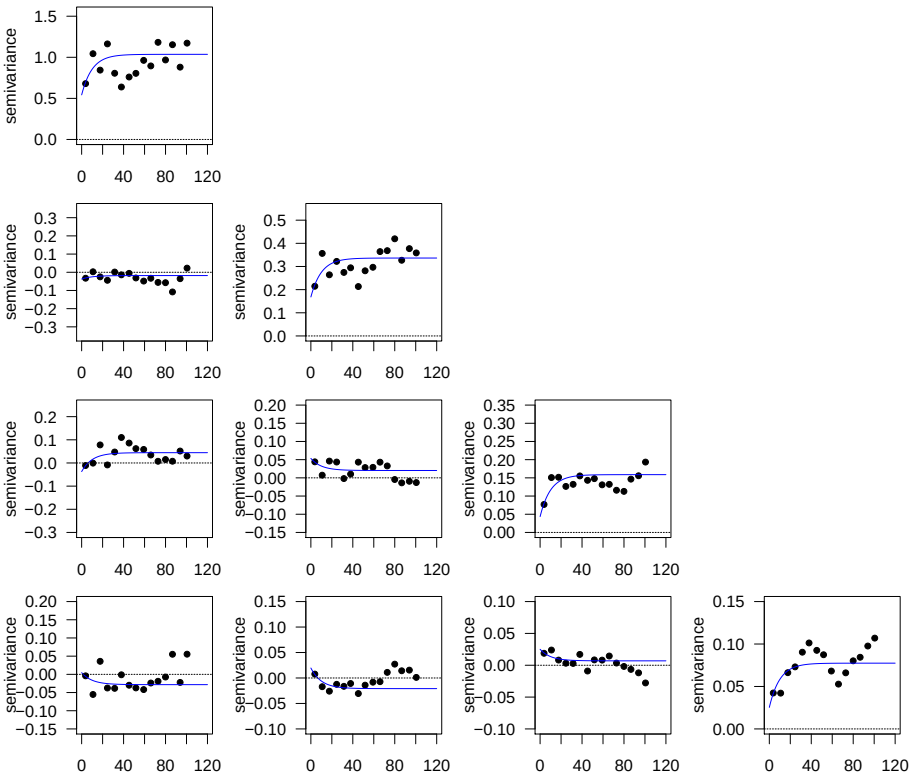

Figure 5: Variogram and cross-variograms estimated from the scores $\widehat{\boldsymbol{\xi}}_{\boldsymbol{s}_1}, ..., \widehat{\boldsymbol{\xi}}_{\boldsymbol{s}_n}$.

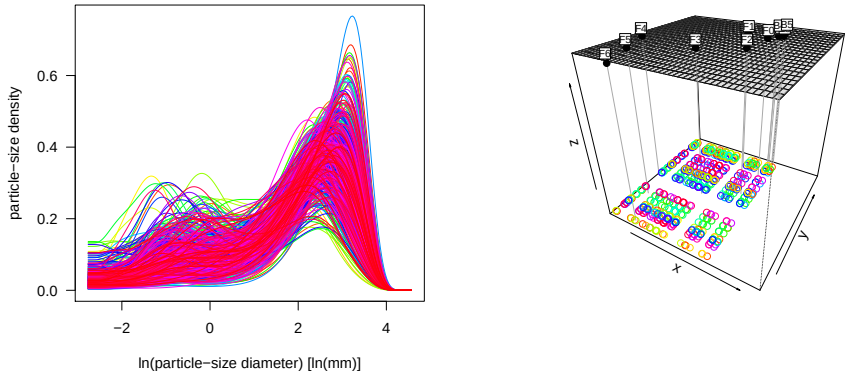

Figure 6: An example of unconditional realization of spatially dependent PSDs (left) and the simulation grid (right).

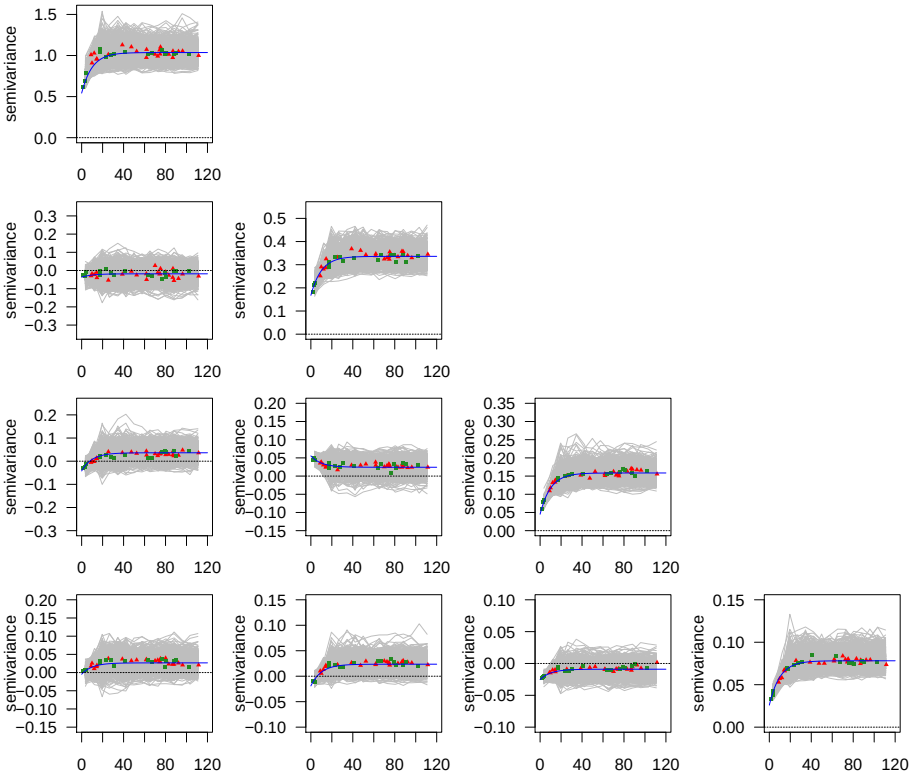

Figure 7: Generating LMC (blue lines), estimated variograms and cross-variogram in 1000 simulations (grey lines), average over 1000 simulation of the variogram estimated at the central point in direction x (red symbols) and y (green symbols).

an ensemble sense. Results of corresponding quality are obtained for other reference points in the system (not shown).

As an additional test, we repeat the same analysis by considering the trace-semivariogram of the field of PSCs, defined in this setting as

$$\gamma_{tr}(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|) = \mathbb{E}[\|\mathcal{Y}_{\boldsymbol{s}_i} \ominus \mathcal{Y}_{\boldsymbol{s}_j}\|^2_{A^2}], \quad \boldsymbol{s}_i, \boldsymbol{s}_j \in D. \tag{15}$$

The trace-variogram is a global measure of spatial dependence undertaking, in the functional context, the same role as its finite-dimensional counterpart [see, e.g., Menafoglio et al., 2013, 2014, and references therein].

The quality of the results of this analysis depicted in Figure 8 further corroborates our conclusions, thus imbuing us with confidence about the potential of the generation method and results.

5.4 Conditional Simulation of Particle-Size Densities at the Lauswiesen field site

Here, we illustrate an example of conditional simulation at the field site. For the purpose of our illustration, we consider a one-dimensional grid $D_1 \subset D$ of

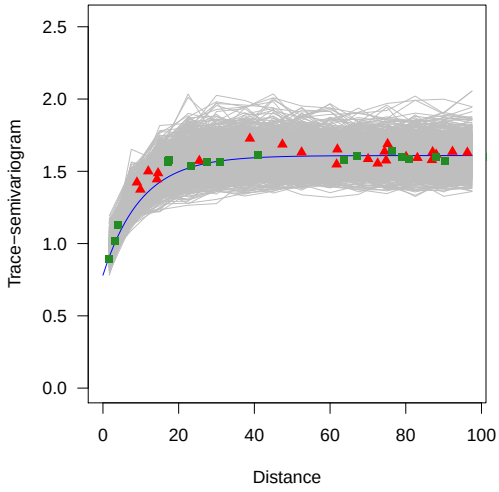

Figure 8: Generating model (blue lines), estimated trace-semivariograms in 1000 simulations (grey lines), ensemble average over 1000 simulation of the trace-semivariogram estimated at the central point in direction x (red symbols) and y (green symbols).

250 points taken along borehole B5 at the site. Simulations are here performed conditional to the set of approximated PSDs obtained in Subsection 5.1.

Figure 9 depicts a selected realization on grid D_1, obtained by conditionally simulating the K-dimensional vectors of coefficients $\widehat{\boldsymbol{\xi}}(\boldsymbol{s}_0)$, for $\boldsymbol{s}_0$ in D_1, according to the LMC of Figure 5. The CPU time for the simulation based on the R package `gstat`, within R version 3.0.2 took approximately 21'53" (CPU time refers to an Intel® Core™ i7-3517U CPU @ 1.90 GHz). It can be noted that, by construction, the simulation interpolates the approximated PSDs $\mathcal{Y}^K_{\boldsymbol{s}_1},...,\mathcal{Y}^K_{\boldsymbol{s}_n}$, rather than the observed PSDs $\mathcal{Y}_{\boldsymbol{s}_1},...,\mathcal{Y}_{\boldsymbol{s}_n}$. We refer to Appendix A for a strategy to honor the smoothed data – i.e., those prior to SFPCA.

To assess the quality of the prediction, we perform 1000 simulations on the grid D_1. We notice that, for each $\boldsymbol{s}_0 \in D_1$, the ensemble average of the simulations at $\boldsymbol{s}_0$, i.e., $\bigoplus_{j=1}^{1000} \mathcal{Y}^{(j)}_{\boldsymbol{s}_0}$, should approximate the conditional expectation $\mathbb{E}[\mathcal{Y}^K_{\boldsymbol{s}_0}|\mathcal{Y}^K_{\boldsymbol{s}_1},...,\mathcal{Y}^K_{\boldsymbol{s}_n}]$, as simulations $\mathcal{Y}^{(j)}_{\boldsymbol{s}_0}$, $j = 1,...,1000$, are draws from the (approximated) conditional distribution of $\mathcal{Y}^K_{\boldsymbol{s}_0}$ given $\mathcal{Y}^K_{\boldsymbol{s}_1},...,\mathcal{Y}^K_{\boldsymbol{s}_n}$. The conditional expectation $\mathbb{E}[\mathcal{Y}^K_{\boldsymbol{s}_0}|\mathcal{Y}^K_{\boldsymbol{s}_1},...,\mathcal{Y}^K_{\boldsymbol{s}_n}]$ can be estimated from available smoothed data $\mathcal{Y}^K_{\boldsymbol{s}_1},...,\mathcal{Y}^K_{\boldsymbol{s}_n}$ as

$$\mathcal{Y}^{*K}_{\boldsymbol{s}_0} = \widehat{m} \oplus \bigoplus_{k=1}^{K} \widehat{\xi}^*_k(\boldsymbol{s}_0) \odot \widehat{e}_k \tag{16}$$

where $\widehat{\boldsymbol{\xi}}^*(\boldsymbol{s}_0) = (\widehat{\xi}^*_1(\boldsymbol{s}_0),...,\widehat{\xi}^*_K(\boldsymbol{s}_0))^T$ is the Simple Cokriging prediction of the score vector at $\boldsymbol{s}_0$, based on $\widehat{\boldsymbol{\xi}}^*(\boldsymbol{s}_1),...,\widehat{\boldsymbol{\xi}}^*(\boldsymbol{s}_n)$ [see, e.g., Menafoglio and Petris, 2015]. Figure 10a-b displays the ensemble average of the 1000 simulated PSDs and the Kriging prediction based on the variography previously estimated, respectively. From the graphical inspection of Figure 10a-b one can appreciate

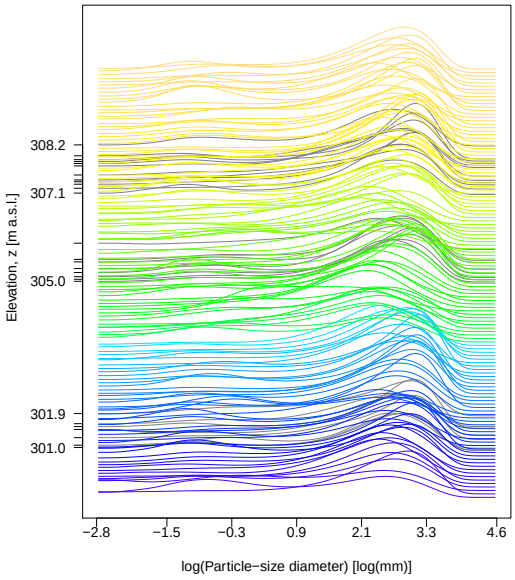

Figure 9: Conditional realization of PSDs at borehole B5 of Lauswiesen field site. Vertical coordinates correspond to the sample/target locations. Elevation is given in meters above sea level (m a.s.l.). Simulated PSDs are plotted as colored curves, data as grey curves.

the high quality of our simulations. This is also confirmed by Figure 10c, which represents, for $J = 1, ..., K$, the minimum, maximum and mean, over $\boldsymbol{s}_0 \in D_1$, of the squared distance $d(\boldsymbol{s}_0; J)^2 = \| \bigoplus_{j=1}^{J} \mathcal{Y}_{\boldsymbol{s}_0}^{(j)} \ominus \mathcal{Y}_{\boldsymbol{s}_0}^{*K} \|^2$ of the partial ensemble averages $\bigoplus_{j=1}^{J} \mathcal{Y}_{\boldsymbol{s}_0}^{(j)}$ from the Simple Kriging prediction $\mathcal{Y}_{\boldsymbol{s}_0}^{*K}$.

6 Conclusions and further research

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

1. A novel strategy has been proposed to address the problem of stochastic simulation of particle-size curves (PSCs) and associated densities (PSDs). The latter constitute a set of (infinite-dimensional) functional data and embedding them within the Bayes Hilbert space of functional compositions is a key feature of the procedure. Our theoretical framework enables us to (a) formulate a Gaussian model for the infinite-dimensional field of PSDs; (b) project the available data onto a truncated orthonormal basis to obtain a finite-dimensional approximation of the (otherwise infinite-dimensional) PSDs via a set of multivariate vectors of coefficients; and (c) perform either unconditional or conditional stochastic simulation, based on the multivariate random field of coefficients. The latter step can be addressed through the use of any of the available techniques for multivariate stochastic simulation (including, e.g., sequential Gaussian cosimulation).

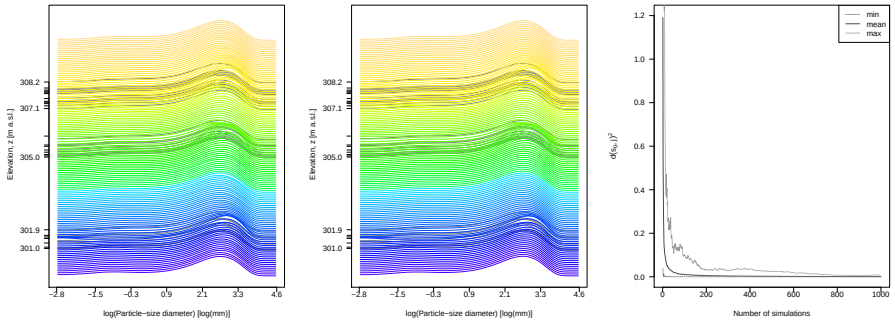

Figure 10: Assessment of the quality of conditional simulations at borehole B5 of Lauswiesen field site. (a) Average of 1000 conditional simulations of PSDs (b) Simple Kriging prediction of PSDs. (c) Squared distance between partial ensemble averages $\bigoplus_{j=1}^{J} \mathcal{Y}_{\boldsymbol{s}_0}^{(j)}$ and Simple Kriging prediction $\mathcal{Y}_{\boldsymbol{s}_0}^{*K}$. In panels (a) and (b), vertical coordinates correspond to the sample/target locations. Elevation is given in m a.s.l.. Simulated PSDs are plotted as colored curves, data as grey curves.

2. We study the way one can set the dimension of the approximating problem and the functional basis onto which these types of functional data can be projected. Our results suggest that an optimal solution is provided upon relying on a simplicial functional principal component analysis (SFPCA). In this context, one may need to set the dimensionality of the approximated problem according to the available computational resources. As such, key challenges associated with future direct implementation of the approach to field scale settings are related to improving the computational efficiency required for the simulation of the spatial field of coefficients, a step which still appears to be quite costly.

3. The stochastic simulation procedure has been demonstrated through an extensive Monte Carlo study based on a set of particle-size curves collected within a shallow alluvial heterogeneous aquifer system. The quality of our results appear to be quite satisfactory in all tested scenario. While we employ a stationary assumption for the purpose of our demonstration, it is possible to extend the technique to nonstationary settings of the kind arising, e.g., when an aquifer is conceptualized as a composite medium, where diverse non-overlapping materials form its internal architecture. Work in this direction is currently under way [Menafoglio et al., 2015]. With reference to practical applications, we note that, in contrast to common approaches relying solely on a few selected features of PSCs (e.g., selected quantiles), our approach yields collections of stochastic realizations of the spatial distribution of the entire PSC, thus contributing to a key improvement of one's ability to characterize the complete information content embedded in PSC data.

Appendix A: interpolating the observations in conditional simulations

By construction, the conditional simulations obtained through the projection strategy of Section 4 are based on the approximated PSDs $\mathcal{Y}^K_{\boldsymbol{s}_1}, ..., \mathcal{Y}^K_{\boldsymbol{s}_n}$, rather than the observed PSDs $\mathcal{Y}_{\boldsymbol{s}_1}, ..., \mathcal{Y}_{\boldsymbol{s}_n}$. Here, we illustrate a strategy to obtain simulations that honor the actual observations at locations where these are collected.

We call $\mathcal{Y}^K_{\boldsymbol{s}_0}$ the simulated PSD at a target location $\boldsymbol{s}_0 \in D$, and denote by $\epsilon^K_{\boldsymbol{s}_i} = \mathcal{Y}_{\boldsymbol{s}_i} \ominus \mathcal{Y}^K_{\boldsymbol{s}_i}$, $i = 1, ..., n$, the residuals of SFPCA. These residuals are neglected when analyzing and simulating PSDs via approximation (13). One can embed these in the (conditional) simulation procedure by interpolating them through an appropriate notion of Kriging, and then sum the result to the simulated realization $\mathcal{Y}^K_{\boldsymbol{s}_0}$.

Menafoglio et al. [2014] introduce the notion of Functional Compositional Kriging (FCK), that allows obtaining the best linear unbiased prediction in the sense of linear combination of the data in A^2. We call $\epsilon^{*K}_{\boldsymbol{s}_0}$ the FCK prediction of the residual at $\boldsymbol{s}_0$. This prediction is obtained as the linear combination $\epsilon^{*K}_{\boldsymbol{s}_0} = \bigoplus_{i=1}^n \lambda^*_i \odot \epsilon^K_{\boldsymbol{s}_i}$ of the residuals $\epsilon^K_{\boldsymbol{s}_i}$, $i = 1, ..., n$, whose weights minimize the prediction mean square error (MSE). Note that no unbiasedness constraint needs to be imposed, as the residuals $\epsilon^K_{\boldsymbol{s}_i}$ are zero mean by construction. Taking advantage of the work of Menafoglio et al. [2013, 2014], it is possible to show that minimization of the MSE is tantamount to solving the FCK system

$$\Gamma^\epsilon \boldsymbol{\lambda} = \boldsymbol{\gamma}^\epsilon_0, \tag{A1}$$

where $\Gamma^\epsilon_{i,j} = \mathbb{E}[\|\epsilon^K_{\boldsymbol{s}_i} \ominus \epsilon^K_{\boldsymbol{s}_j}\|^2_{A^2}]$, $i, j = 1, ..., n$, $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)^T \in \mathbb{R}^n$, $(\boldsymbol{\gamma}_0)_i = \mathbb{E}[\|\epsilon^K_{\boldsymbol{s}_i} \ominus \epsilon^K_{\boldsymbol{s}_0}\|^2_{A^2}]$, $i = 1, ..., n$. Note that (A1) is a Simple Kriging system, consistent with the observation that residuals are zero-mean.

Having computed the prediction $\epsilon^{*K}_{\boldsymbol{s}_0}$, one can finally obtain the desired simulation as $\mathcal{Y}^K_{\boldsymbol{s}_0} \oplus \epsilon^{*K}_{\boldsymbol{s}_0}$.

Acknowledgments. Funding 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 acknowledged. All data used in the paper will be retained by the authors for at least 5 years after publication and will be available to the readers upon request.

References

P. Abrahamsen and F. Benth. Kriging with inequality constraints. *Mathematical Geology*, 33(6):719–744, 2001. doi: 10.1023/A:1011078716252.

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

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

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

V. Bogachev. *Gaussian measures*. American Mathematical Society, 1998.

D. Bosq. *Linear Processes in Function Spaces*. Springer, New York, 2000.

J. P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, New York, 1999.

B. S. Das, N. W. Haws, and P. S. C. Rao. Defining geometric similarity in soils. *Vadose Zone Journal*, 4(2):264–270, 2005.

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

J.J. Egozcue, V. Pawlowsky-Glahn, R. Tolosana-Delgado, M.I. Ortego, and K.G. van den Boogaart. 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, 2013.

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

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

K. Hron, A. Menafoglio, M. Templ, K. Hruzova, and P. Filzmoser. Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis*, 2015. in press.

B. X. Hu, M. M. Meerschaert, W. Barrash, D. W. Hyndman, C. He, X. Li, and L. Guo. Examining the inïňĆuence of heterogeneous porosity ïňĄelds on conservative solute transport. *Journal of Contaminant Hydrology*, 108(3-4): 77–88, 2009.

G. Mariethoz and J. Caers. *Multiple-point Geostatistics: Stochastic Modeling with Training Images*. John Wiley & Sons, Ltd, 2015.

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

A. Menafoglio and G. Petris. Kriging for hilbert-space valued random fields: The operatorial point of view. *Journal of Multivariate Analysis*, 2015. DOI:10.1016/j.jmva.2015.06.012.

A. Menafoglio, P. Secchi, and M. Dalla Rosa. A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics*, 7:2209–2240, 2013.

A. Menafoglio, A. Guadagnini, and P. Secchi. 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, 2014.

A. Menafoglio, P. Secchi, and A. Guadagnini. A Class-Kriging predictor for Functional Compositions with application to particle-size curves in heterogeneous aquifers. *Mathematical Geosciences*, 2015. DOI:10.1007/s11004-015-9625-7.

E. Miller and R. Miller. Physical theory for capillary flow phenomena. *Journal of Applied Physics*, 27(4):324–332, 1956.

P. Nasta, N. Romano, S. Assouline, J. A. Vrugt, and J. W. Hopmans. Prediction of spatially variable unsaturated hydraulic conductivity using scaled particle-size distribution functions. *Water Resources Research*, 49:4219–4229, 2013.

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

Y. Pachepsky, W. Rawls, and H. Lin. Hydropedology and pedotransfer functions. *Geoderma*, 131(3):308–316, 2006.

M. Panzeri, M. Riva, A. Guadagnini, and S.P. Neuman. Enkf coupled with groundwater flow moment equations applied to lauswiesen aquifer, germany. *Journal of Hydrology,*, pages 205–216, 2015. DOI: 10.1016/j.jhydrol.2014.11.057.

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

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

Edzer J. Pebesma. Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, 30:683–691, 2004.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

J. Ramsay and B. Silverman. *Functional data analysis*. Springer, New York, second edition, 2005.

W. Rawls, D. Brakensiek, and K. Saxton. Estimation of soil water properties. *Transactions of the ASAE*, 28(5):1316–1320, 1982.

M. Riva, L. Guadagnini, A. Guadagnini, T. Ptak, and E. Martac. Probabilistic study of well capture zones distributions at the Lauswiesen field site. *Journal of Contaminant Hydrology*, 88:92–118, 2006.

M. Riva, A. Guadagnini, D. Fernández-Garcáa, X. Sánchez-Vila, and T. Ptak. Relative importance of geostatistical and transport models in describing heavily tailed breakthrough curves at the lauswiesen site. *Journal of Contaminant Hydrology*, 101:1–13, 2008.

M. Riva, L. Guadagnini, and A. Guadagnini. Effects of uncertainty of lithofacies, conductivity and porosity distributions on stochastic interpretations of a field scale tracer test. *Stochastic Environmental Research Risk Assessment*, 24: 955–970, 2010.

M. Riva, X. Sanchez-Vila, and A. Guadagnini. Estimation of spatial covariance of log-conductivity from particle-size data. *Water Resources Research*, 50: 5298–5308, 2014.

B. Rogiers, D. Mallants, O. Batelaan, M. Gedeon, M. Huysmans, and A. Dassargues. Estimation of hydraulic conductivity and its uncertainty from grain-size data using glue and artificial neural networks. *Mathematical Geosciences*, 44 (6):739–763, 2012.

J. Rosas, O. Lopez, T. M. Missimer, K. M. Coulibaly, A. H. A. Dehwah, K. Sesler, L. R. Lujan, and D. Mantilla. Determination of hydraulic conductivity from grain-size distribution for different depositional environments. *Ground Water*, 52(3):399–413, 2014.

M. G. Schaap. *Advances in Hydrogeology*, chapter Description, Analysis, and Interpretation of an Infiltration Experiment in a Semiarid Deep Vadose Zone, pages 159–183. Springer Science+Business Media, New York, 2013.

M. G. Schaap, F. J. Leij, and M. T. van Genuchten. Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology*, 251(3-4):163–176, 2001.

A. Tuli, K. Kosugi, and J. Hopmans. Simultaneous scaling of soil water retention and unsaturated hydraulic conductivity functions assuming lognormal pore-size distribution. *Advances in Water Resources*, 24(6):677–688, 2001.

K. G. van den Boogaart, J. J. Egozcue, and V. Pawlowsky-Glahn. Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56:171–194, 2014.

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

T. Vienken and P. Dietrich. Field evaluation of methods for determining hydraulic conductivity from grain size data. *Journal of Hydrology*, 400(1-2): 58–71, 2011.

T. Vogel, M. Cislerova, and J. Hopmans. Porous media with linearly variable hydraulic properties. *Water Resources Research*, 27(10):2735–2741, 1991.

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

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

58/2015 Iapichino, L.; Rozza, G.; Quarteroni, A.
Reduced basis method and domain decomposition for elliptic problems in networks and complex parametrized geometries

54/2015 Canuto, C.; Nochetto, R. H.; Stevenson, R.; Verani, M.
Adaptive Spectral Galerkin Methods with Dynamic Marking

55/2015 Fumagalli, A.; Zonca, S.; Formaggia, L.
Advances in computation of local problems for a flow-based upscaling in fractured reservoirs

56/2015 Bonaventura, L.; Della Rocca, A.
Monotonicity, positivity and strong stability of the TR-BDF2 method and of its SSP extensions

57/2015 Wilhelm, M.; Dedè, L.; Sangalli, L.M.; Wilhelm, P.
IGS: an IsoGeometric approach for Smoothing on surfaces

53/2015 Menafoglio, A; Grujic, O.; Caers, J.
Universal kriging of functional data: trace-variography vs cross-variography? Application to forecasting in unconventional shales

51/2015 Ballarin, F.; Faggiano, E.; Ippolito, S.; Manzoni, A.; Quarteroni, A.; Rozza, G.; Scrofani, R.
Fast simulations of patient-specific haemodynamics of coronary artery bypass grafts based on a POD–Galerkin method and a vascular shape parametrization

52/2015 Giverso, C.; Scianna, M.; Grillo, A.
Growing Avascular Tumours as Elasto-Plastic Bodies by the Theory of Evolving Natural Configurations

50/2015 Grillo, A.; Guaily, A.; Giverso, C.; Federico, S.
Non-Linear Model for Compression Tests on Articular Cartilage

49/2015 Ghiglietti, A.; Ieva, F.; Paganoni, A.M.
Statistical inference for stochastic processes: two sample hypothesis tests