1
Geostatistical analysis in Bayes spaces: probability densities and compositional data

Alessandra Menafoglio, Piercesare Secchi and Alberto Guadagnini

1.1
Introduction and motivations

The availability of complex, high-dimensional and often constrained data has recently fostered new areas of statistical research. These are typically placed at the intersection between functional data analysis (FDA, [1]), geostatistics, and other fields classically devoted to the analysis of constrained data, such as compositional data analysis (CoDa, [2]). In this context, there is a general consensus that modern geostatistical approaches should always consider the nature of the data. In some cases, this would require resorting to a geometry which is not necessarily the one of the space of square-integrable functions (i.e., L^2).

The focus of this Chapter is on Functional Compositions (FCs), that constitute the generalization to the functional setting of multivariate compositional data [3, 2]. The latter are defined as vector data that only provide relative information, i.e., for which the only relevant information is conveyed by the ratios between their components (termed *parts*). Examples of data that can be interpreted as compositional are discrete distributional data (i.e., probability mass functions), or, more generally, data whose components represent *parts* (e.g., proportions, percentages) of a whole (e.g., they sum up to unity) with respect to a given partition of the domain. For instance, concentration of chemicals adsorbed onto soil samples, or distribution of population in age classes are often considered as compositional information.

In this broad context, FCs are functional data which only convey relative information. One can envision FCs as positive data, constrained to integrate to a constant – even as this might not be the case for some applications. Informally, in FCs the ratios between their point evaluations are considered to be informative rather than their absolute values. For instance, probability density functions can be interpreted as FCs, and their point evaluations as infinitesimal parts of a whole, that is the probability of the sample space.

One can readily see that PDFs - as well as FCs in general - cannot simply be con-

sidered as square-integrable functions, because the geometry of L^2 is not appropriate to treat them (e.g., the L^2-sum of two FCs is meaningless). Instead, the Bayes space geometry, introduced in [4, 5, 6] and recalled in Section 1.2, is well-suited for functional compositional data, since it was precisely designed to correctly represent the peculiar features of those data.

Throughout the Chapter we will illustrate the geostatistical methods for FCs developed in [7, 8, 9], and their application to the field setting which firstly motivated those works, which deals with particle-size distributions sampled in a heterogeneous aquifer system. These data describe the local distribution of soil particles sizes and are relevant to problems related to groundwater hydrology, soil science, geophysics, petroleum engineering and geochemistry, with emphasis on applications oriented towards modeling physical and chemical processes occurring in heterogeneous Earth systems. Here, we illustrate methods for the pre-processing, kriging and assessing uncertainty of such data. These methods need to be framed within a space different from L^2.

The remaining of the Chapter is organized as follows. Section 1.2 introduces the Bayes space geometry for FCs, whereas Section 1.3 illustrates the data. The stationary kriging for FCs is addressed in Section 1.4, and the non-stationary approach is addressed in Section 1.5. Section 1.6 concludes the Chapter.

1.2
Bayes Hilbert spaces: natural spaces for functional compositions

The theory of Bayes spaces [4, 5, 10, 6] was introduced as a generalization to density functions of the Aitchison geometry. The latter is commonly employed to deal with compositional data, that are multivariate observations carrying only relative information [e.g., 3, 2, and references therein]. Compositional data are usually collected in the form of constrained objects summing up to a constant, usually set to 1 or 100, in case of proportions or percentages, respectively. Probability density functions can be then considered as compositional vectors with infinitely many parts [4], and with the key properties of compositions [e.g., 11].

We denote by f the density function of an absolutely continuous measure μ with respect to the Lebesgue measure on the Borel space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, with compact support $I \subset \mathbb{R}$. In the following, we will express the properties of μ through those of f. It should be noted that the theory of Bayes spaces was developed in a completely general framework in [5, 10, 6]. Two density functions f, g are considered as equivalent if they are proportional, and we denote such equivalence relation by $f =_{\mathcal{B}} g$. In this setting, the integral constraint $\int_I f(x)\,\mathrm{d}x = 1$ of PDFs singles out a representative within an equivalence class of FCs that are equivalent from the viewpoint of the *relative* information they provide. Indeed, for any other representative $\widetilde{f}$ (i.e., such that $\widetilde{f} = c \cdot f$ for $c > 0$) the relative contribution of Borel subsets of $\mathbb{R}$ w.r.t. the measure of the support is the same. This property is known as *scale invariance*, and is related to the observation that the probability of an event has no meaning *per se* - as noted in [10]. Otherwise, it is clearly framed in a relative context, as it is related

to the probability of the entire sample set, which is set to unity for convenience.

Another relevant feature of FCs is the *relative scale*. The latter indicates that the increase of probability should be understood and measured in a relative sense, rather than on an absolute scale. For instance, the increase of probability over a Borel set from 0.05 to 0.1 (2 multiple) differs from the increase 0.5 to 0.55 (1.1 multiple), although the absolute differences are the same in both cases. This property further motivates the use of the log-ratio approach to deal with density functions.

The above mentioned properties are well-known and recognized in the multivariate setting [e.g., 2], but are completely neglected when considering probability density functions as unconstrained objects. For instance, the notions of sum and product by a constant that would be used for data analysis in L^2 (the space of square-integrable functions) appear to be inappropriate for compositions, their application may yield functions that are no longer compositions. These elements motivated the introduction of a geometry capable to capture and properly incorporate the properties of FCs. Such geometry is that of Bayes Hilbert spaces, that generalize the Aitchison geometry [12] to the functional setting.

For ease of notation, and following [13, 14, 7, 8, 9], we focus here on density functions with compact support. Note that the theory here presented could be extended to general supports, through the use of reference measures different from the Lebesgue one. However, it should be noted that, in several real datasets, finite values for the inferior and superior extremes of the support can be determined without a substantial loss of generality, or working with conditional distributions.

We term $\mathcal{B}^2(I)$ the Bayes space of (equivalence classes of) positive FCs f on I with square-integrable logarithm. In the following, the representative of an equivalence class will be its element integrating to 1; moreover, we only consider continuous FCs on a closed interval $I = [a, b]$, any compact subset of $\mathbb{R}$ being compatible with our framework. Given two FCs $f, g \in \mathcal{B}^2(I)$ and $\alpha \in \mathbb{R}$ we denote by $f \oplus g$ and $\alpha \odot f$ the perturbation and powering operations, defined as, respectively, [4, 6]:

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s)\,\mathrm{d}s}, \quad (\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_I f(s)^\alpha\,\mathrm{d}s}, \quad t \in I.$$

It is then clear that the results of such operations are still probability density functions. Note that $\mathcal{B}^2(I)$ endowed with the operations $(\oplus, \odot)$ is a vector space [4] and that the origin of the space $\mathcal{B}^2(I)$ is $e(t) = 1/\eta$, with $\eta = b - a$. Moreover, the difference between two FCs $f, g \in \mathcal{B}^2(I)$ is obtained as perturbation of f with the reciprocal of g, i.e., $f \ominus g = f \oplus [(-1) \odot g]$.

Figure 1.1 depicts an example considered in [14] of the effect of perturbation and powering operations in $\mathcal{B}^2(I)$, as opposed to standard operations of sum and product by a constant in $L^2(I)$. In [14], the authors considered the restriction to $I = [-5, 5]$ of the Gaussian densities $f =_{\mathcal{B}} \exp\{-t^2/2\}$ and $g =_{\mathcal{B}} \exp\{-(t-m)^2/(2s^2)\}$, with $m = 1$ and $s^2 = 2$. Figure 1.1a juxtaposes the perturbation of f by g ($f \oplus g$) to the sum in L^2 of f and g ($f + g$). Note that the latter sum does not result in a probability density function, while the former does. Further, the perturbation of f by g yields a density function that is more concentrated than f and shifted towards g: this is the consequence of adding to f the information content in g and viceversa.

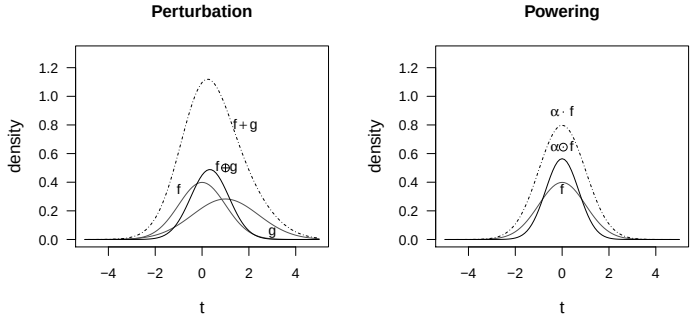

Figure 1.1 Example of perturbation and powering in $\mathcal{B}^2(I)$, compared to the typical operations in $L^2(I)$. Left: Perturbation $f \oplus g$ (solid black curve) of two Gaussian densities f, g restricted to $I = [-5, 5]$ (grey curves), and the sum $f + g$ in the space $L^2(I)$ (dot-dashed curve). Right: Powering of a Gaussian density f restricted to $I = [-5, 5]$ (grey curve) by $\alpha = 2$, $\alpha \odot f$ (solid black curve), and its counterpart $\alpha \cdot f$ in L^2 (dot-dashed curve). Modified from [14]

Notice that the operation of perturbation can be interpreted as a Bayesian update of information, and $\ominus$ as a cancellation of information [10]. As such, all conjugate priors define affine subspaces of $\mathcal{B}(I)$. Within the latter class we mention the Gaussian family and, more generally, the exponential family. Thus, it is not surprising that the result $f \oplus g$ displayed in Figure 1.1 is still a Gaussian density, as shown in [14].

Figure 1.1b depicts the result of the powering operation $\alpha \odot f$ in $\mathcal{B}^2(I)$, as well as the multiplication $\alpha \cdot f$ in $L^2(I)$, for the same f of Figure 1.1a and the scalar $\alpha = 2$. It is noted that $\alpha \cdot f$ is not a density function, and, as an element of $\mathcal{B}^2(I)$, it belongs to the same equivalence class as f itself. Otherwise, the powering of f by $\alpha = 2$ has the effect of increasing the concentration of f around its mean (i.e., it decreases the variance of f by a factor 2). In the Bayesian framework, this is interpreted as the increase of information which is obtained by incrementing the "evidence" in f by the "evidence" in f itself.

The space $(\mathcal{B}^2(I), \oplus, \odot)$ is a separable Hilbert space structure if equipped with the inner product [4]

$$\langle f, g \rangle_{\mathcal{B}} = \frac{1}{2\eta} \int_I \int_I \ln \frac{f(t)}{f(s)} \ln \frac{g(t)}{g(s)} \, \mathrm{d}t \, \mathrm{d}s, \quad f, g \in \mathcal{B}^2(I), \tag{1.1}$$

which induces the following norm

$$||f||_{\mathcal{B}} = \left[\frac{1}{2\eta} \int_I \int_I \ln^2 \frac{f(t)}{f(s)} \, \mathrm{d}t \, \mathrm{d}s \right]^{1/2}.$$

Each element of $\mathcal{B}^2(I)$ can be mapped onto an element of $L^2(I)$, preserving its distance and angle with any other element, that is, isometric isomorphisms exist between $\mathcal{B}^2(I)$ and $L^2(I)$. An example of such isometric isomorphism is defined by

the *centred log-ratio* (clr) transformation [6, 7], which is defined, for $f \in \mathcal{B}^2(I)$, as

$$\mathrm{clr}(f)(t) = f_c(t) = \ln f(t) - \frac{1}{\eta}\int_I \ln f(s)\,\mathrm{d}s. \tag{1.2}$$

One can see that the operations and inner products among the elements in $\mathcal{B}^2(I)$ can be equivalently computed in $L^2(I)$ as

$$\begin{aligned} &\mathrm{clr}(f \oplus g)(t) = f_c(t) + g_c(t), \quad \mathrm{clr}(\alpha \odot f)(t) = \alpha \cdot f_c(t), \\ &\langle f, g\rangle_{\mathcal{B}} = \langle f_c, g_c\rangle_2 = \int_I f_c(t) g_c(t)\,\mathrm{d}t. \end{aligned} \tag{1.3}$$

Note that clr-transform induces, by construction, a zero-integral constraint, which may yield model-singularities. However, this is not the case of the geostatistical methods here presented.

1.3 A motivating case study: particle-size data in heterogeneous aquifers – data description

This section illustrates the key features of the field setting within which our theoretical framework is applied. As a showcase scenario, we consider the Lauswiesen site, which is an experimental test site located near the city of Tuebingen, Germany. The aquifer system under consideration has been the subject of an extensive series of experimental campaigns and modeling studies. Amongst these, the reader is referred to the works of Riva et al. [15, 16, 17, 18], Hoffmann and Dietrich [19], Rein et al. [20], Neuman et al. [21, 22], Lessof et al. [23], Barahona-Palomo et al. [24], Handel and Dietrich [25], and Menafoglio et al. [7, 8, 9]. Characterization of the site has been based on data acquired through detailed geological, hydrogeological, hydraulic, sedimentological and geophysical investigations. The latter have been conducted at the field and laboratory scale.

The lithostratigraphic characterization has been performed through the stratigraphy information stemming from 150 mm-diameter monitoring wells [26, 27]. The aquifer at the site has a saturated thickness of about 5 m and is composed of fluvial geomaterial, overlain by stiff silty clay and underlain by hard silty clay. Available datasets include particle-size curves (PSC), pumping and tracer tests, direct-push injection logging and down-hole impeller flowmeter records. A detailed description of the analyses performed at the site is presented by Riva et al. [15, 16] and Lessof et al. [23], to which the reader is referred for details.

Of particular interest to our application are a collection of more than 400 PSCs collected along 12 vertical boreholes at the site. These indicate the presence of very heterogeneous, highly conductive alluvial deposits and were previously employed in [15, 16, 17] to provide a stochastic Monte Carlo-based numerical study of flow and transport process at the site. These studies considered diverse conceptual geological models of the structural heterogeneity of the system and analyzed their elative skill

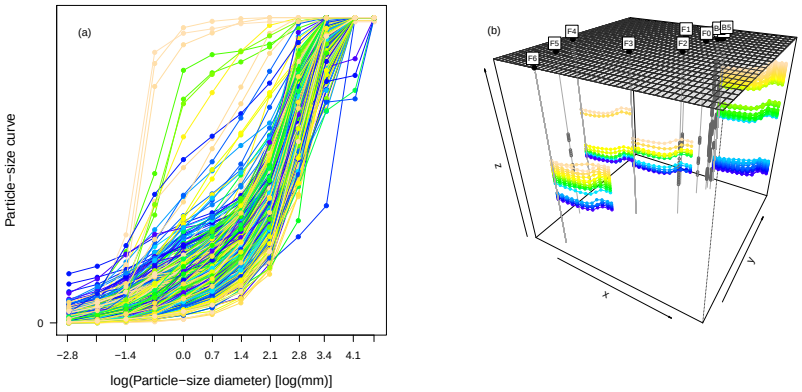

Figure 1.2 Raw particle-size data at the Lauswiesen site. (a) Collection of all avaialble Particle-Size Curves (PSCs) (b) Raw PSCs along boreholes B5, F3, F4 and F6. Colors indicate the depth of the sampling locations. Modified from [8].

to interpret available tracer tests data. The available PSCs were assessed on core samples of characteristic length ranging from 5 to 26.5 cm. They are reconstructed through grain sieve analysis performed with 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). Figure 1.2 depicts the three-dimensional structure of the sampling network at the site.

These PSCs have been employed in [15] to classify the types of geomaterials at the site and to construct geostatistically-based models of the internal architecture of the aquifer. In this context, the latter could then be conceptualized as formed by a collection of regions (or blocks), randomly located in space, each formed by a given material type. Hydraulic properties of each of these blocks can then be estimated through available empirical formulations relating, e.g., permeability and porosity to characteristic diameters of a PSC. For example, Riva et al. [15, 16, 17] relate d_{10} and d_{60} (respectively representing the particle size associated with the 10th and 60th percentile of a given PSC) to permeability through the Beyer's formula [28]. A geostatistical analysis of d_{10} and d_{60} or of the associated permeability can then be employed to characterize the heterogeneous distribution of hydraulic properties within the region occupied by each of the materials identified. The details of these analysis can be found in [15, 17]. Barahona-Palomo et al. [24] analyze the relationship between hydraulic conductivity estimates obtained through particle-size curves and impeller flowmeter measurements, while Riva et al. [18] rely on the available data to demonstrate their analytical study rendering relationships between the spatial covariance of hydraulic conductivity and of representative soil particle sizes and porosity.

1.4
Kriging stationary functional compositions

1.4.1
Model description

We term D the compact subset of $\mathbb{R}^d$ (usually $d = 2, 3$) corresponding to the spatial domain of the study, and denote by $s_1, ..., s_n$ the sampling locations in the test area. We denote by $\chi_{s_1}, ..., \chi_{s_n}$ the dataset collected at those locations, formed by a set of positive probability density functions on a compact domain I, i.e., $\chi_{s_i} : I \to (0, +\infty)$, such that $\int_I \chi_{s_i}(t)dt = 1$. Following Section 1.2, we consider $\chi_{s_1}, ..., \chi_{s_n}$ as objects of the Bayes Hilbert space $\mathcal{B}^2(I)$, and assume these to be a partial observation from a random field $\{\boldsymbol{\chi}_s, s \in D\}$ valued in $\mathcal{B}^2(I)$. For instance, $\chi_{s_1}, ..., \chi_{s_n}$ may be the densities of the particle-size distributions described in Section 1.3. Note that any other PDF can be considered for the application of our theoretical framework, including, e.g., rainfall (precipitation) distributions, or population pyramids [13], or dissolved chemical concentrations in groundwater.

In this section, we assume the process to be globally second order stationary and isotropic, i.e., the following conditions hold:

(i) Spatially constant mean: $\mathbb{E}[\boldsymbol{\chi}_s] = m$ for all $s \in D$;
(ii) Stationary and isotropic trace-covariogram: $\mathbb{E}[\langle \boldsymbol{\chi}_{s_1} \ominus m, \boldsymbol{\chi}_{s_2} \ominus m \rangle] = C(\|s_1 - s_2\|_d)$ for all $s_1, s_2 \in D$, $\|\cdot\|_d$ denoting a metric in $\mathbb{R}^d$.

Here, the mean and the covariogram are expressed in $\mathcal{B}^2(I)$, according to its geometric structure illustrated in Section 1.2. In such a space, under stationarity and isotropy, one may also define the spatial dependence structure through the trace-variogram of the process as

$$2\gamma(\|s_1 - s_2\|_d) = \mathbb{E}\left[\|\boldsymbol{\chi}_{s_1} \ominus \boldsymbol{\chi}_{s_2}\|^2\right].$$

The ordinary kriging predictor at a target location $s_0 \in D$ assumes in this context the form of the best linear combination of the data, linearity being interpreted in $\mathcal{B}^2(I)$ as $\boldsymbol{\chi}^*_{s_0} = \bigoplus_{i=1}^n \lambda_i^* \odot \boldsymbol{\chi}_{s_i}$. Informally, and in light of the example in Figure 1.1, such a linear combination is interpreted as a weighted sum of the information collected at each location, higher precisions (i.e., higher weight) being associated with nearby locations. Note that a zero weight $\lambda_i = 0$ powering a data-object χ_{s_i}, yields a contribution to the predictor in terms of a uniform PDF. This is precisely a null contribution in Bayes spaces, as the uniform PDF is the neutral element of the perturbation.

The ordinary kriging predictor is then found as the Best Linear Unbiased Predictor, whose weights minimize the variance of prediction error, under the unbiasedness constraint, i.e.,

$$\mathbb{E}\left[\left\|\boldsymbol{\chi}_{s_0} \ominus \bigoplus_{i=1}^n \lambda_i \odot \boldsymbol{\chi}_{s_i}\right\|^2\right] \quad \text{subject to} \quad \mathbb{E}\left[\bigoplus_{i=1}^n \lambda_i \odot \boldsymbol{\chi}_{s_i}\right] = \mathbb{E}[\boldsymbol{\chi}_{s_0}]. \qquad (1.4)$$

Similar to the general case discussed in Chapter 2 (under mild assumptions on the sampling design), the optimal kriging weights are found by solving a linear system

$$\begin{pmatrix} \Sigma & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \vec{\lambda} \\ \zeta \end{pmatrix} = \begin{pmatrix} \vec{\sigma}_0 \\ 1 \end{pmatrix}. \tag{1.5}$$

Here, $\Sigma \in \mathbb{R}^{n \times n}$ denotes the variance-covariance matrix of the observations, $\Sigma_{i,j} = C(\|s_i - s_j\|_d)$ for $i, j = 1, ..., n$, $\vec{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$ the vector of weights and ζ a Lagrange multiplier, and $\vec{\sigma}_0 = (C(\|s_1 - s_0\|_d), ..., C(\|s_n - s_0\|_d))^T$ the vector of (trace-) covariances between observations and the random element at the target location.

Whenever the spatial dependence structure is unknown, the trace-covariogram, or the trace-variogram, can be estimated from the data by embedding the general procedure detailed in Chapter 2 in the Bayes Hilbert setting. In particular, the empirical estimator of the trace-semivariogram takes the form of

$$2\widehat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} \|\boldsymbol{\chi}_{s_i} \ominus \boldsymbol{\chi}_{s_j}\|^2, \tag{1.6}$$

where $N(h) = \{(i,j) \,|\, h - \Delta h \leq \|s_i - s_j\|_d \leq h + \Delta h\}$, and $|N(h)|$ is its cardinality.

Although expression (1.1) could be directly used to estimate (1.6), it involves double integrals, which might pose challenges for their accurate numerical evaluation. For the sake of efficiency, one may perform the computations on a transformed dataset, built upon mapping each data-object from $\mathcal{B}^2(I)$ to $L^2(I)$ through the centred log-ratio transformation 1.2. The latter allows expressing operations and inner products in $\mathcal{B}^2$ as operations and inner products in L^2, which markedly simplifies the calculations. We refer the reader to [7] for additional details.

1.4.2 Data pre-processing

Data pre-processing, or data smoothing, is a first key step of almost any (geo)statistical analysis of functional or object data. Even as a wide body of literature has been devoted to smoothing data in L^2, still limited attention has been given to the problem of smoothing FCs. Since PDFs can be interpreted as instances of FCs, all methods apt to smooth PDFs or cumulative distribution functions (CFDs) can be adopted to deal with a range of FCs as well. This approach was considered in [7], where an extension of a smoothing method based on Bernstein polynomials [29] was proposed to deal with the PSDs described in Section 1.3. We briefly review the method, which serves as a basis to smooth the data described in Section 1.3.

Consider the problem of obtaining from raw data a smooth estimate of the j-th curve, $\boldsymbol{\chi}_{s_j}$, that represents the PDF at location s_j ($j = 1, ..., n$). We first note that the underlying distribution can be equivalently represented by the PDF $\boldsymbol{\chi}_{s_j}$ (our target), or by the corresponding CDF $\mathcal{Y}_{s_j}(t) = \int_a^t \boldsymbol{\chi}_{s_j}(\tau) d\tau$. As such, one can

perform the smoothing either on $\boldsymbol{\chi}_{s_j}$ or through the CDF. Bernstein polynomials are here used to provide a smooth estimate of the CDF, a key advantage with respect to other approaches being that these allow to explicitly obtain a smooth estimate also of the PDF.

For convenience of notation, we assume here that $\boldsymbol{\chi}_{s_j}$ is supported on the compact domain $[0,1]$, for $j = 1, ..., n$; the case of a general compact support $[a, b]$ can be obtained through the variable transformation $x = \frac{(t-a)}{(b-a)}$, with $t \in [a, b]$. Recall that, given a sample $\vec{x}_j = (x_{1j}, ..., x_{N_j j})$ of i.i.d. observations from (the distribution whose PDF is) $\boldsymbol{\chi}_{s_j}$, a (discontinuous) non-parametric estimator for the CDF $\mathcal{Y}_{s_j}$ is given by the Empirical Cumulative Distribution Function (ECDF), denoted by $\overline{\mathcal{Y}}_{s_j}(t; N_j)$ and defined as

$$\overline{\mathcal{Y}}_{s_j}(t; N_j) = \frac{1}{N_j} \sum_{i=1}^{N_j} I_{x_{ij}<t}. \tag{1.7}$$

Equation (1.7) renders estimates with jump discontinuities in correspondence of the data. Bernstein Polynomials can then be introduced to obtain a smooth estimate of $\mathcal{Y}_{s_j}$ from $\overline{\mathcal{Y}}_{s_j}(t; N_j)$. In [29], the following estimator was proposed

$$\widehat{\mathcal{Y}}_{s_j}(t; N_j, B_j) = \sum_{k=0}^{B_j} \overline{\mathcal{Y}}_{s_j}(k/B_j; N_j) b_{k,B_j}(t), \tag{1.8}$$

where $b_{k,B_j}(t) = B_j k t^k (1-t)^{B_j - k}$, $k = 0, ..., B_j$, and B_j denotes the number of Bernstein polynomials used to smooth the j-th ECDF. Estimators (1.7) and (1.8) are strongly consistent for $\mathcal{Y}_{s_j}$, but the latter is also continuous, and allows obtaining a smooth estimate of the PDF $\boldsymbol{\chi}_{s_j}$ as

$$\widetilde{\boldsymbol{\chi}}_{s_j}(t; N_j, B_j) = B_j \sum_{k=0}^{B_j - 1} \left(\overline{\mathcal{Y}}_{s_j}((k+1)/B_j; N_j) - \overline{\mathcal{Y}}_{s_j}(k/B_j; N_j) \right) b_{k,B_j-1}(t). \tag{1.9}$$

Unlike the well-known kernel smoothing estimators, estimator (1.9) is suitable to be adopted for compactly supported PDFs. It was adapted to smooth PSCs collected through grain sieve analysis, by considering a modified yet consistent estimator, based on a pre-processing of partially observed ECDF. We note however that other smoothing methods based on Bernstein polynomials have been developed for the same purpose, e.g., [30, 31, 32].

A different approach to smooth FCs was proposed in [33], by combining the approaches of FDA and CoDa. These authors developed a B-spline representation for the clr-transformation of an FC, estimated from a discrete clr-transformation applied to the histogram of raw data. This idea is closer to the typical viewpoint employed in the main literature on FDA [1]. Extending FDA methods to the Bayes space setting is often non-trivial. For instance, in the case addressed in [33] the B-spline representation had to imbue through appropriate conditions the zero-integral constraints characterizing clr-transformations. Basis expansions are however very useful from the

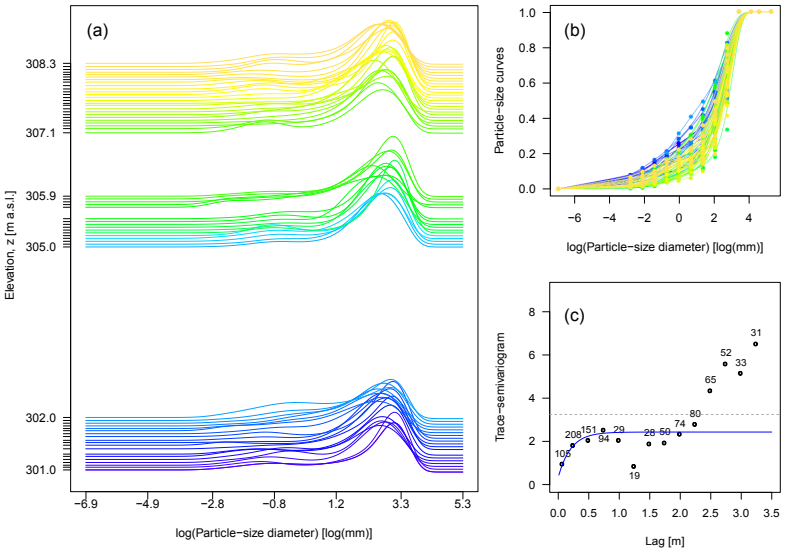

Figure 1.3 (a) Vertical distribution of smoothed densities; (b) raw particle-size curves (symbols) and particle-size curves smoothed by Bernstein Polynomials with $m = 140$ (solid curves); (c) estimated trace-semivariogram of the particle-size densities: empirical trace-semivariogram (symbols), fitted model (solid curve) and sample variance (dotted curve); the number of pairs associated with each lag is reported. Modified from [7].

computational viewpoint: the B-spline representation of [33] was used to markedly simplify computations in [14, 34].

In general, most geostatistical methods for FCs developed in the literature are based on the assumption that the data have been already smoothed. As such, the smoothing procedure is seen as a separate step of the analysis, for which the technique of choice – possibly data driven – can be applied.

1.4.3 An example of application

As an illustration of the approach, we consider here the analysis of the dataset of PSCs illustrated in Section 1.3. Here, we focus on the data observed at borehole B5, as in [7].

Menafoglio et al [7] preprocessed the raw data described in Section 1.3 by smoothing the PSCs through the use of 140 Bernstein polynomials. The density functions of the PSCs were then explicitly computed from the smoothed PSCs (see Section 1.4.2). The latter densities, hereafter called *particle-size densities* (PSDs), were interpreted as functional compositions, and embedded in the Bayes space $\mathcal{B}^2(I)$. In [7], the interval I was set to $I = [\log(0.001), \log(200)]$, considered as the largest range of

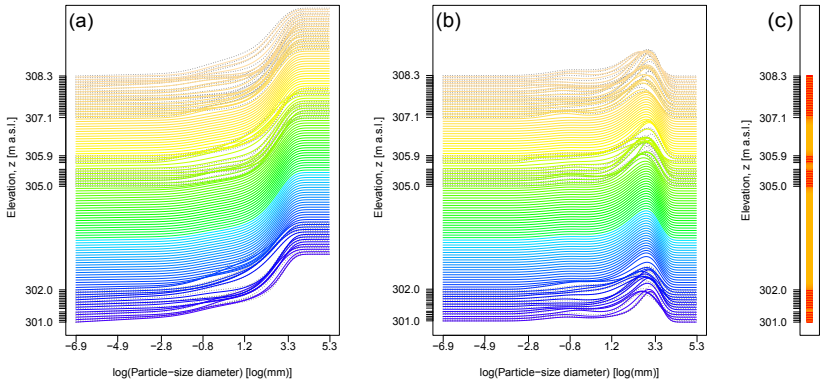

Figure 1.4 Vertical distribution of ordinary kriging predictions results: (a) PSCs: kriged curves (solid curves) and smoothed (dotted curves); (b) PSDs: kriged curves (solid curves) and smoothed (dotted curves); (c) kriging variance, ranging between 0 (darkest shade, corresponding to vertical locations where observations are available) and 2.53 (lightest shade). Modified from [7].

observation consistent with the type of lithology at the site. Other choices are possible: for instance, one may consider the distribution of grain-sizes conditional to the range of observation, as proposed in [8] and discussed in Section 1.5.

In [7], a stationary spatial model was considered for the data. Note that in this case the spatial domain is one-dimensional, as the data at borehole B5 were observed along the vertical coordinate in the range $D = [301.0, 308.3]$ meters above the sea level (m a.s.l.).

Figure 1.3 depicts the smoothed data at borehole B5, together with the empirical estimate of the variogram, estimated through (1.6). The blue curve in Figure 1.3c denotes the exponential structure with nugget which were fitted to the empirical estimate. Note that, although the empirical variogram might show some degree of non-stationarity, prior knowledge on the field site supports adopting a stationary hypothesis at B5, and was thus considered as a basis assumption of the study. From the application viewpoint, the estimated variogram displays a rapid growth up to a lag of about 0.6 m, where it reaches a sill around a value of 2.4. As such, the range of spatial dependence appears quite small if compared with the width of D (7.3 m). This has a direct impact on predictions, as the ordinary kriging sets the predictions to the (GLS) estimated mean when the target location is at a distance higher than the variogram range (in this case 0.6 m) from the closest observed site.

Figure 1.4 shows the results of the ordinary kriging in $\mathcal{B}^2(I)$, for a fine grid of target locations along the vertical direction. One can clearly notice that, consistent with our previous remark, the two widest gaps between the sample locations (i.e., the ranges [302.0,305.0] and [305.9,307.1]) are mostly predicted with the mean PSD. It is noted that, in cases of such short ranges, the experimental design, i.e., the distribution of the sampling points within the domain (here the vertical dimension), is

key to the performance of our predictions. As such, a rigorous assessment of the extent at which the collection of additional information about the system can (a) reduce predictive uncertainty and (b) yield potential benefits in terms of, e.g., reduced sampling cost and/or risk reduction, is key to improve our understanding of complex natural systems such as groundwater reservoirs. The value of additional information can be quantified through a variety of approaches (see, e.g., Neuman et al. [2012] and references therein). An example of these – which is relevant to our application – is the multimodel data worth assessment framework proposed by Neuman et al. [2012] and Xue et al. [2014] and references therein. The approach is based on a Maximum Likelihood version of the Bayesian Model Averaging (MLBMA) and is consistent with modern statistical methods of parameter estimation. Implementations of MLBMA data-worth assessments considered the geostatistical characterization of aquifer hydraulic conductivity fields in the presence of multiple variogram models (and eventually measured values) [Neuman et al., 2012; and Lu et al., 2012].

Dealing with functional data with an approach of the kind we illustrate here is of interest, for example, in the context of the hydrogeological characterization of heterogeneity of aquifers and reservoirs. PSCs are routinely assessed from soil samples in modern laboratories through simple and inexpensive procedures. These typically involve the successive use of a series of sieves of decreasing grid size, which are regulated by appropriate international standards. A variety of other methods are also available to extract PSCs from soil samples, including sedigraph; laser diffraction; dry and wet sieving. The PSCs enable one to characterize a number of effective grain diameters, d_e, defined as the representative particle size diameter in terms of percent in mass, corresponding to the e-th percentile of a measured PSC. Having the ability to treat the whole PSC in a consistent geostatistical framework enables us to transfer information not only on hydraulic, but also on sedimentological and eventually geochemical parameters which can control solute fluxes in the subsurface.

1.4.4 Uncertainty assessment

A kriging prediction is optimal in terms of mean squared error within the class of linear unbiased predictors. However, it does not always represent the natural variability of the process: the field realization is usually much 'rougher' than a typical kriging map. Quantifying the uncertainty associated with predictions is then key to provide a full characterization of the phenomenon. For this purpose, one may employ the kriging variance, that is the variance of prediction error explicitly expressed at a target location s_0 as

$$\sigma^2_*(s_0) = C(0) - \sum_{i=1}^{n} \lambda_i^* C(\|s_i - s_0\|_d) - \zeta^*, \tag{1.10}$$

where (λ^*, ζ^*) are the solutions of the kriging system (1.5). Indeed, on these bases one can provide Chebyshev bands on the norm of the prediction errors by using the

following inequality

$$P\left(\|\boldsymbol{\chi}_{s_0} \ominus \boldsymbol{\chi}^*_{\boldsymbol{s}_0}\| > \kappa \cdot \sigma_*(s_0)\right) < \frac{1}{\kappa^2}. \tag{1.11}$$

Even though expression (1.11) provides a useful bound on the prediction error, it often proves to be very conservative, as shown in [7]. Indeed, in the study presented in [7] and recalled in Section 1.4.3, the authors estimated via cross-validation that the 75% prediction bands constructed through the Chebyshev inequality (1.11) were associated with an empirical level of 98.3 %.

We also remark that the kriging variance does not take into account the uncertainty associated with the estimate of the cross-variogram, as the latter is assumed to be known when formally developing the kriging predictor (see Chapter 2). Hence, prediction bands built on these bases inevitably suffer from being approximate.

Another perspective in assessing the uncertainty of the estimate is that of generating multiple realizations of the field, compatible with the data. This approach was recently pursued in [9], that proposed a methodology for geostatistical simulation in Bayes spaces. The idea upon which the method is grounded is to reproduce the variability of the phenomenon – which is only partially represented by kriging maps – by drawing samples from the conditional distribution of $\boldsymbol{\chi}_{s_0}$ given $\boldsymbol{\chi}_{s_1}, ..., \boldsymbol{\chi}_{s_n}$. Accordingly, if the procedure is performed for multiple target locations in D, one can obtain a set of maps that, although suboptimal, provide an improved representation of the natural variability, and are still 'compatible' with the data, in the sense that they coincide with the data at the measurement locations (as well as kriging maps do).

Before briefly describing the method, we illustrate the results on the field data of Sections 1.3, 1.4.3. Figure 1.5b displays an example of a realization from the conditional field $\{\boldsymbol{\chi}_s|\vec{\boldsymbol{\chi}}, s \in D\}$, with $\vec{\boldsymbol{\chi}} = (\boldsymbol{\chi}_{s_1}, ..., \boldsymbol{\chi}_{s_n})^T$. It is apparent that the spatial variability associated with the realization is much higher than that of the kriged field, displayed in Figure 1.5a.

Performing repeated conditional simulations leads to generate a wide range of scenarios that could have been observed with the same data. As a way of example, Figure 1.5c-d depicts a sample of 1000 conditional simulations at elevations 303.0 and 306.0 m a.s.l, the corresponding prediction being depicted as black curves in Figure 1.5a. The amplitude of the grey shade can be used to qualitatively represent the variability of the predictions at the target location. Note that, although the predicted curves at elevations 303.0 and 306.0 m a.s.l show some similarities, the associated uncertainty is indeed different: at an elevation 303.0 m a.s.l. the variability is much higher, due to the absence of data nearby that location.

From a theoretical viewpoint, generating realizations from the distribution of $\boldsymbol{\chi}_{s_0}|\vec{\boldsymbol{\chi}}$ is a problem of random generation in infinite dimension. It is clear that a strategy based on joint simulations of point-wise values of the curves would not be affordable either from the theoretical or the computational viewpoint. It is also noticed that a global approach as that used for ordinary kriging did not prove to be successful, as the trace-covariogram seems to be insufficient for the characterization of the spatial dependence structure for simulation purposes.

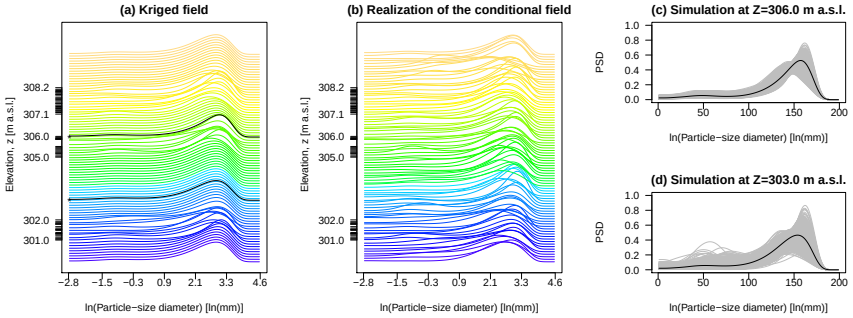

Figure 1.5 Kriged field and conditional realizations. (a) Kriging estimation over a grid along the vertical direction; black curves indicate predictions at elevations 303.0 and 306.0 m a.s.l.. (b) A conditional realization on the same grid considered in panel (a). (c)-(d): Kriging estimation at elevations 303.0 and 306.0 m a.s.l. (black curve) and a sample of 1000 conditional simulations at the same sites (grey curves)

In [9] the authors proposed a simulation strategy based on an optimal dimensionality reduction of the problem in the Bayes Hilbert space. Specifically, to provide a conditional realization at a target location s_0 they proposed to

(i) perform a functional principal component in $\mathcal{B}^2(I)$ [14] and compute the scores along the first K principal components (where K is sufficiently high to represent the data variability);
(ii) model the spatial dependence of the multivariate random field of the scores and perform geostatistical simulation of the latter, through any of the widely employed geostatistical techniques for the simulation of multivariate random fields.

The reason that led the authors to choose a dimensionality reduction step based on principal component analysis is that it provides nested optimal approximations of the observations for any finite order K. The optimal choice for K is critical, because it controls the quality of the approximation of the data through the principal components, and the complexity of step (ii) (thus the computational effort involved in the actual computations). To set K, well-known methods in principal component analysis can be employed, e.g., looking for an elbow in the scree-plot, compatible with the computational power available.

Our methodology and type of results can be readily transferred to the general context of numerically-based Monte Carlo simulations of flow and transport processes in environmentally and industrially relevant scenarios, including, e.g., groundwater systems, oil reservoirs, and shale gas formations. A critical element in these applications is to have at our disposal multiple realizations of (a) the heterogeneous structure of the porous/fractured system (in terms of the spatial arrangements of geomaterials/hydrofacies), and (b) the distribution of properties such as porosity and hydraulic conductivity within each of the identified facies. This enables us to propagate uncertainty associated with the reconstruction of the subsurface onto uncertainties

characterizing target variables of interest such as local composition of soil, pressure heads, dissolved chemical concentrations, reaction rates, and fluid saturations. All of these elements will constitute avenues of future development and exploitation of the approach we present in this work.

1.5 Analyzing non-stationary fields of FCs

In several real cases, the field data cannot be consider either stationary or isotropic. For instance, one may have prior information about possible secondary variables which have an influence on the response. The latter need to be taken into account in the geo-statistical model and exploited for prediction purposes, e.g., in a Universal Kriging setting. A particular case in this broad context is the situation in which data are featured by a grouping structure. This case was addressed in [8], and was motivated by the analysis of the entire dataset of PSCs described in Section 1.3 for which the existence of different soil types was observed. In this case, the field was also found to be anisotropic. In this section, we recall the kriging method of [8] – termed *class-kriging* – for the prediction of anisotropic random fields of grouped FCs.

Throughout the section, we consider the setting in which the field of FCs $\{\boldsymbol{\chi}_s, s \in D\}$, is observed together with a secondary field $\{T_s, s \in D\}$, whose elements represent random labels associated with the grouping structure of the data. For instance, they may represent soil types, in case PSCs are observed in a heterogeneous system, but may also represent climatic regions, if weather data over a large region are concerned instead.

The random elements T_s, $s \in D$, are discrete variables. We call $\tau^{(1)}, ..., \tau^{(K)}$ the K values which may be taken by the T_s (i.e., the labels of the K possible groups), and denote by $(\chi_{s_1}, \tau^{(k_1)}), ..., (\chi_{s_n}, \tau^{(k_n)})$ the pairs of FCs and labels observed at the measurement locations $s_1, ..., s_n$. In [8], the authors proposed to model the field $\{\boldsymbol{\chi}_s, s \in D\}$, conditional to the field of labels $\{T_s, s \in D\}$ as the sum (in $\mathcal{B}^2$) of a drift term dependent on the label at s, and a stationary residual, independent of the grouping structure. Formally,

$$\boldsymbol{\chi}_s | \{T_s = \tau^{(k)}\} = m^{(k)} \oplus \boldsymbol{\delta}_s,$$

where $m^{(k)} = \mathbb{E}[\boldsymbol{\chi}_s | T_s = \tau^{(k)}]$ denotes the drift, and $\{\delta_s, s \in D\}$ is a random field of FCs, with 'zero-mean' in $\mathcal{B}^2$, i.e., with mean coinciding with the neutral element of perturbation $\mathbb{E}[\boldsymbol{\delta}_s] = 0_{\oplus} = 1/\eta$. The random field $\{\delta_s, s \in D\}$ is also assumed to be (i) independent of the field of labels $\{T_s, s \in D\}$, and (ii) globally second-order stationary (possibly anisotropic), with trace-covariogram C and trace-variogram γ:

$$\begin{aligned} C(s_1 - s_2) &= \mathbb{E}[\langle \boldsymbol{\delta}_{s_1}, \boldsymbol{\delta}_{s_2} \rangle], \\ 2\gamma(s_1 - s_2) &= \mathbb{E}\left[\|\boldsymbol{\delta}_{s_1} \ominus \boldsymbol{\delta}_{s_2}\|^2\right], \quad s_1, s_2 \in D. \end{aligned}$$

This model can be framed in the Universal Kriging setting introduced in Chapter 2. Indeed, denote by $\{\psi_k(\boldsymbol{s}), k = 1, ..., K-1\}$ a set of binary variable, which represent indicators associated with the labels: for $k = 1, ..., K-1$, $\psi_k(\boldsymbol{s}) = 1$ if $T_{\boldsymbol{s}} = \tau^{(k)}$, and $\psi_k(\boldsymbol{s}) = 0$ otherwise; if $T_{\boldsymbol{s}} = \tau^{(K)}$ then $\psi_k(\boldsymbol{s}) = 0$ for every $k = 1, ..., K-1$. The drift in $\boldsymbol{s} \in D$ can then be described through a linear model in $\mathcal{B}^2$, with these indicators as regressors

$$\mathbb{E}[\mathcal{Y}_{\boldsymbol{s}} | \, \Pi_{\boldsymbol{s}} = \boldsymbol{\pi}_{\boldsymbol{s}}, T_{\boldsymbol{s}} = \tau^{(k)}] = a_0 \oplus \bigoplus_{l=1}^{K-1} \psi_l(\boldsymbol{s}) \odot a_l, \tag{1.12}$$

where $a_0, ..., a_{K-1}$ are (possibly unknown) deterministic coefficients in $\mathcal{B}^2$. In the light of model (1.12), one has

$$\begin{cases} m^{(k)} = a_0 \oplus a_k, & k = 1, ..., K-1, \\ m^{(k)} = a_0, & k = K. \end{cases} \tag{1.13}$$

Coefficients $a_0, ..., a_{K-1}$ thus represent how different the drift in the k-th group is from that of a reference group, which is here set to the K-th group, without loss of generality.

In [8], the authors relied on the Universal Kriging results introduced in Chapter 2, to propose a class-kriging predictor for $\boldsymbol{\chi}_{s_0}$ at a target location s_0, given the realization of $\{T_s, s \in D\}$ in D (i.e., the grouping structure over the entire spatial domain). The class-kriging predictor is defined as $\boldsymbol{\chi}_{s_0} = \bigoplus_{i=1}^{n} \lambda_i^* \cdot \boldsymbol{\chi}_{s_i}$, whose weights minimize the (conditional) variance of prediction error under the unbiasedness constraint, that is, solve

$$\begin{aligned} \min_{\substack{\lambda_1,...,\lambda_n \in \mathbb{R}: \\ \boldsymbol{\chi}_{s_0}^{\vec{\lambda}} = \bigoplus_{i=1}^{n} \lambda_i \odot \boldsymbol{\chi}_{s_i}}} \quad & \mathbb{E}\left[\left\| \boldsymbol{\chi}_{s_0}^{\vec{\lambda}} \ominus \boldsymbol{\chi}_{s_0} \right\|^2 \middle| \, T_{s_0} = \tau^{(k_0)}, T_{s_i} \in \tau^{(k_i)}, i = 1, ..., n \right] \\ \text{subject to} \quad & \mathbb{E}\left[\boldsymbol{\chi}_{s_0}^{\vec{\lambda}} \middle| \, T_{s_0} = \tau^{(k_0)}, T_{s_i} \in \tau^{(k_i)}, i = 1, ..., n \right] = m^{(k_0)}. \end{aligned} \tag{1.14}$$

The drift being linear, the optimal weights are found by solving the system of $n + K$ linear equations, obtained by embedding in the Universal Kriging setting model (1.12)

$$\begin{pmatrix} \Sigma & \Psi \\ \Psi^T & 0 \end{pmatrix} \begin{pmatrix} \vec{\lambda} \\ \vec{\zeta} \end{pmatrix} = \begin{pmatrix} \vec{\sigma}_0 \\ \vec{\psi}_0 \end{pmatrix}, \tag{1.15}$$

where $\vec{\zeta} = (\zeta_0, ..., \zeta_{K-1})^T$ are K Lagrange multipliers associated with the unbiasedness constraint, whereas $\vec{\sigma}_0 = (C(\boldsymbol{h}_{i,0})) \in \mathbb{R}^n$, and $\vec{\psi}_0 = (\psi_k(\boldsymbol{s}_0)) \in \mathbb{R}^K$.

From the application viewpoint, several critical points may be encountered in class-kriging. Firstly, the estimate of the spatial dependence structure is crucial to solve (1.15). Here, all the methods described in Chapter 2 can be employed. In particular, one may resort to an iterative algorithm to estimate the drift via generalized least

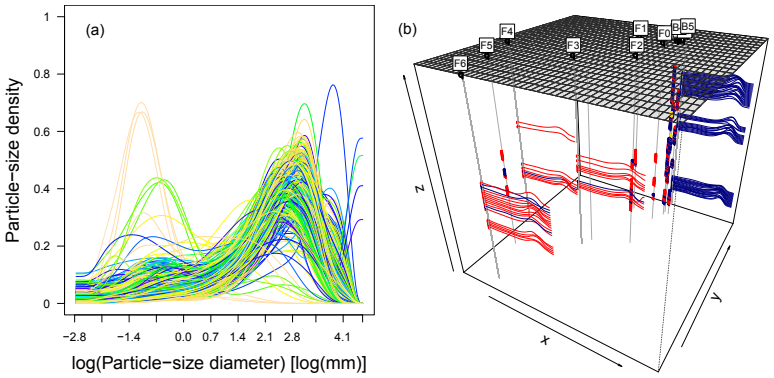

Figure 1.6 Field data: (a) smoothed PSDs; (c) soil types at the field site (denoted with colors) and smoothed PSDs along the boreholes B5, F3, F4 and F6. Modified from [8].

squares and jointly estimate the residual variogram 2γ. More delicate is the case in which the field $\{T_s, s \in D\}$ is only observed at the measurement locations, or if it is completely latent. In [8], methods to deal with both the situations were developed. For the first case (i.e., the labels are only observed at $s_1, ..., s_n$), one needs to formulate a model for the stochastic distribution of $\{T_s, s \in D\}$, and then to employ such a model to predict the T_s at unsampled locations. For instance, the T_s may be modeled as independent realizations from a multinomial variable, and the interpolation can be consistently performed via indicator kriging, as in [8]. When the field is completely latent, one needs additionally to cluster the data. Although several methods are available for spatial clustering of scalar data, little attention has been paid so far to the problem of spatial clustering of FCs. In [8], the authors proposed a spatial K-mean clustering, which is an extension of the K-mean method, tailored on model (1.12). Other methods could be applied to this purpose, for instance the Bagging Voronoi method illustrated in Chapter YYY (Vitelli et al.).

As an illustration of the class-kriging method, we illustrate the results of its application to the dataset described in Section 1.3, following [8]. Unlike the case discussed in Section 1.4.3, the authors focused on these restriction of the PSDs to the actual domain of observation, because for most data no information was available on the left tail of the PSC, due to the sieve measurement procedure. Figure 1.6 displays the full dataset of smoothed PSDs at the site. The colors of the symbols in Figure 1.6c denote the three soil types identified in the study region, associated with as many groups in the data.

A geometric anisotropy was found by the authors when looking at directional variograms. This was corrected by scaling the vertical dimension by a factor $r = 25$, thus working in the modified spatial domain where an isotropic model can be used. The estimated omnidirectional trace-semivariogram of the residuals is displayed in Figure 1.7a, together with the fit of an exponential model with nugget. Similarly as in Section 1.4.3, the range of the variogram appears to be quite short when compared to the extension of the domain. The drift estimated within the groups is reported in

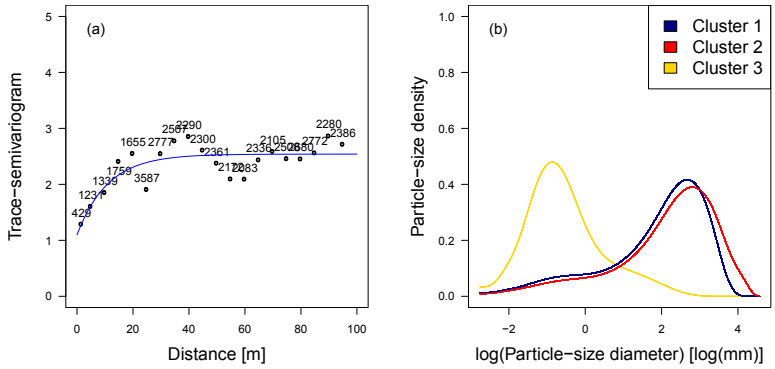

Figure 1.7 (a) Estimated trace-semivariogram of the residuals; (b) Estimated drift for the three soil types. Modified from [8].

Figure 1.7b. Here, the first two groups are interpreted as a characterization of two diverse behaviors within the right tail of the PSD, the first cluster featuring a lighter tail than the other one. The third group, formed by 1% of the sample, is associated with a drift displaying its main peak at a grain size of about 0.4 mm. As shown in Figure 1.6b, the first group is mainly associated with the boreholes B1-B5, and the second group with the boreholes F0-F6. The former group of boreholes is located in an area where the Neckar river displays a bend, and thus favors the accumulation of the finer sediments in this area. The PSDs at borehole B5 – considered in Section 1.4.3 – belongs to the first group, consistent with the stationarity assumption considered before.

Figure 1.8 finally displays the prediction of the field in some unsampled locations. The kriged field is a smooth interpolation of the available data. The outlying observations, such as the blue curve at $z = 305.5$ at borehole F6, influence prediction results only locally. For distances higher than the estimated range, the kriged field is representative of mean particle-size distribution associated with the soil type at the target location.

Uncertainty assessment of such predictions is non-trivial, as it should take into account the prediction variance as well as the uncertainty in parameters estimate (variogram and drift). For this purpose, an extension of the simulation methods discussed in Section 1.4.4 can be considered.

1.6 Conclusions and perspectives

We here illustrated methods for the geostatistical analysis of functional compositions, which combine the perspective of Object Oriented Spatial Statistics (O2S2, [35]), with that of Compositional Data Analysis in Bayes spaces. The main points addressed in this Chapter can be summarized as follows:

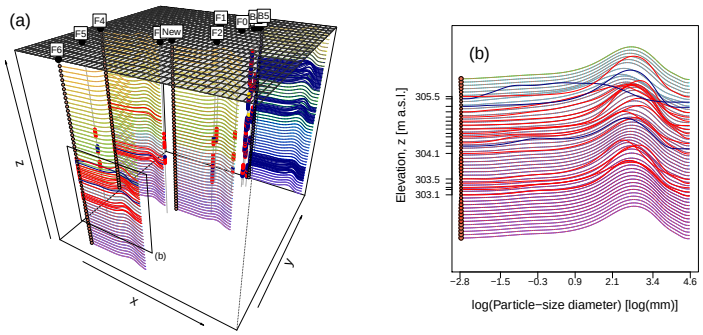

Figure 1.8 Class-kriging of PSDs: (a) results at boreholes B5, F4 and F6 and at an unsampled vertical (denoted as "New"). (b) Vertical distribution of predicted PSDs, for the group of samples at elevations $301 \leq z \leq 306$ m above sea level (a.s.l.), at borehole F6. In both panels: colors of the solid curves indicate depth; colors of the symbols indicate the soil type. Smoothed data are represented with solid curves colored according to the cluster assignment. Modified from [8].

1) Functional Compositions (FCs), such as probability density functions, should not be considered just as data in the space L^2, but one should pay close attention to treat them through an appropriate geometry. A possible sensible geometry is that of Bayes Hilbert spaces. Although we focused on FCs with compact support, the theory of Bayes spaces is available for FCs with support in non-compact set, provided that a reference measures other than the Lebesgue one is considered.
2) Stationary and non-stationary methods are available to predict FCs in Bayes spaces – and particularly PDFs – by relying on the theory of Universal Kriging in Hilbert spaces of [36]. Unlike traditional methods based on selected features of distributional data (e.g., moments or quantiles), predicting the entire PDF allows taking into account the entire information content of the data, and project it to unsampled location in the system.
3) Uncertainty assessment is key for a full characterization of the field under study. Here, we illustrated a method for stochastic simulation grounded upon dimensionality reduction in Bayes spaces. Although only the stationary case was considered, the extension of the strategy to the non-stationary setting can be readily envisaged.
4) Throughout the chapter, we illustrated the methodologies with a real case study, dealing with PSDs. Advancements of the work illustrated here include embedding our theoretical and operational framework in the context of (forward and inverse) stochastic analyses of subsurface flow and transport in heterogeneous media by way of numerical Monte Carlo simulations or groundwater flow and transport Moment Equations (e.g., [37, 38, 39, 40] and references therein).
5) Future perspectives for application of the approach include the analysis of environmental and Earth system variables whose main characteristics can be encapsulated in terms of a functional behavior. In addition to particle-size curves of the kind we examine here, these might comprise, for example, relative permeability curves (which are relevant in multi-phase flow settings), mineralogic composition of rocks (for the geochemical characterization of a host reservoir), as well as breakthrough curves of dissolved chemical migrating in water bodies and/or chemical composition of fluids sampled at multiple locations in an aquifer system (with implication on human exposure and health hazards).

References

1 Ramsay, J. and Silverman, B.W. (2002) *Applied Functional Data Analysis: Methods and Case Studies*, Springer-Verlag, New York.

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

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

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

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

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

7 Menafoglio, A., Guadagnini, A., and Secchi, P. (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.

8 Menafoglio, A., Secchi, P., and Guadagnini, A. (2016) A Class-Kriging Predictor for Functional Compositions with Application to Particle-Size Curves in Heterogeneous Aquifers. *Mathematical Geosciences*, **48**, 463–485.

9 Menafoglio, A., Guadagnini, A., and Secchi, P. (5708–5726) Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a bayes space approach. *Water Resources Research*, **52** (8), 5708—-5726.

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

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

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

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

14 Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., and Filzmoser, P. (2016) Simplicial principal component analysis for density functions in bayes spaces. *Computational Statistics & Data Analysis*, **94**, 330 – 350.

15 Riva, M., Sánchez-Vila, X., Guadagnini, A., Simoni, M.D., and Willmann, M. (2006) Travel time and trajectory moments of conservative solutes in two-dimensional convergent flows. *J. Cont. Hydrol.*, **82**, 23–43.

16 Riva, M., Guadagnini, A., Fernández-Garcáa, D., Sánchez-Vila, X., and Ptak, T. (2008) Relative importance of

geostatistical and transport models in describing heavily tailed breakthrough curves at the lauswiesen site. *Journal of Contaminant Hydrology*, **101**, 1–13.

17 Riva, M., Guadagnini, L., and Guadagnini, A. (2010) 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.

18 Riva, M., Sanchez-Vila, X., and Guadagnini, A. (2014) Estimation of spatial covariance of log-conductivity from particle-size data. *Water Resources Research*. In press.

19 Hoffmann, R. and Dietrich, P. (2004) An approach to determine equivalent solutions to the geoelectrical 2d inversion problem. *Journal of Applied Geophysics*, **56** (2), 79–91.

20 Rein, A., R.H. and Dietrich, P. (2004) Influence of natural time - dependent variations of electrical conductivity on dc resistivity measurements. *Journal of Hydrology*, **285** (1-4), 215–232.

21 Neuman, S.P., Blattstein, A., Riva, M., Tartakovsky, D.M., Guadagnini, A., and Ptak, T. (2007) Type curve interpretation of late-time pumping test data in randomly heterogeneous aquifers. *Water Resources Research*, **43** (10), W10 421.

22 Neuman, S.P., Riva, M., and Guadagnini, A. (2008) On the geostatistical characterization of hierarchical media. *Water Resources Research*, **44** (2), W02 403.

23 Lessoff, S.C., Schneidewind, U., Leven, C., Blum, P., Dietrich, P., and Dagan, G. (2010) Spatial characterization of the hydraulic conductivity using direct-push injection logging. *Water Resour. Res.*, **46**, W12 502. Doi:10.1029/2009WR008949.

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

25 Händel, F. and Dietrich, P. (2012) Relevance of deterministic structures for modeling of transport: The lauswiesen case study. *Groundwater*, **50**, 935–942. Doi:10.1111/j.1745-6584.2012.00948.x.

26 Sack-Kühner, B. (1996) Einrichtung des naturmessfeldes "lauswiesen tübingen", erkundung der hydraulischen eigenschaften, charakterisierung der untergrundheterogenität und vergleich der ergebnisse unterschiedlicher erkundungsverfahren. M.Sc. Thesis, University of Tübingen, Geological Institute.

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

28 Beyer, W. (1964) Zur bestimmung der wasserdurchlässigkeit von kiesen und sanden aus der kornverteilungskurve. *Wasserwirtschaft-Wassertechnik (WWT)*, **14** (6), 165–168.

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

30 Manté, C. (1999) The use of regularization methods in computing Radon–Nikodým derivatives. Application to grain-size distributions. *SIAM Journal on Scientific Computing*, **21** (2), 455–472.

31 Manté, C. (2012) Application of iterated Bernstein operators to distribution function and density approximation. *Applied Mathematics and Computation*, **218**, 9156–9168.

32 Manté, C. (2015) Iterated Bernstein operators for distribution function and density estimation: Balancing between the number of iterations and the polynomial degree. *Computational Statistics & Data Analysis*, **84**, 68–84.

33 Machalová, J., Hron, K., and Monti, G.S. (2016) Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics*, **43** (8), 1419–1435.

34 Talska, R., Menafoglio, A., Machalova, J.,

Hron, K., and Fiserova, E. (2017) Compositional regression with functional response, *Mox report 27/2017*, Politecnico di Milano.

35 Menafoglio, A. and Secchi, P. (2017) Statistical analysis of complex and spatially dependent data: A review of object oriented spatial statistics. *European Journal of Operational Research*, **258** (2), 401 – 410.

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

37 Guadagnini, A. and Neuman, S. (2001) Recursive conditional moment equations for advective transport in randomly heterogeneous velocity fields. *Transport in Porous Media*, **42** (1/2), 37–67.

38 Morales Casique, E., Neuman, S., and Guadagnini, A. (2006) Nonlocal and localized analyses of nonreactive solute transport in bounded randomly heterogeneous porous media: Computational analysis. *Adv. Water Resour.*, **29**, 1399–1418.

39 Xue, L., Zhang, D., Guadagnini, A., and Neuman, S. (2014) Multimodel bayesian analysis of groundwater data worth. *Water Resour. Res.*, **50**, 8481–8496.

40 Panzeri, M., Riva, M., Guadagnini, A., and Neuman, S. (2015) Enkf coupled with groundwater flow moment equations applied to lauswiesen aquifer, germany. *J. Hydrology*, **521**, 205–216.