Object oriented spatial analysis of natural concentration levels of chemical species in regional-scale aquifers

Alessandra Menafoglio[a], Laura Guadagnini[b], Alberto Guadagnini[c], Piercesare Secchi[a,d]

[a] *MOX, Department of Mathematics, Politecnico di Milano*
[b] *Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya*
[c] *Department of Civil and Environmental Engineering, Politecnico di Milano*
[d] *Center for Analysis, Decisions and Society, Human Technopole*

Abstract

We address the problem of characterizing spatially variable Natural Background Levels (NBLs) of concentrations of chemical species of environmental concern in a large-scale groundwater body. Assessment of NBLs is critical to identify significant trends of (possibly hazardous) chemical concentrations in aquifer systems, the latter being typically associated with spatially heterogeneous hydrogeochemical characteristics. Our study considers the entire probability density function (PDF) of the concentration of the chemical species of interest as atom of the statistical analysis. These PDFs are estimated across a network of observation boreholes in the investigated spatial domain, and modeled as random points in a Bayes Hilbert space, in the context of Object Oriented Data Analysis. This approach enables one to take advantage of the entire information content provided by these objects for the purpose of spatial prediction and uncertainty quantification. As a key element of innovation, we investigate the use of depth measures for distributional data with the distinctive aims of (*i*) detecting central and outlying NBL distributions in the dataset, and (*ii*) building prediction regions for NBL distribution at unsampled locations. We illustrate the results of the proposed approach to the analysis of NBLs of a selected chemical species detected at an environmental monitoring network within a large-scale alluvial aquifer in Northern Italy.

Keywords: Kriging, Uncertainty quantification, Bayes Hilbert space, Spatial depth measure

1. Introduction

Natural background levels (NBLs) of a chemical species in a groundwater body are concentrations that can be considered as unaffected by anthropogenic actions (Directive 2006/118/EC GWDD, 2006). An appropriate assessment of these concentrations is key to identify trends of contaminations and/or plan activities to manage unfavorable trends which might be detected. In this context, characterizing the natural chemical signature of a groundwater body is set as a priority in the EU Water Framework Directive (Directive 2000/60/EC WFD, 2000, art. 17) and in the Ground Water Daughter Directive (2014/80/EU, 2014).

NBL estimates are generally grounded on analyses of monitored concentration samples (Edmunds et al., 2003; Wendland et al., 2005; Panno et al., 2006; Walter, 2008; Urresti-Estala et al.,

2013; Kim et al., 2015). One can rely on the so-called Pre-Selection (PS) approach (Hinsby and de Melo, 2006; Coetsiers et al., 2009, and references therein), which consists of (*i*) identifying groundwater samples within an available set of observations that can be defined as pristine and representative of the population of the natural resident concentration, and (*ii*) inferring from these a unique (or bulk) NBL value, which is then taken to represent the environmental status of the examined subsurface reservoir. Concentrations exceeding such an NBL should be ascribed to anthropogenic sources. Even as Directive 2014/80/EC explicitly recognizes the importance of the spatial variability of groundwater chemistry, current regulatory frameworks are based on requirements according to which one needs to estimate only a single threshold value, the latter being then considered as uniform across the target water body and employed as a threshold against which anthropogenic contamination is assessed (Reimann and Garrett, 2005).

While a series of studies (e.g., Reimann and Garrett (2005); Hinsby and de Melo (2006); Edmunds et al. (2003); Wendland et al. (2005); Panno et al. (2006); European Commission (2009)) evidence that the qualitative status of a groundwater body may be defined through a range of concentrations of a given chemical species rather than a unique value, it is nowadays recognized that the assessment of NBLs in large-scale groundwater systems should explicitly account for possible spatial (or temporal) variability of NBLs due to local hydrogeochemical features of the target aquifer. This is also consistent with approaches treating the characterization of spatial arrangement of geomaterials constituting the internal architecture of aquifer systems within a probabilistic context (see, e.g., Winter et al. (2003); Short et al. (2010); Perulero Serrano et al. (2014); Bianchi Janetti et al. (2021) and references therein). As such, modern approaches to the management and assessment of the quality of groundwater resources should take full advantage of refined probabilistic approaches to properly characterize the spatial distribution of NBLs. This is, for instance, the case of the exemplary application we consider in our study, where we focus on a large-scale aquifer body located in the Po Basin fill in Italy and encompassing an area of approximately 200 km^2 (see Figure 1 and Section 5).

Although recent works have proposed methods to characterize the spatial variability of NBL concentrations (Molinari et al., 2012; Ducci et al., 2016; Libera et al., 2017; Molinari et al., 2019), none of these yield a full spatial characterization of NBL distributions. For instance, Ducci et al. (2016) apply indicator kriging to delineate regions with the same probability of exceeding a pre-defined NBL value. Libera et al. (2017) rely on geostatistical tools (i.e., kriging) to provide a zoned map of (piece-wise constant) NBL concentrations. The latter are set to the 90th percentile of predicted NBL values which are then reclassified according to three given concentration ranges. Molinari et al. (2019) provide spatially variable estimates of NBLs through geostatistical analysis of the 90th percentile of sample NBL distributions detected at a collection of observation boreholes. The resulting kriged values and associated variance are then used to assess local probabilities of exceedance of certain threshold concentrations assuming a log-normal distribution for NBL values. In all of these cases, measured natural background concentrations are first represented through scalar summaries (i.e., indicators, probability of exceedance thresholds or quantiles), these summaries being then projected onto unsampled locations in the system. This inevitably yields loss of information and requires resorting to parametric assumptions to obtain further summaries in addition to those considered for the spatial analysis (as in, e.g., Molinari et al. (2019)).

Following our preliminary study presented in Guadagnini et al. (2020), we here rest on the

paradigm of Object Oriented Data Analysis (Marron and Alonso, 2014) and consider the entire distribution function of the NBL concentrations as the object of the spatial analysis. Our approach enables us to provide (a) spatial predictions and uncertainty quantification of the entire distribution of NBL concentration in a fully non-parametric setting, and (b) consistent joint assessment of all summaries of interest (e.g., quantiles, probabilities). We embed NBL distributions in a space whose elements are probability density functions (i.e., a Bayes Hilbert space Egozcue et al. (2006); van den Boogaart et al. (2014)), and rely on the framework of Object Oriented Spatial Statistics (O2S2, Menafoglio and Secchi (2017)) to characterize the spatial dependence among observations, provide spatial prediction through kriging, and quantify the associated uncertainty via stochastic simulation.

As a key element of innovation, here we investigate the use of depth measures for distributional data with the distinctive aims of (i) detecting central and outlying NBL distributions in the dataset, and (ii) building prediction regions for NBL distribution at unsampled locations. Achieving this objective requires extending the approach of Sun and Genton (2009) for the construction of functional boxplots, to comply with the intrinsic constraints of density data. To this end, we rely on the notion of spatial depth measures, that are naturally prone to be used in any Hilbert (and Banach) space (Chakraborty and Chaudhuri, 2014).

The study is organized as follows. Section 2 describes the motivating case study and the available dataset; Section 3 illustrates the modeling approach and introduces the depth measure that is used to provide center-outward ordering of density data. Section 4 introduces the methods used for spatial prediction and uncertainty quantification, detailing the use of depth measures to demarcate prediction bands for NBL distributions at unsampled locations. Section 5 illustrates the results obtained in a field scale context, conclusions being presented in Section 6.

2. Motivating case study

Study area and available data. The area where our study is framed is part of the Po Basin fill, in Italy. Details about the hydrogeological setting of the region are available in Molinari et al. (2012); Farina et al. (2014). The information acquired from sedimentological and hydrogeological analyses supports the identification of three main hydrogeological complexes, i.e., Apennines alluvial fans, Apennine alluvial plain, and alluvial and deltaic Po plain. As an exemplary showcase for the application of our approach, we focus on the groundwater body termed 0630. The latter is located in the upper confined portion of the aquifer system of the Po Basin fill, has an average depth of 65 m, average thickness of 110 m and area of about 1995 km^2; Figure 1 depicts its limits and planar extent. Here, the area colored in red corresponds to the location of the entire Emilia Romagna region within Italy, the inset showing the complete set of groundwater bodies demarcated in the area, including the large scale aquifer (in cyan, denoted as 0630) which is the subject of our study. We consider this aquifer because it is characterized by a significant planar extent (with a representative scale of the order of hundreds of kilometers) and constitutes a stark example of a setting where the need to take into account regional-scale spatially varying NBL distributions for the assessment of groundwater quality is markedly evident.

We ground our analyses on temporal series of concentrations collected at monitoring stations where records of about 20 years of observations are available (collected at a six-month interval between 1987 and 2008, albeit not continuously for some wells). For the purpose of demonstrating

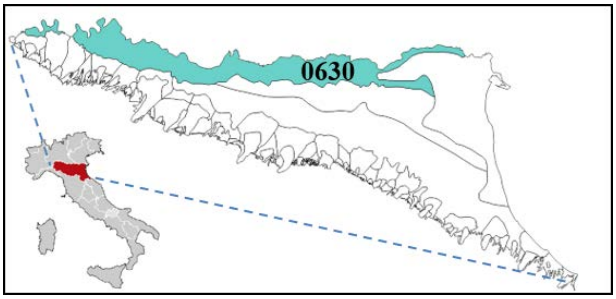

Figure 1: Planar extent of water body 0630.

our methodological workflow and approach, the chemical species considered in this study is ammonium (NH_4). The latter (see also Molinari et al. (2012), to which we refer for further details) is a key element for the assessment of the quality of the chemical status of a groundwater body, according to Italian and European Regulation (D. Lgs. 30/09, i.e., Decreto Legislativo n. 30, 16 March 2009; Directive 2006/118/EC GWDD (2006)) and it is documented to be associated with concentrations which can be markedly higher than reference values provided by Italian regulations.

NBL Estimation. Pre-Selection (PS) was applied to available raw data for the identification of NBL values. The approach is based on the analysis of concentration data collected across the network of monitoring boreholes to select samples that meet certain criteria and can thus be considered unaffected by anthropogenic influence. Criteria typically employed to classify a sample as affected by anthropogenic actions include: (a) chloride concentrations > 1000 mg/l, as indicator of salinity; (b) nitrates (NO_3) concentrations > 10 mg/l, as indicator of human influence caused by, e.g., fertilizers; and (c) ammonium (NH_4) concentrations > 0.5 mg/l, as indicator of human impact under reducing conditions. Additional criteria, such as redox conditions, dissolved oxygen, sulfate concentration, can be considered for sample exclusion (see, e.g., Hinsby and de Melo (2006); Hinsby et al. (2008)). Here, we apply the exclusion criteria described above upon disregarding NH_4 because sample cores collected within the study area provide evidence of natural occurrence of paleo-peats (Amorosi et al., 1996; Cremonini et al., 2008) consistent with large values of NH_4 concentrations in the investigated groundwater body (see also Molinari et al. (2012)). Samples characterized by markers of anthropogenic contamination (e.g., nitrates or pesticides) larger than a given value have been disregarded from the original data set. We recall that Molinari et al. (2012) applied the standard PS procedure to obtain a unique reference NBL value of 5.2 mg/l at the regional scale (i.e., to be considered as constant for the entire groundwater body), such an estimate exceeding the EU Drinking Water Standard for ammonium, which is set to 0.5 mg/l. This result evidences the global critical status with respect to ammonium for the groundwater body we consider. It also highlights the importance of a robust and clear delineation of spatial variability patterns of NBLs to assist decision making under uncertainty.

Data pre-processing. Our preparatory analyses are consistent with Guadagnini et al. (2020), who focus on a nearby aquifer body. In this context, the raw data after PS were considered as observations of naturally occurring (or NBL) concentrations at diverse time instants across the window 1987–2008. On average, 26.48 observations per location were available, with a range of 6 to 43 observations/locations – most locations including more than 15 temporal observations. Note that, while our methodological approach is envisioned to perform best within a context where sample size is relatively high, it can viably assist the analysis of environmental settings where the sample size is moderate, as in the case here considered.

A preliminary analysis of the data showed that most locations (58 out of 60) do not display any autocorrelation in the time series of NBL concentrations (level 1%, a result obtained through a Durbin-Watson test on each time series, the p-value of single tests being corrected via Holm's method), the only exceptions being locations PC80-00 and PR03-01 (marked with crosses in Figure 2). Analogous results are obtained when considering the data on a log-scale. Note that the lack of correlation may be due to the temporal lag taking place between observations (6 months), which might shadow possible correlations at shorter time scales. In this sense, having at our disposal more dense measurement may reveal stronger autocorrelations than those displayed by the currently available data. The estimation of the PDF of NBL concentration was thus performed at each borehole location and upon neglecting the temporal autocorrelation. PDF estimation at a given location is structured according to two steps, i.e., (*i*) histogram computation and (*ii*) representation of histograms through a compositional smoothing B-spline basis (Machalová et al., 2016). The latter method was selected because of its consistency with the modeling framework considered for this study, which is developed in Bayes spaces. We note that the strategy of analysis we propose is not substantially affected by the selection of any other method of smoothing.

Evaluation of histograms is performed by subdividing the common support of the log-transformed observations $\mathcal{T} = [-6, 3]$ into 25 equally spaced classes. The number of non-empty classes was found to range between 2 to 11 across the collection of spatial locations where data are available. To cope with the issue of classes with zero frequency, which poses challenges when using a compositional approach, we followed the Bayesian-multiplicative strategy suggested by (Machalová et al., 2016). The latter represents the standard choice in this literature stream, and is implemented in the R package `robCompositions` (Templ et al., 2011). Alternative strategies for zero-replacement could be used, including, e.g., the model-based approaches proposed by Martín-Fernández et al. (2015). Studying the robustness of the smoothing methods in Bayes spaces to the zero-replacement strategies is still an research challenge, which will be subject to future investigations. Note also that the use of a log-transformation for concentrations enables us to obtain a good reproduction of the PDFs associated with either very low or high concentration values, which are of major interest in environmental applications due to proper management and health implications, depending on the chemical species considered. This pre-processing choice also allows accounting for the relative scale of concentration values, which has been recognized as a key property of concentration data in the context of compositional data analysis (Pawlowsky-Glahn et al., 2015). We further note that it has been suggested that geochemical elements should not be considered singularly but otherwise relative to other elements, a feature which is typically referred to as the problem of single element mapping (McKinley et al., 2016). For the scope of this work – which is mainly devoted to the development of the workflow for the character-

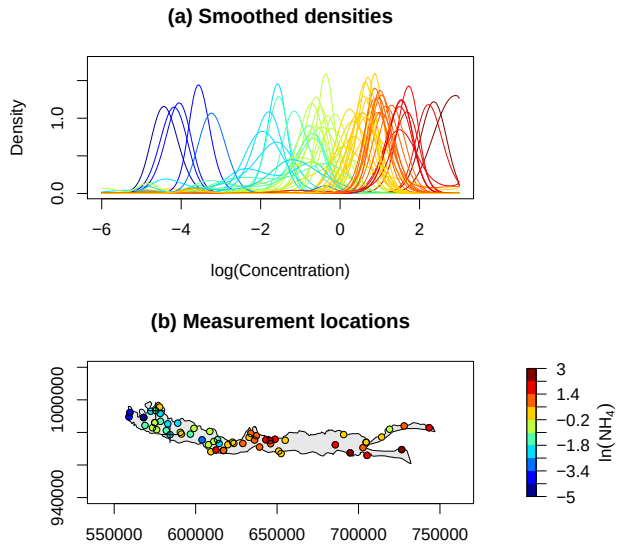

Figure 2: Smoothed data and corresponding spatial locations in the investigated domain. Colors are given according to the mean associated with the corresponding smoothed density. Crosses in panel (b) represent the only two locations characterized by a significant autocorrelation in the time series of NBL concentrations, which was not considered during data smoothing.

ization of NBL spatial variability – we consider the distribution of NBL log-concentrations of the single element NH_4, which is taken as a key environmental performance metric in EU-wide and national-level regulations. We emphasize that the methodology we illustrate could be extended to accommodate the observations of McKinley et al. (2016) upon considering the log-ratio between different available elements by, e.g., evaluating the PDFs of such a log-ratio.

Each pre-processed histogram was then smoothed through a cubic B-spline basis, with 9 equally spaced knots and smoothing parameter $\alpha = 10^3$ (the value of the latter is set on the basis of cross-validation). The smoothing procedure takes into account the occurrence of zero frequency classes by down-weighting the influence of these to one tenth of the weight of those with non-zero frequency. The smoothed data are depicted in Fig. 2a. In the following, we refer to the PDF of the NBL log-concentrations as *NBL densities* or *NBL PDFs* for simplicity.

3. Estimating the centrality of a set of PDFs through spatial depth measures

3.1. Problem setting

We denote by $s_1, ..., s_n$ the n sampling locations in the studied spatial domain D, and term $\mathcal{X}_{s_1}, ..., \mathcal{X}_{s_n}$ the NBL PDFs at these locations. In the following, $\mathcal{X}_{s_1}, ..., \mathcal{X}_{s_n}$ indicate random elements, whereas $x_{s_1}, ..., x_{s_n}$ denote the actual observed data (i.e., realizations). As in Guadagnini et al. (2020), we consider NBL PDFs $\mathcal{X}_{s_1}, ..., \mathcal{X}_{s_n}$ to belong to the Bayes Hilbert space $\mathcal{B}^2$ (Egozcue

et al., 2006; van den Boogaart et al., 2014), whose elements are positive functions defined on the compact set $\mathcal{T} \subset \mathbb{R}$ with square-integrable logarithm. In $\mathcal{B}^2$, two elements f and g are considered equivalent (i.e., $f =_{\mathcal{B}^2} g$ or $f = g$ for short) if they are proportional. The vector structure of $\mathcal{B}^2$ is defined by the following operations

$$(f + g)(t) =_{\mathcal{B}^2} f(t)g(t),\ t \in \mathcal{T}; \quad (\alpha \cdot f)(t) =_{\mathcal{B}^2} (f(t))^{\alpha},\ t \in \mathcal{T}; \tag{1}$$

for $f, g \in \mathcal{B}^2$, and $\alpha \in \mathbb{R}$. The difference operation induced by (1) reads $f - g = f + (-1) \cdot g$, for $f, g \in \mathcal{B}^2$. Note that studies focused on the analysis of distributional data in Bayes spaces often employ an alternative notation for the operations defined in (1) (i.e., the symbols $(\oplus, \odot)$, that are also used in compositional data analysis, Pawlowsky-Glahn et al. (2015)). In this work we rest on the simplified notation $(+, \cdot)$ to favor ease of reading. For clarity, all symbols $\cdot$ are explicitly stated, even when the standard notation would have dropped them.

Given f and g in $\mathcal{B}^2$, their inner product is defined as

$$\langle f, g \rangle = \frac{1}{2|\mathcal{T}|} \int_{\mathcal{T}} \int_{\mathcal{T}} \ln \frac{f(t)}{f(s)} \ln \frac{g(t)}{g(s)} dt ds. \tag{2}$$

The space $(\mathcal{B}^2, +, \cdot, \langle \cdot, \cdot \rangle)$ is a separable Hilbert space (Egozcue et al., 2006). The advantage of relying on a Bayes space approach with respect to unconstrained L^2 methods is well-documented (see, e.g., Delicado (2011); Hron et al. (2016); Menafoglio et al. (2014, 2016b,a, 2018b)). Embedding PDF data within $\mathcal{B}^2$ enables one to take advantage of the geometry of this space to account for the properties of PDFs (in terms of positivity and constraints) while performing the statistical analysis. Benefits of employing the Bayes space context in spatial data analysis are also illustrated in environmental applications, e.g., by Menafoglio et al. (2014, 2016b,a); Álvarez Vázquez et al. (2020); Talská et al. (2020). The approach has also been used to model cointegrated linear processes of densities (Seo and Beare, 2019). Note that the choice of the space within which the analysis is embedded (also termed as the *feature space*) is critical in Object Oriented Spatial Statistics (O2S2) – and, more generally, in Object Oriented Data Analysis – because it has a key impact on the results and on their interpretation. We also note that the Bayes space approach we consider shares some similarities with the approach advocated by Petersen and Müller (2016); Han et al. (2019), who propose a strategy for the analysis of distributional datasets consisting of (a) mapping the data in L^2 through an isometric isomorphism, (b) performing an unconstrained analysis in L^2, and (c) mapping back the results to the space of densities.

We model observations $\mathcal{X}_{s_1}, ...\mathcal{X}_{s_n}$ upon considering the mathematical framework of O2S2 and assuming they are a partial observation of a random field $\{\mathcal{X}_s, s \in D\}$ valued in $\mathcal{B}^2$, whose elements are random NBL densities. In the following, we always assume that the field $\{\mathcal{X}_s, s \in D\}$ is globally second order stationary in $\mathcal{B}^2$ in the sense of Menafoglio et al. (2013), i.e.,

$$\begin{aligned} &\mathbb{E}[\mathcal{X}_s] = m_{\mathcal{X}}, \quad \text{for all } s \in D, \\ &Cov(\mathcal{X}_{s_i}, \mathcal{X}_{s_j}) = \mathbb{E}[\langle \mathcal{X}_{s_i} - m_{\mathcal{X}}, \mathcal{X}_{s_j} - m_{\mathcal{X}} \rangle] = C_{\mathcal{X}}(s_i - s_j), \quad \text{for all } s_i, s_j \in D. \end{aligned} \tag{3}$$

Quantity $m_{\mathcal{X}}$ in (3) denotes the expected value of $\mathcal{X}$, which is defined in the Fréchet sense (Fréchet, 1948), and is an element of $\mathcal{B}^2$, whereas $C_{\mathcal{X}}$ is a real valued, positive-definite function named *trace-covariogram*, which plays the role of the (global) covariance function of the field

(see, Menafoglio et al. (2013)). The stationarity assumption (3) implies that the (marginal) second moment of $\mathcal{X}_s$, $\mathbb{E}[\|\mathcal{X}_s\|^2]$ is constant across the entire domain D, $\|\cdot\|$ denoting for the norm in $\mathcal{B}^2$ induced by the inner product (2).

3.2. Spatial depth measures for density data in Bayes spaces

We now introduce the statistical methods we here propose to define *patterns of centrality* within a set of PDF data. Building a measure of centrality for NBL densities enables one to (*a*) determine a median NBL distribution, similar to the usual practice in the analysis of scalar data, (*b*) identify possible outlying observations which might deserve further attention, and (*c*) build prediction bands, thus enhancing the breadth of the assessment of the uncertainty associated with predictions of NBL densities at unsampled locations.

When data dimensionality exceeds one, these aspects cannot be addressed by simply relying on the natural ordering of real numbers. Rather, they can be faced through an appropriate *depth measure* (e.g., Liu, 1990; Liu and Singh, 1993; Liu et al., 1999; Mosler, 2013; Zuo and Serfling, 2000) to allow ordering any point z in the *feature space* (here set to $\mathcal{B}^2$) according to its centrality (i.e., depth) with respect to a given distribution (e.g., to the empirical distribution of the observations).

Depth measures have been widely used in multivariate statistics (see Cascos, 2009; Serfling, 2006, for a review). While a variety of depth measures for functional data have been defined and successfully used (e.g., Claeskens et al., 2014; López-Pintado and Romo, 2009; Ieva and Paganoni, 2013; Nagy et al., 2017), not all of these are well-suited to be applied in the presence of density data. For instance, Sun and Genton (2009) propose to build functional boxplots based on the modified band depth, an approach which does not lead to readily interpretable results in the presence of density data. Here, we propose to rely on the theory of *spatial depth measures* for Hilbert data proposed by Chakraborty and Chaudhuri (2014). These authors investigate the properties of the so-called *spatial distributions* in infinite dimensional Hilbert and Banach spaces, with emphasis on the properties of the spatial depth and of the spatial median and quantiles, the latter providing finite or infinite dimensional counterparts of univariate median and other quantiles. One should note that the term *spatial* does not refer to the geographical region of the study, but rather to the *feature space* set for the data. Although Chakraborty and Chaudhuri (2014) provide definitions and results under general conditions, we introduce here the key definitions for the feature space $\mathcal{B}^2$.

The spatial depth measure of a point z in $\mathcal{B}^2$ w.r.t. the distribution of a random element $\mathcal{X}$ of $\mathcal{B}^2$ is defined as

$$SpD(z|\mathcal{X}) = 1 - \left\| \mathbb{E}\left[\frac{z - \mathcal{X}}{\|z - \mathcal{X}\|} \right] \right\|. \tag{4}$$

When fixing $z \in \mathcal{B}^2$, the quantity $u(z) = \mathbb{E}\left[\frac{z-\mathcal{X}}{\|z-\mathcal{X}\|}\right]$ represents the expected value of the unit vector rooted in $\mathcal{X}$ and pointing toward z. Thus, element $u(z)$ conveys information on the *degree of outlyingness* of z with respect to the distribution of $\mathcal{X}$ (through the norm $\|u(z)\|$) and the direction along which outlyingness is displayed (through $u(z)/\|u(z)\|$). The former quantity is precisely captured by $SpD(z|\mathcal{X})$: the point z at which $SpD(z|\mathcal{X})$ is maximum is the most central one, and is termed as *spatial median*. Note that the norm of u is non-negative, and lower or equal to one by Jensen's inequality; hence, $SpD(z|\mathcal{X})$ is valued in the closed interval $[0, 1]$.

Given n observations $\mathcal{X}_1, ..., \mathcal{X}_n$, distributed as $\mathcal{X}$, the sample version of (4) can be defined as

$$\widehat{SpD}(z|\mathcal{X}_1, ..., \mathcal{X}_n) = 1 - \left\| \frac{1}{n} \cdot \sum_{i=1}^{n} \left[\frac{z - \mathcal{X}_i}{\|z - \mathcal{X}_i\|} \right] \right\|. \tag{5}$$

Here, $\frac{z-y}{\|z-y\|} = 0_+$ if $z = y$, 0_+ denoting the neutral element of the sum in $\mathcal{B}^2$, i.e., the uniform density. The estimator (5) was proved to be consistent and asymptotically normal in the presence of independent observations (Chakraborty and Chaudhuri, 2014).

In the setting of our study, we use the centrality measure induced by definition (5) for the purpose of (i) estimating the degree of centrality of the observed NBL densities, (ii) quantifying the centrality of an unobserved NBL density, and (iii) enhancing the strength of uncertainty quantification upon building prediction bands based on Monte Carlo (MC) conditional simulations. Details about these issues are offered in Section 4. With reference to (i), we rely on definition (4) and assume that the marginal distribution of the elements $\mathcal{X}_s$ does not depend on s. This implies that $SpD(z|\mathcal{X}_{s_1}) = SpD(z|\mathcal{X}_{s_2})$ for all z in $\mathcal{B}^2$ and s_1, s_2 in D. The above mentioned requirement is stronger than the assumption of global second-order stationarity in (3), and it is otherwise weaker than the assumption of strong second-order stationarity considered, e.g., in Gromenko et al. (2012); Menafoglio and Petris (2016). In practice, given the NBL density data $x_{s_1}, ..., x_{s_n}$, the depth of the j-th data point x_{s_j}, $j = 1, ..., n$ is obtained as

$$\widehat{SpD}(x_{s_j}|x_{s_1}, ..., x_{s_n}) = 1 - \left\| \frac{1}{n} \cdot \sum_{i=1,...,n; i \neq j} \left[\frac{x_{s_j} - x_{s_i}}{\|x_{s_j} - x_{s_i}\|} \right] \right\|. \tag{6}$$

4. Spatial prediction and uncertainty quantification for NBL levels

Here, we illustrate the methods used to make inference (i.e., prediction and uncertainty assessment) at unsampled locations in the geographical domain D (Subsection 4.1), and show how the spatial depth measure can support such inference (Subsection 4.2).

4.1. Object Oriented Kriging and Stochastic Simulation

The spatial prediction of an NBL density at a target location s_0 is performed upon relying on the object-oriented kriging method of Menafoglio et al. (2014). The kriging predictor is evaluated by minimizing the variance of prediction error under unbiasedness, i.e., by solving

$$\underset{\lambda_1,...,\lambda_n \in \mathbb{R}}{\operatorname{argmin}} \; \mathbb{E}\left[\left\| \mathcal{X}_{s_0} - \sum_{i=1}^{n} \lambda_i \cdot \mathcal{X}_{s_i} \right\|^2 \right] \quad \text{subject to} \quad \mathbb{E}\left[\mathcal{X}_{s_0} - \sum_{i=1}^{n} \lambda_i \cdot \mathcal{X}_{s_i} \right] = 0_+. \tag{7}$$

The solution $(\lambda_1^*, ..., \lambda_n^*)$ to problem (7) can be obtained by solving a linear system (see Menafoglio et al., 2014)

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \, C_{\mathcal{X}}(s_i - s_j) + \xi = C_{\mathcal{X}}(s_i - s_0), \quad i = 1, ..., n \\ \sum_{i=1}^{n} \lambda_i = 1, \end{cases} \tag{8}$$

where ξ represents the Lagrange multiplier associated with the unbiasedness constraint.

Solving system (8) yields the kriging predictor $\mathcal{X}^*_{s_0} = \sum_{i=1}^n \lambda_i^* \cdot \mathcal{X}_{s_i}$. Further insights on the prediction can be obtained through the spatial depth measure (5). The latter enables one to quantify how central the NBL density is predicted to be with respect to the estimated distribution of the element $\mathcal{X}_s$ of the field. Prediction $\mathcal{X}^*_{s_0}$ can also be used to provide exhaustive summaries of the NBL distribution at the target locations. Summaries of interest in our study include the mean, standard deviation, and quantiles or probabilities of exceedance of thresholds set, e.g., by regulators. We remark that analyzing NBL densities as entire objects (along the perspective of O2S2) allows obtaining summaries which are mutually consistent and do not violate possible ordering relations (e.g., for quantiles). This would not be necessarily true for multivariate geostatistical methods applied directly on the PDF summaries.

Assessment of uncertainty associated with predictions is here performed along the lines of the conditional geostatistical simulation approach of Menafoglio et al. (2016a). Conditional simulation is widely used in classical geostatistics (Chilès and Delfiner, 1999). As opposed to kriging estimates of a random field, where a single and smooth representation of the system is obtained, simulation allows generating a collection of scenarios, each honoring available data and contributing to reproduce the spatial patterns of the simulated quantity. Alternative approaches for uncertainty assessment have been proposed in the literature. These are mainly based on resampling methods, including e.g., bootstrap (see,e.g., Franco-Villoria and Ignaccolo, 2017; Pigoli et al., 2016), and are typically unconditional, i.e., they do not reproduce observations at data locations.

The method we consider relies on obtaining random realizations by sampling from the approximate conditional distribution of $\mathcal{X}_{s_0}|\{\mathcal{X}_{s_1}, ..., \mathcal{X}_{s_n}\}$, for $s_0 \in D$, hereafter denoted as $\mathcal{X}_{s_0}|\boldsymbol{\mathcal{X}}$ for simplicity. The approximation is based on the truncated empirical Karhunen-Loève expansion of the data $\mathcal{X}_{s_1}, ..., \mathcal{X}_{s_n}$,

$$\widetilde{\mathcal{X}}_s = \overline{\mathcal{X}} + \sum_{k=0}^{K} \mathrm{x}_{k,s} \cdot \phi_k, \quad s \in D,$$

where $\mathrm{x}_{k,s}$ denotes the projection coefficient of the i-th centered observation $(\mathcal{X}_s - \overline{\mathcal{X}})$ along ϕ_k in $\mathcal{B}^2$, $\mathrm{x}_{k,s} = \langle \mathcal{X}_s - \overline{\mathcal{X}}, \phi_k \rangle$, with $\overline{\mathcal{X}} = \frac{1}{n} \cdot \sum_{i=1}^n \mathcal{X}_{s_i}$. The set $\{\phi_k, k = 1, ..., K\}$ represents the eigenfunctions associated with the K largest eigenvalues of the sample covariance operator of the data $\mathcal{C}$, acting on $x \in \mathcal{B}^2$ as

$$\mathcal{C}x = \frac{1}{n} \cdot \sum_{i=1}^{n} \langle \mathcal{X}_{s_i} - \overline{\mathcal{X}}, x \rangle \cdot (\mathcal{X}_{s_i} - \overline{\mathcal{X}}). \tag{9}$$

For simplicity, we denote by $\overline{\mathcal{X}}$ both the estimator and the estimate, the meaning of the notation being clear from the context.

A realization of $\mathcal{X}_{s_0}|\boldsymbol{\mathcal{X}}$ at a location s_0 in D is then obtained by simulating the coefficients $\{\mathrm{x}_{k,s_0}, k = 1, ..., K\}$ of the truncated expansion (9), where K is set to allow capturing a given amount of variability (e.g., 99%). Simulation of $\{\mathrm{x}_{k,s_0}, k = 1, ..., K\}$ can be performed by relying on multivariate geostatistical methods, such as sequential Gaussian co-simulation (e.g., Chilès and Delfiner, 1999; Abrahamsen and Benth, 2001). Note that this simulation procedure is clearly

affected by the curse of dimensionality, as the computational resources (in terms of memory and simulation time) are directly dependent on the dimension K of the Karhunen-Loève expansion. However, the latter representation is optimal, in the sense that, for any given K, it minimizes the reconstruction error in the mean square sense (i.e., the quantity $\|\mathcal{X}_s - \widetilde{\mathcal{X}}_s\|^2$, Menafoglio et al. (2016a)).

We remark that, when using Gaussian co-simulation, the assumption of Gaussianity involves the NBL densities and not the actual NBL log-concentrations, as was typically done in previous studies (e.g., Molinari et al. (2012) and references therein). In this context, we note that the framework proposed in this work is non-parametric in terms of the NBL density.

Spatial prediction and conditional simulations of the field $\{\mathcal{X}_s, s \in D\}$ lead to prediction and simulation of a number of quantities of interest for NBL studies, including, e.g., (i) mean and variance, (ii) quantiles of any given order, and (iii) probability of exceeding thresholds and target environmental performance metrics established by regulators and water management companies. Furthermore, the depth-based method described in Section 3 enables us to identify patterns of centrality in the prediction results (see Subsection 5.2) as well as to identify prediction bands from conditional simulations (see Subsection 5.3).

4.2. Prediction bands through α-central regions based on conditional simulation

Conditional stochastic simulation at an unobserved location s_0 in D yields a collection (i.e., an ensemble) of NBL densities,

$$\mathcal{X}_{s_0}^{*m} = \overline{\mathcal{X}} + \sum_{k=0}^{K} \mathrm{x}_{k,s_0}^{*m} \cdot \phi_k, \quad s \in D,\ m = 1, ..., M \tag{10}$$

from the (approximate) distribution of $\mathcal{X}_{s_0}|\boldsymbol{\mathcal{X}}$, x_{k,s_0}^{*m} denoting the conditional realization at s_0 of the $k-$th score for the m-th simulation of the field, $m = 1, ..., M, k = 1, ..., K$; in the following, we denote $\mathbf{x}_{s_0}^{*m} = (\mathrm{x}_{k,s_0}^{*m}, ..., \mathrm{x}_{k,s_0}^{*m})$. If the feature space were a one-dimensional Euclidean space, the Monte Carlo (MC) collection of realizations $\{\mathcal{X}_{s_0}^{*m}, m = 1, ..., M\}$ could be used to build an approximate prediction region for $\mathcal{X}_{s_0}$, based on the sample quantiles of the collection. Considering the feature space $\mathcal{B}^2$, we ground the construction of the prediction region on the spatial depth measure induced by the distribution of $\mathcal{X}_{s_0}|\boldsymbol{\mathcal{X}}$, i.e.,

$$SpD(z|\{\mathcal{X}_{s_0}|\boldsymbol{\mathcal{X}}\}) = 1 - \left\| \mathbb{E}\left[\frac{z - \mathcal{X}_{s_0}}{\|z - \mathcal{X}_{s_0}\|} \middle| \boldsymbol{\mathcal{X}} \right] \right\|. \tag{11}$$

The latter metric can be estimated from the MC ensemble as

$$\widehat{SpD}(z|\{\mathcal{X}_{s_0}|\boldsymbol{\mathcal{X}}\}) = 1 - \left\| \frac{1}{M} \cdot \sum_{m=1}^{M} \left[\frac{z - \mathcal{X}_{s_0}^{*m}}{\|z - \mathcal{X}_{s_0}^{*m}\|} \right] \right\|. \tag{12}$$

Estimator (12) is consistent for (11) (see (Chakraborty and Chaudhuri, 2014)).

An MC prediction region can be built as the α-central region induced by (11), which in turn is defined as (see, e.g., Sun and Genton (2009))

$$C_\alpha = \{z \in \mathcal{B}^2 : \widehat{SpD}(z|\{\mathcal{X}_{s_0}|\boldsymbol{\mathcal{X}}\}) \geq \alpha\}, \tag{13}$$

i.e., as the subset of points in $\mathcal{B}^2$ whose depth is higher than α. Note that the α-central region is not generally associated with a coverage probability α. However, by setting α to the quantile of order $1-\widetilde{\alpha}$ of the sample spatial depths of the realizations $\{\mathcal{X}_{s_0}^{*m}, m=1,...,M\}$ one obtains an empirical coverage probability $\widetilde{\alpha}$, i.e., to $\alpha = q_{1-\widetilde{\alpha}}\{\widehat{SpD}(\mathcal{X}_{s_0}^{*m}|\{\mathcal{X}_{s_0}|\boldsymbol{\mathcal{X}}\}), m=1,...,M\}$.

We shall now focus on the identification and graphical representation of the α-central region C_α (13) at the target location s_0, based on the MC ensemble $\{\mathcal{X}_{s_0}^{*m}, m=1,...,M\}$. For this purpose, we rely on the following three steps:

(i) construction of a grid $\mathcal{G}_{s_0} \subset \mathcal{B}^2$ of points in the feature space, exploring the portion of feature space around the ensemble of MC simulations $\{\mathcal{X}_{s_0}^{*m}, m=1,...,M\}$ at s_0;

(ii) estimation of the spatial depth of each grid point in $\mathcal{G}_{s_0}$ using (12);

(iii) identification of C_α as the set of points in the grid $\mathcal{G}_{s_0}$ whose depth is higher than α.

Step (i) is aimed to characterize the portion of the feature space $\mathcal{B}^2$ which is explored by the MC ensemble, whereas steps (ii) and (iii) allow to identify, within the latter portion, the region C_α. Clearly, the finer the grid $\mathcal{G}_{s_0}$, the better the identification of C_α. We remark that, in the presence of non-standard data, such as the NBL densities we study here, visualization of the region explored by the ensemble carries a remarkable information content, as we further elucidate in Section 5.

In practice, to build the grid $\mathcal{G}_{s_0}$ and identify C_α, one can rely on the Karhunen-Loève expansion in (10). Indeed, let $\widetilde{\mathcal{G}}_{s_0} \subset \mathbb{R}^K$ be a rectangular grid of points around the simulated score vectors $\mathbf{x}_{s_0}^{*m}$, $m=1,...,M$. The grid $\widetilde{\mathcal{G}}_{s_0}$ clearly induces a grid $\mathcal{G}_{s_0}$, whose points are $z = \overline{\mathcal{X}} + \sum_{k=1}^{K} \mathrm{z}_k \cdot \phi_k$ for $\mathbf{z} = (\mathrm{z}_1,...,\mathrm{z}_K)' \in \widetilde{\mathcal{G}}_{s_0}$. Moreover, due to the orthonormality of the FPCs, (12) can then be reformulated as

$$\widehat{SpD}(z|\{\mathcal{X}_{s_0}|\boldsymbol{\mathcal{X}}\}) = 1 - \left\| \frac{1}{M} \sum_{m=1}^{M} \left[\frac{\mathbf{z} - \mathbf{x}_{s_0}^{*m}}{\|\mathbf{z} - \mathbf{x}_{s_0}^{*m}\|_2} \right] \right\|_2, \tag{14}$$

by virtue of the Parseval identity. This implies that the α-central region can be equivalently identified by selecting either the points z in $\mathcal{G}_{s_0} \subset \mathcal{B}^2$ deeper than α according to (12), or the points $\mathbf{z}$ in $\widetilde{\mathcal{G}}_{s_0} \subset \mathbb{R}^K$ deeper than α according to (14). We remark that evaluation of the depth measure in (6) does not strictly require a dimensionality reduction. The Karhunen-Loève expansion is here used to provide a representation of the α-central region C_α, a task which would be otherwise impossible because of the infinite-dimensionality of the ambient space $\mathcal{B}^2$.

The application of the procedure outlined above is exemplified through the depiction offered in Figure 3, which is related to a simple scenario of a two-dimensional Karhunen-Loève expansion (10). The example is generated by setting $\overline{\mathcal{X}}$ to the null element of the sum operation defined in (1) (i.e., the uniform distribution, see van den Boogaart et al. (2014)), the two basis functions ϕ_k, $k=1,2$ to

$$\phi_k(t) =_{\mathcal{B}^2} \exp\left\{ -\frac{(t-\mu_k)^2}{\sigma_k^2} \right\} I_{[-5,5]}(t),$$

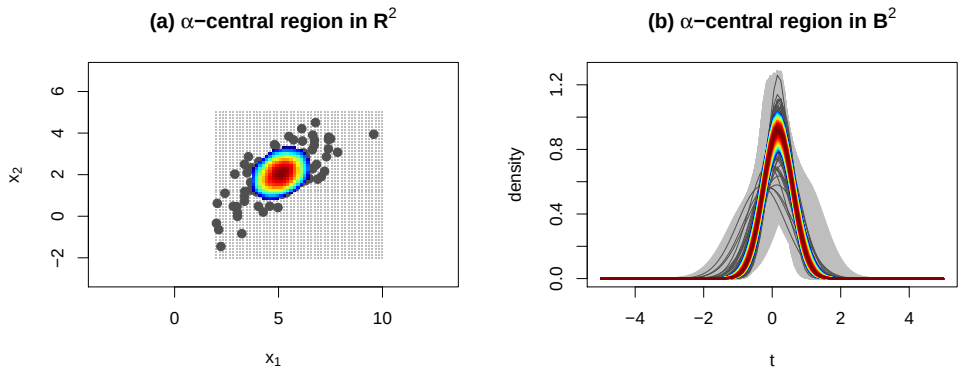

Figure 3: Exemplary illustration of the procedure to identify the α-central region in the presence of a dataset in $\mathcal{B}^2$ expressed over a two-dimensional basis. (a) identification of the α-central region in the space of the scores, $\mathbb{R}^2$; (b) identification of the α-central region in the feature space of the densities $\mathcal{B}^2$. In both panels: dark grey points/curves represent the ensemble of Monte Carlo simulations; light grey points/curves represent the rectangular region $(\widetilde{\mathcal{G}}_{s_0}, \mathcal{G}_{s_0})$ of the feature space explored by the MC ensemble; colored points/curves represent the α-central region (the colors being given according to their deepness).

with $\mu_1 = 0$, $\mu_2 = 4$, $\sigma_1^2 = 1$, $\sigma_2^2 = 9$, and by considering for the target location s_0 the following conditional distribution of the scores

$$\mathbf{x}_{s_0}^{*m} \sim N_2\left(\begin{pmatrix}5\\2\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}\right).$$

The MC ensemble of scores is generated by simulating $M = 100$ samples (depicted as dark grey symbols in Figure 3a) according to the scores' distribution, and by building the corresponding densities in $\mathcal{B}^2$ as in (10) (represented as dark grey curves in Figure 3b). In this setting, the grid $\widetilde{\mathcal{G}}_{s_0} \subset \mathbb{R}^K$ is represented by the light grey symbols in Figure 3a, whereas the associated grid $\mathcal{G}_{s_0} \subset \mathcal{B}^2$ by light grey curves in Figure 3b. The estimated α-central region in $\mathbb{R}^2$, with empirical coverage $\widetilde{\alpha} = 0.5$, is depicted with colored symbols in Figure 3a, and the equivalent α-central region in $\mathcal{B}^2$ is represented with colored curves in Figure 3b. These regions can be interpreted in a way similar to the box between the first and third quartiles in a classic boxplot; their use to support the uncertainty assessment at s_0 is further illustrated in Section 5.

We finally note that the identification of the α-central region through the construction above is clearly affected by the curse of dimensionality, because the size of the $\mathcal{G}_{s_0}$ grows exponentially with the number K of retained principal components $\{\phi_k, k = 1, ..., K\}$ (e.g., its grow rate is $O(\tau^K)$ for an equally spaced grid of τ points along each of the K dimensions). Reliance on the Karhunen-Loève expansion guarantees that, for given computational resources (i.e., for any given K), the dimensionality reduction performed prior to the conditional simulation step is uniformly optimal in the spatial domain D in a mean square sense (for further details see, Menafoglio et al., 2016a).

The following section is devoted to the illustration of the application of the proposed method-

ological framework to the NBL density data described in Section 2.

5. Results on NBL density data

5.1. Centrality patterns in NBL density data

The method proposed in Section 3 is applied to the smoothed data pictured in Figure 2. Results are displayed in Figure 4, the deepest and outermost data being represented with warm and cold colors, respectively. The spatial median (marked with a dashed line in Fig. 4a) is observed at borehole RE63-00 and represents a symmetric unimodal NBL distribution, with a mean of 0.526 [ln(mg/l)] and a standard deviation of 0.527 [ln(mg/l)]. The most outlying observations are generally characterized by lower mean values and higher variances and in some cases display a bimodal behavior. The outermost observation is characterized by a mean of -4.19 [ln(mg/l)] and standard deviation 0.40 [ln(mg/l)] and is observed at well PC56-10. Visual inspection of Figure 4 suggests that the deepest observations (i.e., observations which are closest to the mean) are indeed associated with the central part of the domain, whereas the outermost ones tend to be associated with the western part of the domain. This result is consistent with the observation that the central portion of the groundwater body is associated with depths which are generally higher than those related to the western region and are representative of more confined hydrogeological conditions, thus contributing to enhancing differences between resident concentration values detected in these two areas.

For completeness, Figure 4 also depicts the sample mean of the observations (solid line). The latter is markedly different from the spatial median, being bimodal and characterized by a much higher spread (mean of -0.34 [ln(mg/l)] and standard deviation 1.67 [ln(mg/l)]). The spatial median is clearly less influenced by the outermost observations, which tend to be concentrated at lower values of concentrations. In fact, the spatial median is a *robust* indicator of centrality, in the sense that it is only partially influenced by outlying observations, unlike, e.g., the sample mean (Chaudhuri, 1996; Ding et al., 2007). It also appears to be better representative of the shape of the data in terms, e.g., of the associated dispersion. As such, if one would be interested in providing a unique NBL density as representative of the entire aquifer body, the median NBL distribution, i.e., the one observed at well RE63-00, could be considered as a candidate to this end.

5.2. Spatial prediction of NBL density data

The analyses illustrated in this section are based on the Euclidean distance within the domain of the aquifer body. Although the shape of the domain in its right part (corresponding to the North-East sector of the region) may support the use of a non-Euclidean metric to analyze data dependence, kriging results based on a Random Domain Decomposition approach (Menafoglio et al., 2018a) did not show relevant differences with respect to those presented in the following, and are thus omitted.

The global stationarity assumption appears to be supported by the sample trace-variogram. The latter is reported in Figure 5 and is characterized by a clear asymptote for increasing spatial distances. An exponential model with nugget was used to interpret the empirical variogram, estimated values of its parameters being listed in Figure 5.

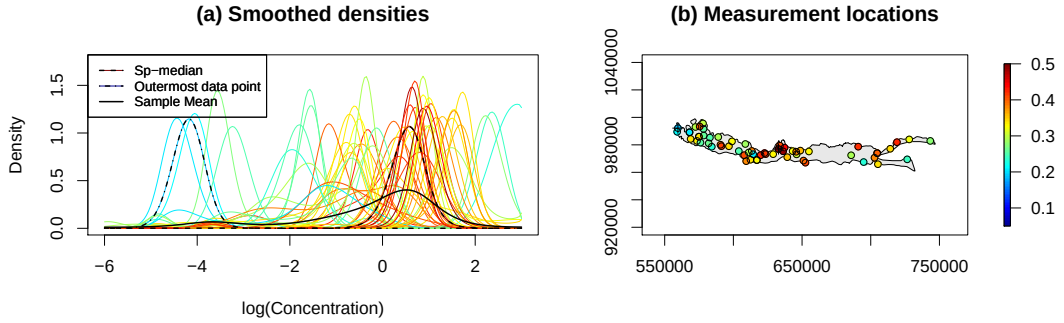

Figure 4: Spatial Depth measure of the data. (a) NBL densities colored according to their estimated depth, (b) geographical locations of the NBL densities, colors being associated the estimated spatial depth of the corresponding data.

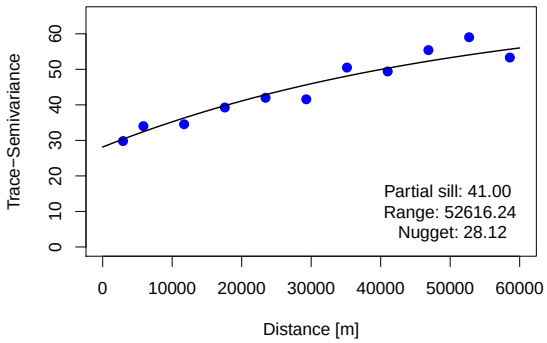

Figure 5: Stationary trace-variogram estimated from the smoothed data

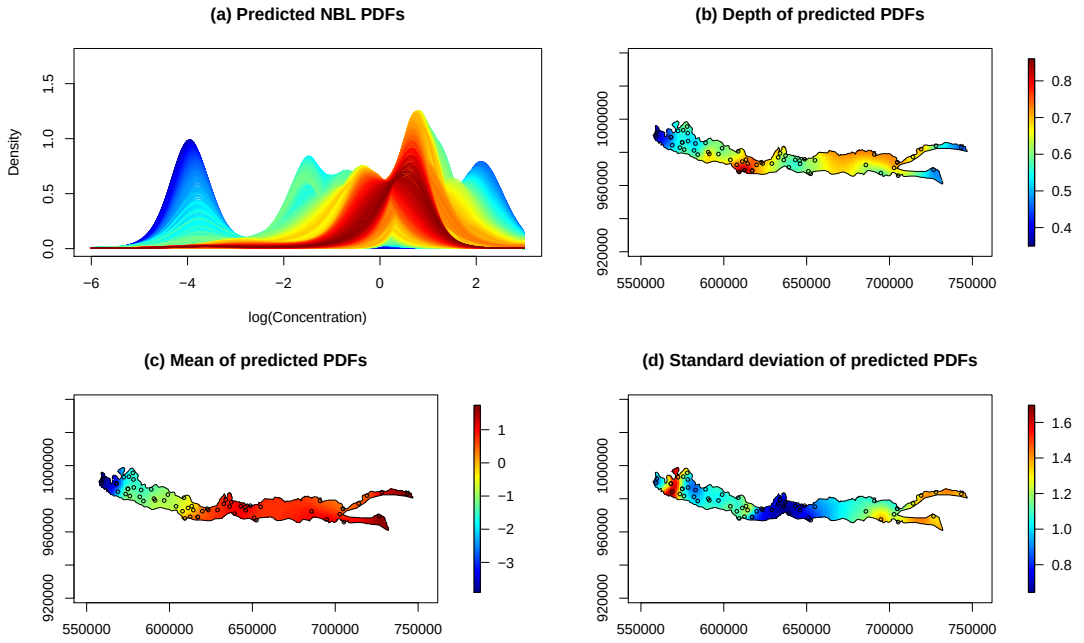

Figure 6: Kriging prediction of NBL densities: (a) Predicted densities, (b) estimated depth of the predicted densities, (c) mean value of NBL concentration estimated from the predicted densities, (d) standard deviation of NBL concentration estimated from the predicted densities.

Kriging prediction is then performed as detailed in Section 4.1 and based on a regular grid $\mathcal{G}$ of 4664 cells according to which we discretize the domain of the aquifer body. Figure 6a displays the predicted PDFs of NBL (log)concentrations. A graphical representation of the estimated depth of the predictions with respect to the sample is shown in Figure 6b, whereas Figure 6c and Figure 6d report the mean and the standard deviation of the predicted NBL densities, respectively.

Analysis of Figure 6 reveals clear spatial patterns in the field of NBL densities. The deepest NBL distributions are mostly located in the central part of the domain (Figure 6b). These areas are associated with moderate values of mean NBL concentrations and of their standard deviation (Figure 6c and d). The outermost predictions are instead clearly associated with the most western and most eastern regions of the domain. These two areas are however associated with differing characteristics, the western region denoting very low mean values and high variability, the eastern region being characterized by high mean values and reduced variability. As such, the analysis of these results enables one to clearly identify segments of the target aquifer body within which the chemical signature of interest differs from the average system behavior. This, in turn, leads to establishing spatially variable NBLs which can be integrated within sustainable groundwater and land management policies.

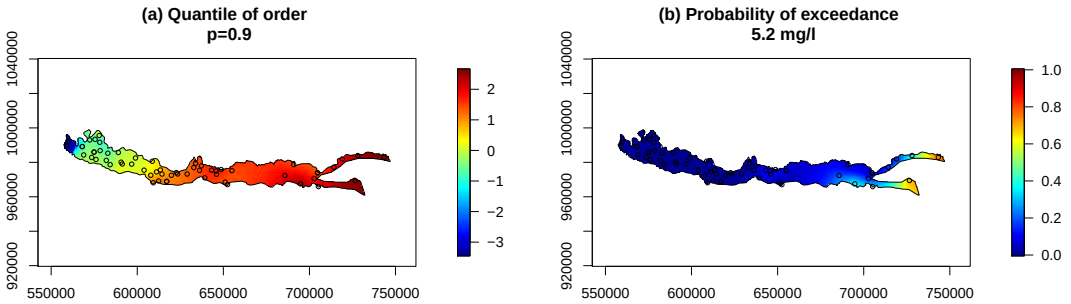

Figure 7: Predicted (a) quantile of order 90% and (b) probability of exceedance of a concentration threshold of 5.2 mg/l

Figure 7a depicts the predicted quantile of order 90% of the NBL concentration, whereas Figure 7b shows the probability of exceeding the reference NBL value of 5.2 mg/l, which has been evaluated by Molinari et al. (2012) as representative of the global chemical status of the system (see also Section 2).

Cross-validation results are included in the supplementary material and support the satisfactory performance of the prediction method. With reference to this point, note that prediction errors associated with cross-validation are assessed through leave-one-out cross-validation (LOO-CV), which is broadly used in the functional kriging context (see, e.g., Giraldo et al. (2010); Caballero et al. (2013)). LOO-CV errors are quantified via the sum of squared error, defined as $SSE = \sum_{i=1}^{n} \|\mathcal{X}_{s_i} - \mathcal{X}_{s_i}^{*(-i)}\|^2$, $\mathcal{X}_{s_i}^{*(-i)}$ being the prediction of the i-th observation when this is left out of the training set. The average SSE is 32.2, which is fully compatible with the estimated variance – the estimated sill of the trace-variogram being 69.12. Our results are also substantially in agreement with those obtained by Molinari et al. (2019), who focused solely on analysing the 90% quantile of the NBL distribution, and estimated the probability of exceeding such a reference value based on a log-normal model. Note that, although the results of Molinari et al. (2019) appear to be overall consistent with those of our study, the assumption upon which our O2S2 method is based are much less restrictive that those of Molinari et al. (2019), being fully non-parametric on the distribution of NBL concentrations.

Our results suggest that values of the 90th percentile can display marked variability within the region. Thus, setting environmental goals related to the quality of the groundwater resource can require differing approaches depending on local context. As an example, it is noted that the Eastern sector of the aquifer is characterized by a high probability of exceeding the NBL value of 5.2 mg/l evaluated on the basis of standard PS procedure and typically considered as uniform across the system (see Figure 7). It is then clear that concentrations exceeding such a threshold value within this region are not necessarily ascribed to anthropogenic actions, thus potentially hampering the effectiveness of measures aiming at restoring groundwater quality on the basis of this goal.

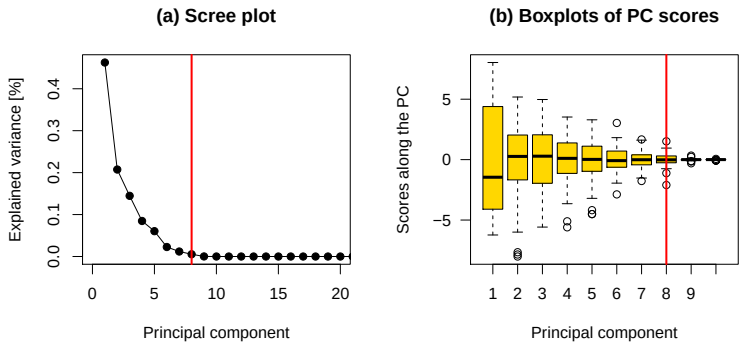

Figure 8: Dimensionality reduction of the density data: (a) scree plot; (b) boxplot of the scores x_{k,s_i}, $i = 1, ..., n$ along the first 10 PCs. The vertical line denotes the threshold value $K = 8$ selected for the dimensionality reduction.

5.3. Conditional stochastic simulation of NBL density data

The approach illustrated in Section 4.1 is here applied to the smoothed density data. The latter are projected on the basis generated by the first $K = 8$ principal components (PCs), that altogether explain 98.3% of the total data variability. Indeed, a graphical inspection of the scree plot (Figure 8a) suggests that the residual variability after the 8th PC would lead to incorporating noise within the analysis, besides being associated with a very small portion of the variability of the sample (e.g., the 9th PC is representative of 0.01% of variability). A similar conclusion is inferred from the boxplot of the corresponding scores x_{k,s_i}, $i = 1, ..., n$, $k = 1, 2, ..., 10$, depicted in Figure 8b.

The scores x_{k_1,s_1} x_{k_2,s_2} are modeled as uncorrelated for $k_1 \neq k_2$ and $s_1 \neq s_2$ in D, as supported by visual inspection of cross-variograms (not shown). An exponential model is calibrated to the empirical variogram for each field of scores. Conditional Gaussian simulations are performed to yield a collection of $B = 100$ realizations. We do so by using sequential Gaussian simulation of Abrahamsen and Benth (2001), implemented within the R package `gstat` (Pebesma, 2004), and setting a local neighborhood of 60km, to lower the computational time. The sample of NBL distributions is then built from the MC ensemble of scores as in (10), i.e.,

$$\mathcal{X}^{\star,m}_{s_0} = \overline{\mathcal{X}} + \sum_{k=0}^{K} \mathrm{x}^{\star,m}_{k,s_0} \phi_k, \quad s_0 \in \mathcal{G}, m = 1, ..., M$$

where $\mathrm{x}^{\star,m}_{k,s_0}$ denotes the m-th conditional simulation of the k-th score at location s_0.

As an example of the type of results one can obtain, Figure 9 depicts a realization of the field of NBL densities (Figure 9a), of the corresponding quantiles of order 90% (Figure 9b) and of the corresponding probability of exceeding the threshold value of 5.2 mg/l (Figure 9c). Further results are presented in the supplementary material, in the context of a leave-one-out cross-validation analysis. These types of results embed a high added value in terms of operational environmental implications. In this context, providing spatial estimates of NBL densities

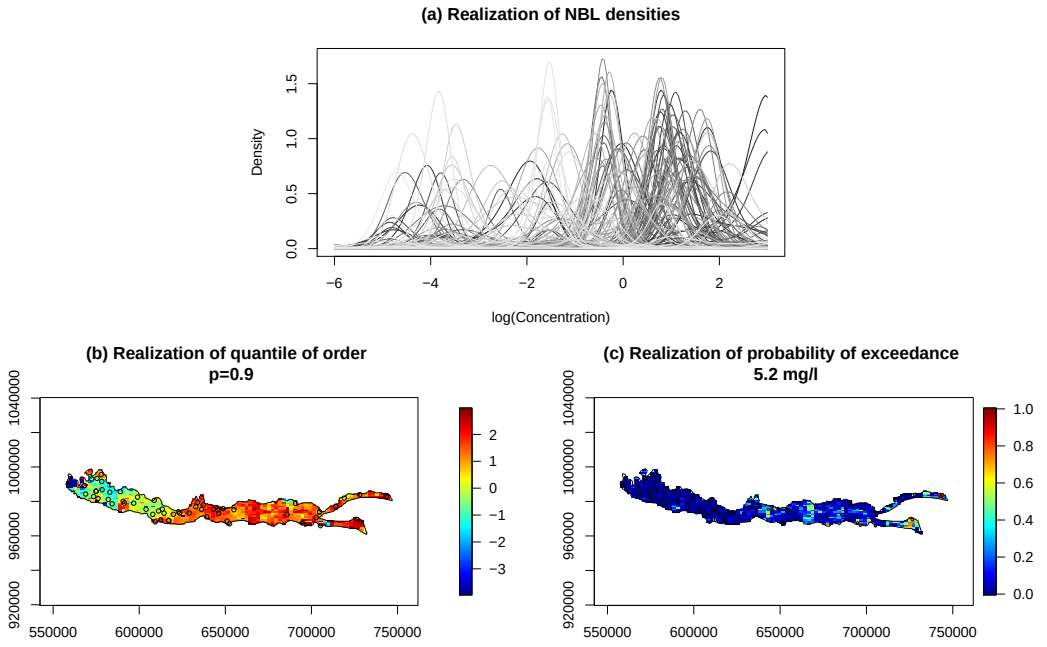

Figure 9: A sample realization resulting from the conditional stochastic simulation of NBL distributions. (a) Simulated NBL densities; (b) Simulated quantiles of order 90%, for the same realization displayed in panel (a); (c) Simulated probability of exceeding the threshold value of 5.2 mg/l, for the same realization displayed in panel (a).

and the ensuing uncertainty is critical to assist (i) the identification of local and diffuse sources of potential anthropogenic contamination, thus enabling to quantify the current environmental status and set achievable clean-up goals for remediation actions, as well as (ii) the proper engineering and management of monitoring networks to optimize the value of information associated with areas characterized by marked uncertainty or probability of exceeding given environmental limits/standards. Our approach can then provide water and environmental agencies enhanced information which can then be used to design and engineer decision-making protocols grounded on an appropriate and reliable quantification of the associated uncertainty.

5.4. Prediction band for NBL densities

We describe here the results associated with the construction of a prediction band based on the conditional simulation in Subsection 5.3. For illustrative purposes, we consider two target locations $s_{0,1}, s_{0,2}$, close to the wells RE63-00 and PC56-10, i.e., the locations associated with the median NBL density and the outermost NBL density, respectively.

Figure 10 depicts the MC ensembles obtained at $s_{0,1}, s_{0,2}$, with colors assigned according to the depth of the simulated densities with respect to the corresponding (approximated) conditional

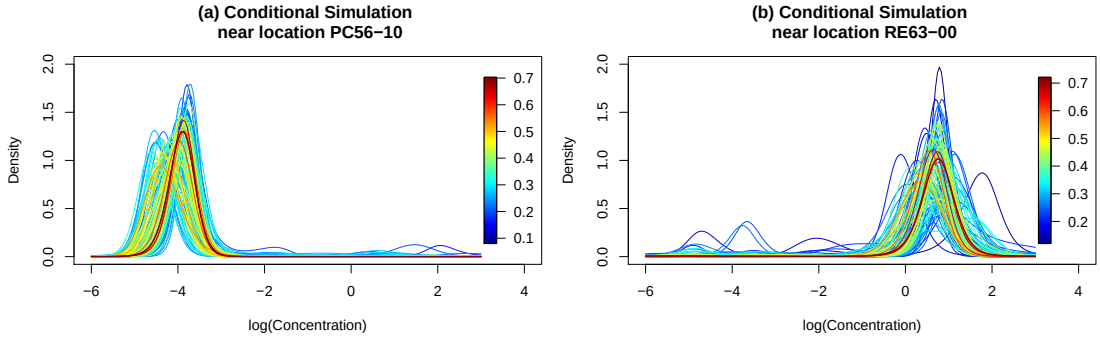

Figure 10: Representation of the Monte Carlo ensemble at two target locations $s_{0,1}, s_{0,2}$. Monte Carlo ensemble at (a) $s_{0,1}$ and (b) $s_{0,2}$. Colors are given according to the depth of the PDF with respect to the corresponding (approximated) conditional distribution $\mathcal{X}_{s_{0,i}}|\boldsymbol{\mathcal{X}}$, $i = 1, 2$.

distribution $\mathcal{X}_{s_{0,i}}|\boldsymbol{\mathcal{X}}$, $i = 1, 2$.

The prediction band is obtained upon building the α-central region with empirical coverage of $\widetilde{\alpha} = 0.5$. Similar to the illustrative example presented in Section 4, for the target location $s_{0,i}$, $i = 1, 2$, we build a uniform grid $\mathcal{G}_{s_{0,i}}$ in $\mathbb{R}^8$, covering the region spanned by the MC ensemble of scores at $s_{0,i}$. The PDFs associated with the grid points in $\mathcal{G}_{s_{0,1}}$ and $\mathcal{G}_{s_{0,2}}$ are depicted as light grey curves in Figures 11a and 11b, respectively. These curves allow visualizing the rectangular envelope of the region explored by the conditional distributions $\mathcal{X}_{s_{0,i}}|\{\mathcal{X}_{s_{0,i}}|\boldsymbol{\mathcal{X}}\}$, $i = 1, 2$ in the feature space. To display the prediction band at the target locations, we select the grid points in $\mathcal{G}_{s_{0,i}}$, $i = 1, 2$, whose associated PDFs are characterized by a spatial depth higher than the median spatial depth.

The α-central regions $C_\alpha(s_{0,i})$ at locations $s_{0,1}, s_{0,2}$ are depicted in Figure 11, colors being assigned according to the depth of the PDFs. This representation provides insights about the variability of the kriging predictors at these target locations, represented as dashed black curves in Figure 11. For instance, NBL densities near location PC56-10 are predicted to be unimodal and to display their peak around a value of log-concentration of -4 [log(mg/l)]. The α-central regions $C_\alpha(s_{0,1})$ reveals that the position of the peak is affected by a certain degree of uncertainty. This is also reflected at the right tail of the NBL distribution, where local (albeit of modest height) peaks may appear when considering the collection of MC realizations and the associated $C_\alpha(s_{0,1})$. Visual inspection of Figure 11b suggests that the kriging prediction at $s_{0,2}$ is associated with highest uncertainty in the position of the main peak as well as in the height of additional peaks appearing at the left tail of the NBL density. These interpretations are consistent with those provided in Figure 10, and strengthen our ability to characterize possible behaviors of NBL densities at locations of environmental concern in the domain.

6. Conclusions

We illustrate a novel approach to the characterization of Natural Background Levels (NBL) concentrations of target chemical species in large-scale groundwater bodies. We leverage on the

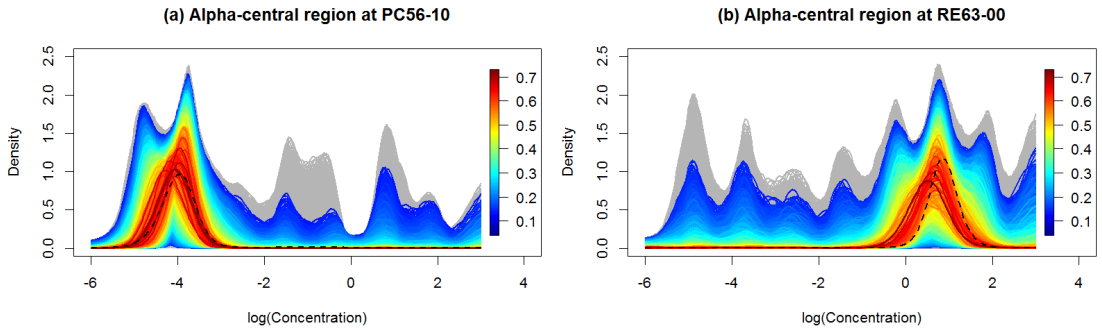

Figure 11: Prediction band for NBL densities based on the evaluation of the α-central region. Prediction band at target locations (a) $s_{0,1}$ and (b) $s_{0,2}$. Colored curves form the approximated α-central region, colors being associated with the corresponding depth; light grey colors represent curves associated with the rectangular region of the feature space explored by the distribution of $\mathcal{X}_{s_{0,i}}|\{\boldsymbol{\mathcal{X}}\}$, $i = 1, 2$. The dashed black curves represent the kriging predictions at locations $s_{0,1}, s_{0,2}$.

theory of Object Oriented Spatial Statistics (O2S2) and consider the entire distribution of NBL concentrations rather than focusing on selected moments or percentiles. Such distributions are represented through their (estimated) densities and are modeled by embedding them in a Bayes Hilbert space. Our approach represents an innovative way to view NBL concentrations and enables one to (i) obtain a complete characterization of the spatial arrangement of NBL PDFs in terms of point prediction and uncertainty quantification (UQ), and (ii) provide consistent predictions and UQ of any summary statistic (e.g., moments, quantiles) derived from the NBL distribution, with significant advantages with respect to traditional approaches employed in engineering and environmental applications. In this framework, our work leads to the following major conclusions.

1. We provide theoretical and operational tools for generating and exploring ensembles of NBL PDFs based on the key concept of *spatial depth measures*. These are employed here to (i) identify central and outlying observations in the distributional dataset and (ii) build prediction regions and distributional boxplots for NBL distributions, leading to enhanced assessment of the uncertainty associated with predictions. The proposed strategy is fully nonparametric and allows overcoming limitations of the statistical methods so far used for the characterizations of NBL distributions.

2. Our approach enables one to ground the assessment of NBLs of a given groundwater body on the entire information content embedded in the complete series of concentration values which are routinely monitored by environmental agencies and water companies across a network of observation boreholes covering the system. While our showcase application is focused on sampled values of a given chemical species, i.e., ammonium, the proposed methodology and workflow are portable and can be applied to analyze other chemical species whose temporal sampling is satisfactorily dense.

3. The approach leads to the identification of sectors of a given groundwater body which can

be characterized by differing behaviors of the considered chemical species and to the assessment of the uncertainty associated with their spatial patterns. From an environmental standpoint, this element is key in large scale groundwater bodies of the kind we analyze, where critical conditions associated with unsampled regions can be identified. As new samples become available over time, the methodology also enables one to detect temporal dynamics or shifts of spatial distributions of NBLs, thus leading to possible refinement or updating of selected management practices.

References

2014/80/EU, D., 2014. amending Annex II to Directive 2006/118/EC of the European Parliament and of the Council on the Protection of Groundwater Against Pollution and Deterioration, OJ L182, 21 June 2014. pp. 52-55.

Abrahamsen, P., Benth, F., 2001. Kriging with inequality constraints. Mathematical Geology 33 (6), 719–744.

Álvarez Vázquez, M., Hošek, M., Elznicová, J., Pacina, J., Hron, K., Fačevicová, K., Talská, R., Bábek, O., Matys Grygar, T., 2020. Separation of geochemical signals in fluvial sediments: New approaches to grain-size control and anthropogenic contamination. Applied Geochemistry 123, 104791.

Amorosi, A., Farina, M., Severi, P., Preti, D., Caporale, L., Di Dio, G., 1996. Genetically related alluvial deposits across active fault zones: an example of alluvial fan-terrace correlation from the upper Quaternary of the southern Po Basin, Italy. Sediment. Geol. 102, 275–295.

Bianchi Janetti, E., Riva, M., Guadagnini, A., 2021. Natural springs protection and probabilistic risk assessment under uncertain conditions. Science of the Total Environment 751 (141430), 1–13.

Caballero, W., Giraldo, R., Mateu, J., 2013. A universal kriging approach for spatial functional data. Stochastic Environmental Research and Risk Assessment, 1–11.

Cascos, I., 2009. New Perspectives in Stochastic Geometry. Clarendon Press, Oxford University Press, Oxford, Ch. Data depth: Multivariate statistics and geometry, pp. 398–423.

Chakraborty, A., Chaudhuri, P., 2014. The spatial distribution in infinite dimensional spaces and related quantiles and depths. The Annals of Statistics 42 (3), 1203–1231.

Chaudhuri, P., 1996. On a geometric notion of quantiles for multivariate data. Journal of the American Statistical Association 91, 862–872.

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

Claeskens, G., Hubert, M., Slaets, L., Vakili, K., 2014. Multivariate functional halfspace depth. Journal of the American Statistical Association 109 (505), 411–423.

Coetsiers, M., Blaser, P., Martens, K., Walraevens, K., 2009. Natural background levels and threshold values for groundwater in fluvial Pleistocene and Tertiary marine aquifers in Flanders, Belgium. Environmental Geology volume 57, 1155–1168.

Cremonini, S., Etiope, G., Italiano, F., Martinelli, G., 2008. Evidence of possible enhanced peat burning by deep-origin methane in the Po River delta Plain (Italy). J. Geol. 116, 401–413.

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

Ding, Y., Dang, X., Peng, H., Wilkins, D., 2007. Robust clustering in high dimensional data using statistical depths. BCM Bioinformatics 8 (7), S8.

Directive 2000/60/EC WFD, 2000. Directive of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy, OJ L327, 22 Dec 2000, pp 1-73.

Directive 2006/118/EC GWDD, 2006. Directive of the European Parliament and of the Council of 12 December 2006 on the protection of groundwater against pollution and deterioration, OJ L372, 27 Dec 2006, pp 19-31.

Ducci, D., de Melo, M. C., Preziosi, E., Sellerino, M., Parrone, D., Ribeiro, L., 2016. Combining natural background levels (NBLs) assessment with indicator kriging analysis to improve groundwater quality data interpretation and management. Sci. Total Environ 569-570, 569–584.

Edmunds, W., Shand, P., Hart, P., Ward, R., 2003. The natural (baseline) quality of groundwater: a UK pilot study. Sci. Total Environ. 310, 25–35.

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

European Commission, 2009. Guidance on groundwater status and trend assessment, guidance document no 18. Tech. rep., European Communities, technical Report 2009, ISBN 978-92-79-11374-1, Luxembourg.

Farina, M., Marcaccio, M., Zavatti, A., 2014. Experiences and perspectives to monitoring of groundwater resources: the contribution of Emilia Romagna (Esperienze e prospettive nel monitoraggio delle acque sotterranee: il contributo dell'Emilia-Romagna). Pitagora Editrice. Bologna, (in Italian).

Franco-Villoria, M., Ignaccolo, R., 2017. Bootstrap based uncertainty bands for prediction in functional kriging. Spatial Statistics 21, 130 – 148.

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

Giraldo, R., Delicado, P., Mateu, J., 2010. Continuous time-varying kriging for spatial prediction of functional data: An environmental application. Journal of Agricultural, Biological, and Environmental Statistics 15 (1), 66–82.

Gromenko, O., Kokoszka, P., Zhu, L., Sojka, J., 2012. Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. Annals of Applied Statistics 6 (2), 669–696.

Guadagnini, L., Menafoglio, A., Sanchez-Vila, X., Guadagnini, A., 2020. Probabilistic assessment of spatial heterogeneity of natural background concentrations in large-scale groundwater bodies through functional geostatistics. Science of The Total Environment 740, 140139.

Han, K., Müller, H.-G., Park, B., 2019. Additive functional regression for densities as responses. Journal of the Americal Statistical AssociationDOI: 10.1080/01621459.2019.1604365.

Hinsby, K., de Melo, M. C., 2006. Application and evaluation of a proposed methodology for derivation of groundwater threshold values-a case study summary report. Tech. rep., In Deliverable D22 of the EU project "BRIDGE", available at: `www.keriel.org/BIB/Publ_UNESCO/SOG_BRIDGE/Deliverables/WP4/D22.pdf`.

Hinsby, K., de Melo, M. C., Dahl, M., 2008. European case studies supporting the derivation of natural background levels and groundwater threshold values for the protection of dependent ecosystems and human health. Sci. Total Environ. 401 (1–3), 1–20.

Hron, K., Menafoglio, A., Templ, M., Hruzova, K., Filzmoser, P., 2016. Simplicial principal component analysis for density functions in Bayes spaces. Computational Statistics & Data Analysis 94, 330–350.

Ieva, F., Paganoni, A., 2013. Depth measures for multivariate functional data. Communications in Statistics - Theory and Methods 42 (7), 1265–1276.

Kim, K., Yun, S., Kim, H., Kim, J., 2015. Determination of natural backgrounds and thresholds of nitrate in South Korean groundwater using model-based statistical approaches. J. Geochem. Explor. 148, 196–205.

Libera, N. D., Fabbri, P., Mason, L., Piccinini, L., Pola, M., 2017. Geostatistics as a tool to improve the natural background level definition: an application in groundwater. Sci. Total Environ. 598, 330–340.

Liu, R. Y., 1990. On a notion of data depth based on random simplices. The Annals of Statistics 18 (1), 405–414.

Liu, R. Y., Parelius, J., Singh, K., 1999. Multivariate analysis by data depth: descriptive statistics, graphics and inference. The Annals of Statistics 27 (3), 783–858.

Liu, R. Y., Singh, K., 1993. A quality index based on data depth and multivariate rank tests. Journal of the American Statistical Association 88 (421), 252–260.

López-Pintado, S., Romo, J., 2009. On the concept of depth for functional data. Journal of the American Statistical Association 104 (486), 718–734.

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

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

Martín-Fernández, J., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J., 2015. Bayesian-multiplicative treatment of count zeros in compositional data sets. Statistical Modelling 15 (2), 134–158.

McKinley, J., Hron, K., Grunsky, E., Reimann, C., de Caritat, P., Filzmoser, P., van den Boogaart, K. G., Tolosana-Delgado, R., 2016. The single component geochemical map: Fact or fiction? Journal of Geochemical Exploration 162, 16 – 28.

Menafoglio, A., Gaetani, G., Secchi, P., 2018a. Random domain decompositions for object-oriented kriging over complex domains. Stochastic Environmental Research and Risk Assessment.

Menafoglio, A., Grasso, M., Secchi, P., Colosimo, B., 2018b. Profile Monitoring of Probability Density Functions via Simplicial Functional PCA with application to Image Data. Technometrics 60 (4), 497–510.

Menafoglio, A., Guadagnini, A., 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.

Menafoglio, A., Guadagnini, A., Secchi, P., 2016a. Stochastic Simulation of Soil Particle-Size Curves in Heterogeneous Aquifer Systems through a Bayes space approach. Water Resources Research 52, 5708–5726.

Menafoglio, A., Petris, G., 2016. Kriging for Hilbert-space valued random fields: The operatorial point of view. Journal of Multivariate Analysis 146, 84–94.

Menafoglio, A., 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.

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

Menafoglio, A., Secchi, P., Guadagnini, A., 2016b. A Class-Kriging predictor for Functional Compositions with application to particle-size curves in heterogeneous aquifers. Mathematical Geosciences 48(4), 463–485.

Molinari, A., Guadagnini, L., Marcaccio, M., Guadagnini, A., 2012. Natural background levels and threshold values of chemical species in three large-scale groundwater bodies in Northern Italy. Sci. Total Environ. 425, 9–19.

Molinari, A., Guadagnini, L., Marcaccio, M., Guadagnini, A., 2019. Geostatistical multimodel approach for the assessment of the spatial distribution of natural background concentrations in large-scale groundwater bodies. Water Research 149, 522–532.

Mosler, K., 2013. Robustness and Complex Data Structures, Festschrift in Honour of Ursula Gather. Springer, Berlin, Ch. Depth statistics, pp. 17–34.

Nagy, S., Gijbels, I., Hlubinka, D., 2017. Depth-based recognition of shape outlying functions. Journal of Computational and Graphical Statistics 26 (4), 883–893.

Panno, S., Kelly, W., Martinsek, Hackley, K., 2006. Estimating background and threshold nitrate concentrations using probability graphs. Groundwater 44 (5), 697–709.

Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, R., 2015. Modeling and analysis of compositional data. Wiley, Chichester.

Pebesma, E. J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences 30, 683–691.

Perulero Serrano, R., Guadagnini, L., Riva, M., Giudici, M., Guadagnini, A., 2014. Impact of two geostatistical hydro-facies simulation strategies on head statistics under non-uniform groundwater flow. J. Hydrology 508, 343–355.

Petersen, A., Müller, H., 2016. Functional data analysis for density functions by transformation to a Hilbert space. Annals of Statistics 44 (1), 183–218.

Pigoli, D., Menafoglio, A., Secchi, P., 2016. Kriging prediction for manifold-valued random field. Journal of Multivariate Analysis 145, 117–131.

Reimann, C., Garrett, R., 2005. Geochemical background: concept and reality. Sci. Total Environ. 350, 12–27.

Seo, W., Beare, B., 2019. Cointegrated linear processes in Bayes Hilbert space. Statistics & Probability Letters 147, 90 – 95.

Serfling, R., 2006. Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications. American Mathematical Society, Ch. Depth functions in nonparametric multivariate inference, pp. 1–16.

Short, M., Guadagnini, L., Guadagnini, A., Tartakovsky, D., Higdon, D., 2010. Predicting vertical connectivity within an aquifer system. Bayesian Analysis 5 (3), 557–582.

Sun, Y., Genton, M., 2009. Functional boxplots. Journal of Computational and Graphical Statistics 20 (2), 316–334.

Talská, R., J. Machalová, P. S., Hron, K., 2020. A comparison of seed germination coefficients using functional regression. Applications in Plant Sciences 8 (7), e11366.

Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: an R-package for robust statistical analysis of compositional data. John Wiley and Sons.

Urresti-Estala, B., Carrasco-Cantos, F., Vadillo-Pérez, I., Jiménez-Gavilán, P., 2013. Determination of background levels on water quality of groundwater bodies: A methodological proposal applied to a Mediterranean River basin (Guadalhorce River, Málaga, southern Spain). J. Environ. Manage 117, 121–130.

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

Walter, T., 2008. Determining natural background values with probability plots. EU Groundwater Policy Developments Conference, UNESCO, Paris, France, 13-15 Nov 2008.

Wendland, F., Hannappel, S., Kunkel, R., Schenk, R., Voigt, H., Wolter, R., 2005. A procedure to define natural groundwater conditions of groundwater bodies in Germany. Water Sci. Technol. 51 (3–4), 249–257.

Winter, C., Tartakovsky, D., Guadagnini, A., 2003. Moment differential equations for flow in highly heterogeneous porous media. Surveys in Geophysics 24 (1), 81–106.

Zuo, Y., Serfling, R., 2000. General notions of statistical depth function. The Annals of Statistics 28, 461–482.