Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-018-1596-z

ORIGINAL PAPER

Random domain decompositions for object-oriented Kriging over complex domains

Alessandra Menafoglio[1] · Giorgia Gaetani[1] · Piercesare Secchi[1]

Abstract
We propose a new methodology for the analysis of spatial fields of object data distributed over complex domains. Our approach enables to jointly handle both data and domain complexities, through a *divide et impera* approach. As a key element of innovation, we propose to use a random domain decomposition, whose realizations define sets of homogeneous sub-regions where to perform simple, independent, weak local analyses (*divide*), eventually aggregated into a final strong one (*impera*). In this broad framework, the complexity of the domain (e.g., strong concavities, holes or barriers) can be accounted for by defining its partitions on the basis of a suitable metric, which allows to properly represent the adjacency relationships among the complex data (such as scalar, functional or constrained data) over the domain. As an illustration of the potential of the methodology, we consider the analysis and spatial prediction (Kriging) of the probability density function of dissolved oxygen in the Chesapeake Bay.

Keywords Object oriented data analysis · Spatial dependence · Local stationarity · Variogram kernel estimator · Bayes spaces

1 Introduction

The analysis of complex data distributed over large or highly textured regions poses new challenges to spatial statistics. Methods developed so far to deal with spatial complex data often rely upon global models to capture variability and spatial dependence (e.g., Menafoglio et al. 2013; Menafoglio and Secchi 2017). A common assumption regards the stationarity of the process generating the data, i.e., the homogeneity of its distributional properties over the whole domain. However, in many applications, observed data are not compatible with a global stationarity assumption. Moreover, the features of the domain may even prevent the definition of a globally stationary model.

Several approaches exist to handle non-stationary spatial fields (see Fouedjio 2017, for a recent review). Of particular interest for the scope of this work are the methods based on local models which describe the spatial dependence only within subregions of the spatial domain, where stationarity is taken to be a viable assumption (e.g., Fuentes 2001, 2002; Fouedjio et al. 2016; Heaton et al. 2015; Haas 1990; Harris et al. 2010; Kim et al. 2005). For instance, Kim et al. (2005) propose a Bayesian hierarchical model, that identifies an optimal partition of the domain in disjoint and independent stationary subregions. Other authors, e.g., Fouedjio et al. (2016), developed estimation methods for non-stationary covariance models, based on the key assumption of local stationarity.

All these methods are model-based, they often require strong assumptions on the distribution generating the data and provide estimation procedures which can rarely be extended to high- or infinite-dimensional data, like curves, surfaces or images. When these are the data at hand, algorithmic approaches, possibly based on computationally intensive yet simple techniques, are usually preferred.

In this work we propose a new computational method for the analysis of spatial data distributed over a possibly complex domain. The latter may consist in a very large domain, or in a region with natural or artificial constraints, such as holes, barriers, irregular boundaries.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00477-018-1596-z) contains supplementary material, which is available to authorized users.

✉ Alessandra Menafoglio
alessandra.menafoglio@polimi.it

[1] MOX-Department of Mathematics, Politecnico di Milano, Milan, Italy

Published online: 17 August 2018

The basic idea is to use simple, local and repeated analyses instead of an unique global and complex one. Our line of attack is based on a *divide et impera* strategy. During the *divide* step, the spatial domain is randomly partitioned into a set of disjoint sub-regions, within which local geostatistical analyses are performed. These local and weak analyses are repeated for different realizations of the random domain decomposition and then aggregated into a final strong analysis during the *impera* step.

The non-parametric nature of our approach is open to handle both the complexity of the data and that of the domain. We pursue the viewpoint of object oriented spatial statistics [O2S2, Menafoglio and Secchi (2017)] and represent data as spatially dependent atoms embedded in a suitable *feature space* whose geometry should be founded on, and should elicit the data characteristics that the researcher deems to be essential for the goal of the analysis. Instead, the complexity of the spatial domain will control the metric upon which it is partitioned. Here, we will argue in favor of graph-based metrics, and accordingly evaluate the distance between two sites as the length of the shortest path linking them on a given undirected graph representing the actual spatial closeness.

A precursor of our method is the Bagging Voronoi algorithm of Secchi et al. (2013, 2015); Abramowicz et al. (2016). The idea founding the Bagging Voronoi approach to the analysis of spatially dependent data, is to consider a target statistical method developed for independent data—say, a method for classification, regression, dimensional reduction—and to apply it to local representatives built upon a random partition of the spatial domain—i.e., across the elements of the partition. In the methodology developed in this paper, the target statistical method explicitly incorporates spatial dependence and generates different local analyses, one within each cell of the random partition of the spatial domain. Although this work focuses on the problem of spatial prediction (i.e., Kriging), the novel strategy is entirely general and can be successfully employed to tackle several geostatistical problems (e.g., classification or spatial regression and smoothing).

As an illustrative case study, we consider the problem of spatial prediction of dissolved oxygen (DO) in the Chesapeake Bay, that is the largest, most productive and biologically diverse estuary in North America (Fig. 1). An estuarine system develops on a complex, non-convex and highly irregular domain where the areas of land between adjacent tributaries act as barriers for many aquatic variables. Here, the use of the Euclidean distance is inappropriate for describing the adjacency relation between observations in different sites. Moreover, the variable of interest may not be scalar, as when the data object observed in each spatial location is the distribution of DO. Although methods to treat the domain complexity are known (e.g., Sangalli et al. 2013, and references therein), and have been inspirational for the domain representation used in this work, the proposed methodology has the advantage of being able to handle jointly both data and domain complexities and of being much simpler to implement while providing accurate predictions.

The remaining part of the paper is organized as follows. Section 2 describes the key idea of the proposed methodology, that is the use of random domain decompositions of the spatial domain, to allow for local analyses. Section 3 explores via simulations the performance of the method. Section 4 illustrates the case study, where the goal is the prediction of the dissolved oxygen in the Chesapeake Bay.

2 Kriging via random domain decompositions

2.1 A locally-stationary model for object data

Set $(\Omega; \mathcal{F}; \mathbb{P})$ to be a probability space and $\mathcal{H}$—the *feature* space—to be a separable Hilbert space, with operations $(+, \cdot)$, inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$.

Given the spatial domain $D \subseteq R^d$ and the sampled locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ in D, we denote by $\mathcal{X}_{\mathbf{s}_1}, \ldots, \mathcal{X}_{\mathbf{s}_n}$ the random elements whose realizations are the available data; they are assumed to be generated by a partial observation of a random field $\{\mathcal{X}_{\mathbf{s}}, \mathbf{s} \in D\}$ defined on $(\Omega; \mathcal{F}; \mathbb{P})$ and with values in $\mathcal{H}$. The first and (global) second order properties of the field can be defined in terms of mean (or drift) and trace-covariogram. We thus call $m_{\mathbf{s}} = \mathbb{E}[\mathcal{X}_{\mathbf{s}}]$ the mean of the field at $\mathbf{s}$ in D, and $C : D \times D \to \mathbb{R}$ the *trace-covariogram* of the field (Menafoglio et al. 2013), defined, for $\mathbf{s}_1, \mathbf{s}_2$ in D, as

$$C(\mathbf{s}_1, \mathbf{s}_2) = \mathbb{E}[\langle \mathcal{X}_{\mathbf{s}_1} - m_{\mathbf{s}_1}, \mathcal{X}_{\mathbf{s}_2} - m_{\mathbf{s}_2} \rangle]. \quad (1)$$

The trace-covariogram (1) plays the role of the classical covariogram, which is widely-used in geostatistics to represent the second-order properties of the field (e.g., Cressie 1993). The trace-covariogram is a *global* measure of spatial dependence, whereas the family of cross-covariance operators—of which the trace-covariogram represents the trace—describes the full dependence structure of the data. Nevertheless, the global viewpoint is sufficient for the purpose of Kriging prediction with scalar weights (for further details refer to Menafoglio et al. 2013; Menafoglio and Petris 2016). As in real-valued geostatistics, one may also define the *trace-variogram*,

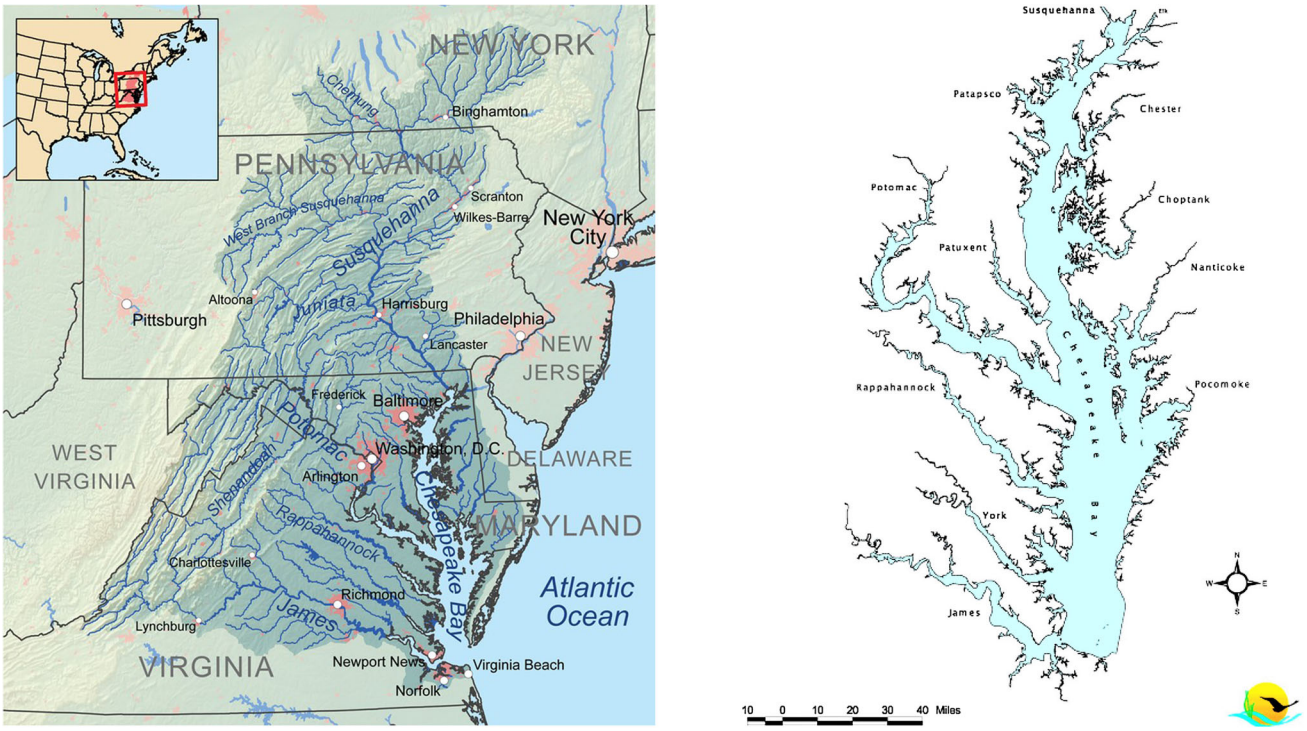

Fig. 1 Chesapeake Bay and its main rivers.. *Source*: (**a**) KMusser on Wikipedia, (**b**) modified from http://www.chesapeakebay.net/

$$2\gamma(\mathbf{s}_1, \mathbf{s}_2) = \mathbb{E}[\|\mathcal{X}_{\mathbf{s}_1} - \mathcal{X}_{\mathbf{s}_2}\|^2] - \|m_{\mathbf{s}_1} - m_{\mathbf{s}_2}\|^2, \quad \mathbf{s}_1, \mathbf{s}_2 \in D, \tag{2}$$

which is the object-oriented counterpart of the variogram.

On the basis of (1) and (2), (global) second-order stationarity can be formulated, by requiring that the mean of the field is spatially constant over D (i.e., $m_{\mathbf{s}} = m$ for all the $\mathbf{s}$ in D) and that the trace-covariogram depends only on the increment between locations [i.e., $C(\mathbf{s}_1, \mathbf{s}_2) = C(\mathbf{s}_1 - \mathbf{s}_2)$, for $\mathbf{s}_1, \mathbf{s}_2$ in D]. Under these assumptions, Ordinary Kriging methods can be developed (e.g., Menafoglio and Secchi 2017, and references therein).

In this work, we consider a more general setting, which is the one of local stationarity, also known as quasi-stationarity. This notion of stationarity is well-known in the literature on scalar geostatistics and was introduced by Matheron (1971). In this framework, stationarity is assumed to hold true only in neighborhoods of a given radius around any location $\mathbf{s}$ in D. That is, an $\mathcal{H}$-valued random field $\{\mathcal{X}_{\mathbf{s}}, \mathbf{s} \in D\}$ is said to be *locally stationary* if it is characterized by a mean $m_{\mathbf{s}}$ and a covariance function $C(\mathbf{s}_1, \mathbf{s}_2)$ such that (1) for any location $\mathbf{s} \in D$, the mean $m_{\mathbf{s}}$ of the field is approximately constant in a neighborhood $V_{\mathbf{s}}$ of $\mathbf{s}$; and (2) for any location $\mathbf{s} \in D$, the covariance function can be approximated via a stationary model in the neighborhood $V_{\mathbf{s}}$ of $\mathbf{s}$. Although this definition may appear vague, a number of authors (e.g Fouedjio et al. 2016, and references therein) have recently embedded it in a formal modeling framework within which they propose geostatistical methods for the spatial prediction of scalar random fields. In the context of object data distributed over complex domains these methods cannot be used, as (a) they do not account for the complexity of the data objects, and (b) they cannot properly deal with non-Euclidean domains. In the following subsections, we thus illustrate the computational method we propose to cope with the latter issues, that employs the quasi-stationary assumption to perform stationary analyses within each of the neighborhoods identified by a suitable random domain decomposition.

2.2 Defining partitions of the domain

We now consider the problem of generating a decomposition of the study domain, suitable for generating locally stationary analyses.

We first note that only in a few applications a unique optimal domain partition exists, and, among these, very seldom the information about it is available a priori [see, e.g., Fuentes (2001, 2002); Kim et al. (2005)]. Moreover, the

existence of a sharp partition in sub-domains [as in Kim et al. (2005)] is in contrast with local stationarity as defined above. This, in turn, needs a system of neighborhoods to be substantiated; final predictions will strongly depend on this system, which therefore becomes a crucial issue of the analysis. We here follow the intuition that, when local stationarity is assumed, the domain partition implementing the system of neighborhood defining it is auxiliary to model estimation, rather than a founding element of the model. Accordingly, we introduce, in a simple yet effective way, a system of neighborhoods generated by a random partition of the domain, driven by the domain features, and only weakly by the data.

Let us first start by setting $\mathcal{P} = \{D_1, \ldots, D_K \subset D : D = \bigcup_{k=1}^{K} D_k \text{ and } D_i \cap D_j = \emptyset, \forall i, j = 1, \ldots, K, i \neq j\}$ to be a partition of the domain D into $K \geq 1$ disjoint sub-regions. If $\mathcal{P}$ is sufficiently fine, we will take each element of $\mathcal{P}$ to be representative of a neighborhood in D where one can approximately assume stationarity. In practice, we will suppose that the following conditions hold:

1. $\mathbb{E}[\mathcal{X}_{\mathbf{s}}] = m_k$, for all $\mathbf{s} \in D_k$ and for all $k = 1, \ldots, K$;
2. $\mathbb{E}[\langle \mathcal{X}_{\mathbf{s}_1} - m_k, \mathcal{X}_{\mathbf{s}_2} - m_k \rangle] = C(\mathbf{s}_1 - \mathbf{s}_2; k)$, for all $\mathbf{s}_1, \mathbf{s}_2$ in D_k and for all $k = 1, \ldots, K$.

Since we do not assume to have a definite prior knowledge of the system of neighborhoods which provides support to local stationarity, we will use a random partition $\mathcal{P}$ to generate it. For simplicity and ease of exposition, hereafter we focus on random Voronoi tessellations of the domain D induced by a metric d. Like other modeling choices described below, this is not exclusive and is in part driven by the characteristics of the case studies we are going to illustrate in this paper. In fact, one may modify this and other modeling specifications, to accommodate a different prior knowledge guiding the analysis.

A Voronoi tessellation is defined by a set of sites (*nuclei*) and a metric function $d(\cdot, \cdot)$. In the following, the set of nuclei of the Voronoi tessellation is generated by randomly selecting K points among the n sampled locations. Other more refined sampling schemes—or even different systems of neighborhoods—can be selected according to the information available for the case study at hand; for instance, one could appeal to a non-homogeneous Poisson process, with intensity related to the prior knowledge on the stochastic process generating the data. Let $\Phi_K = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ be the set of locations of the K nuclei in D. The kth Voronoi cell is defined as

$$V(\mathbf{c}_k|\Phi_K) = \{\mathbf{s} \in D : d(\mathbf{s}, \mathbf{c}_k) \leq d(\mathbf{s}, \mathbf{c}_j), \text{ for all } \mathbf{c}_j \in \Phi_K, j \neq k\}. \quad (3)$$

The random partition $\mathcal{P}$ is then defined as $\mathcal{P} = \{V(\mathbf{c}_k|\Phi_K),\ k = 1, \ldots, K\}$.

Note that no restriction is imposed on the metric d, but we observe that different metrics produce different partitions of the same domain, even when the set of nuclei is the same. Although typical Voronoi tessellations are based on the Euclidean metric, a non-Euclidean metric might be more suitable to capture *closeness* between sites when the domain is complex in terms of boundary shape or for the presence of holes and barriers. For instance, in the example of Fig. 1, the distance between two sites lying in the tributaries should be computed as a 'water distance'—using the terminology of Rathbun (1998)—rather than via Euclidean distance. To formalize this idea, we propose to map the sampled locations on a neighborhood relational graph, properly representing the spatial adjacency of the observed objects. This allows to use a graph-based metric d to define homogeneous regions within the complex domain, while accounting for its geometrical properties. The neighborhood relational graph is generated by a triangulation of the domain D, using as vertices the set of sampled sites, $\mathbf{s}_1, \ldots, \mathbf{s}_n$. Among several possible available triangulation methods, in this work we consider the Delaunay triangulation (Hjelle and Dæhlen 2006) which is closely related to Voronoi tessellations, besides maximizing the triangles' angles. In point of fact, to account for boundaries with complex shape, holes or barriers, we use a *constrained Delaunay triangulation* (CD-T), illustrated, e.g., in Lin et al. (2013). The distance between two sites belonging to the neighborhood relational graph is then defined as the (Euclidean) length of the shortest path on the graph connecting the two sites: this is computed by the Dijkstra's algorithm (Dijkstra 1959). More generally, the distance between a site $\mathbf{s}_0 \in D$ and a site $\mathbf{s}_j$ belonging to the graph is computed by first connecting $\mathbf{s}_0$ to the closest graph vertex, and then by measuring the length of the shortest path connecting $\mathbf{s}_0$ and $\mathbf{s}_j$. This is the distance d that will be used to generate our Voronoi tessellation of the domain D.

2.3 Estimation and prediction

Given a realization of the random partition $\mathcal{P} = \{D_1, \ldots, D_K\}$, we now focus on estimation methods for the local trace-variogram, and the associated Kriging prediction. The field being locally stationary, one can employ stationary methods to estimate the trace-variogram within the cell $D_k \in \mathcal{P}$, for $k = 1, \ldots, K$. We recall that in this case, as in classical geostatistics (Cressie 1993), most methods for trace-variogram estimation consist of two stages, namely (a) computing an empirical estimate from the data, and (b) fitting a parametric valid model via least

square (LS) or maximum likelihood (ML). Within cell D_k, one may define an empirical estimator of the corresponding trace-semivariogram via the method of moments as (Menafoglio and Secchi 2017, and references therein)

$$\hat{\gamma}(\mathbf{h};k) = \frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})} \|\mathcal{X}_{\mathbf{s}_i} - \mathcal{X}_{\mathbf{s}_j}\|^2, \tag{4}$$

where $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) \in D_k \times D_k : \mathbf{h} - \Delta\mathbf{h} \leq \mathbf{s}_i - \mathbf{s}_j \leq \mathbf{h} + \Delta\mathbf{h}\}$ collects the set of pairs in D_k separated approximately by a vector $\mathbf{h}$, and $|N(\mathbf{h})|$ is its cardinality.

One should note that estimator (4) suffers from a bias-variance trade-off. Indeed, to guarantee that stationarity is a viable assumption within each cell D_k, one should strive for a fine partition $\mathcal{P}$ entailing a low bias for (4). However, fine partitions inevitably yield cells with very few data, thus inflating the variance of (4).

To cope with such trade-off, we consider a kernel-weighted empirical estimator of the local variogram, of which (4) is a particular case. We here generalize the approach proposed by Fouedjio et al. (2016) in a scalar setting. We set $K_\epsilon : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ to be a kernel function, i.e., a non-negative symmetric function, where $\epsilon > 0$ is its bandwidth parameter. An instance of such a kernel is the Gaussian kernel, defined as

$$K_\epsilon(\mathbf{s}_1, \mathbf{s}_2) = \exp\left\{-\frac{1}{2\epsilon^2} d^2(\mathbf{s}_1, \mathbf{s}_2)\right\}. \tag{5}$$

Here, the distance d is the same distance appearing in the definition of the Voronoi cells in (3). We remark that the role of the Gaussian kernel (5) is not essential; one could use indeed a different kernel, more suitable for the case under study. For instance, one may give a constant weight to all the data pairs belonging to D_k and a weight proportional to the distance from the center c_k to the others.

Given a kernel K_ϵ, we then consider the following estimator for the local trace-semivariogram

$$\hat{\gamma}_\epsilon(\mathbf{h};k) = \frac{\sum_{N(\mathbf{h})} K_\epsilon(\mathbf{c}_k, \mathbf{s}_i) K_\epsilon(\mathbf{c}_k, \mathbf{s}_j) \|\mathcal{X}_{\mathbf{s}_i} - \mathcal{X}_{\mathbf{s}_j}\|^2}{2\sum_{N(\mathbf{h})} K_\epsilon(\mathbf{c}_k, \mathbf{s}_i) K_\epsilon(\mathbf{c}_k, \mathbf{s}_j)}, \tag{6}$$

where
$N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) \in D \times D : \mathbf{h} - \Delta\mathbf{h} \leq \mathbf{s}_i - \mathbf{s}_j \leq \mathbf{h} + \Delta\mathbf{h}\}$.
Intuitively, the kernel appearing in (6) down-weights the contribution of data "far apart"—according to the metric d—from the center of the cell, where the range of influence of neighbour locations is controlled by the bandwidth parameter ϵ. Unlike estimator (4), estimator (6) allows to borrow strength from data outside the cell D_k. Whenever one may assume isotropy, (6) reads

$$\hat{\gamma}_\epsilon(\|\mathbf{h}\|_d;k) = \frac{\sum_{N(\|\mathbf{h}\|_d)} K_\epsilon(\mathbf{c}_k, \mathbf{s}_i) K_\epsilon(\mathbf{c}_k, \mathbf{s}_j) \|\mathcal{X}_{\mathbf{s}_i} - \mathcal{X}_{\mathbf{s}_j}\|^2}{2\sum_{N(\|\mathbf{h}\|_d)} K_\epsilon(\mathbf{c}_k, \mathbf{s}_i) K_\epsilon(\mathbf{c}_k, \mathbf{s}_j)} \tag{7}$$

with $N(\|\mathbf{h}\|_d) = \{(\mathbf{s}_i, \mathbf{s}_j) \in D \times D : \|\mathbf{h}\|_d - \Delta h \leq \|\mathbf{s}_i - \mathbf{s}_j\|_d \leq \|\mathbf{h}\|_d + \Delta h\}$, and $\|\mathbf{h}\|_d$ denoting the *Euclidean* norm of $\mathbf{h}$.

Note that in (7) two metrics on the domain D are considered: (1) the metric d embedded in the kernel K_ϵ and generating the random partition $\mathcal{P}$ and (2) the Euclidean metric implied by the norm argument of the trace-semivariogram. This apparent ambiguity is needed to guarantee that the commonly-employed parametric families for variogram estimation (e.g., spherical, Matérn) are valid, and thus that the geostatistical prediction yields sensible results. Indeed, this may not be the case when using a non-Euclidean metric (Huang et al. 2011; Jensen et al. 2006; Rathbun 1998, and references therein).

Having estimated the trace-semivariogram within the cell D_k according to (6) [or (7)], a parametric model can be fitted, e.g., via least squares. In the following, we shall denote by $\hat{\gamma}(\cdot;k,\epsilon)$ the fitted trace-semivariogram model for D_k, $k = 1, \ldots, K$. Each of these semivariograms is the cornerstone for the local object oriented Kriging (OOK).

Call $\mathcal{X}_{\mathbf{s}_0}$ the random element of the field $\{\mathcal{X}_\mathbf{s}, \mathbf{s} \in D\}$ at an unsampled location $\mathbf{s}_0 \in D$. Let D_k be the cell of $\mathcal{P}$ containing $\mathbf{s}_0$ and let $\hat{\gamma}(\cdot;k,\epsilon)$ be the corresponding estimated trace-semivariogram. To predict $\mathcal{X}_{\mathbf{s}_0}$ we look for the OOK predictor within the cell D_k, that is the Best Linear Unbiased Predictor (BLUP)

$$\mathcal{X}^*_{\mathbf{s}_0} = \sum_{i=1}^{n} \lambda_i^* \cdot \mathcal{X}_{\mathbf{s}_i} \mathbb{1}\{\mathbf{s}_i \in D_k\}, \tag{8}$$

where $\mathbb{1}$ is the indicator function and the weights $\lambda_i^*, \ldots, \lambda_n^* \in \mathbb{R}$ minimize

$$\mathbb{E}\left[\left\|\mathcal{X}_{\mathbf{s}_0} - \sum_{i=1}^{n} \lambda_i \cdot \mathcal{X}_{\mathbf{s}_i} \mathbb{1}\{\mathbf{s}_i \in D_k\}\right\|^2\right] \text{ subject to}$$

$$\mathbb{E}\left[\sum_{i=1}^{n} \lambda_i \cdot \mathcal{X}_{\mathbf{s}_i} \mathbb{1}\{\mathbf{s}_i \in D_k\}\right] = m_k,$$

over $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. The OOK problem can be explicitly solved through a linear system, which is the very counterpart of the classical Kriging system. We refer the reader to Menafoglio and Secchi (2017) for further details.

We remark that predictor $\mathcal{X}^*_{\mathbf{s}_0}$ is a function of the data and of the realization of the random partition $\mathcal{P}$. Indeed, in (8), data observed in sites lying outside the cell D_k get zero weight. Moreover we also note that, for a given realization of $\mathcal{P}$, the Kriging prediction of the field $\{\mathcal{X}_\mathbf{s}, \mathbf{s} \in D\}$ over the domain D could be discontinuous, with discontinuities

located at the boundary of adjacent cells of $\mathcal{P}$. However, this field prediction is only a weak auxiliary step of the overall analysis, as we shall explain in the next subsection.

2.4 Bootstrap and aggregation

We now incorporate the elements introduced in the previous subsections within a bagging algorithm (Breiman 1996) made of two steps: (1) a first *bootstrap* stage, where the same analysis is performed several times on different learning samples, and (2) a final *aggregating* stage where the weak analyses generated at step (1) are aggregated into a final strong analysis.

2.4.1 Bootstrap

At each iteration of this step: (1) the domain D is decomposed into K cells according to independent realizations of the random partition $\mathcal{P}$; (2) K local analyses are performed, one for each cell in $\mathcal{P}$. That is, for $k = 1, \ldots, K$, the trace-semivariogram is estimated in D_k as in (6), and the value of the random element $\mathcal{X}_{\mathbf{s}_0}$ at a target location $\mathbf{s}_0 \in D_k$ is predicted as in (8). At the end of B iterations of the bootstrap step, one obtains a collection of Kriging predictors $\{\mathcal{X}^{*b}_{\mathbf{s}_0}\}_{b=1}^{B}$ for each target location in D.

2.4.2 Aggregation

The Kriging predictions $\{\mathcal{X}^{*b}_{\mathbf{s}_0}\}_{b=1}^{B}$ at the target site $\mathbf{s}_0$ need eventually to be aggregated into a final prediction $\mathcal{X}^{*}_{\mathbf{s}_0}$. To this end, one may employ the average of the predictors obtained along the bootstrap iterations, $\mathcal{X}^{*}_{\mathbf{s}_0} = \frac{1}{B}\sum_{b=1}^{B} \mathcal{X}^{*b}_{\mathbf{s}_0}$. Note that the final predictor $\mathcal{X}^{*}_{\mathbf{s}_0}$ will depend not only on the data, but also on the B independent realizations of the random partition $\mathcal{P}$. Indeed $\mathcal{X}^{*}_{\mathbf{s}_0}$ is the sample version of

$$\mathbb{E}_{\mathcal{P}}[\mathcal{X}^{*}_{\mathbf{s}_0}(\mathcal{X}_{s_1}, \ldots, \mathcal{X}_{s_n}; \mathcal{P})] = \int_{\mathfrak{P}} \mathcal{X}^{*}_{\mathbf{s}_0}(\mathcal{X}; p) d\mu_{\mathcal{P}}(p)$$

where $\mathfrak{P}$ is the set of all possible partitions of the domain, and $\mu_{\mathcal{P}}$ the distribution over $\mathfrak{P}$ of $\mathcal{P}$. One can readily envision other types of aggregations, e.g., weighted averaged kernel-based, or based on the Kriging variance. Figure 2 contains a pseudo-code summarizing our proposal.

As noted in Sect. 2.3, each *weak* Kriging map may have discontinuities at the boundary between adjacent cells. Being randomly generated, at each bootstrap iteration of the algorithm these boundaries are differently arranged in the space domain. In the final aggregation step, each weak map contributes to the *strong* map with a small weight and this is also true for the associated discontinuities due to the cells boundaries. The possible discontinuities of the weak maps are thus eventually smoothed away during the aggregation step. We illustrate this point through the simulation study of Sect. 3.

2.5 On the model parameters

The algorithm summarized in Fig. 2 requires to initialize a few parameters: the number of auxiliary analyses B, the number K of cells of the random partition $\mathcal{P}$ of D, the kernel K_ϵ and its bandwidth ϵ.

Conditionally on computational time, the parameter B should be chosen as large as possible to ensure that the algorithm reaches a desired accuracy. It controls the robustness of the final result: the higher the number of B *weak* analyses performed, the *stronger* the basis upon which the final result is obtained.

Object Oriented Kriging via RDD

Initialization.
Set the parameters $1 \leq K \leq n$, $B \geq 1$, a kernel K_ϵ with its bandwidth ϵ, a valid variogram model, a metric d for the spatial domain D and the target location $\mathbf{s}_0$.

Bootstrap step.
for $b := 1$ to B do
Step 1. Draw a realization of $\mathcal{P}$.
Randomly generate a set of nuclei $\Phi_K = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ among the observed sites $\mathbf{s}_1, \ldots, \mathbf{s}_n \in D$; define the Voronoi cells $\{V(\mathbf{c}_k|\Phi_K)\}_{k=1}^{K}$ by assigning each site $\mathbf{s}$ to the nearest nucleus $\mathbf{c}_k$, according to the metric d.
Step 2. For each Voronoi cell D_k: estimate the semivariogram $\hat{\gamma}_\epsilon(\mathbf{h}; k)$, by means of (6) or (7); fit the parametric valid model to the empirical estimate and obtain $\hat{\gamma}(\cdot; k, \epsilon)$.
Step 3. For $\mathbf{s}_0 \in D_k$, obtain the OOK prediction $\mathcal{X}^{*b}_{\mathbf{s}_0}$, as in (8).
end for.

Aggregation step.
For $\mathbf{s}_0 \in D$, compute the final prediction (RDD-OOK predictor) by aggregating the B predictions as their average $\mathcal{X}^{*}_{\mathbf{s}_0} = \frac{1}{N} \cdot \sum_{b=1}^{B} \mathcal{X}^{*b}_{\mathbf{s}_0}$.

Fig. 2 Pseudocode scheme of the algorithm for object oriented Kriging through random domain decomposition (RDD-OOK)

The parameter K should be carefully evaluated since it has a great influence on the algorithm performances. Indeed, K affects the Kriging bias-variance trade-off: if K is small, the predictor at an unsampled location will be based on large sub-samples. This tends to minimize the variance of the Kriging predictor, but also to increase its bias since local stationarity within each of the few cells of the partition $\mathcal{P}$ may not be verified. The limiting case is $K = 1$: here we would assume stationarity over the whole domain, although we might not even be able to formulate a clear assumption of stationarity due to, e.g., domain complexities, such as holes or irregular boundaries. On the other hand, if K increases, the partition of the domain will become more and more refined, being able to accurately define the boundaries of different homogeneous sub-regions. This tends to minimize bias; at the same time, the sample size pertaining to each cell of the partition will decrease with the effect of increasing the variance of the Kriging predictor. The limiting case is $K = n$, when the prediction of $\mathcal{X}_{\mathbf{s}_0}$ is based on a single observation, the closest datum to the location $\mathbf{s}_0$.

As mentioned before, we can include within the general RDD methodology, a geographically weighted approach by using a kernel function. This is an optional choice which contributes to the flexibility and robustness of our proposal. However, the use of a kernel function for estimating the local semivariograms can be unimportant in some cases, e.g, when large sample sizes in each cell of the partition $\mathcal{P}$ are guaranteed. Using a kernel may become necessary if one has very few observations, since each local analysis would be then performed on a subset of the data of very small size. In the numerical illustrations shown in this work, we consider Gaussian kernels; nevertheless, other kernels are possible, and their choice should be driven by prior knowledge. As regards the bandwidth parameter ϵ, it controls the range of influence of the observations on the estimate of the local semivariogram. Once again, a small value of ϵ implies a variogram estimate based on too few observations, and therefore highly uncertain, while a large value of ϵ assigns considerable weights to observations very far away from the cell of interest, against the assumption of local stationarity.

To avoid limiting cases in the random generation of the partitions, one may consider to select only partitions which guarantee a minimum number of observations within each cell. For instance, in the simulation study illustrated in the following section, we set the threshold to $n_k = 5$ data (i.e., in case one or more cells of the partition contain less then n_k data, the partition is discarded and a new one is randomly generated). However, it should be noted that all the data pairs are used to build the kernel-weighted variogram estimator, allowing for more reliable and robust estimates even in elements of the partition with a very limited number of data points. Here, the number of data affects mostly the uncertainty of the Kriging predictor at a given iteration of the algorithm. Note that the contribution of each single *weak* prediction to the final *strong* one is small when compared to the overall weight of the ensemble of predictions produced by the B bootstrap iterations of the algorithm. In fact, a site belonging to a cell with a small number of observations at a given iteration, most likely will belong to a cell with a higher number of observations in other iterations.

When a Euclidean domain is concerned, a sensible working strategy may consist of (1) performing a global analysis ($K = 1$) to provide initial evidence of non-stationarity and anisotropy, and (2) using local models at a spatial scale compatible with stationarity assumption. The latter assumption can be assessed—at a given spatial scale—via standard tools (e.g., variography) or via indices of heterogeneity, such as the indicator of spatial stratified heterogeneity proposed by Wang et al. (2016). The latter working strategy may not be applicable in the general framework of complex domains, as one should pay close attention in the interpretation of a possible global model estimated for spatial dependence. Indeed, valid covariance models for general non-Euclidean domains are yet to be developed, and they are likely to be strongly dependent on the very specific type of domain under study. In all these cases, one might not be able to assess the assumption of stationarity on a global scale, since even its definition becomes somehow problematic. Instead, it is more difficult to deny the appropriateness of the usual Euclidean assumptions when stated at the local level and it is them which drive the local analyses. In this sense, the *locality* of the analysis—controlled by the parameters K and ϵ—should attain a balance between our ability to (1) fairly represent the domain, (2) assume local stationarity, and (3) estimate the local model (i.e., enough data). To support the choice of the parameters, one may also consider data-driven methods, e.g., cross-validation.

3 Simulation study

In this section, we explore, through a simulated example, the performances of our random domain decomposition approach for Kriging object data (hereby named RDD-OOK) on complex domains. For ease of exposition and representation of the results, we here focus on the case of scalar data. In Sect. 4 we will illustrate an application to the analysis of more complex object data, in the form of distributional data. An additional simulated example is provided in the supplementary material.

3.1 A locally stationary field over a complex domain

This simulation example concerns a spatial random field distributed over a C-shaped domain. Figure 3 displays the test function, which is the same test function used by Sangalli et al. (2013). We observe the test function in $n = 250$ locations randomly and uniformly selected within the domain.

Embedding the C-shaped domain into a larger rectangular domain and then assuming global spatial stationarity defined in terms of the Euclidean metric does not seem to be a suitable modeling approach, because of the apparent drift in the data and due to the shape of the domain, characterized by the presence of the thin space which separates the two branches of the C. Nonetheless, local stationarity holds thanks to the smooth variation of the field. Note however that the Euclidean distance is not appropriate as a measure of adjacency, because it crosses over the space separating the two branches of the C. Hence, as detailed in Sect. 2.2, a constrained Delaunay triangulation was built to represent the domain (Fig. 3) and, on this basis, a graph-based metric d was defined.

For the simulations, we employed a Gaussian kernel with bandwidth $\epsilon = 1$. We remark that this kernel is isotropic with respect to d, but anisotropic with respect to the Euclidean metric. Figure 4 displays an example of kernel weights for both the Euclidean metric and the graph-based metric. One can appreciate that points on one branch do not contribute to the estimate of variograms for the other branch, when d is used.

At each bootstrap iteration, a realization of a random Voronoi partition of the domain was generated according to d, and for each cell of the partition the empirical variogram was estimated and fitted with an exponential model with nugget, as detailed in Sect. 2.3. Given the fitted variograms, OOK was performed on a fine grid D_0 of points covering the C.

Figure 5 reports the results obtained for a particular bootstrap iteration, when the number of Voronoi cells is $K = 16$. Figure 5a displays the realization of the RDD, Fig. 5b–d the estimated variogram parameters for each Voronoi cell, Fig. 5e the weak Kriging map and Fig. 5f the map of Kriging variances. Figures 5b–d show that the variograms estimated within each subregions are indeed different. The estimated variogram parameters indicate that the area where the C displays its bend is the one characterized by the most complex spatial structure, as demonstrated by the high values for the sills and the practical ranges. All the estimated variograms are characterized by much lower sills than that estimated from a global Euclidean model, the latter having sill $\sigma^2 = 385.89$, practical range $R = 163.38$ and nugget $\tau^2 = 0$. This preliminary analysis suggests that the use of a small number of Voronoi cells might call into question the viability of the local stationary assumption within the cells. Finally, from the graphical inspection of Fig. 5b one may notice that, as expected, the weak Kriging map appears discontinuous at the boundary between subregions.

The number B of bootstrap iterations was set equal to 100, and the results aggregated through a simple average. Figure 6a displays RDD-OOK predictions obtained when the number of Voronoi cells is $K = 16$. For comparison, Fig. 6b reports the OOK predictions when the analysis is performed without the perturbation introduced by the RDD (i.e., RDD-OOK setting $K = 1$) and measuring distances in the C through the Euclidean metric. Figure 6c shows predictions obtained by using RDD-OOK and a random Voronoi partition of the domain with $K = 16$ cells, but based on the Euclidean metric to measure distances in the C instead of d. Graphical inspection of Fig. 6 suggests that the use of a graph-based metric generates better predictions, especially in the central part of the C where the RDD-OOK is not induced to borrow information from the wrong branch of the domain, as it happens when the Euclidean metric is in force. Further, no apparent

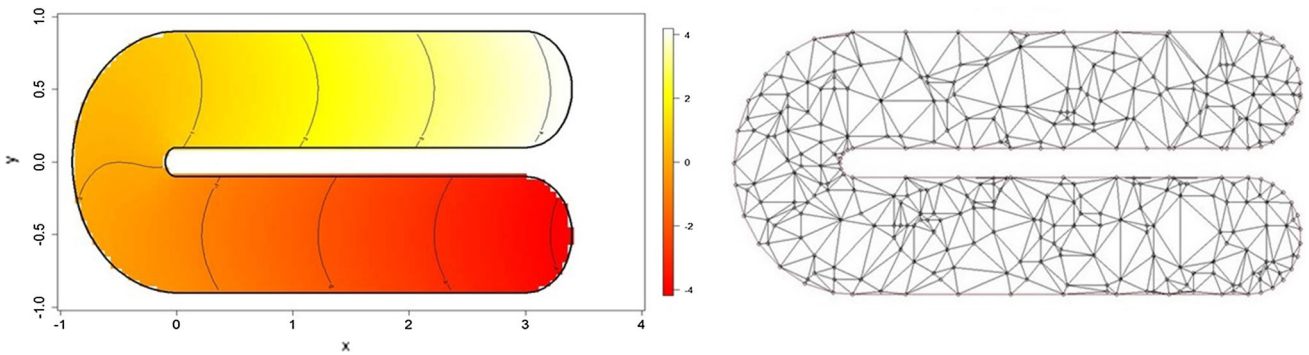

Fig. 3 C-shaped domain: smooth test function and Delaunay triangulation

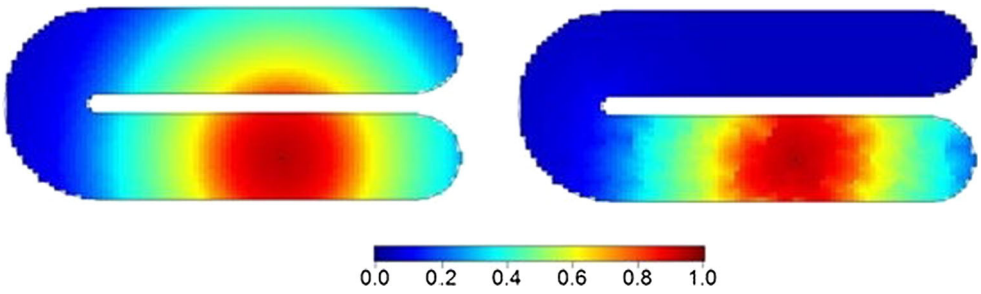

Fig. 4 C-shaped domain: example of kernel weights for different metric $d(\cdot,\cdot)$. For a fine grid of points $\mathbf{s}_0$ in D, the plots refer to the value of the kernel $K_\epsilon(\mathbf{s}_0,\mathbf{c})$, where $\mathbf{c}$ is the center (symbol) and $\epsilon = 1$

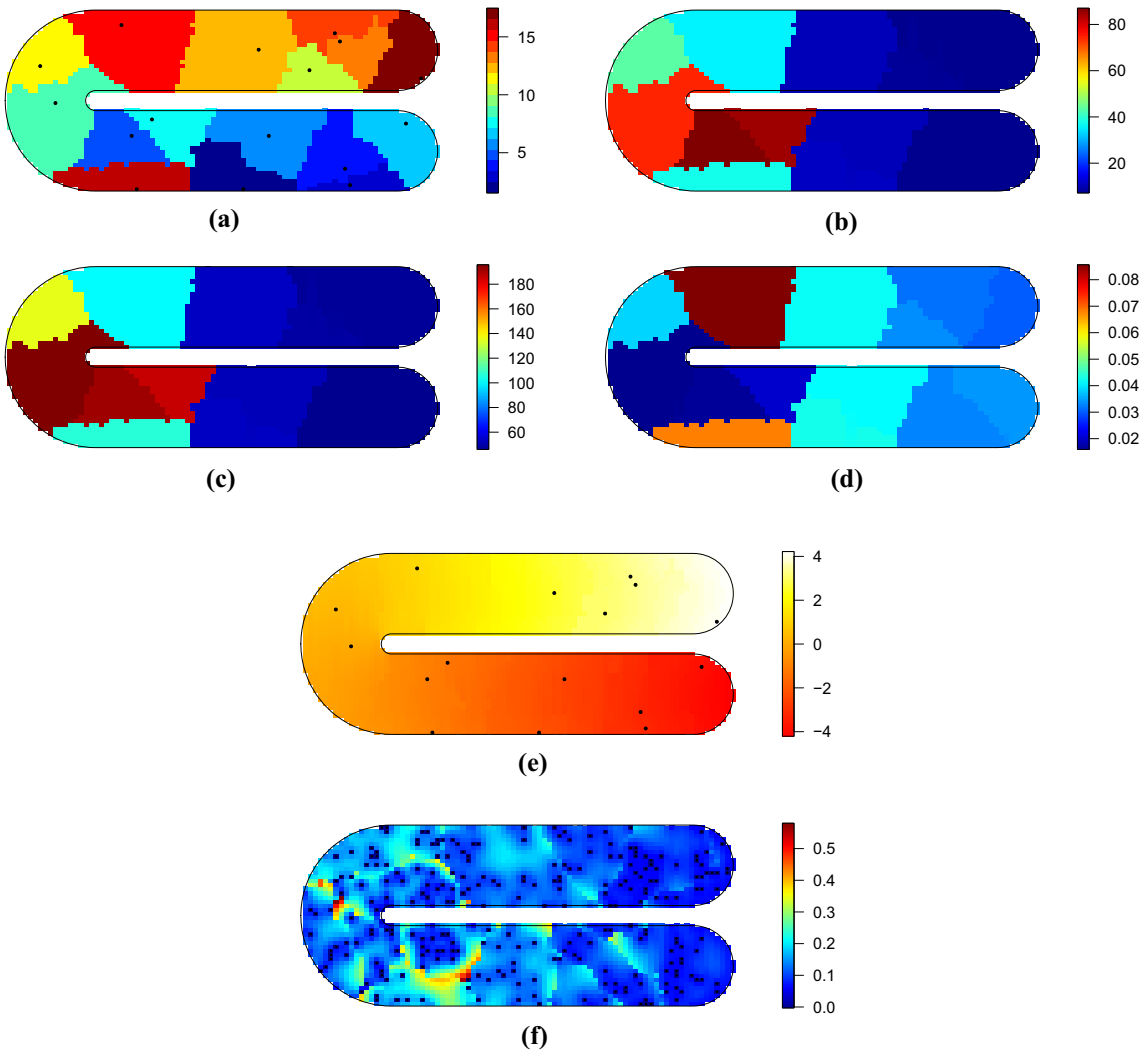

Fig. 5 C-shaped domain: example of results obtained at a bootstrap iteration of the algorithm. In **a**, **e** black dots indicate the position of the Voronoi nuclei c_k, $k = 1,\ldots,16$. In **f** black dots indicate the data locations. **a** Realization of the RDD. **b** Estimated sill. **c** Estimated practical range. **d** Estimated nugget. **e** Weak Kriging map. **f** Weak Kriging variance map

discontinuities are visible in the RDD-OOK map, as these are smoothed away during the aggregation step.

To better illustrate the effect of the parameter B over the final prediction, Fig. 7 displays the prediction error $\mathcal{X}^*_{\mathbf{s}_0} - \mathcal{X}_{\mathbf{s}_0}$ of the graph-based RDD-OOK when the parameter B is set to 1, 10, 100 (Fig. 7a, b, c, respectively). One may notice that the errors appear overall larger when only one realization of the RDD is considered ($B = 1$, SD = 0.04—Fig. 7a), and they are discontinuous at the boundary between subdomains. When increasing B to $B = 100$, the errors decrease (SD = 0.015—Fig. 7c), as the final Kriging

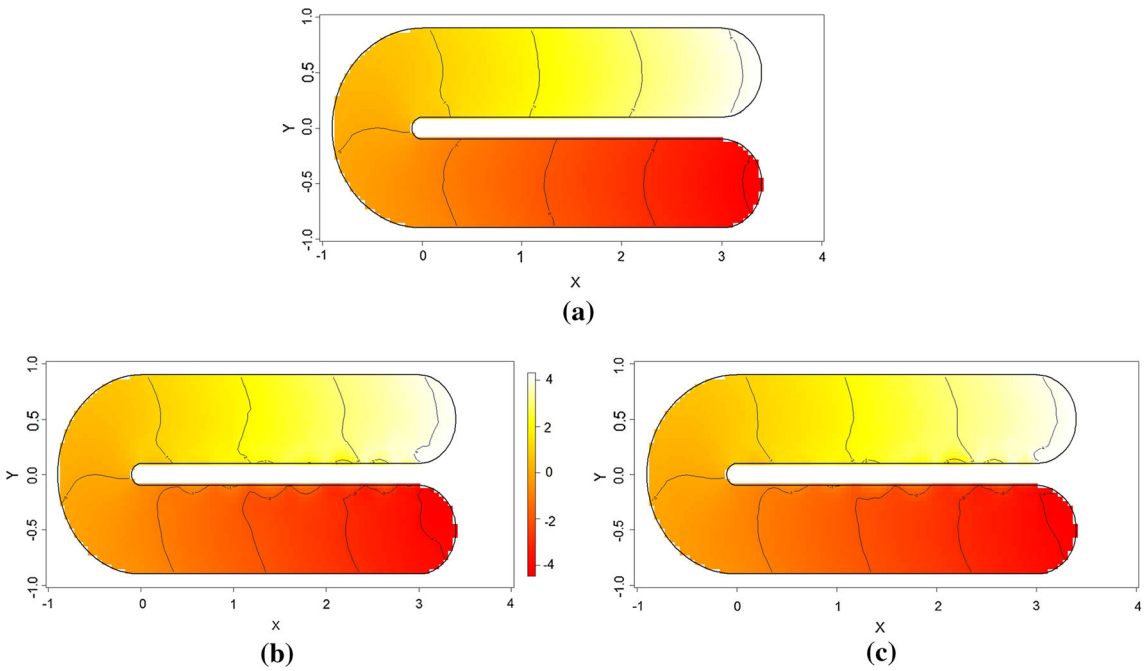

Fig. 6 C-shaped domain: Prediction of the field. **a** RDD-OOK, graph-based distance ($K = 16$). **b** RDD-OOK, Euclidean distance ($K = 1$). **c** RDD-OOK, Euclidean distance ($K = 16$)

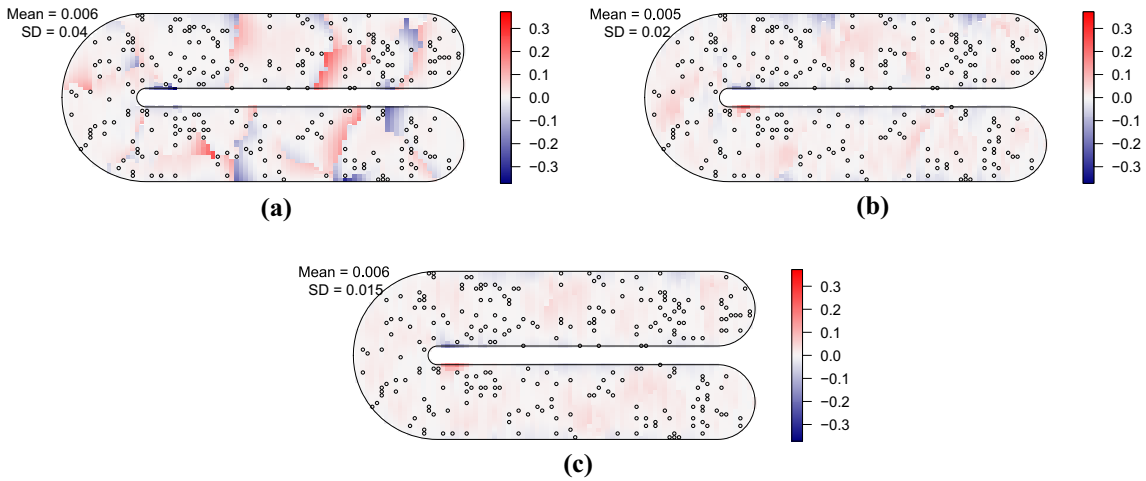

Fig. 7 C-shaped domain: prediction error $\mathcal{X}^*_{\mathbf{s}_0} - \mathcal{X}_{\mathbf{s}_0}$ of the RDD-OOK for B = 1, 10, 100, with K = 16. Dots indicate the data locations. **a** Prediction error with B = 1. **b** Prediction error with B = 10. **c** Prediction error with B = 100

map is built upon a larger ensemble of bootstrap repetitions.

To test the performance of the method, we repeated the analysis $M = 50$ times, for different values of K in $\{1, 2, 4, 8, 16\}$, $K = 1$ meaning object oriented Kriging with no random domain partition and Euclidean metric for measuring distances in the C. Note that, in the latter case, the kernel is not used for the estimation of the only variogram involved in the analysis. At each of the M repetitions of the analysis, a different set of $n = 250$ of locations was uniformly sampled out of the C and the corresponding value of the field observed. To compare the performances of the method for different parameter settings and metrics, for $m = 1, \ldots, M$, we computed the (relative) mean square prediction error (MSPE), defined as

$$\mathrm{MSPE}_{\mathrm{m}} = \frac{\sum_{\mathbf{s}_0 \in \mathcal{G}} \|\mathcal{X}^*_{\mathbf{s}_0} - \mathcal{X}_{\mathbf{s}_0}\|^2}{\sum_{\mathbf{s}_0 \in \mathcal{G}} \|\mathcal{X}_{\mathbf{s}_0}\|^2} \tag{9}$$

where $\mathcal{G}$ is the set of target points in D_0, and $\|\cdot\|$ denotes the norm on $\mathbb{R}$ (i.e., the absolute value). Results are reported in Table 1 and Fig. 8 which show the improvements in predictions obtained by increasing the value of

Table 1 C-shaped domain. RDD-OOK predictions with different distances: average of the MSPE over the $M = 50$ repetitions

Distance	K				
	1	2	4	8	16
Euclidean	3.287×10^{-3}	3.365×10^{-3}	3.307×10^{-3}	3.207×10^{-3}	3.029×10^{-3}
Graph-based	3.287×10^{-3}	0.4807×10^{-3}	0.1168×10^{-3}	0.0603×10^{-3}	0.0551×10^{-3}

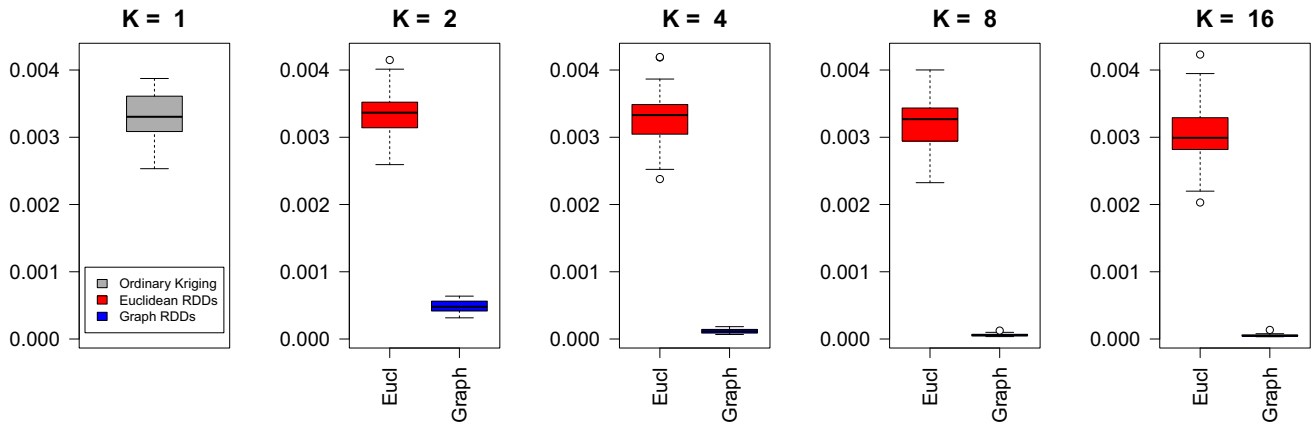

Fig. 8 C-shaped domain: MSE boxplots for different values of K in $\{1, 2, 4, 8, 16\}$ ($K = 1$ meaning ordinary Kriging) and for the Euclidean and graph-based distance

K and by moving from the Euclidean metric to the graph-based metric d.

A second simulated example is provided as supplementary material. It is aimed to test the performance of the method when its hypotheses are not met, based on the model of Kim et al. (2005). Those simulations suggest that a major role in evaluating the validity of the model assumptions—particularly the local-stationarity—is played by the *bootstrap* variance, defined, for any grid point s_0 in target set D_0, as

$$\sigma_B^2 = \frac{1}{B}\sum_{b=1}^{B} \|\mathcal{X}_{\mathbf{s}_0}^{*b} - \mathcal{X}_{\mathbf{s}_0}^{*}\|^2, \tag{10}$$

where $\mathcal{X}_{\mathbf{s}_0}^{*b}$ is the prediction at $\mathbf{s}_0$ at the bth iteration, $\mathcal{X}_{\mathbf{s}_0}^{*}$ is the final prediction obtained by aggregation of the B bootstrap replicates, and $\|\cdot\|$ represents the norm on the feature space. Intuitively, a large value of the bootstrap variance indicates large deviations of the predictions obtained along the B repetitions from the final one. Instabilities in the map of the bootstrap variance may indicate a violation of the above-mentioned assumptions, and thus serve as a driver, e.g., in the definition of the RDD (for further details, see the supplementary material).

4 A case study: analysis of the distribution of dissolved oxygen in the Chesapeake Bay

4.1 Problem setting

The Chesapeake Bay is the largest estuary in the United States and the third largest in the world. This estuarine ecosystem is approximately 300 km long, from Havre de Grace, Maryland (on the North) to Virginia Beach, Virginia (on the South). Its width ranges between 5 km (the mean width of the mainstream) and 30 km, if one considers the lateral tributaries. The total shoreline, including tributaries, is 18,804 km long, and circumnavigates a surface area of 11,601 km^2.

The Bay is one of the most productive and complex ecosystems in the US, besides being a very important economic resource for the zone. The extreme use of the land around the estuary and, in particular, the pollution due to the close farms and cities, changed the Bay over the years. Human activities caused a drastic reduction of oxygen, which must be present underwater, in dissolved form, to guarantee the life of most marine species. The most critical areas of the Bay—i.e., those with the lowest values of dissolved oxygen (DO)—are called *Dead zones*. These are the areas of the estuary where the presence of oxygen in the water is below 2 mg/l. In these areas most of the marine species cannot move quickly enough and, consequently, they usually suffocate.

The Bay's degradation problem motivated the constitution of the *Chesapeake Bay Program* (CBP) in 1983, which is a regional partnership aimed to provide a support for the restoration and protection activities for the Bay. Monitoring DO is crucial for the purpose of determining the areas that deserve more attention. For this reason, the values of DO are collected at monitoring stations in the Bay, on a regular basis. Nonetheless, such observations provide only a partial picture of the distribution of DO in the Bay. As such, its spatial prediction is of key importance.

4.2 The data

We consider the DO data at the 110 measurement locations in the Bay for which data are available along the period 1990–2006 [source: US Environmental Protection Agency Chesapeake Bay Program (US EPA-CBP)]. Note that the spatial sample size is relatively small compared to the covered area, which represents a critical issue when applying local models. Figure 9 shows the mean of the summer values of DO, collected at the 110 measurement locations in the Bay for which data are available along the period 1990–2006. Figure 9a displays the sampling scheme, while Fig. 9b represents the average values of DO recorded during the summer season of each year.

A preliminary analysis of the data showed that no significant autocorrelation exists, along the years, for the time series of DO (level 1%, result obtained through a Durbin-Watson test on each time series, the p value of single tests being corrected via Holm's method). Further, no evident trend is displayed by the observations (Fig. 9b).

We here consider as data objects the probability density functions of DO in the sampling locations. Considering the whole information content provided by the distribution of DO allows one to provide predictions not only of some selected data features (e.g., a few moments or quantiles), but of all the moments and quantiles jointly, as well as the probability of events of interest (e.g., observing a DO lower than the attention limit of 2 mg/l). Note that the joint analysis of multiple quantiles would require the construction of a model for a vector of ordered components—due to the ordering of quantiles—which is highly non-trivial. For the sake of brevity, in the following we limit to show the results in terms of selected features (mean or median). Additional plots related with further quantiles of the distribution are provided in the supplementary material.

4.3 A feature space for PDFs

Probability density functions (PDFs) are an instance of data objects which can be analyzed in the setting of O2S2 through the embedding within an appropriate feature space. Several authors (Egozcue et al. 2006; Delicado 2011; Hron et al. 2016; Menafoglio et al. 2014, 2016a, b) suggested that PDFs can be considered as the generalization to the functional setting of multivariate compositional data, i.e., vectors whose components represent parts of a given total (e.g., 1 or 100, if proportion or percentages are considered). A Bayes Hilbert space (van den Boogaart et al. 2014) is the natural feature space for PDFs, as it was precisely built as a generalization to infinite-dimension of the Aitchison geometry for multivariate compositions (Pawlowsky-Glahn and Egozcue 2001).

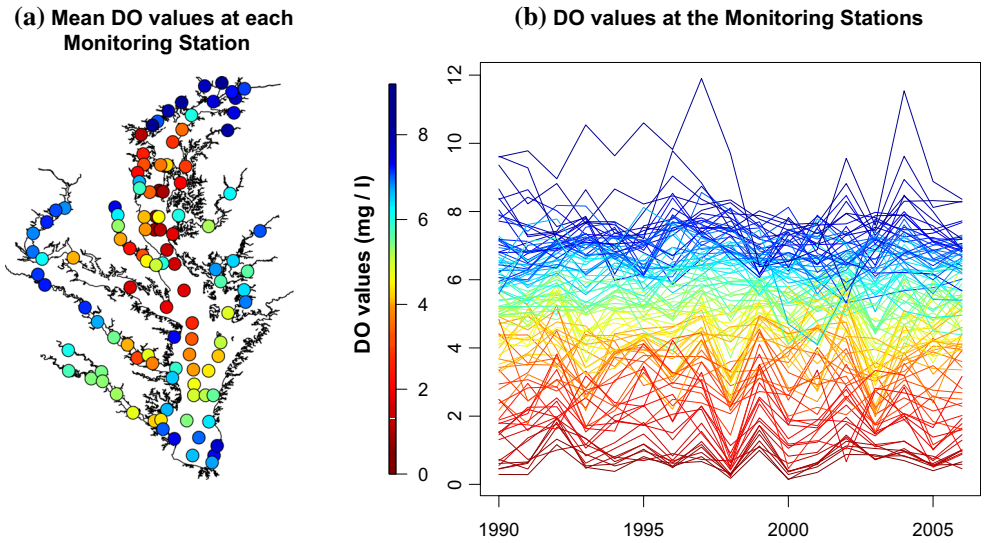

Fig. 9 Data at the Chesapeake Bay. **a** Scheme of the sampling locations. **b** Average of values of DO recorded during the summer season of each year between 1990 and 2006. In both panels, colors are given according to the mean of DO values along the years 1990–2006

The Bayes Hilbert space $\mathcal{B}^2(I)$ is the space of real valued positive functions on $I \subset \mathbb{R}$, whose logarithm is squared-integrable, i.e.,

$$\mathcal{B}^2(I) = \{f : I \to (0, +\infty), \int_I \log[f(\tau)]^2 d\tau < \infty\}.$$

In $\mathcal{B}^2$, two functions are equivalent if they are proportional, i.e., $f \sim g$ if $f = \alpha g$ for $\alpha > 0$, f, g in $\mathcal{B}^2$. The theory of the Bayes space $\mathcal{B}^2$ is well developed, and interesting interpretations of its geometric structure are given in the literature. We here limit to mention the geometric structure of $\mathcal{B}^2$—which shall be used in the present case study—and refer the interested reader to Egozcue et al. (2013); Hron et al. (2016) and references therein for further details. The space $\mathcal{B}^2$ can be equipped with a separable Hilbert structure, when endowed with the appropriate operations and inner product. The operations $(+, \cdot)$ in this setting are named *perturbation* and *powering*, and defined respectively as (Egozcue et al. 2006; van den Boogaart et al. 2014):

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

The inner product is defined as (Egozcue et al. 2006)

$$\langle f, g\rangle_{\mathcal{B}}^2 = \frac{1}{2\eta}\int_I \int_I \ln\frac{f(t)}{f(s)}\ln\frac{g(t)}{g(s)}\,\mathrm{d}t\,\mathrm{d}s, \quad f, g \in \mathcal{B}^2(I). \tag{11}$$

Having embedded a dataset of PDFs in this space, one can perform the analysis as detailed in Sect. 2. Nonetheless, observations are rarely given in the form of smoothed PDFs, hence pre-processing of the data is usually needed. To estimate the smooth PDFs of DO from available data we followed the approach proposed by Machalová et al. (2016). These authors developed a constrained B-spline basis in $\mathcal{B}^2$ (to fulfill the integral constraint), and a smoothing procedure for histogram data, consistent with the geometry of $\mathcal{B}^2$. At each measurement station, we thus used the 17 DO values to compute a histogram, which was built by considering 15 equally spaced classes partitioning the common domain $I = [0, 11.95]$. At each location, the number of non-empty classes ranged from 3 to 6, in agreement with the Sturges' rule. To cope with the problem of classes with zero frequency—which represent an issue when using a compositional approach—we followed the Bayesian-multiplicative strategy advocated by Machalová et al. (2016); Martín-Fernández et al. (2015) and based on a uniform prior. Each pre-processed histogram was then smoothed by using a B-spline basis of order 2, with 13 equally spaced knots and smoothing parameter $\alpha = 0.98$. In the smoothing procedure we accounted for the pre-processing of the zero frequency classes by downweighting their influence to one fifth of the weight of the those with positive frequency. All the above-mentioned parameters were set as to guarantee a good fitting to the data, yet avoiding overfitting. The smoothed data are displayed in Fig. 10.

We finally remark that one may consider transformations of DO concentrations (e.g., a log-transformation) which could better reflect the possible relative scale of the data while symmetrizing their distributions. However, this would introduce methodological issues related to the use of the Bayes space geometry in the presence of (1) unbounded domains and (2) disjoint PDFs supports (problems of zeros). Facing these issues would require further technical arguments that are outside the scope of this work. For these reasons, in the following we consider the DO values on their original scale.

4.4 Modeling the domain and its metric

We note that the highly irregular boundaries of the Bay prevent the use of a globally stationary model, and of the Euclidean metric on the domain. Indeed, even if few kilometres separate two points lying on adjacent and *parallel* tributaries, long and narrow areas of land could be separating them (i.e., a high *water distance*). These land areas represent barriers for the distribution of many aquatic variables, as was also recognized by Jensen et al. (2006), who analyzed the distribution of blue crab in the Bay.

To model the *water distance* between points of the estuary, we thus built a graph-based metric, using a constrained Delaunay triangulation of the domain (Fig. 10c). For illustrative purposes, we employed a simplified description of the Bay's border in order to define its triangulation; the use of more refined meshes and its impact on the prediction could be the scope of future work. The boundaries were thus defined through straight segments, able to approximate the real boundary of the estuary, with particular emphasis on the main tributaries and on the land areas separating them. The short and tight channels outgoing the tributaries and the central unit of the Bay were partially neglected. We note that the definition of the boundary required the addition of vertices appearing only for the Delaunay triangulation, but not accounted for in the analysis (red symbols in Fig. 10c). Additional points (empty symbols in Fig. 10c) were added for the purpose of refining the quality of the constrained Delaunay triangulation.

4.5 Prediction results

We applied the procedure detailed in Sect. 2, by setting the number of bootstrap iterations to $B = 150$, and the

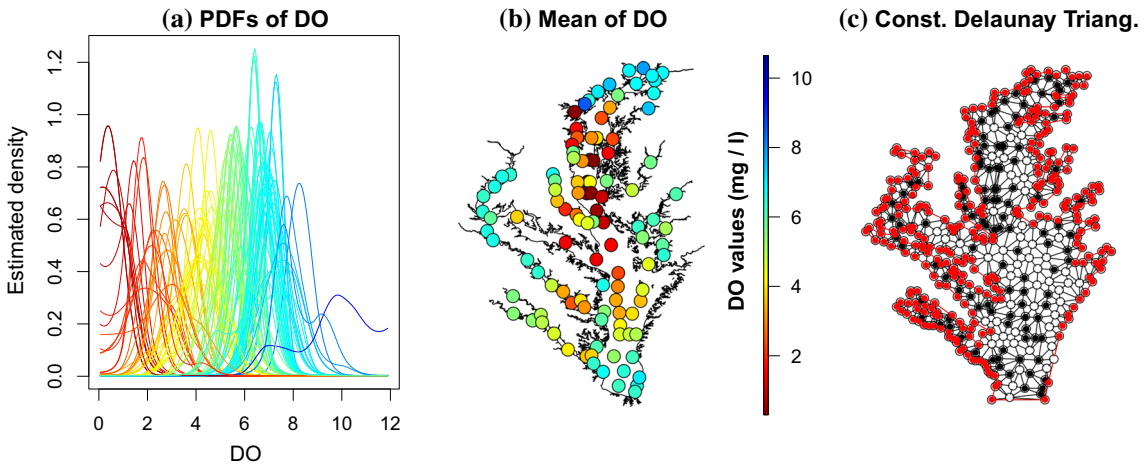

Fig. 10 Smoothed data and domain representation. **a**, **b** Smoothed PDFs and mean of DO at the sampled locations; colors given on the same scale. **c** Constrained Delaunay triangulation of the domain; red symbols indicate points defining the domain boundary, black symbols the measurement locations, empty symbols the additional points used to build the triangulation

bandwidth parameter to $\epsilon = 0.75$. The latter value was chosen as to balance the trade-off between the locality of the variogram estimate and its stability along the realization of the RDD. The analysis was performed for values of K in $\{1, 2, 4, 8, 16\}$, $K = 1$ representing the case of object oriented Kriging under a globally stationary model based on the Euclidean distance. We here focus on the results obtained for $K = 1$ and $K = 16$.

Figure 11 reports the predicted medians using RDD-OOK with a Voronoi random domain decomposition with $K = 16$ cells (Fig. 11a) and using RDD-OOK when $K = 1$ (Fig. 11b). Although we recognize a general agreement in the predicted patterns, the case $K = 16$ is characterized by more localized features, compatible with the peculiar morphology of the domain. This is more evident in the central part of the main branch of the estuary and in its main left tributary. Similar patterns are visible from the maps of other quantiles, provided in the supplementary material.

With regard to the monitoring program of the Bay, of particular interest are the *Dead zones*. Figure 12 represents the predicted probability p of observing a DO value below

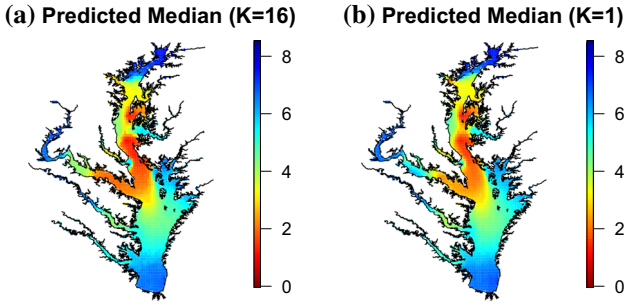

Fig. 11 Medians of the predicted distribution, by using the RDD-OOK with $K = 16$ or $K = 1$

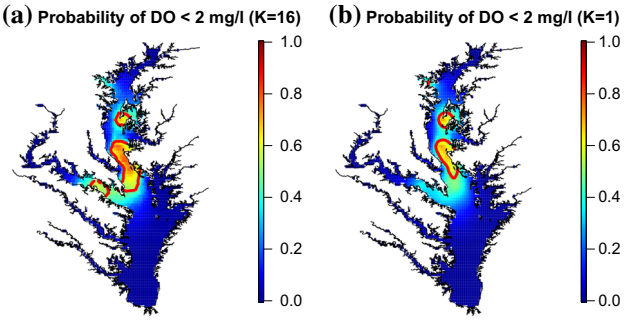

Fig. 12 Probability of observing a DO value below the attention limit of 2 mg/l. The contour line $p = 0.5$ is indicated by the thick solid line

the attention limit of 2 mg/l, i.e., of being a *Dead zone*. Figure 12a represents the predictions obtained through RDD-OOK when $K = 16$, Figure 12b when $K = 1$. For ease of comparison between the probability maps, Fig. 12 also reports the contour line at $p = 0.5$. The latter may be used to identify the dead zones as those regions with probability $p > 0.5$ of observing DO < 2 mg/l. When $K = 16$ these areas appear larger, and localized not only in the main branch, but also in the main left tributary. In fact, the values in the tributaries are likely to suffer from the smoothing effect of neighboring locations when the Euclidean distance is in force. Such effect is instead partially mitigated by the use of a random domain decomposition approach. The supplementary material contains a cross-validation study to support the RDD-OOK results. The study shows in particular that the predictions obtained via RDD-OOK in the tributaries tend to be slightly more accurate as K increases.

Finally, Fig. 13 reports the bootstrap variance [as defined in (10)], and the aggregated Kriging variance, defined as the average, along the bootstrap replicates, of the

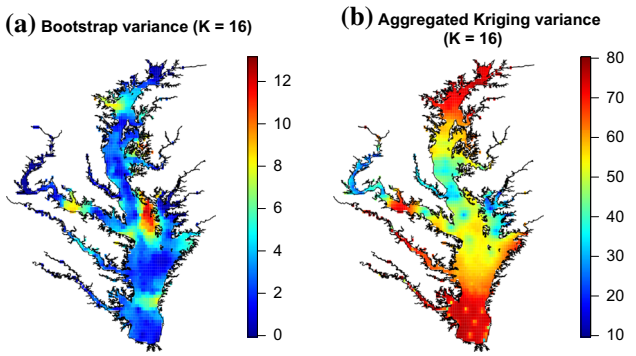

Fig. 13 Bootstrap and Kriging variances for RDD-OOK

OOK variance, i.e., $\sigma^2(\mathbf{s}_0) = \frac{1}{B}\sum_{b=1}^{B} \sigma_{OK}^{2,b}(\mathbf{s}_0)$, with $\sigma_{OK}^{2,b}(\mathbf{s}_0)$ being the Kriging variance at the target location $\mathbf{s}_0$, at the bth iteration, when $K = 16$. For completeness, we also report the Kriging variance obtained when a global stationary model is assumed and the Euclidean distance is used (i.e., for $K = 1$) (Fig. 14).

We note from Fig. 13 that the highest uncertainty is associated with the areas in the central left tributary, where only few observations are available, and at the conjunction between the latter and the main branch. This latter area displays a high local variability, and is located in a region of the domain with a high degree of non-convexity. Regarding Fig. 13b, we note that the Kriging variances do not appear to be homogeneous in the region. This suggests that the local variogram estimated from the data are indeed different in different locations. As a way of example, Fig. 15 reports the variogram parameters estimated for an iteration of the bootstrap step. It can be appreciated that the estimated variograms appear to be different, with particular reference to the areas in the Northern part of the Bay and in its central part. These areas seem to be associated with high values of sills and nuggets across the bootstrap repetitions, providing indication of areas of strong spatial variability.

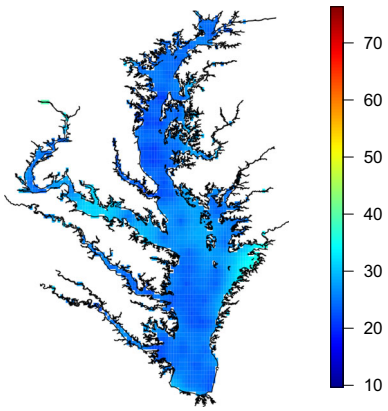

Fig. 14 Kriging variances for OOK (i.e., RDD-OOK, with $K = 1$). The color scale is the same as in Fig. 13b

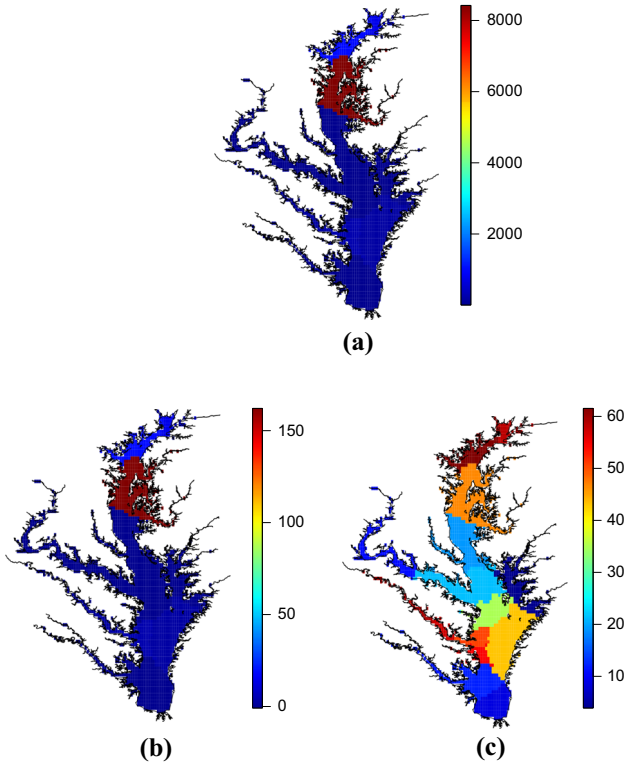

Fig. 15 Estimated variogram parameters for a bootstrap iteration. **a** Estimated sill. **b** Estimated range. **c** Estimated nugget

5 Conclusions and discussion

We proposed a methodology for the analysis of spatial random fields of object data, when the use of a global model for the observations is not appropriate either because of data non-stationarities or due to domain complexities. Our approach is based upon the idea of repeatedly and randomly partitioning the domain—through a random domain decomposition (RDD)—and then accordingly estimating multiple locally stationary models, instead of a unique and globally non-stationary one.

Although RDD is here considered for the purpose of performing Kriging predictions (RDD-OOK), the approach is entirely general and may be employed for different types of analysis, e.g., classification or estimation of a drift. Here, the model for the local mean may be extended to account for possible (scalar or object) covariates. For instance, for the study of dissolved oxygen one may consider the auxiliary information provided by the water temperature, salinity and dissolved nutrient concentrations [e.g., Prasad et al. (2011)]. In this context, our method may be interpreted as alternative to geographically-weighted regression, in an object-oriented setting.

Amongst the approaches to the analysis of non-stationary spatial fields of scalar data, moving window Kriging (MWK, Haas 1990; Harris et al. 2010) is particularly

related with the class of methodologies here proposed. As in RDD-OOK, MWK is based on multiple, locally estimated variograms and solves multiple Kriging systems (virtually, one for each point belonging to the grid partitioning the space domain), without providing a global model for the structure of spatial dependence. However, a strong plus of the RDD-OOK approach, not shared by moving window methods, is to provide an ensemble of simple learning methods (the weak Kriging maps generated at each bootstrap iteration of the algorithm). The rich information regarding the data and its spatial dependence is captured by this ensemble and its (bootstrap) distribution, of which the strong predictor produced in the aggregation step of the algorithm is just a further summary. Moreover, MWK may be associated with discontinuities in the prediction maps, while such discontinuities do not appear in RDD-OOK maps.

Amongst the critical issues associated with the use of partitions and local models, we mentioned the problem of the bias-variance trade-off. To borrow strength from data outside the cells and lower the variance of the estimates within the cells, we proposed a geographically-weighted approach to estimate the trace-variogram. Here, the use of a kernel greatly enhances the flexibility and robustness of the analysis, particularly for those sub-regions with few observations. In this setting, data-driven approaches may be developed to improve the selection of the parameters related with the kernel and the tessellation. For instance, the parameters ϵ—controlling the bandwidth of the kernel—could be chosen together with K—the number of cells—and both may be selected locally to accommodate for possible non-homogeneous sampling designs [see, e.g., Tavakoli et al. (2016)].

The RDD-OOK method, together with the use of kernel weighted variogram estimates, open new venues for the use of directional data to enhance the prediction power in the presence of complex phenomena. Indeed, even though for the examples here discussed we always employed a simple isotropic kernel, one can readily envision more complex kernel functions (e.g., anisotropic), able to capture and take advantage of the prior knowledge on the phenomenon under investigation (e.g., directional dependence). As a way of example, the problem of DO depletion within the Chesapeake Bay is known to be influenced by the summertime wind direction [see, e.g., Scully (2010)] and such information could be used to enhance its modeling based on anisotropic kernels. In this context, one may readily envision extensions of the proposed methodology to include local anisotropic variographic models [e.g., Fouedjio et al. (2016)], to be possibly associated with the use of anisotropic kernels.

The use of anisotropic (possibly locally varying) kernels may also be considered as a driver for the definition of the partition. In this work, we considered Voronoi tessellations that are consistent with the use of an isotropic kernel. Indeed, assigning a location $\mathbf{s}_0$ to the nearest center is equivalent to its assignment to the center associated with the highest value for the kernel, i.e., $s \in V(\mathbf{c}_k|\Phi_K)$ iff $K_\epsilon(\mathbf{c}_k, \mathbf{s}) > K_\epsilon(\mathbf{c}_j, \mathbf{s})$ for $j \neq k$. However, once a kernel has been identified, the RDD may be consistently accommodated to account for peculiar non stationarities, as well as the design of the experiment.

We remark that the use of a non-Euclidean metric as a measure of the adjacency relations among the locations is in general incompatible with the valid covariance structures commonly used in geostatistics. This motivated us to locally consider an Euclidean metric, even though both the partitions and the kernel are indeed based on a non-Euclidean metric. An interesting yet challenging future direction of research will concern the development of a general theory of Kriging for random field defined on non-Euclidean spatial domains, possibly represented through an undirected graph. This is likely to require the development of novel covariance classes, valid over textured, irregularly shaped domain. The advantage of such developments would be relevant, as the existence of a global model for the phenomenon would allow for Kriging in a unique neighborhood, which is currently not possible in the presence of non-Euclidean domains.

On the other hand, the RDD could be used as a support for the exploration and estimation of global non-stationary models—when the domain characteristics allow to formulate one. Indeed, a relatively large body of literature has been recently focused on developing globally non-stationary and anisotropic models for scalar and vector data over Euclidean or spherical domains [e.g., Fouedjio (2017) and references therein]. In this context, RDDs may drive the selection of neighborhoods where to estimate locally stationary models [in the same flavor as in Fouedjio et al. (2016)], or even used to obtain ensembles of variographic parameters (sill, range, nugget at a set of bootstrap iterations), to be then aggregated in a *strong* estimate of a global model.

Finally, a remarkable open issue regards the uncertainty associated with the final Kriging prediction. Simulation results suggest that the bootstrap variance may play a major role in identifying areas of the field in which the local stationary assumption may not be viable. Further research will be however needed to combine the latter with the average Kriging variance. Here, one should decouple the endogenous and exogenous variability, the former due to the natural variability of the phenomenon (thus of the prediction), the latter to the bagging algorithm. To this end, a general theoretical framework should be established to

formalize the relation between the model for the field and the generation scheme of the random partitions.

References

Abramowicz K, Arnqvist P, Secchi P, de Luna SS, Vantini S, Vitelli V (2016) Clustering misaligned dependent curves applied to varved lake sediment for climate reconstruction. Stoch Environ Res Risk Assess 31(1):71–85

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

Cressie N (1993) Statistics for spatial data. Wiley, New York

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

Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1:269–271

Egozcue JJ, Díaz-Barrero JL, Pawlowsky-Glahn V (2006) Hilbert space of probability density functions based on Aitchison geometry. Acta Math Sin Engl Ser 22(4):1175–1182

Egozcue J, Pawlowsky-Glahn V, Tolosana-Delgado R, Ortego M, van den Boogaart K (2013) Bayes spaces: use of improper distributions and exponential families. Rev Real Acad Cienc Exactas Fis Nat Ser A Matematicas 107(2):475–486

Fouedjio F (2017) Second-order non-stationary modeling approaches for univariate geostatistical data. Stoch Environ Res Risk Assess 31(8):1887–1906

Fouedjio F, Desassis N, Rivoirard J (2016) A generalized convolution model and estimation for non-stationary random functions. Spat Stat 16:35–52

Fuentes M (2001) A high frequency Kriging approach for non-stationary environmental processes. Environmetrics 12:469–483

Fuentes M (2002) Interpolation of nonstationary air pollution processes: a spatial spectral approach. Stat Model 2:281–298

Haas TC (1990) Kriging and automated variogram modeling within a moving window. Atmos Environ Part A Gen Top 24:1759–1769

Harris P, Charlton M, Fotheringham AS (2010) Moving window Kriging with geographically weighted variograms. Stoch Environ Res Risk Assess 24:1193–1209

Heaton MJ, Christensen WF, Terres MA (2015) Nonstationary Gaussian process models using spatial hierarchical clustering from finite differences. Technometrics 59:93–101

Hjelle Ø, Dæhlen M (2006) Triangulations and applications. Mathematics and visualization. Springer, Berlin

Hron K, Menafoglio A, Templ M, Hrůzová K, Filzmoser P (2016) Simplicial principal component analysis for density functions in Bayes spaces. Comput Stat Data Anal 94:330–350

Huang C, Zhang H, Robeson SM (2011) On the validity of commonly used covariance and variogram functions on the sphere. Math Geosci 43(6):721–733

Jensen OP, Christman MC, Miller TJ (2006) Landscape-based geostatistics: a case study of the distribution of blue crab in Chesapeake bay. Environmetrics 17:605–621

Kim HM, Mallick BK, Holmes CC (2005) Analyzing nonstationary spatial data using piecewise gaussian processes. J Am Stat Assoc 100:653–668

Lin J, Chen C, Wu J (2013) CD-graph: planar graph representation for spatial adjacency and neighbourhood relation with constraints. Int J Geogr Inf Sci 27:1902–1923

Machalová J, Hron K, Monti GS (2016) Preprocessing of centred logratio transformed density functions using smoothing splines. J Appl Stat 43(8):1419–1435

Marron JS, Alonso AM (2014) Overview of object oriented data analysis. Biom J 56:732–753

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

Matheron G (1971) The theory of regionalized variables and its applications. Centre de Morphologie Mathématique Fontainebleau: Les cahiers du Centre de Morphologie Mathématique de Fontainebleau. École national supérieure des mines

Menafoglio A, Petris G (2016) Kriging for Hilbert-space valued random fields: the operatorial point of view. J Multivar Anal 146:84–94

Menafoglio A, Secchi P (2017) Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. Eur J Oper Res 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. Electron J Stat 7:2209–2240

Menafoglio A, Guadagnini A, Secchi P (2014) A Kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. Stoch Environ Res Risk Assess 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 Resour Res 52(8):5708–5726

Menafoglio A, Secchi P, Guadagnini A (2016b) A class-kriging predictor for functional compositions with application to particle-size curves in heterogeneous aquifers. Math Geosci 48:463–485

Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis in the symplex. Stoch Environ Res Risk Assess 15:384–398

Prasad MBK, Long W, Zhang X, Wood RJ, Murtugudde R (2011) Predicting dissolved oxygen in the Chesapeake Bay: applications and implications. Aquat Sci 73:437–451

Rathbun SL (1998) Spatial modelling in irregularly shaped regions: Kriging estuaries. Environmetrics 9:109–129

Sangalli LM, Ramsay JO, Ramsay TO (2013) Spatial spline regression models. J R Stat Soc Ser B (Stat Methodol) 75:681–703

Scully ME (2010) Wind modulation of dissolved oxygen in Chesapeake bay. Estuaries Coasts 33:1164–1175

Secchi P, Vantini S, Vitelli V (2013) Bagging voronoi classifiers for clustering spatial functional data. Int J Appl Earth Obs Geoinf 22:53–64 **(Spatial statistics for mapping the environment)**

Secchi P, Vantini S, Vitelli V (2015) Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan (with discussion). Stat Methods Appl 24(2):279–300

Tavakoli S, Pigoli D, Aston JAD (2016) Spatial modeling of object data: analysing dialect sound variations across the UK. arXiv:1610.10040. https://arxiv.org/pdf/1610.10040v1.pdf

van den Boogaart KG, Egozcue JJ, Pawlowsky-Glahn V (2014) Bayes Hilbert spaces. Aust N Z J Stat 56:171–194

Wang J-F, Zhang T-L, Fu B-J (2016) A measure of spatial stratified heterogeneity. Ecol Indic 67:250–256