Stoch Environ Res Risk Assess (2017) 31:87–103
DOI 10.1007/s00477-016-1291-x

ORIGINAL PAPER

Modified half-region depth for spatially dependent functional data

Antonio Balzanella[1] · Elvira Romano[1] · Rosanna Verde[1]

Published online: 21 July 2016

Abstract In this paper, we address the problem of getting order statistics for georeferenced functional data by means of depth functions. To reach this aim, we introduce the concept of spatial dispersion function for functional data in a specific location of the geographic space. Then we generalize the notion of modified half-region depth to spatial dispersion functions. Through the use of spatial dispersion functions we show how the data ordering criterion depends not only on the functional but also on the spatial component. The proposal is applied to two wide simulation studies and to real data coming from sensors.

Keywords Georeferenced functional data · Depth functions · Spatial dependence · Geostatistics

1 Introduction

Multivariate ordering has attracted particular interest over the years. To generalize the ranking process to the multivariate setting several and different definitions of data depth have been introduced. A depth function is a non parametric tool for ordering multivariate data according to their centrality in the data cloud. It measures how deep a point in Euclidean d-space is, that is, how close it is to the center of the data cloud.

The main notions of multivariate data depth proposed in the literature start from Tukey (1975) and Liu (1990). The former defines what is known as half-space depth (also called Tukey depth or location depth), the latter introduced a list of desirable proprieties that the depth definitions should meet. Later a vast theory of depth functions has been developed in R^d by Dyckerhoff (2002), Mosler (2002) and Zuo and Serfling (2000).

Recently with the aim to extend the concept of data ordering to functions, depth notions were proposed in Functional Data Analysis (FDA) Ramsay and Silverman (2005), as well.

FDA deals with situations in which the observations are functions by nature, such as temporal curves or spatial surfaces, where the basic unit of information is the entire observed function.

In this framework, depth functions provide a natural ordering of curves, which makes it possible to define ranks and order statistics.

There have been several alternative notions of depth for functional data: in Fraiman and Muniz (2001), it is given a concept of depth for functional observations based on the integrals of univariate depths; in Cuevas et al. (2007) it is proposed a projection-based depth for functions; in Cuesta-Albertos and Nieto-Rayes (2008) the depth corresponds to the univariate depth of the function values randomly taken at several instants; in Lopez-Pintado and Romo (2009) it is proposed a definition of depth for functional observations based on the graphic representation of the curves; in Lopez-Pintado and Romo (2011) it is defined an Half-Region Depth (HRD) and its modified version (MHRD) based on the notions of *hypograph* and *epigraph* for functional data, with computational advantages when compared to the depths previously proposed in the literature; in Claeskens

✉ Antonio Balzanella
antonio.balzanella@unina2.it

Elvira Romano
elvira.romano@unina2.it

Rosanna Verde
rosanna.verde@unina2.it

[1] Second University of Naples, Caserta, Italy

et al. (2014) it is proposed a generalization of the previous proposals by constructing a depth function for K-variate samples of curves named Multivariate Functional Depth (MFD). Recently in Nieto-Reyes and Battey (2016) a formal definition of statistical depth for functional data on the basis of six properties, recognizing topological features such as continuity, smoothness and contiguity, is provided.

Most of the notions of depth in infinite dimensional spaces have been analyzed, in terms of their properties under various stochastic models, in Chakraborty and Chaudhuri (2014).

In particular, it has been shown that infinite dimensional extensions of most of the depth functions, including HRD, have degenerate behavior while the modified version of HRD, does not suffer from such issue.

Continuous monitoring mechanisms in time and space for several applicative domains has opened up new frontiers that merge Functional Data Analysis and spatial data analysis with the aim of analyzing data which have both a functional and spatial component. Consistently with Delicado (2010) and Romano et al. (2015), we refer to them as spatially dependent functional data.

Let's think for instance to remotely sensed data observed over a number of years across the surface of the earth. Remote sensors collect functional data by detecting the energy that is reflected from the Earth (the space). One could be interested to search for the spatial distribution of reflectance by characterizing it by a center-outward ordering.

In this and other applicative frameworks, the classic ordering, defined by one of the depth measures for functional data, could lead to a define a median curve as representative of an area without considering the effect of its spatial dependence with the others.

This median curve will account for a unique source of variability, the time. However a curve that has a different shape but is geographically far from the median curve will have the same incidence in the depth computation.

For this reason, for each site, we define a function describing the spatial dependence between a curve located in a point of the geographic space and all the other curves located at several spatial distances. We name such function *spatial dispersion function*.

Following the modified Half-Region Depth definition in Lopez-Pintado and Romo (2011), we propose its generalization to spatial dispersion functions with the aim of ordering the georeferenced functional data.

By applying this graph-based approach on the spatial dispersion functions, we include both, the spatial and the functional component, in the ordering criterion for georeferenced functional data.

Until now, only the proposal Balzanella and Elvira (2015) addresses this issue. It introduces the spatial dependence among the curves in the definition of the band depth through the spatial covariance function which plays the role of weight among georeferenced functional data. The band depth is, then, constructed by evaluating the inclusion of the whole curve inside several possible bands graphically obtained by the curves.

However, this last approach could fail for two main reasons. Firstly, the considered spatial covariance function measures the spatial dependence of all the curves in the space and does not consider each single contribute that a curve gives to the whole spatial variability. Secondly, the proposed band depth suffers from degeneracy for some standard probability models in function spaces as stated in Chakraborty and Chaudhuri (2014).

The approach proposed in this paper overcomes these problems.

It allows to characterize the distribution of the georeferenced functions by defining order statistics. We get, for instance, a median spatial dispersion function which assumes the role of representative of the spatial variability in the area. It is a a different concept than the classic variogram since the latter corresponds to the average of the dispersion functions. Moreover, the variogram does not support the detection of a representative site in the geographic space. Instead, being the spatial dispersion function linked to a georeferenced functional data, it is possible to detect a representative of the geographic space in terms of the spatial and the functional component.

The organization of the paper is the following: Sect. 2 introduces the Georeferenced functional data. Section 3 introduces the concept of spatial dispersion functions and illustrates the modified Half-Region Depth definition for spatial dispersion functions.

Section 4 shows an application on simulated and real data. A discussion of the results is provided in the conclusions Sect. 5.

2 Georeferenced functional data

Let $\boldsymbol{\chi}_s(t), \mathbf{s} \in D \subset \mathbb{R}^d$ and $t \in T \subset \mathbb{R}$, be a spatial functional random variable. Following Bohorquez et al. (2016) and Bohorquez et al. (2016a), a spatial functional dataset $\left(\chi_{s_1}(t), \ldots, \chi_{s_i}(t), \ldots, \chi_{s_n}(t)\right)$ is the observation of n functional variables $\left(\boldsymbol{\chi}_{s_1}(t), \ldots, \boldsymbol{\chi}_{s_i}(t), \ldots, \boldsymbol{\chi}_{s_n}(t)\right)$ at the n sites $(\mathbf{s}_1, \ldots, \mathbf{s}_i, \ldots, \mathbf{s}_n) \in D$

Each function, defined on $T = [a, b] \subset \mathbb{R}$, is assumed to belong to an Hilbert space

$$L_2(T) = \left\{ \chi_{s_i} : T \to \mathbb{R}, \text{ such that } \int_T \chi^2_{s_i}(t)dt < \infty \right\}.$$

with the inner product $\langle \chi_{s_i}, \chi_{s_j} \rangle = \int_T \chi_{s_i}(t)\chi_{s_j}(t)dt.$

Especially, for a fixed site s_i, the observed functions can be expressed according to the following model:

$$\chi_{s_i}(t) = \mu_{s_i}(t) + \epsilon_{s_i}(t), \quad i = 1, \ldots, n \qquad (1)$$

where $\epsilon_{s_i}(t)$ are zero-mean residuals and $\mu_{s_i}(\cdot)$ is the mean function.

For each t, we assume that the process is a second-order stationary: that is, the mean and variance functions are constant and the covariance depends only on the distance between sampling sites.

Formally we have:

$$\mathbb{E}(\chi_s(t)) = m(t) \text{ and } \mathbb{V}(\chi_s(t)) = \sigma^2(t), \ \forall\, t \in T,\ s \in D.$$
$$\mathbb{C}ov(\chi_{s_i}(t), \chi_{s_{i+h}}(u)) = \mathbb{C}(h,t,u) \quad \text{where} \quad h = \|s_{i+h} - s_i\|$$
$$\forall\, s_i, s_{i+h} \in D \ \text{ and } \forall\, t,u \in T$$

These assumptions imply that there exists a mapping $h \rightarrow \gamma(h,t,u)$, called variogram, such that:

$$Var(\chi_{s_{i+h}}(t) - \chi_{s_i}(u)) = \mathbb{E}((\chi_{s_{i+h}}(t) - \chi_{s_i}(u))^2) = \gamma(h,t,u)$$
$$= \mathbb{C}ov(0;t,u) - \mathbb{C}ov(h;t,u)$$

By considering the integral on T of the expression above, using Fubini's theorem and following Delicado (2010), a measure of spatial variability can be considered as:

$$\gamma(h) = \frac{1}{2}\mathbb{E}\left[\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt\right], \text{ for } s_i, s_j \in D$$

It corresponds to the trace-variogram introduced in Delicado (2010), estimated as:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt, \qquad (2)$$

where: $N(h) = \{(s_i, s_j) : \|s_i - s_j\| = h\}$, and $|N(h)|$ is the number of distinct elements in $N(h)$.

For irregularly spaced data there are generally not enough observations exactly separated by h so, $N(h)$ is modified to $\{(s_i, s_j) : \|s_i - s_j\| \in (h - \varepsilon, h + \varepsilon)\}$, with $\varepsilon > 0$.

The estimation of the trace-variogram using (2), consistently with Delicado (2010), involves the computation of integrals that can be simplified by considering that the functions are expanded in terms of some basis functions.

The trace-variogram, as the classic variogram for purely spatial data, is used to describe the spatial variability among functional data across an entire spatial domain and not related to a specific location of the space.

Moreover since it has been integrated for every pair of curves, it is scalar and modeled with usual spatial variogram models which allow to include geometric anisotropy Bohorquez et al. (2016a).

3 Modified half-region depth for spatial dispersion functions

We introduce a depth function with the aim of providing an ordering of the georeferenced functional data on the basis of the spatial dependence of each georeferenced curve with the others.

To address this challenge, we introduce the concept of spatial dispersion function $\delta^{s_i}(h)$ Romano et al. (2016) and use half-region depth to construct order statistics for georeferenced functional data.

A spatial dispersion function is a transformation of $\chi_{s_i}(t)$ (for $i = 1, \ldots, n$) which allows to know how the data recorded at a site contribute to the definition of the spatial variability of the whole geographic area.

For each curve $\chi_{s_i}(t)$, at a pivot spatial location s_i, the *spatial dispersion* function around s_i can be defined as:

$$\delta^{s_i}(h) = \sum_{s_i, s_j \in N^{s_i}(h)} \left[\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt\right] \qquad (3)$$

for each $s_j \neq s_i \in D$.

From the previous expression, we still define the *normalized spatial dispersion* function as:

$$\bar{\delta}^{s_i}(h) = \frac{1}{|N^{s_i}(h)|} \sum_{s_i, s_j \in N^{s_i}(h)} \int_T \left(\chi_{s_i}(t) - \chi_{s_j}(t)\right)^2 dt \qquad (4)$$

with $N^{s_i}(h) = \left\{(s_i, s_j) : \|s_i - s_j\| = h\right\} \subset N(h)$ be the number of couple of curves at each lag h.

The *normalized spatial dispersion function* shows how the dissimilarity between the phenomenon recorded at s_i and the phenomenon recorded at each $s_j \neq s_i$ changes with the growing of the spatial distance h. If there is some spatial dependence, the sites closest to the reference one have low values of normalized spatial dispersion whereas sites far away from s_i have high values of normalized spatial dispersion.

The shape and intensity of $\bar{\delta}^{s_i}(h)$ reveals the behavior of the georeferenced curve recorded at s_i as part of the global monitored phenomenon. In this sense the normalized spatial dispersion function captures the information about the effect of the spatial correlation function.

Once we have estimated the spatial dispersion functions for a sequence of k values h_k, these values reflect a smooth variation, thus, we fit data using a basis system by least squares estimation. Note that the fitted dispersion functions are always a valid normalized spatial dispersion functions since we search only for their functional approximation.

The spatial dispersion functions, quantify the spatial variability in the geographic space due to $\chi_{s_i}(t)$ as we can observe by looking at the properties these functions have.

Since we have assumed that the functional random process is second-order stationary, the spatial dispersion function and its normalized form have some interesting characteristics. The first one is related to the monotony of the normalized dispersion functions.

In particular, if $\forall h, h' \in R^+$ such that $h \geq h' \longrightarrow \bar{\delta}^{s_i}(h) \geq \bar{\delta}^{s_i}(h') \quad \forall s_i \in D$. This still involves that $\gamma(h) \geq \gamma(h')$.

Moreover (through straightforward algebraic operations) it is possible to show that

$$\mathbb{E}(\delta^{s_i}(h)) = \gamma(h)\ s_i \in D \tag{5}$$

and

$$\gamma(h) = \frac{1}{2|N(h)|}\sum_{i=1}^{n} \bar{\delta}^{s_i}(h) 2|N^{s_i}(h)|. \tag{6}$$

where $N(h) = \cup N^{s_i}(h)$ and $|N(h)| = \sum_i |N^{s_i}(h)|$.

That is, the variogram function is the weighted average of the normalized spatial dispersion functions $\bar{\delta}^{s_i}(h)$, with $N^{s_i}(h)$ weights as well as the variogram is the average of the dispersion functions.

Our proposal consists in using the normalized spatial dispersion functions as core tool for providing a ranking of the of the georeferenced functional data and, thus, for getting appropriate order statistics.

Especially, we define the notion of modified half-region depth generalizing the half-region depth Lopez-Pintado and Romo (2011) to the normalized spatial dispersion functions.

Let Δ be the set of normalized spatial dispersion functions, for all $s_i \in D$, the graph of a function $\bar{\delta}^{s_i}(h)$ is the subset of the plane $G(\bar{\delta}^{s_i}(h)) = \{(h, \bar{\delta}^{s_i}(h)) : h \in \Re^+\}$.

The hypograph and the epigraph of a function $\bar{\delta}^{s}(h)$ can be defined as:

$$hyp(\bar{\delta}^s) = \{(h, \bar{\delta}) \in \Re^+ \times \Re^+ : \bar{\delta} \leq \bar{\delta}^s(h)\}$$

$$epi(\bar{\delta}^s) = \{(h, \bar{\delta}) \in \Re^+ \times \Re^+ : \bar{\delta} \geq \bar{\delta}^s(h)\}$$

These correspond, respectively, to the region over and under the normalized spatial dispersion function $\bar{\delta}^s(h) \in \Delta$, as shown in Fig. 1.

Following these notions, we define the fraction of functions in the hypograph and in the epigraph of $\bar{\delta}^s(h)$, with respect to the set Δ, as follows:

$$S_1(\bar{\delta}^s) = \frac{\sum_{i=1}^{n} I(G(\bar{\delta}^{s_i}) \subset hyp(\bar{\delta}^s))}{n}$$

$$S_2(\bar{\delta}^s) = \frac{\sum_{i=1}^{n} I(G(\bar{\delta}^{s_i}) \subset epi(\bar{\delta}^s))}{n}$$

where $I(.)$ is the indicator function.

The half-region depth (SB) for the spatial dispersion function $\bar{\delta}^s(h)$, with respect to Δ , is:

$$SB(\bar{\delta}^s) = min\{S_1(\bar{\delta}^s), S_2(\bar{\delta}^s)\}$$

The computation of the Half-Region Depth for the normalized spatial dispersion function $\bar{\delta}^{s_i}(h)$ allows to get an ordering according to decreasing values of $SB(\bar{\delta}^s)$. Thus, we obtain $\bar{\delta}^{s[1]}, \ldots, \bar{\delta}^{s[i]}, \ldots, \bar{\delta}^{s[n]}$ order statistics, where $\bar{\delta}^{s[1]}$ is the median, while $\bar{\delta}^{s[n]}$ is to the most outlying normalized spatial dispersion function. Based on this ordering, we can still generalize the classic box-plot representation by defining the first and third quartile corresponding to the 25 and 75 % central region, as $\bar{\delta}^{s[0.25n]}$ and $\bar{\delta}^{s[0.75n]}$

The definition of the hypograph and the epigraph above, use an indicator function which evaluates if a normalized spatial dispersion function is wholly under or over $\bar{\delta}^s$. Moreover, a more flexible definition is proposed in Lopez-Pintado and Romo (2011): the modified Half-Region Depth (MHRD), which considers functions only partially included into the hypograph or in the epigraph. Such definition can be still extended to normalized spatial dispersion functions by introducing two measures $SL(\bar{\delta}^s)$ and $IL(\bar{\delta}^s)$ that evaluate, respectively, the proportion of lag values h for which $\bar{\delta}^{s_i}$ (with $(i = 1, \ldots, n)$) is greater or smaller than $\bar{\delta}^s$. Formally:

$$SL(\bar{\delta}^s) = \frac{1}{n\lambda(I)}\sum_{i=1}^{n} \lambda\{h \in I : \bar{\delta}^s(h) \leq \bar{\delta}^{s_i}(h)\}$$

$$IL(\bar{\delta}^s) = \frac{1}{n\lambda(I)}\sum_{i=1}^{n} \lambda\{h \in I : \bar{\delta}^s(h) \geq \bar{\delta}^{s_i}(h)\}$$

where λ is the Lebesgue measure on $\Re$ and $I \subset \Re^+$.

The logic behind this modified Half-Region Depth definition is the following: if a normalized dispersion function presents a shape strongly different from the shape of other normalized dispersion functions, few or no spatial dispersion functions would be included in its epigraph or hypograph and a low depth will be associated to it. In this case the normalized spatial dispersion function allows to detect a georeferenced functional outlier.

On the contrary, a normalized spatial dispersion function that is representative of the sample of dispersion functions, is characterized by a high proportion of inclusion in its epigraph and hypograph. This involves that the depth value is the highest and the spatial dispersion function is the median.

In this context a crucial role is played by the shape of the normalized spatial dispersion functions.

The shape of dispersion function contains information related to how a curve located in a point of the geographic

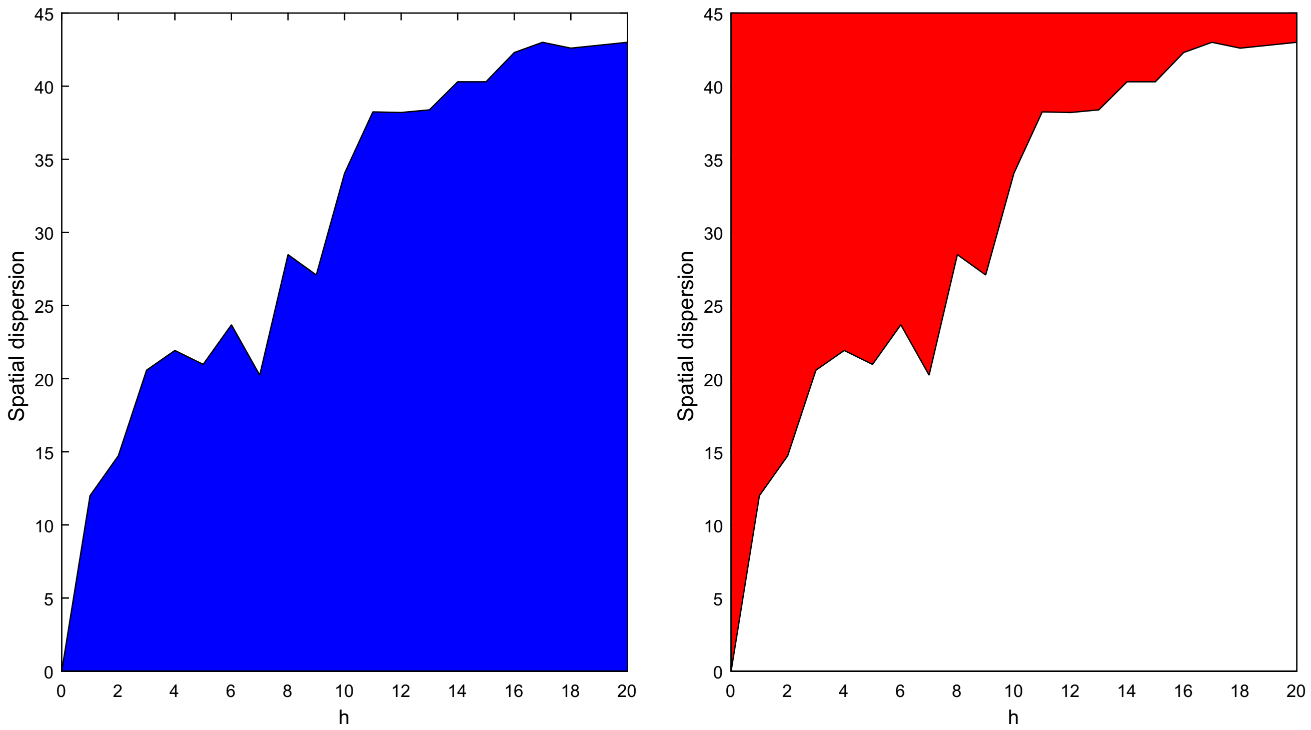


Fig. 1 Hypograph and epigraph of a normalized spatial dispersion function $\bar{\delta}^s$

space differs from the other curves in the space. Thus, looking at normalized spatial dispersion function shape, it is possible to define order statistics in terms of both functional and spatial components.

In particular, we can summarize the main advantages of using this normalized spatial dispersion functions in the ordering of georeferenced curves as follows:

- They allow to define a distribution of the functional data and robust location estimates, such as the median and the quartile functions;
- They provides a measure of the centrality of function located in a site with respect to others related to the other sites, since they provide an ordering starting from the degree of spatial dependence of a curve in a site with the others. The rank-order can be defined from the deepest to the least deep. Where the deepest corresponds to the most representative of the spatial dispersion functions, and the lowest as the most extreme observation, that may be considered outlier. For example, if we observe high values of normalized spatial dispersion for each lag distance h, the curve $\chi_{s_i}(t)$ is strongly different to all the the other curves in the dataset. This allows to select it as outlier. Still, if focusing on the slope of $\bar{\delta}^{s_i}$ we detect an anomalous behavior, we can highlight that the curve $\chi_{s_i}(t)$ has a spatial dependence structure which differs from the other curves in the dataset.

  While the outliers of the first case should be also discovered by traditional dissimilarity based approaches, the second case shows that spatial dispersion functions, which depend on the location of the curves, allow to capture anomalies hidden to traditional distance based approaches.
- The function $\bar{\delta}^{s_{[1]}}$ can be considered as a representative of the spatial dependence in a geographic area since it is such to satisfy:

$$\bar{\delta}^{s_{[1]}} = argmax_{\bar{\delta}^{s_1},\dots,\bar{\delta}^{s_n}} SB(\bar{\delta}^{s}) \qquad (7)$$

- It has the main advantage of including in the ordering process the two data components: the functional and the spatial ones;

The properties of Distance invariance, Maximality, Strictly decreasing w.r.t. the deepest point, Upper semi-continuity, Receptivity to convex hull width across the domain, Continuity, valid for the modified Half-Region Depth for functional data, are still valid for spatial dispersion functions.

4 Applications

In this section we are going to apply the proposed method to three different situations. We propose, at first, two applications on simulated data, in order to show the performance of the method on datasets generated according to a widely simulation scheme for spatio-temporal data and for spatial functional data. Then, we perform an application on a real world dataset in which temporal data, recorded by a sensor network, have a spatial dependence.

4.1 Depth measurement on spatio-temporal simulated data

The first testing of the proposed method has been performed generating the data following the simulation scheme introduced in Sun and Genton (2011) and used in Romano et al. (2015).

Data are drawn from a zero-mean, stationary spatial-temporal Gaussian random field, $\chi_s(t)$ (with $s \in D$ and $t \in T$) whose covariance function $C(h,u) = cov\left\{\chi_{s_i}(t_1), \chi_{s_j}(t_2)\right\}$ depends on the spatial distance $h = ||s_i - s_j||$ and on the functional distance $u = ||t_1 - t_2||$, for any couple of s_i,s_j and t_1, t_2. The spatial temporal variability is defined starting from the following four covariance models (consistently with the setup in Sun and Genton (2011)):

1. Purely spatial covariance function, defined for two generic locations s_i, s_j, that are apart by $h = ||s_i - s_j||$,. presents the form:

$$C_s(h) = (1-v)\exp(-c|h|) + vI\{h=0\} \tag{8}$$

where $c > 0$ controls the intensity of the spatial correlation, and $v \in (0,1]$ is the nugget effect. We set $v = 0$ for all the cases, so that there is no nugget effect.

2. Space-time separable correlation function of the form:

$$C_{SEP}(h,u) = cov\left\{\chi_{s_i}(t_1), \chi_{s_j}(t_2)\right\} = C_s(h)C_T(u) \tag{9}$$

where $C_s(h)$ is expressed by (8) and $C_T(u)$ is a stationary, functional covariance function, of the Cauchy type, having the form:

$$C_T(u) = \left(u + a|u|^{2\alpha}\right)^{-1} \tag{10}$$

with a time span $u = ||t_1 - t_2||$. Here $a > 0$ is the scale parameter in time, that is fixed to $a = 1$ in all the datasets, and α, is the parameter that controls the strength of the functional variability.

3. Symmetric but generally non-separable correlation function of the form:

$$C_{Sim}(h,u) = \frac{1-v}{1+au^{2\alpha}}\left[\exp\left\{-\frac{c||h||}{(1+au^{2\alpha})^{\frac{\beta}{2}}}\right\} + \frac{v}{1-v}I\{h=0\}\right] \tag{11}$$

where the parameter $0 \leq \beta \leq 1$ controls the degree of non-separability. $\beta = 0.9$ has been set for all the datasets, to have the maximum possible non-separability.

4. General stationary correlation model of the form:

$$C(h,u) = (1-\lambda)C_{Sim}(h,u) + \lambda\left(1 - \frac{1}{2v}h_1 - vu\right)_+ \tag{12}$$

where h_1 is the first component of the spatial separation vector h and $0 \leq \lambda \leq 1$ controls the asymmetry. For the datasets generated according to this correlation model, λ has been set to 0.5

For each covariance function, we have generated a dataset made by 196 curves, each one of 50 time points in [0, 1]. The Table 1 shows the parameters used for the simulation of each dataset.

Aim of the analysis of these data is, at first, to evaluate how the introduction of the spatial component impacts on the depth measurement. To tackle this problem, we compare the modified Half-Region Depth for normalized spatial dispersion functions with the weighted band depth for georeferenced functional data (Balzanella and Elvira 2015) and with the modified Half-Region Depth on the time functions (Lopez-Pintado and Romo 2011).

In order to perform the comparison some preliminary steps are needed.

The first step is to get a functional description of the spatio-temporal data by fitting them through a B-spline and using a least squares fitting criterion. We use cubic splines basis functions and equally spaced knots in the interval [0; 1].

As introduced in Sect. 3, the proposed strategy, still requires to compute a normalized spatial dispersion function $\bar{\delta}^{s_i}(h)$ for each site s_i (for $i = 1, \ldots, n$). Such computation, has to be performed ensuring that at each lag distance h there is a sufficient number of curves to compare thus, it is a common practice to set non overlapping intervals $[h-\epsilon,\ h+\epsilon]$ and to calculate the dispersions using the sites whose spatial distance is included in such interval. In order to set ϵ we adopt the widely used rule of the thumb for variogram estimation (Journel and Huijbregts 2004) which suggests to ensure that the minimum number of curve pairs falling into each interval is 30. The application of this criterion to our data involves that $\epsilon = 0.13$ and the number of intervals is 10.

Table 1 Parameters for simulated datasets

Dataset id	Correlation model	Value of c	Value of α
1	Purely spatial	0.01	
2	Separable	0.6	0.2
3	Non-separable	0.4	0.5
4	General stationary	0.1	0.3

As stated in Sect. 3, in order to ensure continuity in the normalized spatial dispersion functions, we have still performed a linear interpolation of each function.

The normalized spatial dispersion functions for the first simulated dataset and the corresponding variogram function are shown in Fig. 2. Looking at the curve shape, we can see that there is a spatial dependence in the data. This is because both, the variogram and the most of the normalized spatial dispersion functions grow with the increase of h.

At the same time, the approach proposed in Balzanella and Elvira (2015) still needs the computation of the variogram function as a weight function in the depth measurement. As for the computation of the normalized spatial dispersion functions, we set non overlapping intervals $[h - \epsilon, h + \epsilon]$ which ensure that the minimum number of curve pairs falling into each interval is 30.

The depth values for the three compared approaches are shown in Fig. 3.

The left column of the figure shows the depth values obtained through the proposed strategy on the four datasets, the central column shows the depths obtained by the Band Depth weighted by the spatial covariance, finally, the third column illustrates the results for the modified Half-Region Depth computed directly on the curve data.

In the figure, we have used a color map through which each site is represented as a colored box in the squared spatial grid with higher temperature colors corresponding to higher depths. Thus, the blue cells correspond to sites having a very low depth while yellow cells have a high depth value.

Looking at the results, the method proposed in this paper and the modified half-region depths get depth values in the range $[0-0.5]$ for all the datasets. The depths obtained by the Band Depth weighted by the spatial covariance are, instead, in the range $[0-0.8]$. In all the cases there is no site which emerges for its high centrality (depth). Still, there is only a partial agreement between the compared methods. This is due to the different way of including the spatial information for the first two methods and because the third one does not consider it.

It is still interesting to note that if we consider a low depth value as a measure outlyingness, only in few cases, the same potential outliers are discovered by all the methods.

A further aid to the evaluation of our proposal comes from using the functional boxplot introduced in Sun and Genton (2011). We will use it in order to highlight the differences among the data in terms of normalized spatial dispersion.

We recall that this tool extends the classic boxplot to functional data by visualizing five descriptive statistics: the envelope of the 50 % central region (in magenta) which extends the concept of interquartile range, the median curve (in black), the maximum e minimum curves (in blue). In order to make our analysis, we still plot (in red) on the plane the average spatial dispersion, which corresponds to the variogram, as well as the whole set of normalized spatial dispersion functions. The results of our method, for all the datasets, are available in Fig. 4.

A first analysis focuses on the behavior of the normalized spatial dispersions. We can see that the sites in the dataset 1 have a higher spatial dependence than the other datasets since the slope of the normalized spatial dispersion functions is higher. This is confirmed by the median curve which grows faster than in the other datasets over the whole set of lag distances. This information is difficult to obtain by simply looking at the original curves in the time domain.

A further difference among the analyzed datasets, is the range of the normalized spatial dispersion functions. Especially, the top and bottom curves in blue, highlight that the third dataset is the one with the widest range. Still, the

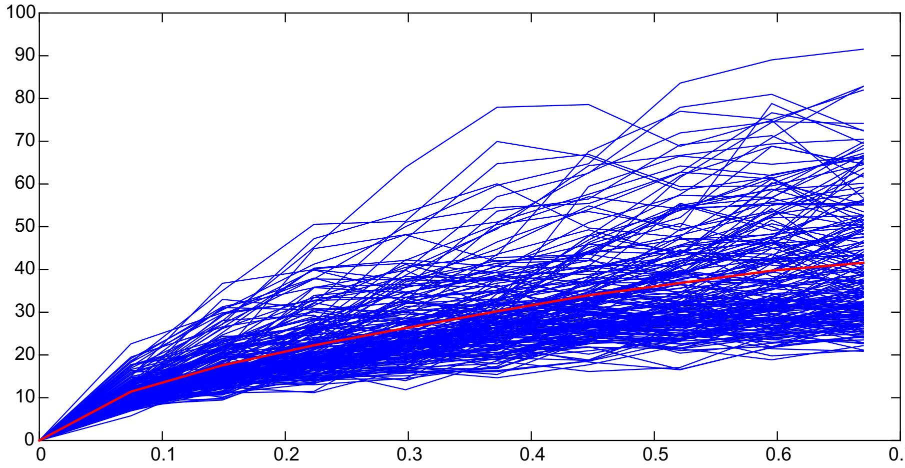
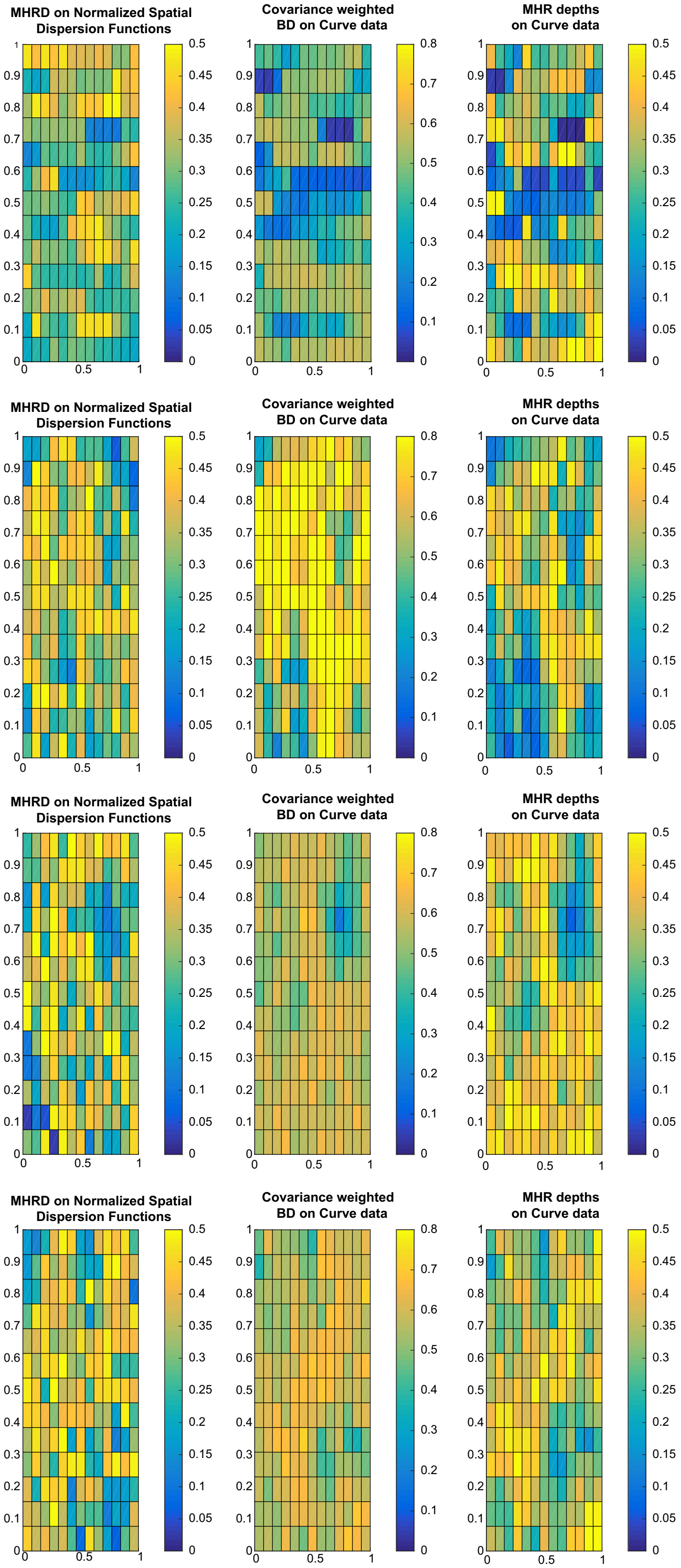


Fig. 2 Normalized spatial dispersion functions for the purely spatial simulated data

Fig. 3 Depth values on the spatial grid for simulated data



interquartile region, which represents the 50 % of the highest depth values, is wider in the first dataset than the other three ones.

Some specific consideration is needed if we look at the behavior of the normalized spatial dispersion function having the highest depth. Since it can be interpreted as a representative of the spatial variability in the area, it is useful to perform a comparison with the variogram plotted in red.

A first aspect to note is that the variogram is sensitive to outliers. This is because it is a weighted average of the normalized spatial dispersion functions. Similarly to the univariate case, the median does not shows this issue. Moreover, through the median function, it is possible to set the site $s_{[1]}$ as representative of the area. On the contrary, the variogram is not associated to a spatial location, thus, it is not possible to obtain a representative site.

As shown in Fig. 4, it is possible to note that the for all the datasets, the median differs from the variogram. Especially, this is evident in the first and second dataset for high values of h.

We have performed a further test in order to highlight the influence of the spatial dependence on the depth measurement by introducing an outlier in the data. Especially, we have generated a function made by 50 observations obtained by a Gaussian random process with parameters $\mu = 0$ and $\sigma = 4$. Unlike to the classic depth measurement of Lopez-Pintado and Romo (2009) and (2011) our strategy measures the depth of a curve incorporating the information about the spatial location where it has been recorded. In particular, the spatial location of a curve influences the computation of the corresponding spatial dispersion function.

In Fig. 5, we show the depth value for the outlier by moving the generated curve on the geographic spatial grid. The results are related to the dataset 1.

As before, the x and y axis represent the spatial coordinates, while the colormap allows to show the depth values for each site. The color of each cell illustrates the depth value of the outlier when it takes every possible position of the plane. In this case the depth value is always in the range [0−0.2], however if the outlier is placed on the boundaries, its depth is lower than if it is placed in the center. This confirms the impact of the spatial location on the depth value.

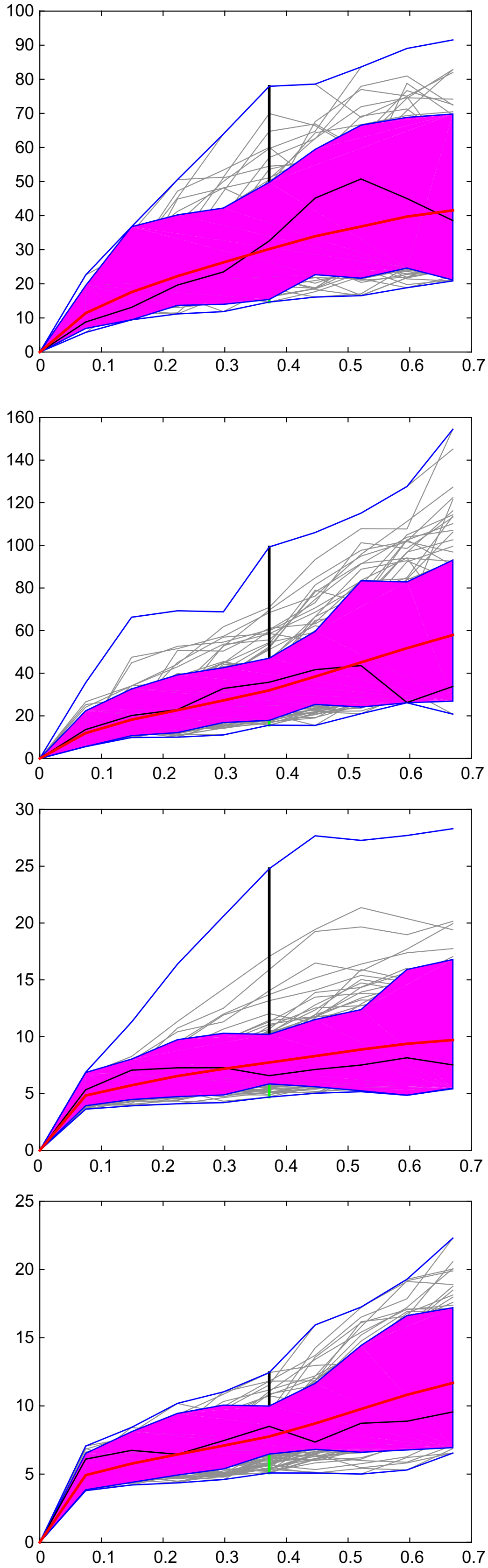


Fig. 4 Functional *boxplots* for simulated data

4.2 Depth measurement on simulated spatially dependent functional data

In this section we evaluate the performance of the depth measurement strategy on a simulated dataset made by spatially dependent functional data. The testing procedure

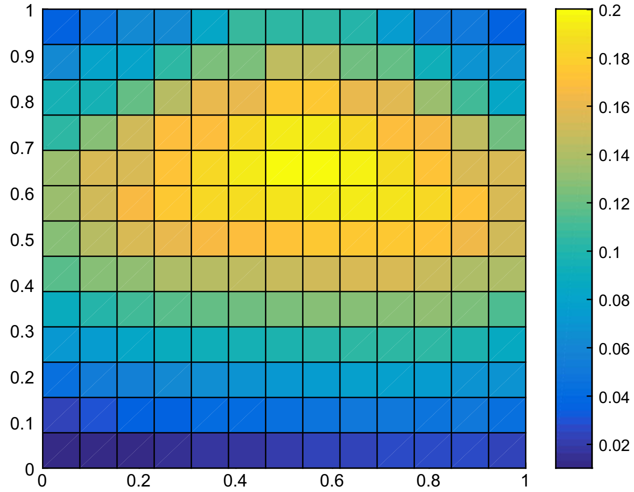


Fig. 5 Depth values for simulated data with outlier on the geographic spatial grid

has the aim of evaluating the effect of ordering the normalized spatial dispersion functions rather than directly the observed functional data. In particular, we want to show that using the information about the spatial dependence among curves affects the ordering of the georeferenced curves. It is an extreme case study, where the data are periodic, with the same frequency and having zero average so that the MHRD are not able to provide an ordering of functional data.

The process from which samples have been generated is the following:

$$\chi_{s_i}(t) = f_{s_i}\sin(\mathrm{t} \times 1000)\mathrm{g}_{\mathrm{s_i}}\cos(\mathrm{t} \times 10) \qquad \mathrm{s_i} \in \mathrm{D} \tag{13}$$

The coefficients f_{s_i} and g_{s_i} are realization of a multivariate Gaussian distribution with zero mean and covariance structure capturing the spatial dependence in the data. We have generated four datasets according to the following models of spatial dependence: Spherical, Bilinear, Exponential, Gaussian. For each model we assume that the range is 1.4, the sill is 200 and the nugget is 0.

In Fig. 6 we show the variogram models from which we have obtained the covariance structure in each dataset.

Similarly to the test above, each dataset is made by 196 curves on a squared spatial grid with range [0–1] [0–1]. The curves of the first dataset are shown in Fig. 7.

We compare the results of the proposed strategy on normalized spatial dispersion functions with the modified Half-Region Depth on the curve data.

In order to get the results, we must set the input parameters. We need, at first, to set the lag distances h where to measure the spatial dispersion. As for the previous application on spatio-temporal data, we set non overlapping intervals $[h - \epsilon, h + \epsilon]$ and calculate the dispersions using the sites whose spatial distance is included in such interval. In order to ensure that the minimum number of curve pairs falling into each interval is 30, we set $\epsilon = 0.13$ and the number of intervals to 10.

A further preprocessing step to be performed in order to get the output of our procedure, is interpolation of the normalized spatial dispersion functions in order to ensure continuity. To this aim, we have a used a linear interpolation which makes each spatial dispersion function a piecewise function.

The first result we show in Fig. 8 is the depth value associated to each site of the spatial region, for each one of the four simulated datasets. Similarly to the previous application on spatio-temporal data, we have used a color map through which each site is represented as a colored box in the squared spatial grid with higher temperature colors corresponding to higher depths.

We can note that the modified Half-Region Depth on the curve data is not able to provide an ordering of the data. Every curve has the same depth value since every box has the same color. This is due to specific characteristics of the data: the periodicity, the constant frequency and the average set to 0 for all the functional data involves that there is always a compensation between the portion of the curves in the epigraph and in the hypograph. Our method is, on the contrary, able to highlight the impact of the spatial covariance so that on all the datasets the locations on the boundary of the spatial region tend to have lower depth values.

A more detailed analysis of the results of our ordering strategy can be performed by looking at the functional boxplots in Fig. 9.

Focusing on the magenta area representing the interquartile region, we can see that for low values of h the normalized spatial dispersion functions have similar values. This highlights that the simulated functional data tend to be similar to the observations in their spatial neighborhood. This is a common feature of the four datasets. With the increase of h, the magenta area becomes wider so that there is more variability in the similarity between functional data at s_i and the functional data observed at far locations. In this sense, we can see that the four datasets behave differently since the first and fourth dataset keep, until $h = 0.45$, a low-width interquartile region while the second and third dataset have at $h = 0.45$ wide interquartile regions.

Another aspect worthy of attention is the behavior of the median function which corresponds, as stated above, to the deepest spatial dispersion curve. Still, the comparison with the average spatial dispersion function (variogram) allows to highlight the capability of the median to provide a robust representative of the spatial variability of the geographic area.

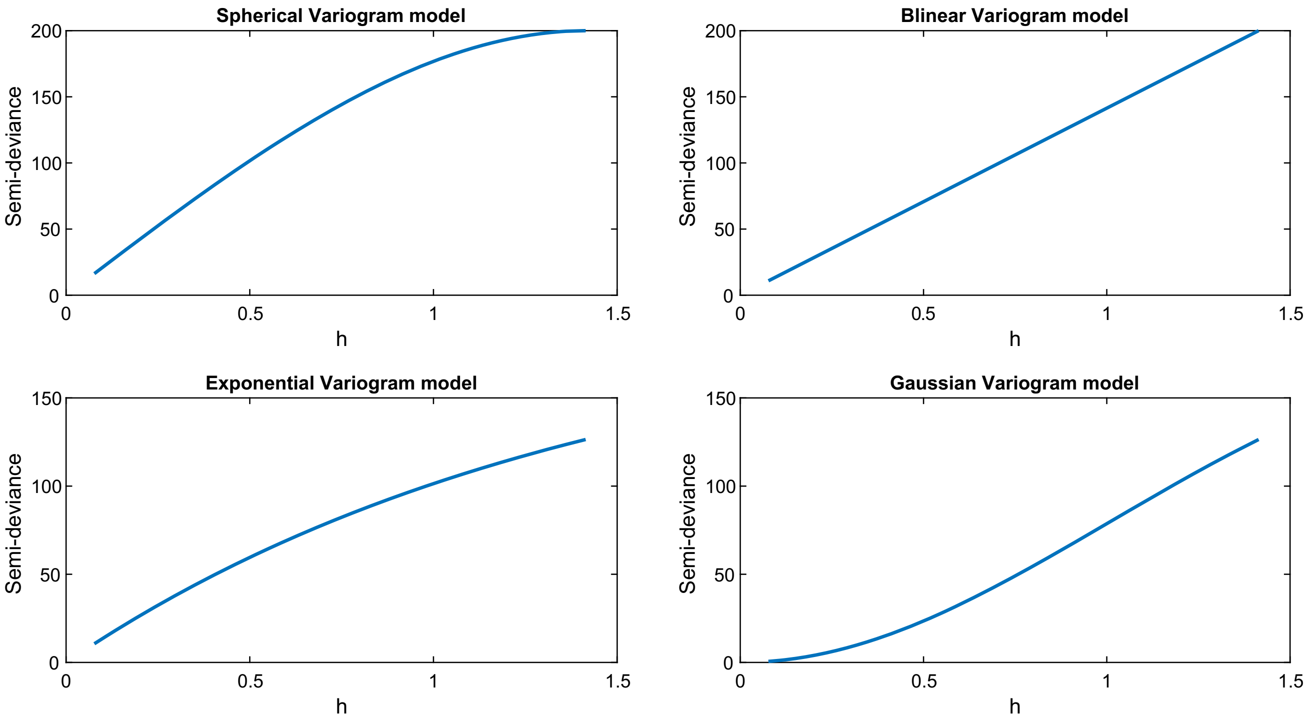


Fig. 6 Variogram models used for generating the spatial functional dataset

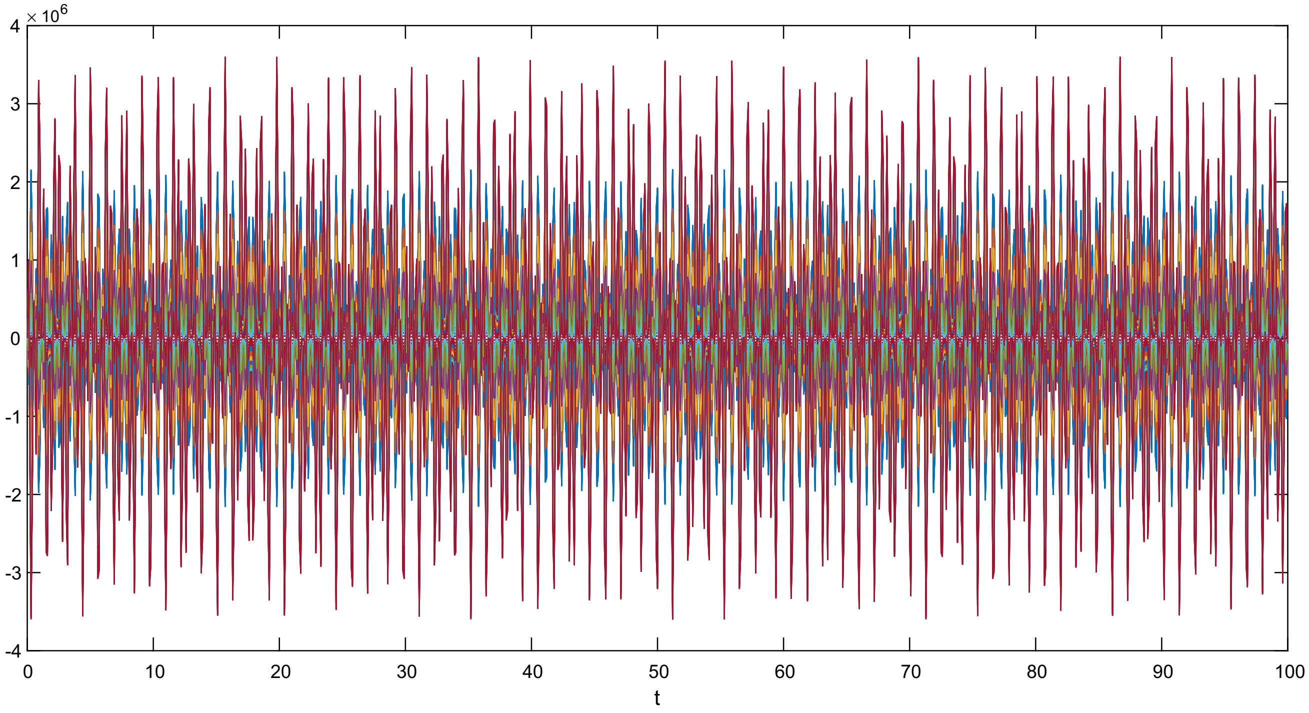
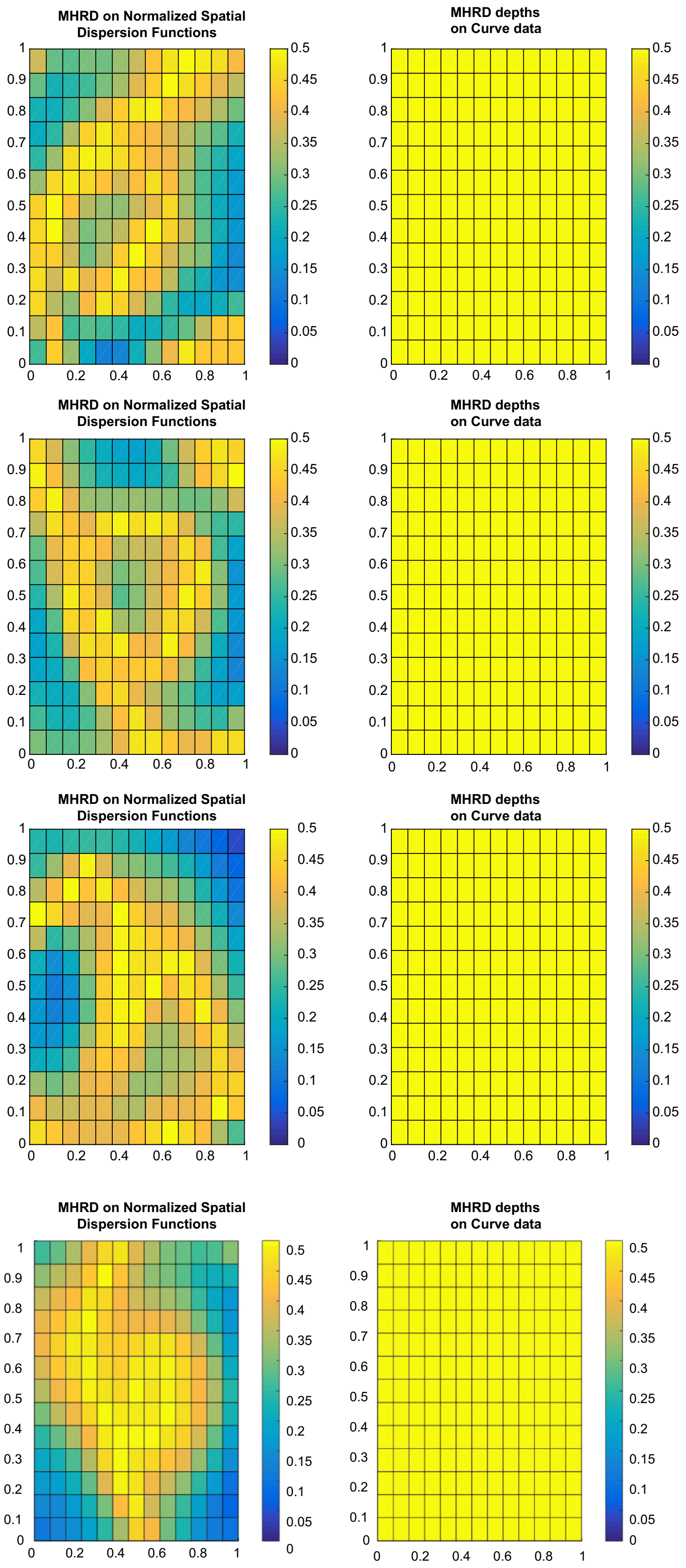


Fig. 7 *Curve* dataset generated according to the gaussian variogram model

Looking at the median (in black) and at the variogram (in red) for values of $h < 0.25$, the two functions have very similar values for all the datasets so they provide the same information about the phenomenon. With the growing of h, the differences between the variogram and the median emerge. In particular, in the first dataset the median has

Fig. 8 Depth values on the spatial grid for the spatial-functional simulated datasets



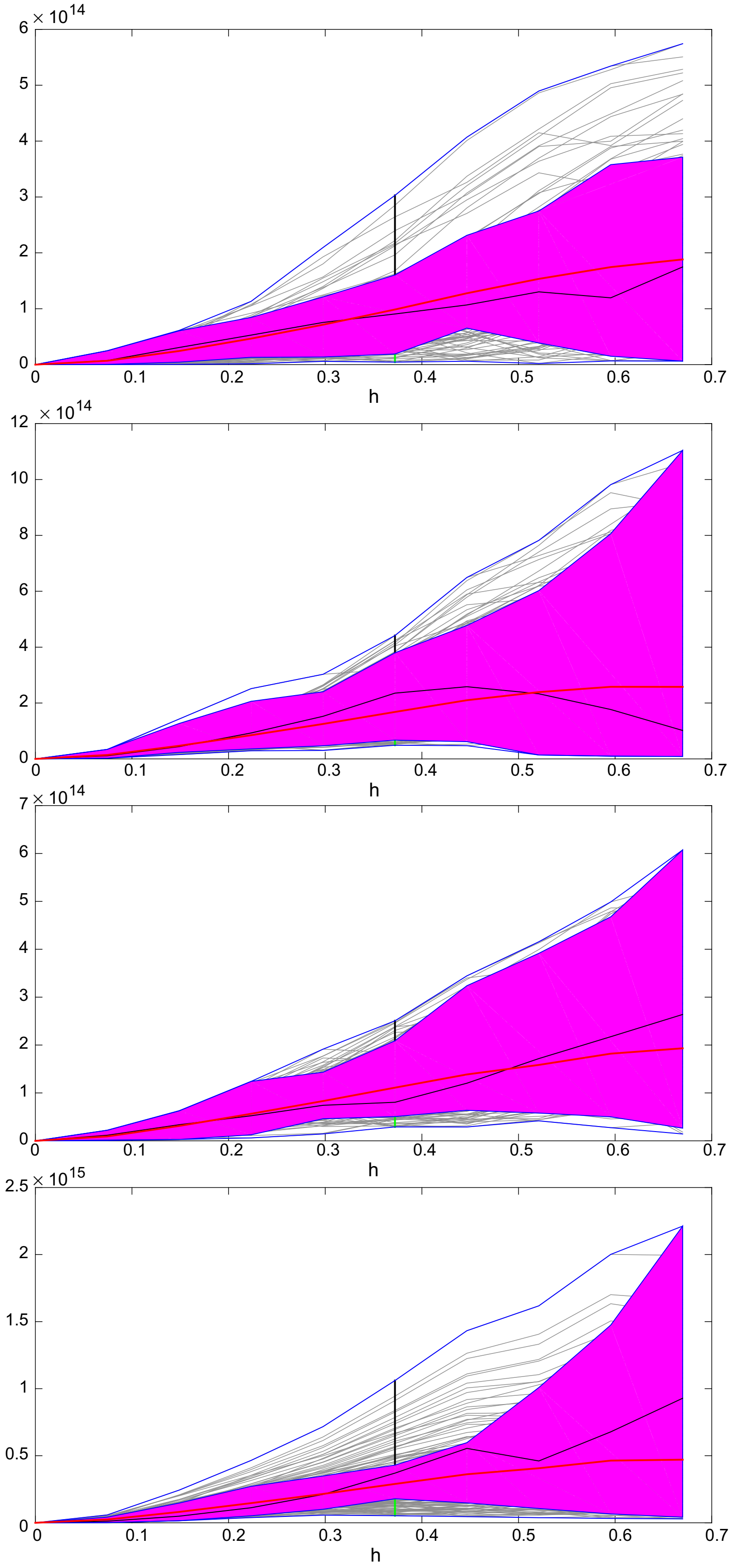


Fig. 9 Functional *boxplots* for the four spatial-functional simulated datasets

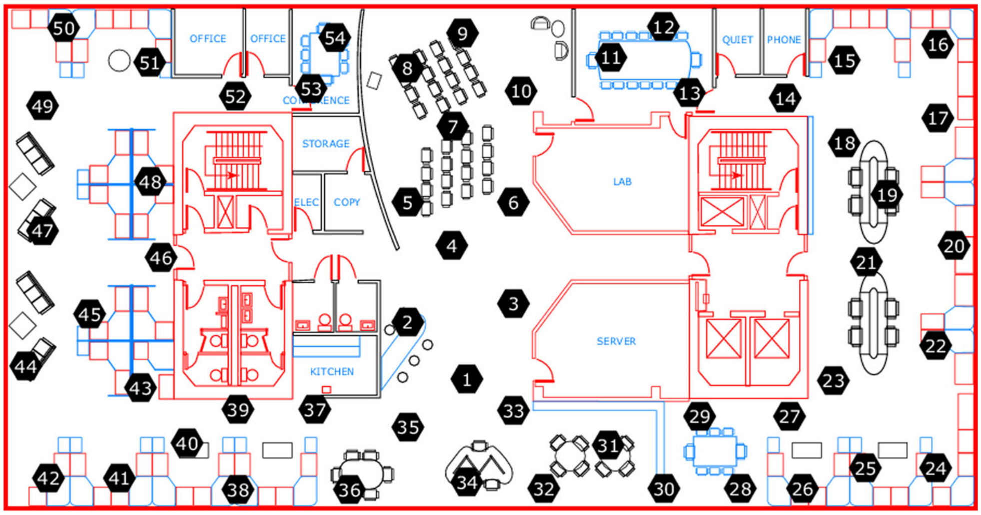


Fig. 10 Diagram of the sensors lab

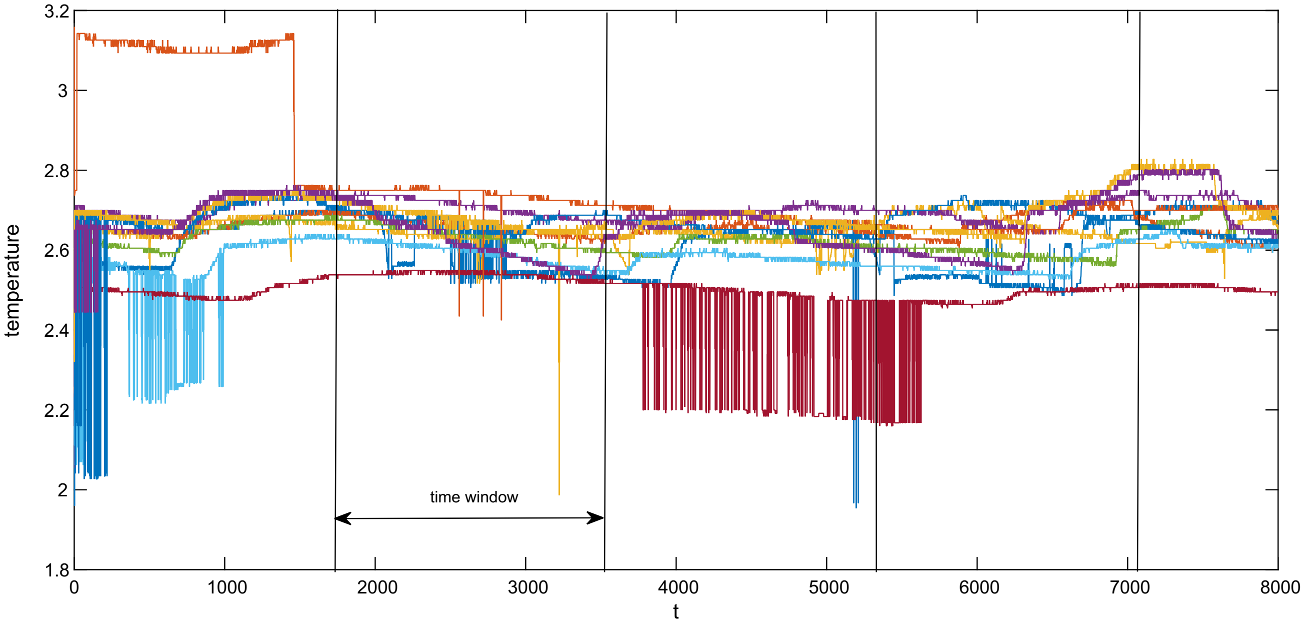


Fig. 11 Temperature records of 54 time series

lower values of spatial variability for $h > 0.3$. This is because the median is not influenced by the highest spatial dispersion curves. In the second dataset, at $h = 0.52$ the median intersecate the variogram so that there is a decrement of the spatial variability for very high values of h. In the third dataset we can see that the median is very similar to the variogram for all the values of h so that both the function provide the same summarization of the spatial dipendence. Finally, the fourth dataset highlights a strong difference between the median and the variogram with the median higher than the variogram for $h > 0.3$, still, it is interesting to note that the median is very near to the boundary of the interquartile region.

4.3 Depth measurement on real data

In this section, we will show how the proposed strategy can be used for monitoring the evolution of sensor data. The test has been performed on a public dataset of real data, available at http://db.csail.mit.edu/labdata/labdata.html.

The dataset collects the records of 54 sensors placed at the Intel Berkeley Research lab between February 28th and April 5th, 2004. Mica2Dot sensors with weather boards collected timestamped topology information, along with humidity, temperature, light and voltage values once every 31 $\sim$ s. Data was collected using the TinyDB in-network query processing system, built on the TinyOS platform.

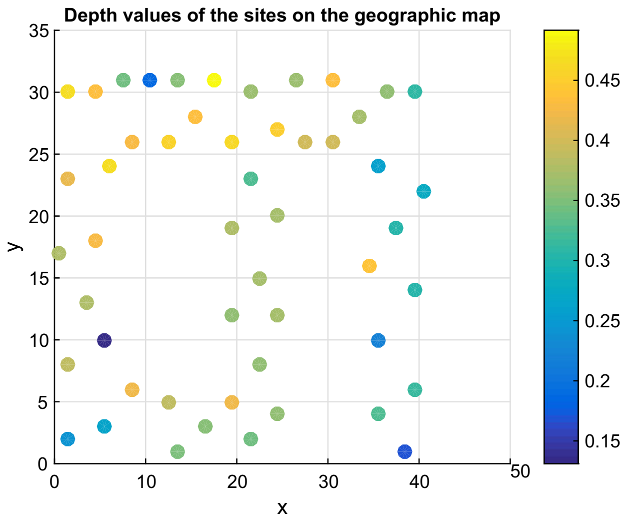


Fig. 12 Depth values of sites

The dataset includes the x and y coordinates of sensors (in meters relative to the upper right corner of the lab). The sensors were arranged in the lab according to the diagram in Fig. 10.

We have analyzed the temperature records of each sensor so that we have a set of 54 time series each one made by 65,694 observations.

Our idea is to split the stream of data into non overlapping temporal windows so that the depth measurement is performed on each window independently.

In order to get windows which collect daily data, we have set their size to 1775 observation as shown in Fig. 11 for a data subset.

On the data of each window we have run our strategy which provides an ordering of the subsequences which accounts for the spatial location of each sensor.

At first, we will focus on the results for the first window. We show in Fig. 12 the depth of each site as a colored dot on the map. In this case, the sites having the highest depth are located in the upper-left part of the geographic space. On the right, there are the most of the sites having low depth values.

We still show in Fig. 13 the functional boxplot for the first data window. This allows to analyze the distribution of the normalized spatial dispersion functions.

The first issue to note is that the functions having the lowest depth do not show any spatial dependence. Especially, the curve on the top has some spatial dependence only for very low values of h while the curve on the bottom has a constant behavior with no spatial dependence at all.

Still, the magenta area which highlights the interquartile region, shows that its wideness is higher for low values of h than for high values. This explains that the 50 % of the deepest spatial dispersion curves tend to be similar at high spatial lags.

Looking at the deepest curve in black (median) and at the average one in red (variogram), we can note that there is a spatial dependence in the data. This is because both the curves grow for increasing values of h. However, as shown on simulated data, the median normalized spatial dispersion function is not sensitive to outliers.

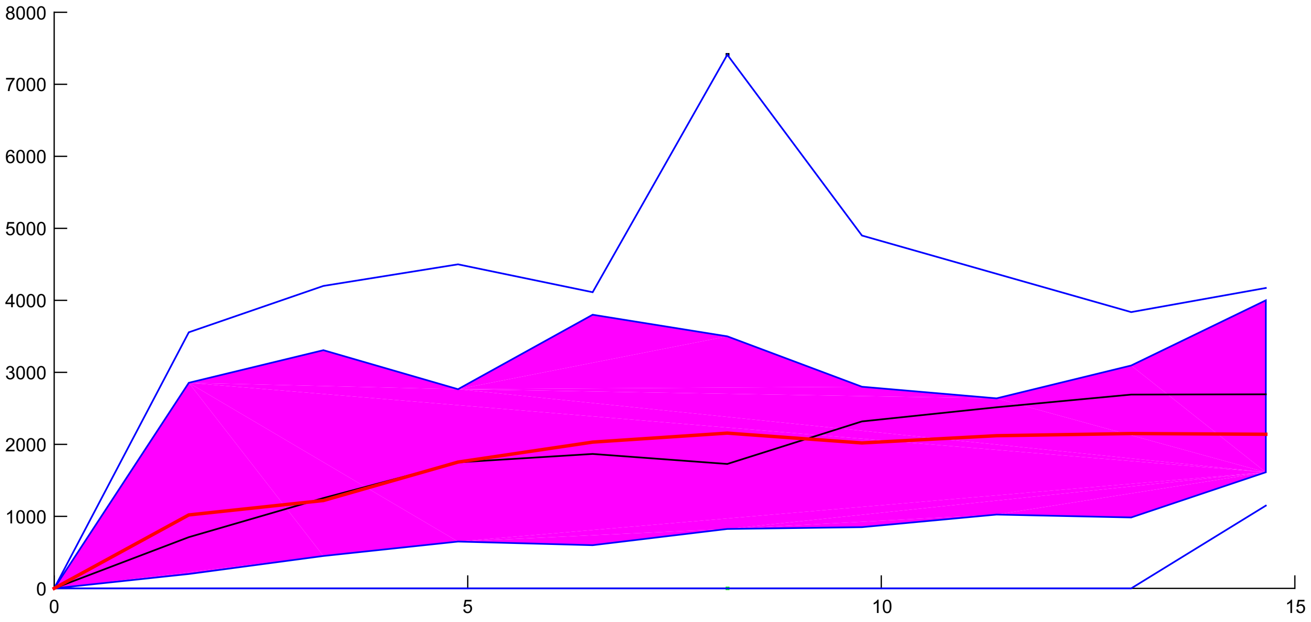


Fig. 13 Functional *boxplot* for the sensor data in the first time window

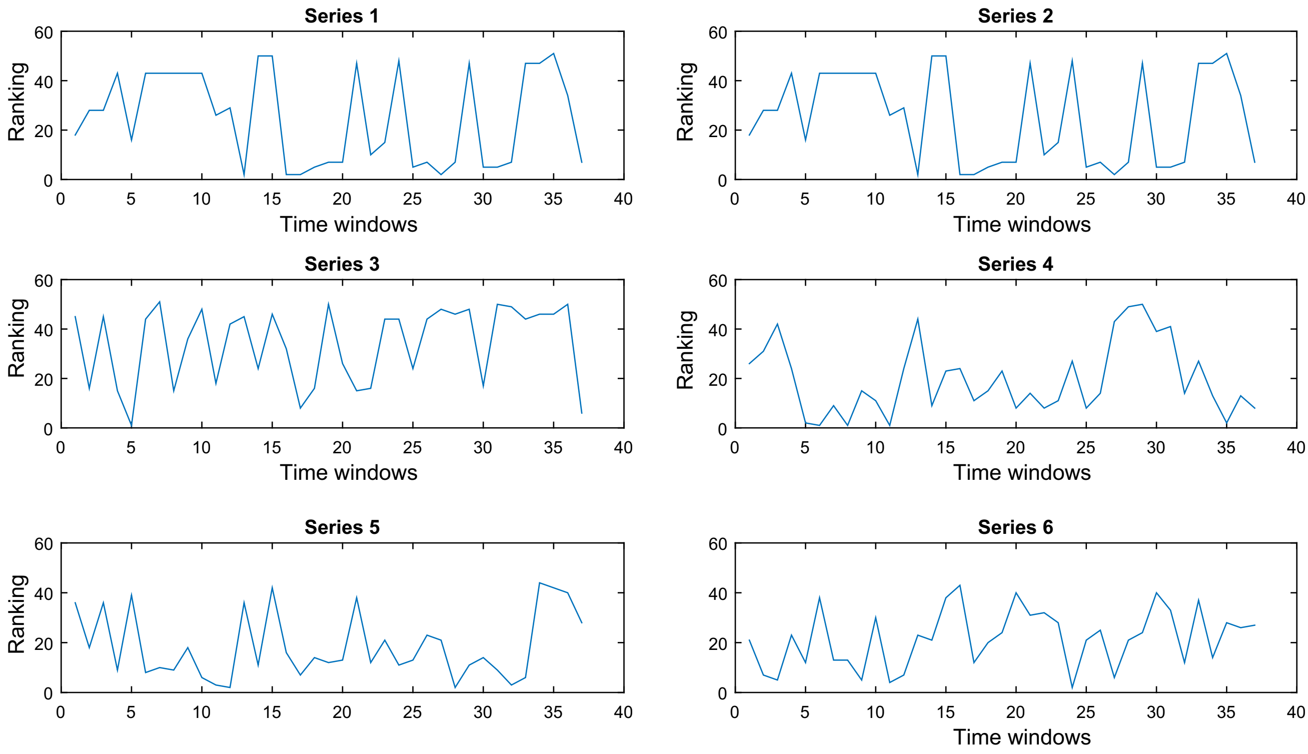


Fig. 14 The ranking of the time series in the dataset over the windows

The execution of the ordering procedure on each window allows to get, for each site, a function in which each observation is the rank of the site in a window. In Fig. 14, we show the ranking of the first six time series in the dataset over the windows. It is interesting to note how the position in the graded list induced by the depth computation changes over the windows so that the data recorded at a site can be highly central in a window and still have a low depth for the following windows.

If we look at the changes over the time of the position in the ranking, we can get an overview of the evolution of the monitored phenomenon. In this sense, the power of this approach can be found also in its application to streaming settings.

5 Discussion and conclusion

Spatial Functional Data Analysis is a developing field in statistics that has emerged in the last decade.

In many applications, the basic underlying observation is a georeferenced curve. New challenges arise when the functions are spatially dependent. In this paper, we have introduced a generalization of the modified-region depth for spatially dependent functional data. These new notions account for one of the hottest challenges in this field: the concept of ranking. By introducing the concept of spatial dispersion function as a transformation of the functional data, our proposal has the advantages of furnishing a criterion for ranking simultaneously the spatial and the functional component of the data. In addition it allows to define a distribution of the spatial dispersion functions characterized by robust location estimates, such as the median spatial dispersion and the quartile functions. This method has been illustrated by assuming stationarity in the data, however it can be applied also when the the more general case of non-stationarity in the data is assumed. As further research we are going to investigate the introduction of directional spatial dispersion functions in order to deal with covariance structures which change according to the spatial direction.

References

Albertos-Cuesta JA, Nieto-Rayes A (2008) The random Tukey depth. Comput Stat Data Anal 52(11):4979–4988

Balzanella A, Elvira R (2015) A depth function for geostatistical functional data. In: Morlini I, Minerva T, Vichi M (eds) Advances in statistical models for data analysis, Springer International Publishing, pp 9–16. doi:10.1007/978-3-319-17377-1_2

Bohorquez M, Giraldo R, Mateu J (2016) Optimal sampling for spatial prediction of functional data. Stat Methods Appl 25(1):3954

Bohorquez M, Giraldo R, Mateu J (2016) Multivariate functional random fields: prediction and optimal sampling. Stoch Environ Res Risk Assess 1–18. doi:10.1007/s00477-016-1266-y

Chakraborty A, Chaudhuri P (2014) On data depth in infinite dimensional spaces. Annal Inst Stat Math 66(2):303–324

Claeskens G, Hubert M, Slaets L, Vakili K (2014) Multivariate functional halfspace depth. J Am Stat Assoc 109:411–423

Cuevas A, Febrero M, Fraiman R (2007) Robust estimation and classification for functional data via projection-based depth notions. Comput Stat 22:481–496

Delicado P, Giraldo R, Comas C, Mateu J (2010) Statistics for spatial functional data: some recent contributions. Environmetric 21:224–239

Delicado P, Giraldo R, Comas C, Mateu J (2010) Spatial statistics for functional data: some recent contributions. Environmetrics 21:224–239

Dyckerhoff R (2002) Data depths satisfying the projection property. Adv Stat Anal 88:163–190

Ferraty F, Vieu P (2006) Nonparametric functional data analysis theory and practice. Springer, New York

Fraiman R, Muniz G (2001) Trimmed means for functional data. Test 10:419–440

Journel AG, Huijbregts Ch J (2004) Mining geostatistics. The Blackburn Press, Caldwell

Liu R (1990) On a notion of data depth based on random simplices. Annal Stat 18:405–414

Lopez-Pintado S, Romo J (2009) On the concept of depth for functional data. J Am Stat Assoc 104:718–734

Lopez-Pintado S, Romo J (2011) A half-region depth for functional data. Comput Stat Data Anal 55(4):1679–1695

Mosler K (2002) Multivariate dispersion, central regions and depth: the Lift Zonoid approach. Springer, New York

Nieto-Reyes A, Battey H (2016) A topologically valid definition of depth for functional data. Stat Sci 31:61–79

Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer, New York

Romano E, Mateu J, Giraldo R (2015) On the performance of two clustering methods for spatial functional data. AStA Adv Stat Anal 99(4):467–492

Romano E, Balzanella A, Verde R (2016) Spatial variability clustering for spatially dependent functional data. Stat Comput 1–14. doi:10.1007/s11222-016-9645-2

Sun Y, Genton MG (2011) Functional boxplots. J Comput Graph Stat 20:316–334

Tukey J (1975) Mathematics and picturing data. In: Proceedings of the 1975 International Congress of Mathematics. 2, 523–531

Zuo Y, Serfling R (2000) General notions of statistical depth function. Annal Stat 28:461–482