Optimally weighted L_2 distances for spatially dependent functional data

Elvira Romano[a,], Andrea Diana[a], Claire Miller[b], Ruth O'Donnell[b]

[a]*Department of Mathematics and Physics, Universitá della Campania "Luigi Vanvitelli", Caserta, Italy*
[b]*School of Mathematics and Statistics, University of Glasgow, Glasgow, UK*

Abstract

In recent years, in many application fields, extracting information from data in the form of functions is of most interest rather than investigating traditional multivariate vectors. Often these functions have complex spatial dependences that need to be accountied for using appropriate statistical analysis. Spatial Functional Statistics presents a fruitful analytics framework for this analysis. The definition of a distance measure between spatially dependent functional data is critical for many functional data analysis tasks such as clustering and classification. For this reason, and based on the specific characteristics of functional data, several distance measures have been proposed in the last few years. In this work we develop a weighted L_2 distance for spatially dependent functional data, with an optimized weight function. Assuming a penalized basis representation for the functional data, we consider weight functions depending also on the spatial location in two different situations: a classical georeferenced spatial structure and a connected network one. The performance of the proposed distances are compared using standard metrics applied to both real and simulated data analysis.

Keywords: spatially dependent functional data, distance, clustering, spatial dependence

Preprint submitted to Journal of LaTeX Templates *September 18, 2020*

1. Introduction

Frequently in many applied contexts, data are curves available at points in space, and in those cases a challenge is to develop useful multivariate approaches that allow us to pool together information from curves at all locations. Let's think, for example, to the problem of monitoring similarities in temporal patterns of water quality parameters by considering the spatial correlation across networks, this can provide information to feed into future monitoring strategies [9] or to the case in which the aim is to investigate similarities in water profiles like salinity by considering the spatial component [24]. The notion of proximity is one of the most important definitions to provide in such a context. However, in an infinite dimensional space, where the equivalence between the norms fails, it is too restrictive to consider the natural measure of distance induced by the Hilbert space. Thus it becomes crucial to define an appropriate metric for distance.

Developing such a metric is non-trivial because functional data can manifest itself in a variety of ways, with the data subject to a large degree of variability due to the nature of the spatial component. The emerging characteristics for recently developed methods are mainly related to modelling correlated functional data using spatial structures (geostatistical data, point patterns and areal data) that can be combined with functional data [17]. For this reason, and based on the specific characteristics of the spatial component related to the functional data, the scientific community has focused on developing methods based on suitable measures of distance, or similarity, related to clustering ([7],[8],[9],[24], [28]), to the definition of depth [1] and to kriging prediction problems [2].

In this work, we introduce an optimally weighted distance for spatially dependent functional data. Spatially dependent functional data come in many forms. Rougly speaking, such curves, spatially located, refer either to curves observed on points, lines or areal spatial units. The definition of the covariance structure among the functions depends on the spatial structure we observe. For instance, the tracevariogram function [8] enables estimation of the interactions among functions observed on a regular grid in terms of variability; whereas the spatial covariance of Haggarty et al. [9] quantifies the interactions between functions on a connected network. The distance we introduce is a generalization of the distance proposed by [4] for spatially dependent functional data, by considering a regular grid and a connected network.

To illustrate when it is appropriate for each covariance measure to be used we compare our proposed distance measure with differnt metrics proposed in the literature from a practical and theoretical point of view. In particular we will focus on distances measuring explicitly differences in terms of spatial dependence such as the distances proposed by [1], [8], [9].

We will do so in steps, first introducing the nature of functional data of interest here, geostatistical functional data and spatial variability measures in general in section 2, then describing a variety of distances and their applications in section 3. We introduce our proposed metric in section 4 and show the main results by an extensive simulation study in section 5 and applied to two different real data monitoring the approximate weekday volume of passengers between each pair of stations on the London network subway and evapotranspiration problem in the italian pensinsula in section 6.

2. Geostatistical functional data and spatial variability measures

Let $(\chi_{s_1}(t), \dots, \chi_{s_i}(t), \dots, \chi_{s_n}(t))$ be a set n of geostatistical functional data. The n points $(s_1, \dots, s_i, \dots, s_n)$ in $D \subseteq \mathbb{R}^d$ identify the n locations where the random functions $\chi_s\,(t)$ are located.

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

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

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

For a fixed site s_i, it is assumed that the observed functions can be expressed according to the following model:

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

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

For each $t, t \in T$, the random process is assumed to be second order stationary and isotropic: that is, the mean and variance functions are constant and the covariance depends only on the distance between sampling sites.

Formally we have:

$\mathbb{E}(\chi_s(t)) = m(t)\ \forall\, t \in T,\ s \in D$,

$\mathbb{V}(\chi_s(t)) = \sigma^2(t),\ \forall\, t \in T,\ s \in D$, and

$\mathbb{C}ov(\chi_{s_i}(t), \chi_{s_j}(t)) = \mathbb{C}(h, t)$, where $h = \|s_i - s_j\|\ \forall\, s_i, s_j \in D$.

2.1. Trace-variogram and Spatial dispersion function for functional data

It is assumed that the mean function is constant over D and that the semi-variogram function $\gamma(h,t) = \gamma_{s_i s_j}(t) = \frac{1}{2}\mathbb{V}(\chi_{s_i}(t) - \chi_{s_j}(t))$, according to [2], can be expressed by:

$$\gamma(h,t) \ = \ \gamma_{s_i s_j}(t) \ = \ \frac{1}{2}\mathbb{V}(\chi_{s_i}(t) \ - \ \chi_{s_j}(t)) \ = \ \frac{1}{2}\mathbb{E}\left[\chi_{s_i}(t) - \chi_{s_j}(t)\right]^2. \quad (2)$$

By considering the integral on T of this expression, using Fubini's theorem and following [5], 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, estimated as:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt, \qquad (3)$$

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, indeed 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 $\varepsilon > 0$.

Consistently with [5], the estimation of the trace-variogram using (3) involves the computation of integrals that can be simplified by considering that the functions are expanded in terms of basis functions.

It is the continuous version of the variogram for spatio-temporal data, which provides a helpful framework to do spatial prediction and, with particular relevance for this paper, to provide a mechanism to incorporate spatial weights in the computation of distance metrics for spatially dependent curves. The trace-variogram, as the classical variogram for purely spatial data, is used to describe the spatial variability among functional data which is distributed across an entire spatial domain and not related to a specific location of the space. However in many application areas, it can be useful to know how the data recorded at each site contribute to the definition of the spatial variability of the geographic area. To reach this aim, a spatial dispersion function has been defined on a specific location of the space [26].

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

$$\delta^{s_i}(h) = \sum_{s_i, s_j \in N^{s_i}(h)} \left[\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt \right] \tag{4}$$

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

Let F be the set of the *spatial dispersion* functions $\delta^{s_i}(h)$. Thus, the mean of the spatial dispersion functions, at the lag h, is defined by the following function:

$$\bar{\delta}^{s_i}(h) = \frac{1}{|N^{s_i}(h)|} \sum_{s_i \in N^{s_i}(h)} \delta^{s_i}(h). \tag{5}$$

That is:

$$\bar{\delta}^{s_i}(h) = \frac{1}{|N^{s_i}(h)|} \sum_{s_i, s_j \in N^{s_i}(h)} \int_T \left(\chi_{s_i}(t) - \chi_{s_j}(t)\right)^2 dt \tag{6}$$

where $N^{s_i}(h) = \{(s_i, s_j) \colon \|s_i - s_j\| = h\} \subset N(h)$,

and it is such that $N(h) = \cup N^{s_i}(h)$ and $|N(h)| = \sum_i |N^{s_i}(h)|$.

Through straightforward algebraic operations, it can be shown that the average of the dispersion functions is a variogram function, expressed by:

$$\gamma(h) = \frac{1}{2\left|N(h)\right|} \sum_{i=1}^{n} \bar{\delta}^{s_i}(h) 2 \left|N^{s_i}(h)\right| . \tag{7}$$

2.2. Correlation based distance for spatially dependent functional data

The trace-variogram and the spatial dispersion function (eqs. 3 and 4) do not provide a measure of the covariance between functions . The latter is given by the spatial covariance function defined by [9], which provides a measure to describe the relative variability between functions.

If the set $(\chi_{s_1}(t), \ldots, \chi_{s_i}(t), \ldots, \chi_{s_n}(t))$ of curves are estimated by using basis functions, then each estimated curve $\hat{\chi}_i(t)$ can be expressed as the product of a row vector of length p of coefficients $\boldsymbol{c}_i^T$ and column vector of length p of basis functions $\phi(t)$ as follows:

$$\hat{\chi}_i(t) = \boldsymbol{c}_i^T \phi(t)$$

Consequently the functional mean of the n curves is defined by the mean of the basis coefficients representing the set of n curves at space point t as follows:

$$\overline{\chi}_i(t) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{c}_i^T \phi(t)$$

that is

$$\overline{\chi}_i(t) = \overline{\boldsymbol{c}}^T \phi(t)$$

The functional covariance is mainly the product of the difference of two areas computed in relation to a reference curve. An area $\overline{A}$ below the mean curve and an area A_i below a generic estimated curve χ_i with respect to a reference curve χ_l. Given a curve of reference χ_l corresponding to the horizontal line which is below the minimum value of the set of the curves, this can be expressed as set of coefficients $\boldsymbol{c}_l$. The area between the reference line and the mean curve is defined as:

$$\text{Area}(\overline{\chi}_i(r), \chi_l) = \int \left\{\overline{\chi}_i(r) - \chi_l\right\}^2 dr = \left(\overline{\boldsymbol{c}} - \boldsymbol{c}_l\right)^T W \left(\overline{\boldsymbol{c}} - \boldsymbol{c}_l\right) = \overline{A}.$$

In the same way, the area between curve $\hat{\chi}_i(r)$ and reference line χ_l is

$$\text{Area}(\hat{\chi}_i(r), \chi_l) = \int \{\hat{\chi}_i(r) - \chi_l\}^2 \, dr = (\boldsymbol{c} - \boldsymbol{c}_l)^T \, W \, (\boldsymbol{c} - \boldsymbol{c}_l) = A_i.$$

Thus a quantification of the difference between a generic function $\hat{\chi}_i(r)$ and a median curve can be expressed by the difference in terms of magnitude between $\hat{\chi}_i(r)$ and the mean curve $\overline{\chi}_i(r)$, which is the difference of their area as $A_i - \overline{A}$. As stayed in [9], the area between the mean curve and a reference line can be used both to reflect the direction of the difference between a given location and the overall mean. In addition it can be used to standardize the areas so that the measures of covariance are on a most suitable scale. We have considered all issues mentioned in the reviewers' comments carefully, now the paper is The estimated functional covariance between the two estimated functions can thus be defined as:

$$\widehat{\mathbb{C}ov} \left(\hat{\chi}_i(r), \hat{\chi}_j(r)\right) := \frac{(A_i - \overline{A})(A_j - \overline{A})}{\overline{A}^2}. \qquad (8)$$

It is a single covariance value between two functions over a space of interest that can then be used to create an adjusted covariogram cloud.

3. Distances

The most simple distance used between two spatially dependent functional data objects has been proposed in the pioneering work [2]. It is a weighted dissimilarity metric among the geo-referenced curves expressed by

$$d_g(\chi_{s_i}(t), \chi_{s_j}(t)) = d(\chi_{s_i}(t), \chi_{s_j}(t))\gamma_{s_i s_j}(h) \qquad (9)$$

where $d\left(\chi_{s_i}(t), \chi_{s_j}(t)\right) = \sqrt{\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt}$ is the distance between the curves without considering the spatial component, and $\gamma_{s_i s_j}(h)$ corresponds to the trace-variogram function calculated for the distance between sites s_i and s_j. Once the trace-variogram has been estimated, a parametric model is fitted following classical geostatistical estimation procedures [8]. The distance between two curves $\hat{\chi_i}(t)$,$\hat{\chi_j}(t)$ can be calculated from the distance between the coefficients of the basis functions $\boldsymbol{c}_i, \boldsymbol{c}_j$ [8] by:

$$d_{ij} = (\boldsymbol{c}_i - \boldsymbol{c}_j)^T \boldsymbol{W} (\boldsymbol{c}_i - \boldsymbol{c}_j)$$

where $\boldsymbol{W} = \int \phi(r)\phi(r)^T$ is a $p \times p$ square matrix, p representing the number of basis functions.

This distance does not consider the spatial covariance among the functional data, whereas the proposal of [9] is a correlation based distance which groups functions together regardless of the amplitude of their functional variation. It is defined as:

$$d_{ij}^c = d_{i,j} \mathbb{C}ov(\chi_{s_i}(t), \chi_{s_j}(t)). \tag{10}$$

where the covariance function is defined as in Eq.(8). It provides differences in terms of relative magnitude, and summarizes in a single value the correlation between two functions over the spatial domain of interest.

A distance among spatially dependent functional data has been defined as

$$d_{i,j}^2 = \sum_{h \in H} (\delta^{s_i}(h) - \bar{\delta}_k(h))^2, \tag{11}$$

that is the Euclidean distance among two spatial dispersion functions in a site.

4. Optimally weighted distances for spatially dependent functional data

As we have shown, measures of distance between spatially dependent functional data can be distinguished according to the nature of space on which these are defined. Georeferenced functional data, also known as geostatistical functional data, are the basic pieces of information needed to identify the geographic location of phenomena across the space. In general, georeferenced functional data consist of curves taken at specific locations (points referenced by latitude and longitude).

One can distinguish between simple and complex spatially dependent functional data types , depending on the spatial complexity. Simple spatially dependent functional data types provide functions observed on simple object structures like single points, observed on a regular grid. From a formal perspective, they do not cover all the variety and complexity of geographic reality. Complex spatially dependent functional data types like functional data observed on a network enable us to provide a fuller treatment of geographical complexity. The spatial domain is thus a spatial network. It can be viewed as a spatially embedded graph which consists of a set of point objects representing its nodes and a set of line objects describing the geometry of its edges.

Spatial applicative domains are highways, rivers, public transport systems, power lines, and phone lines on which it is possible to register variables varying on the continuum. Our main aim is to introduce a distance for spatially dependent functional data considering (i) the simple (georeferenced) and then (ii) the more complex spatial domains (e.g. a connected directed network), as described above.

Using the idea of [4], we define an optimally weighted distance for functional data spatially dependent.

Assuming a basis function representation for functional data we propose to consider weight functions including both the spatial and functional component. It is a generalization of [4] to the spatial functional framework for two different spatial domains: the georeferenced and the directed network. As in [4] we define a weighted L^2 distance as follows:

$$d_{\omega_s}\left(\chi_{s_i}(t), \chi_{s_j}(t)\right) = \sqrt{\int_T \omega_s(t)(\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt} \tag{12}$$

where the weight ω_s satisfies $\omega_s \geq 0$ and $\int \omega_s dt = 1$.

The problem is choosing a weight function $\omega_s(t)$, such that the seminorm is defined by $||\,(\cdot)\,||_{\omega_s} = \sqrt{\int \omega_s \theta(t)^2} dt$. We define a spatio-functional smooth function $\omega_s(t) = \left[\mathbf{b}_{\omega_s}^T(t)\mathbf{q}\right]^2$ where $\mathbf{b}_{\omega_s}(t)$ is a vector of associated basis functions and $\mathbf{q}$ is the vector of coefficients. Conventionally, these spatio-functional smooth function are expressed in the same basic functions as the observed one, however different basis functions could be chosen.

The spatio-functional smooth function is obtained by the following minimization problem:

$$\omega_s(t) = argmin_{||\omega_s||=1} \frac{\sum_{1\leq i<j\leq n} \mathbb{V}(||\theta_{i,j})||^2_{\omega_s})}{\sum_{1\leq i<j\leq n} [\mathbb{E}(||\theta_{i,j})||^2_{\omega_s})]^2}; \tag{13}$$

with $\theta_{i,j}(t) = a_{i,j}x_i(t) - a_{j,i}x_j(t)$, where $a_{i,j}$ and $a_{j,i}$ are obtained starting from the structure of the spatial domain of interest.

The coefficient $a_{j,i}$ is the element reflecting the spatial dependence among functional data and changes according to the spatial grid on which the functional data are observed. When $a_{i,j} = a_{j,i} = 1$, we have a weighted distance d_ω defined by [4] for functional data without spatial dependence.

In the case of spatially dependent functional data observed on a regular grid, we introduce a weight function depending on the spatial variability expressed by a trace-variogram function. Formally we define:

$$a_{i,j} = a_{j,i} = \hat{\gamma}(h_{i,j}) \tag{14}$$

where $\hat{\gamma}(h)$ is the estimated by (Eq.3)

The introduced distance could be viewed in broad terms as a generalization of the dissimilarity measure defined in (9) with the advantage that the distance is optimally calibrated from the functional and spatial point of view.

In the case of spatially dependent functional data observed on a directed network we introduce a weight as a covariance function depending on a structured oriented graph. In particular denote by $\widehat{\mathbb{C}ov}$ the matrix of estimated spatial covariance between the knots of a net (as in Eq.8), $\Gamma = \text{diag}(\widehat{\mathbb{C}ov_j})$ the diagonal covariance matrix and D the matrix of the L_2 functional distance $d\left(\chi_{s_i}(t), \chi_{s_j}(t)\right) = \sqrt{\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt}$, we define:

$$a_{i,j} = L_i \cdot \Gamma \cdot D_j^T;$$

where L_i is a row vector of matrix of contiguity L (identifying the presence and direction of edges) defined as follows:

$$l_{i,j} = \begin{cases} 1 & \text{if } \exists\, edge \, \text{from} \, s_j \, to \, s_i \, or \, i = j \\ \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

and D_j is the j-th row vector of matrix D.

The above measure can be seen as a generalization of the distance introduced by [9] to a directed network by considering the complex spatial interrelationship between curves.

According to this distinction we can rewrite our distance (Eq. 12) as

$$d_{\omega_s}\left(\chi_{s_i}(t), \chi_{s_j}(t)\right) = \begin{cases} d_{\omega_\gamma}\left(\chi_{s_i}(t), \chi_{s_j}(t)\right) & s = \gamma \\ d_{\omega_C}\left(\chi_{s_i}(t), \chi_{s_j}(t)\right) & s = C \end{cases} \tag{16}$$

Where $s = \gamma$ and $s = C$ correspond to weight functions for spatially dependent functional data observed respectively on a regular spatial grid and on a directed network.

In both of the cases above the defined function satisfies all the properties to ensure it is still a distance (as shown in the supplementary material).

5. Simulation study

To analyse and compare the performances of the proposed distances, a simulation study has been performed. Our main aim is to evaluate distances, as defined above, in terms of their ability to enable identification of proposed clustering structure while taking into account the spatial relationship. Inspired by the simulation scheme proposed in [4], we simulate data from four groups of functions by considering two spatial domains: a regular grid and a network one. We thus have two different scenarios depending on the spatial structure we consider, respectively a regular and a network grid. For each scenario, 4 groups of

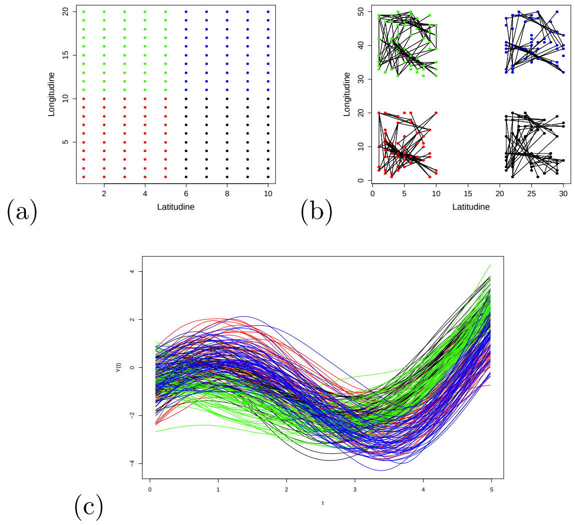


Figure 1: Simulated data on regular grid (a), network grid (b). Four groups of simulated curves (c).

functions were simulated using the same mean functions of [13] as follows:

$$\mu_1(t) = -2\text{sin}(t-1)\log\left(t+\frac{1}{2}\right), \tag{17}$$

$$\mu_2(t) = 2\cos(t)\log\left(t+\frac{1}{2}\right), \tag{18}$$

$$\mu_3(t) = \frac{1}{2} - \frac{1}{5}\text{cos}\left(\frac{1}{2}(t-1)\right)t^{\frac{3}{2}}\sqrt{5t^{\frac{1}{2}}+\frac{1}{2}}, \tag{19}$$

$$\mu_4(t) = \frac{6}{5}\cos(t)\log\left(t+\frac{1}{2}\right)\sqrt{t+\frac{1}{2}}. \tag{20}$$

Where we set:

- $n_k = 50$ curves for mean function μ_k for $k = 1, \ldots, 4$, so we have 200 curves;
- each curve is defined for all $t \in [0, 5]$;

The simulations were built following the procedure used in [4]. As error term we have used function of spatial covariance function introduced in [24]. The data were generated as following:

$$\chi_i(t) = \mu_k(t) + \nu_i(t) + \epsilon_i(h), \ \ k \in \{1, 2, 3, 4\}, \ i = 1, \dots, n,$$

where $\nu_i(t)$ is a subject-specific random process and $\epsilon_i(h)$ is a measurement error process. The subject-specific $\nu_i(t)$, is generated from a Gaussian process with mean 0 and covariance $\sigma_\nu^2(t) = 2\mathrm{e}^{\frac{-(t-2.5)^2}{4}}$. The measurement errors, $\epsilon_i(h)$, is related to the spatial correlation and is generated from $Cov_S(h) = (1-\alpha)\mathrm{e}^{-c|h|} + \alpha\delta_{h=0}$ (introduced in [24]), where $c > 0$ controls the spatial correlation intensity, and $\alpha \in (0, 1]$ is the nugget effect, we fixed $\alpha = 0.04$, $c = 0.01$. In the case of regular grid $\epsilon_i(h) = \sum_{s_j}(1-\alpha)\mathrm{e}^{-c||s_i-s_j||} + \alpha\delta_{||s_i-s_j||=0}$. In the case of network $\epsilon_i(h) = \sum_{s_j}\left((1-\alpha)\mathrm{e}^{-c||s_i-s_j||} + \alpha\delta_{||s_i-s_j||=0}\right)\delta_{l_{i,j}=1}$, where $l_{i,j}$ was definited in equation 15.

The graf structure of the network was built considering a non-regular spatial grid, G. On G, the links were created using the following function: there is a link between sites $s_i(x_i, y_i)$ and $s_j(x_j, y_j)$ of cluster k if and only if $(x_i + y_i + x_j + y_j) \equiv_a b$, based on the choice of parameters a and b, denser networks can be constructed. We fixed $a = 20$ and $b = 5$

Figure 1 shows curves in the two different scenarios, regular and network grid. In the first scenario, the 4 groups of curves observed on a grid of 100 equally spaced time points on a regular spatial grid $[0, 10] \times [0, 20]$ were considered. Figure 1 shows a representation of the set of 200 locations, 50 for each of the four considered clusters (identified with color blue, green, red and violet). In the second scenario, the same groups of curves located on a directed network of 50 equal vertices, with random link, were simulated. Given the generated data, a hierarchical clustering method was evaluated on the two defined scenarios. In order to investigate the effect of the introduction of the spatial weight, for the first scenario we compared the distance d_{ω_γ} of (16) with the original d_ω proposed by [4]. Distances (16) are then compared with the previous defined distances (9) and (10) for the first and second scenario.

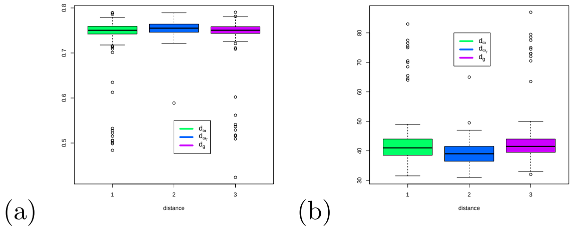


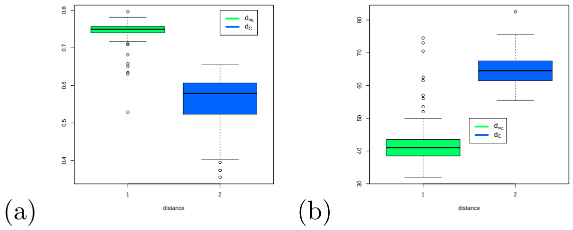
Figure 2: (a) Box plot of one Rand index for 100 simulations for distances d_ω,d_{ω_γ} and d_g. (b) Box plot of MC% (miss-classification %) values of 100 simulation for distances d_ω, d_{ω_γ} and d_g

Figure 2 compares clustering performances with distances d_ω, d_{ω_γ} and d_g on the first scenario. Results coming from the classifications procedure was compared by the Rand index ([21]), and the number of incorrectly classified points, which we refer to as MC%, are displayed in each figure (Fig.2a and 2b). High values of the Rand index imply more accurate classification results. In each figure, boxplots of the Rand Index and the percent of miss-classified curves, between the three hierarchical procedures, for 100 replicates are presented. From this, it becomes clear that the combination between the spatial aspect and the functional optimization enables an improvement in the classification results and better performance is achieved by using the optimized defined distance.

Distance $d_{\omega C}$ was compared with the distance d_C for the second scenario. In this case results illustrate that a network which only considers the correlation between two sites is not enough, incorporating connectivity and direction enables superior results. In Figure 3 we show the boxplot of the Rand index and the MC % values, respectively, for the classification of the curves with the distances d_{ω_C} and d_C on a network structure. It is easy to observe that:

- the distance d_C has the highest MC% values and the lowest Rand index values of the two network structures;
- d_{ω_C} has the highest Rand index values and the lowest MC% values of the two network structures;

A highly evident difference between the results of the two procedures is derived



(a) (b)

Figure 3: (a) Box plot of Rand index for 100 simulations for distances d_{ω_C} and d_C. (b) Box plot of MC% (miss-classification %) values for 100 simulations for distances d_{ω_C} and d_C.

by the assumption in the simulation scheme, where the spatial locations are connected through linkage to more than a couple of sites.

6. Real data Analysis

In this section we introduce two different applications on real data with the aim to illustrate performances of the proposed distances.

6.1. A Meterological study on evapotranspiration in Italy

In this section we show an application of the distance introduced in section 4 to a non-regular grid. We focus on a hierarchical classification of the meteorological time series of evapotranspiration for 12 months, from December 2016 to November 2017, in 103 provinces of Italy. One aim of our analysis was to obtain groups of stations which are similar in terms of evapotranspiration of the determinand of interest. The analyzed meteorological-climatic data are estimated with the data of the daily meteorological historical series of the stations of the Rete Agrometeorologica Nazionale of the Meteorological Service (RAN) of the Air Force and of the Italian regional services. The estimation of the weather-climatic data of the areas (or geographic domains of interest) is performed with a non-stationary geostatistical model that takes into account the location of the stations, the trend and geographical correlation of the meteorological quantities.

The RAN, developed in 1990, is a precious tool for detecting agro-meteorological quantities. The latter are used for the reconstruction of meteorological events

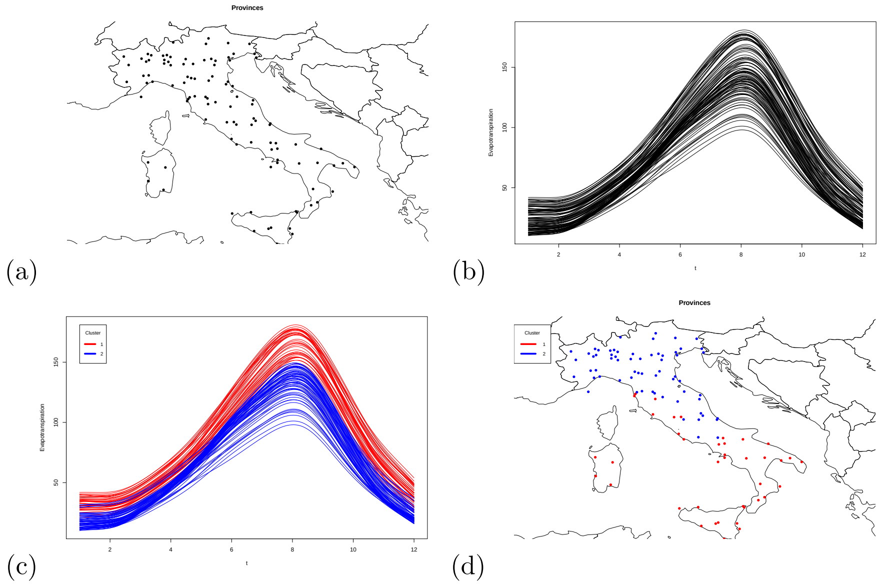


Figure 4: (a) The 103 provinces of Italy. (b) The meteorological time series of evapotranspiration from December 2016 to November 2017.(c) Classification of evapotranspiration curves using the d_{ω_γ}. The curves of cluster 1 are in red and the curves of cluster 2 are in blue. (d) Classification of the 103 provinces of Italy.

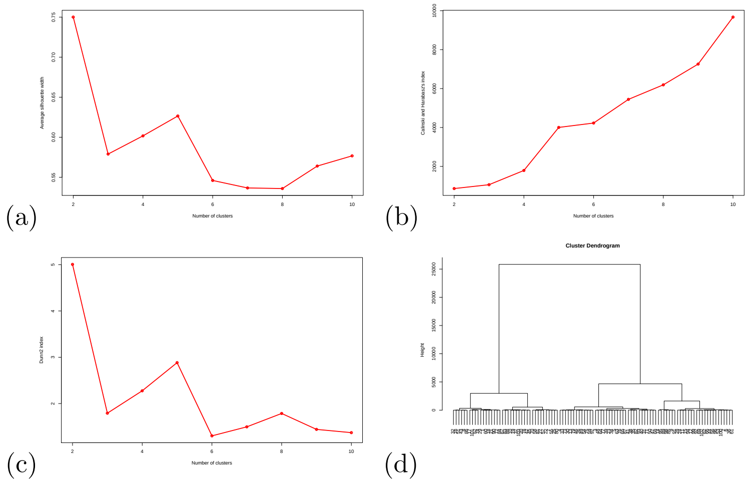


Figure 5: (a) Average silhouette width. (b) Calinski and Harabasz index. (c) Dunn2 index. (d) Dendogram obtained with distance d_{ω_γ}.

(temperature, precipitation, relative humidity, etc.) and for monitoring the agricultural season. The collected data are acquired on an hourly basis and subjected to systematic checks of correctness and physical and meteorological consistency before being stored in the SIAN National Agrometeorological Database and used for agro-meteorological monitoring.

In Figure 4 it is possible to observe the evapotranspiration curves for 103 Italian provinces and the results of a hierachical classification using the distance d_{ω_γ}.

We identify the number of clusters by considering three indices among many proposed in the literature, that is the Average silhouette width [19], Calinski and Harabasz index [3] and Dunn2 index [10]. As can be seen from Figure 5, these indices led to 2 being chosen as the optimal number of clusters. The configuration af the clusters was confirmed by observing the dendogram in Figure 5 (d).

The two groups of curves respectively represented in red and blue reflect the geographic conformation of the provinces (Figures 4(c), 4(d)); the blue curves are associated with the northern Italian provinces while the red curves are as-

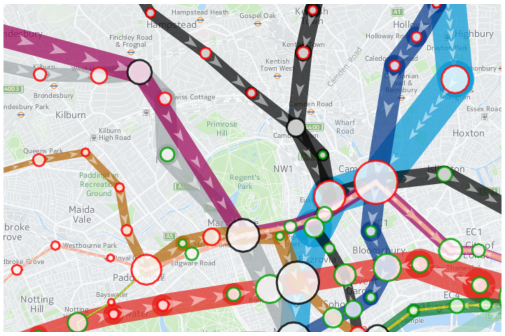


Figure 6: The network configuration.

sociated with the Southen ones. Looking at the two families, we can say that the first cluster, from a functional point of view, is characterized by high values of water evapotranspiration and, from a geographical point of view, covers all the provinces of central-southern Italy; the second cluster, from the functional point of view, is characterized by lower values of water evapotranspiration and, geographically, covers all the provinces of central-northern Italy. It could be expected that the clusters have this configuration, and are coherent by considering the spatial correlation. The reasons for which this could be expected is that locations that are close to the north are more similar in term of temperature.

6.2. Clustering of passenger flows in London's underground railways

A hierarchical clustering of the approximate weekday volume of passengers entering the stations of the network of the London underground railway is proposed. In addition, in order to show how the distance d_{ω_C} introduced here is able to discriminate the clustering structure in the data, by means of spatial component, we compared the obtained results with hierarchical clustering based on the distance of [4]. The data are available from the NUMBAT dataset (http://crowding.data.tfl.gov.uk) which represents the travel demand on a typical autumn weekday, Saturday and Sunday at all stations and lines of the London Underground, London Overground, Docklands Light Railway, TfL Rail / Elizabeth Line and London Trams.

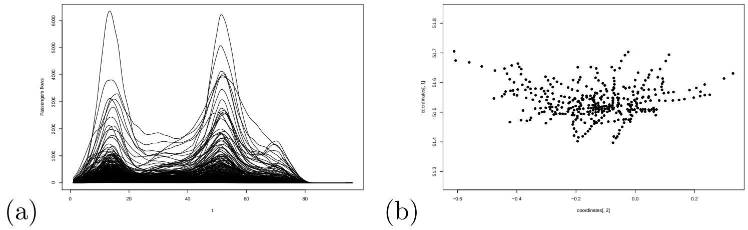


Figure 7: *a*) Passenger flows in London's underground. (*b*) Station distribution on the map.

The data that are used in this work are, in fifteen-minute intervals throughout a weekday, the volume of tube passengers entering the stations of the entire 413 London tube stations. Figure 6 shows the Tube map (also called the London Underground Map or the TfL Services Map). It is a schematic transport map of the lines, stations and services of the London Underground, known colloquially as the Tube, hence the map's name. The individual passenger movements within a subway network include the passengers flowing along links and passengers entering stations, which involve passenger flows through a network. These individual passenger movements within a subway system can be represented by a georeferenced process, where passengers enter/leave their origin/destination stations or nodes, and the resulting flows along links are based on train services. The aim of the analysis was to obtain groups of stations which are similar in terms of flow patterns, while taking into account the spatial links and directions. Figure 7 shows the analyzed data: (*a*) the curves of passenger flows in London's underground and (*b*) the spatial distribution of stations.

Hierarchical clustering results with the distance d_{ω_C} in Figure 8 shows clearly the presence of three clusters. As in the previuos application we observe three different indices for selecting the number of clusters: Average silhouette with, Calinski and Harabasz index, and Dunn index, we thus choose 3 clusters. The dendogram in Figure 8 (*a*) shows a big cluster and two little structures corresponding respectively to the centre of the city where a large and more similar flow of people every day spend their time mainly to go to work and the neighboring areas outside the centre where the London Underground terminates. The

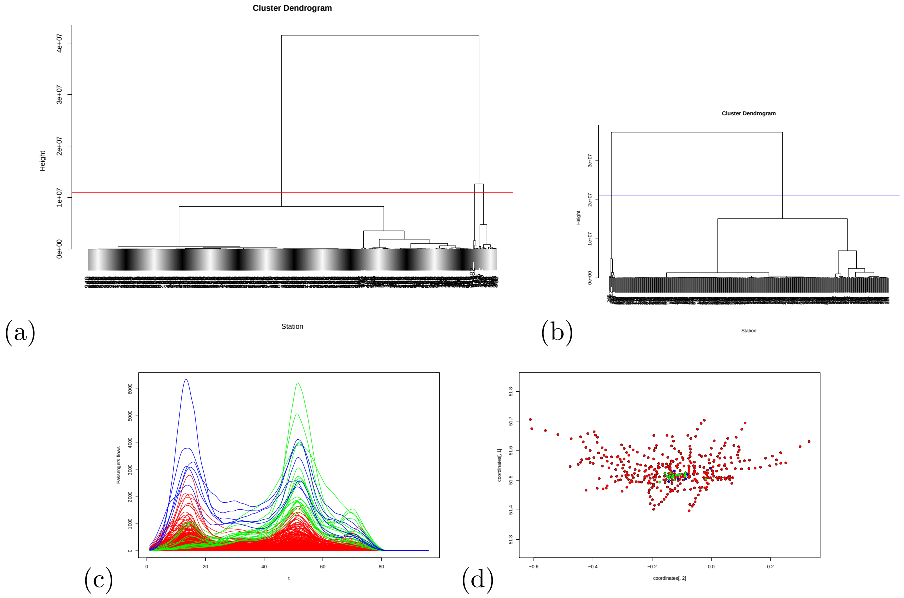


Figure 8: (a) Dendogram by the distance d_{ω_C}. (b) Dendogram by the original Chen distance without the spatial component, (c) Passenger flows clusters in London's underground. (d) Clustering configuration on the map by the distance d_{ω_C}.

clustering structure is clearly shown with the spatial location in Figure 8 (d) and the corresponding clusters of flows in Figure 8 (c). Clusters red and green are the ones detecting the centre of the city while the blue cluster contains the principal stations of the city like, King's Cross-St. Pancras, Liverpool Street LU, London Bridge LU, Stratford, Victoria LU, Waterloo LU. This structure is more confused by observing results obtained by a hierarchical clustering method with the adaptive distance of [4] in Figure 8 (b), where one cluster is masked. Our newly proposed distance measure reveals hidden structure which is masked by standard metrics such as [4] as illustrated by 8. This dendogram only suggests 2 clusters with one informative clusters masked.

7. Discussion

We have introduced a weighted L_2 distance for spatially dependent functional data, with an optimized weight function. Assuming a penalized basis representation for the functional data, a weight function depending also on the spatial location in two different cases a classic georeferenced spatial structure and a connected directed network based one has been proposed. The performance of the proposed distances shows promising results on simulated and real data analysis. We compared them by means of a hierarchical clustering method in two different scenarios covering an exstensive simulation plan, by showing how the inclusion of spatial information in the adaptive distance is informative. The results from the applications of our distances to real data highlighted the effect of including the spatial covariation in the network. This suggests that methods based on a adaptive spatially weighted distances perform better and reveal hidden structure. We have provided two illustrative examples. However, our proposed distance metrics are general and widely applicable for incorporation in different methods and application domains.

The proposed distances in this paper are defined for functional data without reference to a possible stochastic process in the time domain, while the space is treated as a bidimensional argument of a stochastic stationary process with a covariance function defined among functions. A direction for the future research will consist in extending the defined wighted function for functional data with reference to a temporal covariance structure.

Kriging methods and other geostatistical procedure will be object of further analysis by using our proposed distances.

References

[1] Balzanella A., Romano E., Verde R.: Modified half-region depth for spatially dependent functional data, Stochastic Environmental Research and Risk Assessment, 31: 87-103, (2017)

[2] Caballero, W., Giraldo, R., Mateu, J.: A universal kriging approach for spatial functional data. Stochastic Environmental Research and Risk Assessment. Volume 27, Issue 7, pp. 1553-1563, (2013)

[3] Calinski, T., Harabasz, J., A dendrite method for cluster analysis, Communications in Statistics, 3, no. 1:1-27, (1974)

[4] Chen H., Reiss P.T., Tarpey T.:Optimally Weighted L2 Distance for Functional Data.Biometrics, 70(3): 516–525, (2014).

[5] Delicado, P., Giraldo, R., Comas, C. and Mateu, J.: Statistics for spatial functional data: some recent contributions. Environmetric, 21: pp.224-239, (2010)

[6] Ferraty F., Vieu P. NonParametric Functional Data Analysis Theory and Practice. Springer Series in Statistics, Springer, New York. (2006)

[7] Fortuna F., Di Battista T.: Functional unsupervised classification of spatial biodiversity, Ecological Indicators, 111, doi: https://doi.org/10.1016/j.ecolind.2019.106027. (2020)

[8] Giraldo, R., Delicado, P., Comas, C., Mateu, J.: Hierarchical clustering of spatially correlated functional data. Statistica Neerlandica, (2011)

[9] Haggarty, R., Miller, C., Scott, E.M.: Spatially Weighted Functional Clustering of River Network Data. Journal of the Royal Statistical Society, Series C.,(2015)

[10] Halkidi, M., Batistkis, Y., Vazirgiannis M., On Clustering Validation Techniques, Journal of Intelligent Information Systems, 17:2/3, 107-145 (2001)

[11] Hennig, C., Lin, C-J.: Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. Statistics and Computing , 25 (4) pp. 821-833. 10, (2015)

[12] Hennig, C., Liao, T. F., How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification, Royal Statistical Society, Appl. Statist. 62 Part 3, 309-369 (2013)

[13] Hitchcock, D. B., Booth, J. G., and Casella, G., The effect of pre-smoothing functionaldata on cluster analysis. Journal of Statistical Computation and Simulation 77, 1043-1055 (2007)

[14] Jiang,H.,Serban,N.: Clustering Random Curves Under Spatial Interdependence: Classification of Service Accessibility. Technometrics (2010)

[15] Journel A.G., Huijbregts Ch. J.: Mining Geostatistics, The Blackburn Press.(2004)

[16] Lieb, H.,E, Loss, M., Analysis, American Mathematical Society (2001)

[17] Mateu J., Romano, E.: Advances in Spatial functional Statistics. Stochastic Environmental Research and Risk Assessment, Vol. 31, 1, pp 1–6, (2017)

[18] Milligan, G. W., Cooper, M. C., An examination of procedures for determining the number of clusters in a data set, Psychometrika, 50, 159-179 (1985)

[19] Rousseeuw, P. J., Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis, Computational and Applied Mathematics. 20: 53-65 (1987)

[20] Ramsay, J.E., Silverman, B.W.: Functional Data Analysis, (Second ed.) Springer (2005)

[21] Rand, W.M.: Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association. Vol. 66, No. 336, pp.846-850, (1971)

[22] Romano E., Balzanella A., Verde R, Clustering Spatio-functional data: a model-based approach. Studies in Classification, Data Analysis, and Knowledge Organization Springer, Berlin-Heidelberg, New York, (2010)

[23] Romano E., Balzanella A., Verde R., A new regionalization method for spatially dependent functional data based on local variogram models: an application on environmental data. In: Atti delle XLV Riunione Scientifica della Societá Italiana di Statistica Universitá degli Studi di Padova Padova. CLEUP, Padova (2010)

[24] Romano, E., Mateu, J., Giraldo, R.: On the performance of two clustering methods for spatial functional data. AStA Advances in Statistical Analysis, 99 (4), pp. 467-492, (2015).

[25] Romano E., Verde R.: Clustering geostatistical functional data. Advanced Statistical Methods for the analysis of large data-sets, series Studies in Theoretical and Applied Statistics, A. Di Ciaccio, M. Coli, J.M. Angulo eds., Springer, Berlin (2011)

[26] Romano E., Balzanella A., Verde R., Spatial variability clustering for spatially dependent functional data, Statistics and Computing, 27, 3,(2016)

[27] Rouseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. Vol. 20, No. 1, pp. 53-65, (1987)

[28] Secchi P., Vantini S., Vitelli V.: Bagging Voronoi classifiers for clustering spatial functional data.International Giornal of Applied Earth Observation and Geoinformation, pp.53-64, (2012)

[29] Sun, Y., and Genton, M. G.: Adjusted Functional Boxplots for Spatio-Temporal Data Visualization and Outlier Detection. Environmetrics, 23, pp.54-64, (2012)

[30] Sun, Y., Li, B., and Genton, M. G.: Geostatistics for large datasets. In: Space-Time Processes and Challenges Related to Environmental Problems, E. Porcu, J. M. Montero, M. Schlather (eds), Springer, Vol. 207, Chapter 3, pp.55-77, (2012)