Spatial Principal Component Analysis and Moran Statistics for Multivariate Functional Areal Data

Dharini Pathmanathan[1], Issa-Mbenard Dabo[2], Tzung Hsuen Khoo[1], Alaa Ali-Hassan[3], and Sophie Dabo-Niang[4]

[1]Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

[2]Institut de mathématiques de Bordeaux, University of Bordeaux, France

[3]Centre de Recherche et d'Innovation en Intelligence Énergétique (CR2ie) 175, rue De La Vérendrye Sept-Îles (Québec) G4R 5B7

[4]CNRS, UMR 8524-Laboratoire Paul Painlevé, INRIA-MODAL, Université Lille, F-59000 Lille, France

Abstract

In this article, we present the bivariate and multivariate functional Moran's I statistics and multivariate functional areal spatial principal component analysis (mfasPCA). These methods are the first of their kind in the field of multivariate areal spatial functional data analysis. The multivariate functional Moran's I statistic is employed to assess spatial autocorrelation, while mfasPCA is utilized for dimension reduction in both univariate and multivariate functional areal data. Through simulation studies and real-world examples, we demonstrate that the multivariate functional Moran's I statistic and mfasPCA are powerful tools for evaluating spatial autocorrelation in univariate and multivariate functional areal data analysis.

1 Introduction

Functional data analysis (FDA) is widely applied in numerous scientific disciplines. This technique involves modelling discrete observations as functions and analyzing data across a curve, surface, or continuum. FDA

encompasses areas such as dimension reduction, clustering, regression, and classification. Ramsay and Silverman (2005) offered essential information on the use of FDA methods in various analytical scenarios, especially to reduce the dimensions of functional data. Principal component analysis (PCA) along with its functional version (FPCA) is widely used to reduce dimensions. The advent of new types of functional data has led to the development of novel PCA methods for both univariate and multivariate functional data (Bali and Boente, 2014; Hörmann et al., 2015; Happ and Greven, 2018). The conventional FPCA method used for independent and identically distributed functional variables is being progressively expanded to encompass temporal or spatial dynamics within functional series. This introduces a significant challenge in modeling temporal or spatial dependencies in functional data.

Studying spatially functional data entails integrating techniques from functional data analysis and spatial statistics, with the goal of developing versatile methods for modeling relationships while maintaining flexibility and computational efficiency. This approach enables the exploration of intricate spatial or spatio-temporal data with high dimensionality. Spatial functional data analysis considers collections of functions observed at different locations within a region, commonly known as spatially correlated functional data (Mateu and Romano, 2017). Delicado et al. (2010) introduced an effective method that combines classical types of spatial data structures (including geostatistical data, point patterns, and areal data) with functional data. Driven by the growing availability of spatial functional data in various fields, there have been significant advancements in spatial-functional data analyses (Mateu, Koner and Staicu (2023)). Among these advancements, PCA techniques are particularly noteworthy in this work.

Principal component analysis has been utilized within a geospatial framework (e.g., Li and Guan (2014), Liu et al. (2017), Kuenzer et al. (2021)). Concerning spatial areal data, the basis of our previous work (Hassan, 2021), which includes functional areal data, lays the foundation for our present study. Areal or lattice data arise when a study region is partitioned into a limited number of areas, with outcomes being aggregated or summarized within those areas. Spatial functional data refer to data comprising curves or functions that are recorded at each spatial location. Delicado et al. (2010) offered a comprehensive analysis of the context of functional areal data, which is the foundation for this article. Spatial principal component analysis (sPCA) for multivariate areal data, as introduced by Jombart et al. (2008), was designed to explore the spatial patterns of genetic variability using allelic frequency data from individuals or populations. Compared to the classical PCA, sPCA is more effective in uncovering spatial linkages within spatial data. The concept of sPCA (Jombart et al., 2008), which aims to uncover spatial patterns by explicitly incorporating spatial information, was extended to multivariate functional spatial areal data in our study to identify spatial autocorrelation. Specifically, in this study, we extend the work of Jombart et al. (2008) and Hassan (2021) to investigate both Moran's I and sPCA for multivariate spatial functional data. To the best of our knowledge,

this is the first proposal to address this issue within a multivariate context. The functional Moran's I statistic introduced here is novel within the multivariate functional data framework, with no prior work from this perspective. Romano et al. (2022) introduced a univariate functional Moran's I by enhancing a local spatial association index, the local Moran's I, to evaluate spatial dependence among density functions. Their proposed indices were applied to an areal dataset derived from official statistics in the United States. This work was not noted when (Hassan, 2021), which follows a similar direction as a thesis, was published. The univariate Moran's I from this latter work was utilized by (Khoo et al., 2023) to examine spatial dependency in complex spatial data, specifically to analyze the spatial autocorrelation of global stock indices during the 2015-2016 global market sell-off.

An alternative in the realm of PCA is the method proposed by Krzyśko et al. (2023), who introduced spatio-temporal principal component analysis (STPCA) using a partially functional data framework. This approach involves a two-step interdependent process: initially, the original multivariate time series are transformed into raw coefficients of a basis function expansion, followed by the construction of the principal spatiotemporal components based on a classical Moran's I statistic of the matrix of basis coefficients, in contrast to the functional embeddings used in our study. Furthermore, we embedded the data into a multivariate function space, treating them as fully observed functions. We then developed novel bivariate and multivariate versions of the functional Moran's I and introduced a multivariate functional areal spatial principal component analysis (mfasPCA) method for the areal spatial multivariate framework. This methodology assists in detecting spatial autocorrelation within multivariate functional areal spatial data. Moreover, our work significantly advances the spatial-functional PCA framework. Our approach addresses the limitation in Krzyśko et al. (2023) by introducing functional Moran statistics from univariate, bivariate, and multivariate perspectives within a fully functional framework, extending the univariate functional Moran's I and PCA approach of Romano et al. (2022) and (Hassan, 2021) to the multivariate case.

We compare the recently introduced mfasPCA with STPCA and the multivariate fPCA (MFPCA) by Happ and Greven (2018), which does not account for spatial autocorrelation. This study goes beyond just presenting the multivariate functional Moran's I statistic; it also introduces a novel spatial PCA method tailored for univariate and multivariate functional areal data, thereby enhancing the understanding of spatial dependencies. The mfasPCA framework that we propose is intended to be compatible with both global and local structures, which are effectively identified using the functional Moran's I statistic.

The remainder of the paper is organized as follows: Section 2 details the methodology for univariate, bivariate, and multivariate mfasPCA, along with the corresponding multivariate functional Moran's indices. Section 3 presents simulation studies demonstrating the proposed methods' performance. Section 4 applies these methods to real data for both univariate and multivariate PCA. Section 5 concludes.

2 Methodology

2.1 Multivariate functional principal component analysis on areal spatial data

Consider n spatial points $s_i \in \mathcal{I}$, simplified to i. At each location $s_i \in \mathcal{I} \subset \mathbb{Z}^2$ within a lattice region V, we have one observed d-dimensional measurement $Y_{i,x} = (Y_{i,x_1}^1, \ldots, Y_{i,x_d}^d)^\top$ where $d \geq 1$. Here, $x = (x_1, \ldots, x_d)^\top \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d \subset \mathbb{R}^d$. These data points $Y_{i,x}$ are noisy observations of a smooth areal stochastic multivariate functional process $\{S_i = (S_i^1, \ldots, S_i^d)^\top\}_{i \in \mathcal{I}}$:

$$Y_{i,x} = \mu(x) + S_i(x) + \epsilon_{i,x} = X_i(x) + \epsilon_{i,x}. \tag{1}$$

Here, $\mu(.) = (\mu^1(.), \ldots, \mu^d(.))^\top$ is the mean function. The unobserved variables $\{\epsilon_{i,x}, i = 1, \ldots, n\}$ are independent and identically distributed with zero mean Gaussian measurement errors and variance σ^2. The n multivariate functions $S_i(.)$ are centered spatio-temporal square-integrable functional random variables on the spatial domain $\mathcal{I}$. Namely, we consider that at the n spatial units located on $\mathcal{I}$, we observe a multivariate spatial functional process $\{S_i(.) = (S_i^1(.), \ldots, S_i^d(.))^\top\}$, where $i = 1, \ldots, n$, $S_i^j = \{S_i^j(x_j), x_j \in \mathcal{X}_j\}$. For $1 \leq j \leq p$, let $\mathcal{X}_j$ be a compact set in $\mathbb{R}$, with finite (Lebesgue-) measure and such that $S_i^j : \mathcal{X}_j \longrightarrow \mathbb{R}$ is assumed to belong to $\mathcal{L}^2(\mathcal{X}_j, \mathbb{R})$, the space of real-valued square-integrable functions on $\mathcal{X}_j$. In the following let $\mathcal{L}^2(\mathcal{X}_j, \mathbb{R}) = \mathcal{L}^2(\mathcal{X}_j)$. Note that the special case $d = 1$ corresponds to the univariate spatial-functional case (Hassan, 2021).

We denote by $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$, the p-Fold Cartesian product of $\mathcal{X}_j$. So, S_i is a multivariate functional random variable indexed by $\mathbf{t} = (t_1, \cdots, t_d) \in \mathcal{X}$ and taking values in the p-Fold Cartesian product space $\mathcal{H} := \mathcal{L}^2(\mathcal{X}_1) \times \cdots \times \mathcal{L}^2(\mathcal{X}_d)$. Let the inner product $\langle\langle \cdot, \cdot \rangle\rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, for $f, g \in \mathcal{H}$:

$$\langle\langle f, g \rangle\rangle := \sum_{j=1}^d \langle f_j, g_j \rangle = \sum_{j=1}^d \int_{\mathcal{T}_j} f_j(t_j) \bar{g}_j(t_j) dt_j.$$

Then, $\mathcal{H}$ is a Hilbert space with respect to the scalar product $\langle\langle \cdot, \cdot \rangle\rangle$ Happ and Greven (2018).

The focus is on a multivariate functional PCA investigation, wherein the classical PCA is substituted with its spatial counterpart to consider spatial autocorrelation on the the functional variable of interest at the sampling locations. This autocorrelation may be quantified by a weight matrix depending on the neighboring locations.

We postulate in the following a Karhunen-Loève expansion (Ash and Gardner, 1975):

$$S_i(x) = \sum_{k=1}^\infty \beta_{k,i} \phi_k(x), \tag{2}$$

4

where ϕ_k's are the orthonormal eigenfunctions (functional principal components, FPC) and $\beta_{k,i}$ are auto-correlated scores (see Happ and Greven (2018) in the geostatistical case). In practice, the sum is truncated to a finite integer, K which is to be chosen.

To compute the FPCs, let us express the sample data $(S_i)_{i=1,...,n}$ by means of a basis expansion:

$$S_i(x) = \sum_{m=1}^{\infty} c_{i,m} B_m(x) \approx \sum_{m=1}^{p} c_{i,m} B_m(x), \quad x \in \mathcal{X}, \tag{3}$$

where $B_m(.) = (B_m^1(.), ..., B_m^d(.))^\top$ is some collection of multivariate basis functions of $\mathcal{H}$, $c_{i,m} = \langle\langle S_i, B_m \rangle\rangle$ have zero-mean. In practice, the first p functions are used where a sufficiently large p is good for approximation. Ramsay and Silverman (2005) presented two main basis systems for building functions. The Fourier basis system is commonly used for periodic data while the B-spline basis system is preferable for non-periodic data (Ramsay et al., 2009; Happ and Greven, 2018).

We present a multivariate functional Moran statistic, which, to our knowledge, is the first of its kind in the literature, in Section 2.1.1.

2.1.1 Multivariate Functional Principal Components and Multivariate Functional Moran's I Statistic

The univariate form of the functional Moran's I, as previously mentioned, has been extensively covered in Hassan (2021) and Romano et al. (2022), where detailed derivations are presented. The well-known Moran's statistic has been generalized to the multivariate functional context. This generalization takes into account spatial dependence in PCA to evaluate the degree of spatial autocorrelation among observations within the geographic space $\mathcal{I}$, (Jombart et al., 2008). Let $W = (W_{ij})$ represent a weighted spatial matrix where W_{ij} signifies the neighbouring relation between locations i and j. Assume W is standardized so that the rows sum to one. The functional Moran's index for the n row vector $S_i(x)_{i=1,...,n}$ is then introduced as follows:

$$I_n(\mathbf{S}(x)) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} S_i^\top(x) S_j(x)}{\sum_{i=1}^{n} S_i^\top(x) S_i(x)} = \frac{C_n(\mathbf{S}(x))}{\sigma_n(\mathbf{S}(x))}, \tag{4}$$

where

$$C_n(\mathbf{S}(x)) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} S_i^\top(x) S_j(x) \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{m=1}^{p} \sum_{l=1}^{p} W_{ij} c_{i,m} c_{j,l} B_m^\top(x) B_l(x) = \frac{1}{n} \sum_{j=1}^{d} \mathbf{B}^j(x)^\top \mathbf{X}^\top W \mathbf{X} \mathbf{B}^j(x), \tag{5}$$

and

$$\sigma_n(\mathbf{S}(x)) = \frac{1}{n}\sum_{i=1}^{n} S_i^\top(x)S_i(x) \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{p}\sum_{l=1}^{p} c_{i,m}c_{i,l}B_m^\top(x)B_l(y) = \frac{1}{n}\sum_{j=1}^{d}\mathbf{B}^j(x)^\top\mathbf{X}^\top\mathbf{X}\mathbf{B}^j(x). \tag{6}$$

$\mathbf{X}$ is the $n \times p$ matrix composed of the scores $(c_{i,m})_{i=1,...,n;m=1,...p}$ of S_i, $\mathbf{B}^j(x)$ is the $p \times 1$ vector of components $B_m^j(x)$, $m = 1,...,p$, $j = 1,...,d$, $\mathbf{S}(x)$ is the vector of functions $S_i(x)$.

The trace functional Moran's index is then introduced as:

$$I_n(\mathbf{S}) = \int_X I_n(\mathbf{S}(x))dx. \tag{7}$$

The classical univariate Moran's index (Eckardt and Mateu, 2021; Jombart, 2008) of the n raw vector $\mathbf{X}_m$ of components $\{c_{i,m}\}_{i=1,...,n}$ is

$$\tilde{I}(\mathbf{X}_m) = \frac{\mathbf{X}_m^\top W \mathbf{X}_m}{\mathbf{X}_m^\top \mathbf{X}_m}.$$

Let

$$V(\mathbf{X}_m) = \frac{1}{n}(\mathbf{X}_m^\top\mathbf{X}_m)\tilde{I}(\mathbf{X}_m) = \frac{1}{n}\mathbf{X}_m^\top W \mathbf{X}_m.$$

It is highly positive when $\mathbf{X}_m$ has a large variance and shows a global spatial structure and is negative in a situation with high variance and gives a local structure.

The purpose of the multivariate functional areal spatial principal component (mfasPCA) proposed here is based on scaled $\mathbb{R}^p$ vectors $\mathbf{u}$ (loadings) ($\|\mathbf{u}\|$=1) such that the n raw vectors $\chi = \mathbf{X}\mathbf{u}$ (where $\mathbf{X}$ are the scores defined in (5) are scattered and spatially autocorrelated. In other words, this aims to find the extreme values (Jombart et al., 2008) of

$$C(\mathbf{u}) = V(\mathbf{X}\mathbf{u}) = \frac{1}{n}\mathbf{u}^\top\mathbf{X}^\top W \mathbf{X}\mathbf{u}. \tag{8}$$

The solutions (Jombart et al., 2008) are the eigenvectors $\mathbf{u}_k$ of $\frac{1}{2n}\mathbf{X}^\top(W + W^\top)\mathbf{X}$ associated with the largest and smallest eigenvalues $\alpha_k = var(\chi_k)\tilde{I}(\chi_k)$ (where $\chi_k = \mathbf{X}\mathbf{u}_k$, $var(\chi_k)$ the variance of χ_k). Note that some eigenvalues α_k may be negative since $\tilde{I}(\chi_k)$ is not always positive.

By the help of orthonormal vectors $\mathbf{u}_k$ and their eigen-values α_k, the estimated functional loading (eigenfunction), $\hat{\phi}_k(x)$ of the functional spatial areal PCA is introduced.

In fact, approximating $\mathbf{X}$ by

$$\mathbf{X} \approx \widehat{\mathbf{X}} = \sum_{k=1}^{K}\chi_k\mathbf{u}_k^\top,$$

based on K (sufficiently large) relevant scores χ_k corresponding to the K largest (in absolute values) eigen-

values, lead to

$$\mathbf{S}(x) \approx \widehat{\mathbf{X}}\mathbf{B}(x) = \sum_{k=1}^{K} \chi_k \mathbf{u}_k^\top \mathbf{B}(x),$$

where $\mathbf{B}$ is the $p \times d$ matrix with d columns composed of the functions $\mathbf{B}^j$, $j = 1, ..., d$. The functional multivariate spatial PCA is then obtained by letting the estimated eigen-functions as $\hat{\phi}_k(x)^\top = \mathbf{u}_k^\top \mathbf{B}(x)$.

Then the mfasPCA decomposition is obtained using (3) where the orthonormality of the vectors $\mathbf{u}_k$ and the functions B_m gives:

$$S_i(x) \approx \sum_{k=1}^{K} \hat{\beta}_{k,i} \hat{\phi}_k(x), \tag{9}$$

$$X_i(x) \approx \hat{\mu}(x) + \sum_{k=1}^{K} \hat{\beta}_{k,i} \hat{\phi}_k(x), \tag{10}$$

where $\hat{\mu}(x) = \frac{1}{n} \sum_{i=1}^{n} X_i(x)$, is the empirical mean with $\hat{\beta}_{k,i} = \langle\langle S_i, \hat{\phi}_k \rangle\rangle$.

The principal component (PC) scores derived from the multivariate spatial principal component analysis (mfasPCA) exhibit two distinct types of patterns, classified as global and local structures (Jombart et al., 2008). The global pattern distinguishes between two spatial groups or illustrates a cline (or any intermediate state), whereas the local pattern captures more pronounced genetic differentiation among neighboring entities compared to random pairs Jombart et al. (2008). The global pattern is indicative of positive spatial autocorrelation, while the local pattern signifies negative spatial autocorrelation Jombart et al. (2008). In the following section, mfasPCA is implemented utilizing the `fda` (Ramsay et al., 2020), `adegenet` (Jombart, 2008), `ade4` (Chessel et al. (2004); Dray and Dufour (2007); Dray et al. (2007); Bougeard and Dray (2018)), and `adespatial` (Dray et al., 2019) packages available in the R software.

Note that $I_n(\mathbf{S}(x))$ does not take into account the interrelation between the measurements of two distinct components of S_i at the same spatial location s_i (Eckardt and Mateu, 2021). To overcome this limitation, we extend the bivariate Moran's I statistic of (Eckardt and Mateu, 2021) to the functional case:

$$I_{kl}(\mathbf{S}(x)) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} S_i^k(x) S_j^l(x)}{\sqrt{\sum_{i=1}^{n} S_i^k(x)^2} \sqrt{\sum_{i=1}^{n} S_i^l(x)^2}}, \; k, l = 1, ..., d. \tag{11}$$

The steps involved in developing mfasPCA, starting from the presentation of the new multivariate functional Moran's index in (4) through to our method's derivation of $\hat{\phi}_k(x)$, close the gap in Krzyśko et al. (2023) regarding a fully functional data-oriented framework.

7

2.2 Implementation of the functional Moran's I statistic on spatial weight matrices

The Moran's I statistic has been augmented and utilized within the multivariate functional framework. For robustness of the finite sample properties of the Moran's I and the mfasPCA, spatial weight matrices categorized into distance-based and boundary-based weights are employed. In this study, we use the k-nearest neighbors (KNN) weight matrix to signify the distance-based weights. A KNN matrix encompassing the n spatial locations under consideration, representing the five nearest neighbors of each location, is developed. The weighted spatial matrix W_{ij} can be distinguished into distance-based weights and boundary-based weights. The distance d_{ij} between every pair of spatial units (such as regions, cities, centroids, ...); i and j is utilized to formulate spatial weight matrices for distance-based weights. The k-nearest neighbour weights are given as

$$w_{ij} = \begin{cases} 1 & \text{if } j \in N_k(i); \\ 0 & \text{otherwise}; \end{cases}$$

where $N_k(i)$ is the set of the k closest units or regions to i for $k \in \{1, \ldots, n-1\}$.

For weight matrices based on boundaries, the spatial contiguity is often used to specify neighbouring locations that share a common boundary. There are various types of spatial contiguities but the classical cases are those known as the rook's contiguity (where two cells of a matrix which share a common boundary are neighbours), the bishop's contiguity (where two cells of a matrix share a common vertex) and the queen's contiguity (neighbours by either the rook's or the Bishop's contiguity). The contiguity weights are given as

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are contiguous}; \\ 0 & \text{otherwise}. \end{cases}$$

A Shapiro-Wilk test to assess multivariate normality on the data is required to investigate if the data fulfills the normality assumption. If the data fulfils this assumption, the computation of the functional Moran's I statistic for the KNN- and contiguity-based neighbourhoods can be executed to show spatial autocorrelation of the data of interest. If the data violate the normality assumption, a Monte Carlo simulation on the Moran I statistic is required. It is essential to investigate the global autocorrelation, which measures the degree of clustering, as well as the local indicators, which allows the decomposition of the Moran's I global indicator into the contribution of each observation. The spdep (Bivand et al., 2023a) package in the R software can be used to aid the use of spatial weights, as well as the Moran I statistic to measure spatial autocorrelation.

3 Simulation studies

We conducted simulation studies to assess the effectiveness of the newly proposed fasPCA compared to its univariate and multivariate counterparts.

3.1 Univariate case

We consider a grid consisting of 50×50 locations with random allocation of $N = 100, 250$ and 500 spatial units. The data is generated according the following model:

Model 1:

$$X_i(t) = t\alpha_i + \epsilon_i(t), \; \alpha_i \sim \mathcal{U}(-3, 3),$$

$\{\epsilon_i(t)\}$ is a Gaussian process with exponential covariance, where $i \in [1, N]$, denotes a spatial location $\mathbf{s} = (s_1, s_2)$, sampled in the grid. The curve X_i is observed at 101 evenly spaced times of $[0, 1]$.

$$Y_i(t) = \rho \sum_{j=1}^{n} \omega_{i,j} Y_j(t) + X_i(t), \; t \in [0, 1].$$

We compare the performance of fasPCA with fPCA using data generated from Model 1.

The results of the simulation study are presented in Figure 1 which depict the variance explained (%) for the top four principal components in both fPCA and fasPCA based on the kNN weight matrix. The eigenvalues for every iteration are stored in increasing order. The model in this simulation study has been designed to retain the global and local structures. The variance explained by fasPCA is less dispersed as the spatial autocorrelation ρ increases. In general, the findings indicate that fasPCA outperforms fPCA even in cases of low spatial autocorrelation because of fasPCA's ability to effectively capture spatial autocorrelation. Contiguity weights consistently yield the same outcome, favouring fasPCA. Two additional models were employed, and their results are presented in the Supplemental Material.

3.2 Multivariate case

We consider a grid consisting of locations from the second-level administrative division of France (Figure 4 using the `raster` package (Hijmans, 2023) in R. Subsequently, data is generated according to the following model:

$$X_i^j(t) = t\alpha_i^j + u_i^j(t), \; t \in \mathcal{X}_j = [0, 1]$$

$$\alpha_i^j \sim \mathcal{U}(-3, 3), \; j = 1, ..., d$$

$\{u_i^j(t)\}$ is a Gaussian process with exponential covariance, where $i \in [1, N]$, denotes a spatial location $\mathbf{s} = (s_1, s_2)$, sampled in the grid. The curve X_i is observed at 101 evenly spaced time points of $[0, 1]$.

$$Y_i^j(t) = \rho \sum_{k=1}^{n} \omega_{i,k} Y_k^j(t) + X_i^j(t) + \varepsilon_i^j(t), \ t \in [0, 1], j = 1, ..., d$$

where the ε_i^j are i.i.d centered functional Gaussian random variables with $var(\varepsilon_i^j(t)) = \sigma^2$, $cov(\varepsilon_i^j(t), \varepsilon_i^j(u)) = 0$, $t \neq u$. The spatial weight matrix $W = (\omega_{i,j})_{i,j=1,...,N}$ was constructed using the nine weight matrices employed by Krzyśko et al. (2023) with $\rho = 0.3, 0.5, 0.7, 0.9$. The multivariate data generated for the simulation study consists of 10 variables, $d = 10$. The spatio-temporal principal components corresponding to positive eigenvalues (global principal components) and negative eigenvalues (local principal components) are expressed as percentages as given in Krzyśko et al. (2023). Our newly proposed mfasPCA was compared with MF-PCA and STPCA for global PCs and with STPCA for local PCs using this model over 50 iterations. MFPCA does not account for local structures. Functional principal components focus on outcomes associated with functional areal spatial data while spatio-temporal principal components encompass two types of data that reflect the spatial variability of the studied objects (Krzyśko et al., 2023).

Figure 2 shows the variance explained for the top three positive principal components for MFPCA, mfasPCA, and STPCA using rook contiguity weights. The proposed mfasPCA approach outperforms both mfPCA and STPCA. Since mfPCA does not account for spatial autocorrelation, we evaluated the performance of STPCA and mfasPCA with respect to local spatial autocorrelation, and mfasPCA emerged as the superior approach. Eight other weight matrices, as described in Krzyśko et al. (2023), were used to compare the three approaches, and in general, mfasPCA consistently outperformed its counterparts (see Supplemental Text). It is not surprising that MFPCA outperforms STPCA for rook contiguity spatial weights Figure 3 and most other weights when global autocorrelation is of concern, although MFPCA does not take spatial autocorrelation into account. It could outperform STPCA due to the extension of the concept of multivariate functional data of different (dimensional) domains for different elements, which STPCA does not consider. Our proposed mfasPCA performs well due to its ability to overcome the shortcomings of STPCA and MFPCA. Figure 5 shows 50 simulated bivariate and multivariate functional Moran's I indices for kNN-based neighborhoods, which demonstrate spatial autocorrelation. The functional Moran's I indices for both the bivariate and multivariate cases, based on the model described in Section 3.2 where $\rho = 0.7$, indicate significant positive spatial autocorrelation. Similar behavior is observed for all the other weight matrices. The bivariate functional Moran I statistic was computed using the second and third variables of the data.

4   Application to real data

By applying our proposed method to both scenarios: age-based for univariate data and time-based for multivariate data, we demonstrate its versatility and robustness in detecting spatial dependency in different contexts.

4.1   Univariate functional areal spatial principal component analysis

The mortality rates for males in 28 European countries available on the HMD (Human Mortality Database, 2024) for ages 0 to 110 (where ages above 100 are grouped as 100+) are considered to investigate spatial dependency at a fixed time, the year 2010. The analyses for the years 1990 and 2000, along with the analyses conducted for females for the years 1990, 2000, and 2010, are available in the supplemental material. We have also performed similar analyses using contiguity weights, which are detailed in the supplemental material, while the discussion here focuses on KNN weights. To adjust mortality rates and avoid the undefined calculation of logarithms for zero values, a constant was added to each death rate value. This constant was determined as the lowest death rate among the 28 countries studied, ensuring that all values remain positive before applying the natural logarithm. In the context of equation (1), x is an age between 0 and $T = 100$ and $Y_{i,x}$ is a mortality rate observed in a given year, for a country of location i. Lee and Carter (1992) modeled the logs of the age-specific death rates as a linear function of an unobserved period-specific intensity index, with parameters depending on age. This work followed suit. Figure 6(a) represents the smoothed log death rates data using B-splines for the male population in year 2010. The presence of spatial autocorrelation among the data is then investigated.

For the first part of the study, fPCA is performed on the smoothed data (Figure 6). To address the absence of spatial information in fPCA, fasPCA is enhanced by implementing multivariate (non-functional) spatial PCA (sPCA) as proposed by Jombart et al. (2008). The log death rates data for males in the 28 countries do not satisfy the normality assumption. The Shapiro-Wilk test, conducted using the `mvnormtest` R package (Jarek, 2012), was employed to assess multivariate normality of the log death rates data. The results indicated that this data violated the normality assumption. Hence, permutation tests for the Moran's I statistics are calculated for these data using 999 random permutations of the log death rates for all cases studied based on the KNN weights. Table 2 shows the presence of significant spatial autocorrelation for male log death rates for the years 1990, 2000, and 2010 across the 28 countries. In this table, a classical Moran's I index is calculated for each year using the raw data matrix treated as a panel data set. These statistics are somehow aggregations of the functional index (equation 4) as the functional trace index defined in (equation 7). The Moran's I statistics reported values close to +1, suggesting that neighboring locations

exhibit strong positive autocorrelations in male mortality rates for the year 2010. Figure 9a shows the smoothed functional Moran's I statistics for ages 0 to 100+, using B-splines. The spatial dependency in the logarithmic death rates increases between the ages of approximately 20 and 80. However, for individuals aged around 80 and above, spatial autocorrelation declines due to a decrease in the number of mortality cases.

Table 3 shows that for fPCA, the first principal component reveals autocorrelation. The results also indicate that the first fPC alone accounts for more than 99.67% of the total variability of the data in all cases where the data can be visualized in one dimension. However, this does not provide a clear picture of the data due to the absence of the spatial factor in the method. We perform fasPCA on the basis functions of the male log death rates data using KNN weights.The fasPCA described was executed for the cases studied by considering the top three positive and top two negative eigen-values (Table 3). The percentage of variability explained by the functional principal components of fasPCA are given in Table 3. The functional Moran's I statistics calculated based on these functional principal components show significant spatial autocorrelation for the principal components reported in Table 3. Spatial autocorrelation can effectively be detected from the spatial principal components of the functional data which explains 97.48% of the percentage of variability.

The data for each case was reconstructed by using the first two positive and first negative PCs (6(c,d)). These PCs were mapped onto geographic spaces (representing 28 European countries generated using the `maps` (Becker and Wilks, 2022) package and the `rgdal` (Bivand et al., 2023b) package where the black and white squares of the variable size represent positive and negative scores of the PCs respectively. The large black squares are well differentiated from the large white squares, while the small squares are less differentiated. The area of the square is proportional to the absolute value of the score. These graphical representations are applied to the significant PCs.

The first positive PC (Figure 7(a)) shows spatial connectivity between the states, which are split into two clusters: one in the west and one in the east. The projection for the second positive PC (Figure 7(b)) also shows spatial connectivity, with two clusters formed in the northern and southern regions. The first negative PC (Figure 7(c)) does not seem to display a particular spatial pattern. This outcome is anticipated because the negative principal components are associated with local structures that highlight dissimilarities on the geographical map at neighbouring locations. In general, spatial patterns are noticeable for the first and second PCs. Our findings demonstrate that the newly proposed fasPCA, in conjunction with the functional Moran's I statistic, effectively identifies spatial autocorrelation in spatial functional data.

4.2 Multivariate functional areal spatial principal component analysis

We will utilize a comprehensive multivariate dataset sourced from the study conducted by Krzyśko et al. (2023) to assess the reliability of mfasPCA. A detailed description of the multivariate dataset can be found in their paper. Furthermore, we will conduct a direct comparison with the STPCA method introduced by them, using the same multivariate dataset for consistency. This comparative analysis will provide valuable insights into the performance and reliability of our mfasPCA method. The dataset spans from 2002 to 2018 and provides descriptions for 16 regions in Poland based on 12 variables that are proxies for for the socioeconomic development of regions in Poland. The list of variables used in Krzyśko et al. (2023) is given in Table 1. To assess the reliability of the results, computations were performed using both the kNN and contiguity weight matrices. The calculations for all the weight matrices by Kryzsko et al. (2023), are detailed in the supplemental text. Our method was compared with MFPCA and STPCA (Krzyśko et al., 2023). In contrast to the simulation study, where mfasPCA outperformed its counterparts for both positive and negative spatio-temporal principal components, in this case study, mfasPCA tends to report a higher percentage of variability explained for positive components, while STPCA fares better for negative components (see Table 4 and Figure 8). This difference could potentially stem from the dataset containing fewer regions and being analyzed over a shorter time-frame compared to the simulated dataset in Figure 4. The bivariate and multivariate functional Moran I statistics were computed for this data (See Figure 9). In this analysis, we specifically chose variables 2 and 3 (see Table 1) to represent the bivariate functional Moran's I statistic. This selection demonstrates how the spatial dependency between these two variables changes over time. The findings show that the spatial dependency between variables 2 and 3 increases and stabilizes as time progresses. On the other hand, the multivariate functional Moran's I statistic values, ranging from approximately 0.73 to 0.79, display a cyclical pattern with significant declines in 2009 and 2017. Nonetheless, the range of values for this statistic remains relatively narrow.

5 Conclusion

In this study, we presented multivariate spatial functional areal data analysis using Moran's I and functional PCA. Simulation studies and empirical applications have demonstrated the efficacy of integrating multivariate functional Moran's I with PCA in identifying spatial autocorrelation in both univariate and multivariate functional areal datasets. In the real data application for the multivariate scenario, the analysis was performed with a limited number of locations. To gain a more comprehensive understanding, a simulation study with a larger number of locations was conducted. The results of the simulations corroborated the

findings of the multivariate functional principal component analysis (mfasPCA), offering clearer insights into the results. For future work, we plan to enhance the developed methodology for forecasting purposes, such as predicting mortality rates, by developing a spatio-temporal predictive model for neighboring countries with limited or sparse data. This approach can be applied within both univariate and multivariate frameworks.

Tables

Table 1 List of variables characterising different spheres of economy and natural environment of the regions in Poland

Number	Variable
1	Population per km^2
2	Students per 10,000 inhabitants
3	Libraries per 1,000 inhabitants
4	Production sales of industry, total per capita
5	Retail sales of goods per capita
6	Forestry and logging
7	Regional income
8	Regional expenditure
9	Targeted grants received from the state budget for indigenous tasks per km^2
10	Fees and impact on the fund for environmental protection and water management
11	Communal waste-water treatment plants
12	Devasted and degraded land, remediated and developed

Table 2 Moran's test for spatial autocorrelation, based on the log of death rates for males in 28 European countries using KNN weights

	Moran's I (Year 1990)	Moran's I (Year 2000)	Moran's I (Year 2010)
KNN	0.9770***	0.9814***	0.9831***

Note: "***" $p < 0.001$.

Table 3 Moran's test on principal components using fPCA and fasPCA, based on KNN weights for males of 28 European countries in 2010.
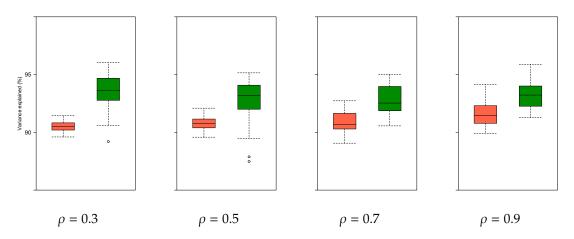
	Moran's I	Variability (%)
Classical FPCA		
1st score	0.5482***	99.67
2nd score	−0.0873	0.14
3rd score	0.2687**	0.08
4th score	−0.0548	0.04
Total	99.96	99.93
fasPCA(KNN (3,2))		
1st score positive	0.6021***	85.14
2nd score positive	0.3239**	2.54
3rd score positive	0.1821*	0.86
2nd score negative	−0.1575†	2.29
1st score negative	−0.1516†	6.65
Total	96.77	97.48

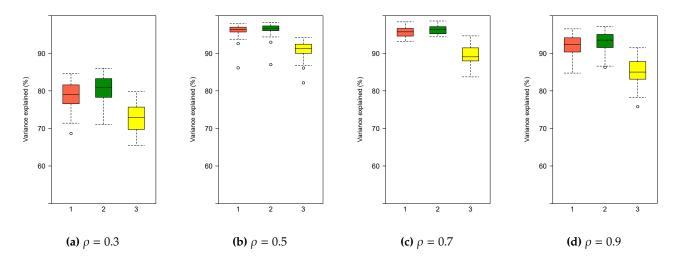Note: † $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 4 Variability (%) of multivariate Polish data explained by the the first two positive and negative principal components

| | Variability (%) | | |
	mfasPCA	STPCA	**M**FPCA
First two positive principal components			
KNN	99.16	95.25	65.70
Rook's contiguity	98.89	86.68	64.88
First two negative principal components			
KNN	57.84	76.10	-
Rook's contiguity	57.38	81.88	-

Figures

$\rho = 0.3$ $\qquad$ $\rho = 0.5$ $\qquad$ $\rho = 0.7$ $\qquad$ $\rho = 0.9$

Figure 1. Variance explained by the top four principal components in 50 simulated data sets with $n = 500$ from Model 1, using fPCA and fasPCA (positive and negative scores).
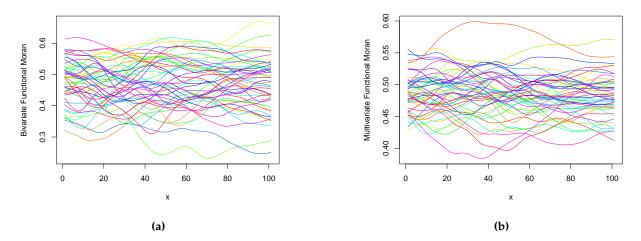
Figure 2. Variance explained by the top three principal components (positive) in 50 simulated data sets for rook contiguity weights: 1. red=MFPCA 2. green=mfasPCA 3. yellow=STPCA.
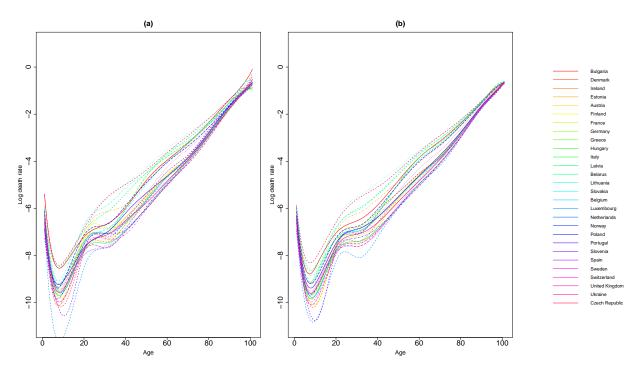

Figure 3. Variance explained by the top three principal components (negative) in 50 simulated data sets for rook contiguity weights: 1. green=mfasPCA 2. yellow=STPCA.
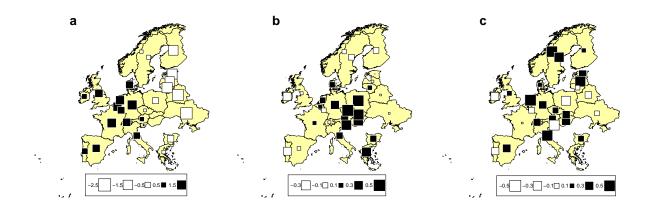
(a)

(b)

Figure 4. Second-level administrative divisions in France based on data generated using the model in Section 3.2 for: (a) positive spatial autocorrelation and (b) negative spatial autocorrelation (rook's contiguity weight matrix) over 101 evenly spaced time points with $\rho = 0.7$ for 10 variables.
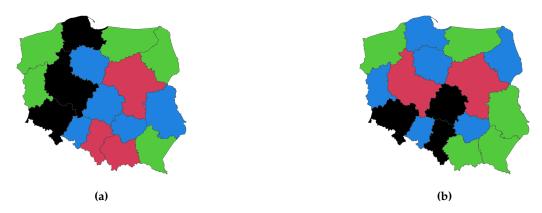

(a)

(b)

Figure 5. Simulated (a) bivariate and (b) multivariate functional Moran's I indices for KNN weights based on $\rho = 0.7$.
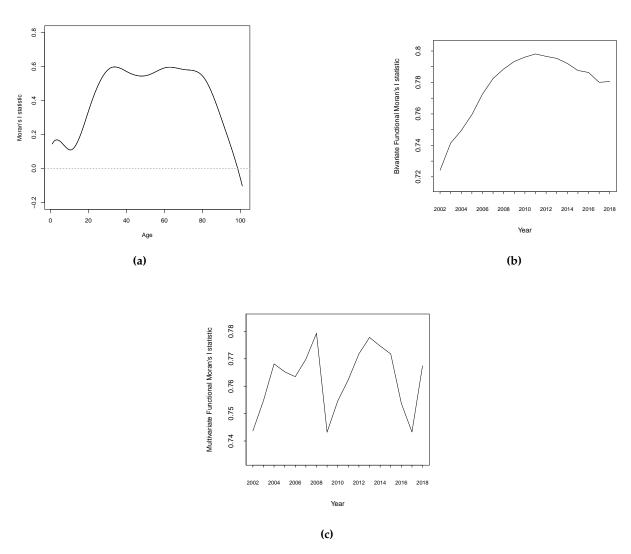
Figure 6. The male log death rates in 2010 for 28 European countries: (a) smoothed, (b) reconstructed using the matrix multiplication of the scores and principal components based on KNN weights.
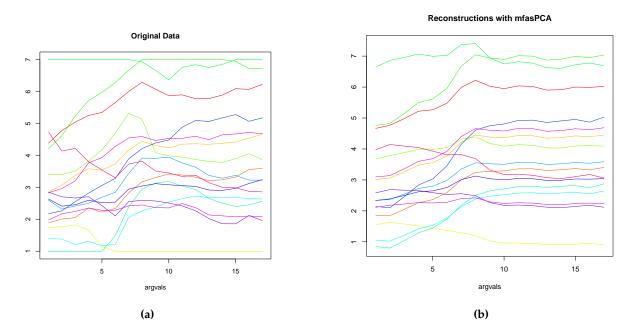

Figure 7. The log death rates scores of the (a) first positive, (b) second positive, and (c) first negative eigenvalues of the fasPCA based on KNN weights for males from 28 European countries in 2010.

Figure 8. Spatial typology of Polish regions as an effect of full (a) positive and (b) negative spatial autocorrelation (rook's contiguity weight matrix)

(a)

(b)

(c)

Figure 9. Functional Moran's I statistic curves: (a) Univariate case: log death rates for males aged 0 to 100+ in 2010 using the KNN weight matrix; (b) Bivariate case (variables 2 and 3 from Table 1) using the rook's contiguity weight matrix; (c) Multivariate case using the rook's contiguity weight matrix. Note that (b) and (c) use the data in Krzyśko et al. (2023).

Original Data

Reconstructions with mfasPCA

(a)

(b)

Figure 10. Polish data (variable 2 in Table 1): (a) Original data (b) Reconstruction with mfasPCA (rook's contiguity).

References

Ash, R. and Gardner, M. F. (1975). Topics in stochastic processes, acad. *Press, New York*.

Bali, J. L. and Boente, G. (2014). Robust functional principal component analysis. *New Advances in Statistical Modeling and Applications*, pages 41–54.

Becker, R. A. and Wilks, A. R. (2022). *maps: Draw Geographical Maps*. R package version 3.4.1.

Bivand, R., Keitt, T., and Rowlingson, B. (2023a). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. R package version 1.6-7.

Bivand, R., Keitt, T., and Rowlingson, B. (2023b). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. R package version 1.6-7.

Bougeard, S. and Dray, S. (2018). Supervised multiblock analysis in r with the ade4 package. *Journal of statistical software*, 86:1–17.

Chessel, D., Dufour, A. B., Thioulouse, J., et al. (2004). The ade4 package-i-one-table methods. *R news*, 4(1):5–10.

Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics: The official journal of the International Environmetrics Society*, 21(3-4):224–239.

Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G., et al. (2019). adespatial: Multivariate multiscale spatial analysis. r package version 0.3–7. 2019.

Dray, S. and Dufour, A.-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software*, 22:1–20.

Dray, S., Dufour, A. B., and Chessel, D. (2007). The ade4 package-ii: Two-table and k-table methods. *R news*, 7(2):47–52.

Eckardt, M. and Mateu, J. (2021). Partial and semi-partial statistics of spatial associations for multivariate areal data. *Geographical Analysis*, 53(4):818–835.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659.

Hassan, A. A. (2021). *Spatial data analysis: applications to population health*. PhD thesis, Université de Lille.

Hijmans, R. J. (2023). *raster: Geographic Data Analysis and Modeling*. R package version 3.6-23.

Hörmann, S., Kidziński, Ł., and Hallin, M. (2015). Dynamic functional principal components. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(2):319–348.

Human Mortality Database (2024).

Jarek, S. (2012). *mvnormtest: Normality test for multivariate variables*. R package version 0.1-9.

Jombart, T. (2008). adegenet: a r package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11):1403–1405.

Jombart, T., Devillard, S., Dufour, A.-B., and Pontier, D. (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, 101(1):92–103.

Khoo, T. H., Pathmanathan, D., and Dabo-Niang, S. (2023). Spatial autocorrelation of global stock exchanges using functional areal spatial principal component analysis. *Mathematics*, 11(3).

Koner, S. and Staicu, A.-M. (2023). Second-generation functional data. *Annual Review of Statistics and Its Application*, 10:547–572.

Krzyśko, M., Nijkamp, P., Ratajczak, W., Wołyński, W., and Wenerska, B. (2023). Spatio-temporal principal component analysis. *Spatial Economic Analysis*, pages 1–22.

Kuenzer, T., Hörmann, S., and Kokoszka, P. (2021). Principal component analysis of spatially indexed functions. *Journal of the American Statistical Association*, 116(535):1444–1456.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671.

Li, Y. and Guan, Y. (2014). Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance. *Journal of the American Statistical Association*, 109(507):1205–1215.

Liu, C., Ray, S., and Hooker, G. (2017). Functional principal component analysis of spatially correlated data. *Statistics and Computing*, 27:1639–1654.

Mateu, J. and Romano, E. (2017). Advances in spatial functional statistics.

Ramsay, J., Graves, S., and Hooker, G. (2020). fda: Functional data analysis. r package version 5.5. 1. *URL: https://CRAN. R-project. org/package= fda*.

Ramsay, J., Hooker, G., Graves, S., Ramsay, J., Hooker, G., and Graves, S. (2009). Introduction to functional data analysis. *Functional data analysis with R and MATLAB*, pages 1–19.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer.

Romano, E., Irpino, A., and Mateu, J. (2022). Spatial functional data analysis for probability density functions: Compositional functional data vs. distributional data approach. *Geostatistical Functional Data Analysis*, pages 128–153.