Big Data Research 39 (2025) 100504

Contents lists available at ScienceDirect

Big Data Research

journal homepage: www.elsevier.com/locate/bdr

# Principal component analysis of multivariate spatial functional data

Idris Si-ahmed [a,*], Leila Hamdad [b], Christelle Judith Agonkoui [c,d], Yoba Kande [e], Sophie Dabo-Niang [d,f,g]

[a] Ecole Nationale Supérieure d'Informatique (ESI) laboratoire LCSI, and PROXYLAN, a Cerist subsidiary, Oued Smar 16309 and PROXYLAN, and Q253+WQV, Rue Frères Aissou, Ben Aknoun 16028, Alger, 16000, Algeria
[b] Ecole Nationale Supérieure d'Informatique (ESI) laboratoire LCSI, Oued Smar 16309, Alger, 16000, Algeria
[c] Institut de Mathématiques et de Sciences Physiques (IMSP), Université d'Abomey-Calavi, 01 B.P.: 613, Porto-Novo, Benin
[d] MODAL team, Centre Inria de l'Université de Lille, Villeneuve d'Ascq-59650, France
[e] Université de Lille, CNRS, UMR 8524 Laboratoire Paul Painleve, 42 rue Paul Duez 59000, Lille, France
[f] CNRS - Université de Montréal, CRM - CNRS, Montréal, Québec, Montréal, Canada
[g] CNRS UMR 8524-Laboratoire Paul Painlevé, Université de Lille, Villeneuve d'Ascq-59650, 59650, Villeneuve d'Ascq, France

ARTICLE INFO

*Keywords:*
Spectral analysis
Functional data analysis
Functional principal component analysis
Spatial-functional principal component analysis
Multivariate analysis

ABSTRACT

This paper is devoted to the study of dimension reduction techniques for multivariate spatially indexed functional data and defined on different domains. We present a method called Spatial Multivariate Functional Principal Component Analysis (SMFPCA), which performs principal component analysis for multivariate spatial functional data. In contrast to Multivariate Karhunen-Loève approach for independent data, SMFPCA is notably adept at effectively capturing spatial dependencies among multiple functions. SMFPCA applies spectral functional component analysis to multivariate functional spatial data, focusing on data points arranged on a regular grid. The methodological framework and algorithm of SMFPCA have been developed to tackle the challenges arising from the lack of appropriate methods for managing this type of data. The performance of the proposed method has been verified through finite sample properties using simulated datasets and sea-surface temperature dataset. Additionally, we conducted comparative studies of SMFPCA against some existing methods providing valuable insights into the properties of multivariate spatial functional data within a finite sample.

## 1. Introduction

Analysis of complex data structures, such as multivariate and spatially indexed functional data, has become more prevalent in recent years due to its significance and application across various fields. Functional data constitute a data type in which each observation is represented by a function, a shape or a complex structure rather than a conventional vector of values. These data are of eventually of infinite dimension, as it is defined across the entire continuum of points. Within the framework of Functional Data Analysis (FDA), Functional Principal Component Analysis (FPCA) holds significant importance for exploratory analysis and dimensionality reduction. Functional Principal Component Analysis (FPCA) is frequently utilized to transform a functional-dimension space, which often presents challenges in computation, into a space with reduced dimensionality. Furthermore, FPCA allows to derive more informative attributes from the data, identify hidden patterns in a dataset, and discover correlations among variables [1–5].

In this work, we explore the field of geostatistical FPCA, focusing on multivariate functional spatially indexed data, where multiple interdependent functional variables are observed across spatial locations organized on a regular grid [6]. The data in question may encompass both spatial variations and temporal dynamics, assuming weak stationarity as considered in the study. This condition is less stringent than strict stationarity, yet it is extensively utilized in time series and spatial statistics, encompassing dependence and forecasting models. Despite the difficulties associated with its verification, weak stationarity is prevalently applied within the literature, owing to its flexibility and applicability across diverse contexts. For instance, [7] demonstrate its use in traffic data analysis, while [8] apply it to spatial functional prediction of sea surface temperature. In practice, the use of weak stationarity can naturally result from directed processes, such as the common formulation of time series that we use here for weakly stationary spatio-temporal data. Considering a regular grid is indispensable for numerous practical applications of spatial data. In domains such as meteorology, climatol-

* Corresponding author.
*E-mail addresses:* i_siahmed@esi.dz (I. Si-ahmed), l_hamdad@esi.dz (L. Hamdad), christelle.agonkoui@imsp-uac.org (C.J. Agonkoui), kandeyoba@gmail.com (Y. Kande), sophie.dabo@univ-lille.fr (S. Dabo-Niang).

https://doi.org/10.1016/j.bdr.2024.100504
Received 30 December 2023; Received in revised form 1 November 2024; Accepted 3 December 2024

Available online 16 December 2024

ogy, and remote sensing, data are commonly structured in regular grids to facilitate their analysis and visualization. For example, sea surface temperature data are systematically gathered and disseminated on regular grids, allowing uniform analysis and facilitating the comparison of different regions and time periods [9]. Employing a regular grid additionally streamlines data processing by enabling the direct application of interpolation and smoothing methods, thereby mitigating the complexity and potential errors associated with irregular data.

Interest in spatial functional data analysis has been growing in the literature. The work by [10] examines progress in spatial functional statistics, with a particular emphasis on analyzing spatially distributed functional data. Similarly, the recent review by [11] address the innovations in second-generation functional data analysis in dealing with complex spatial functional data. In alignment with this increasing interest, efforts are underway to modify the traditional FPCA for univariate functional data characterized by spatial, time or space-time dependencies.

The work of [6] delves into the transition from multivariate to univariate functional data analysis, spotlighting essential ideas and recent methodological advancements. They emphasize novel research areas and underline the complexities of integrating diverse data structures across domains, including spatially indexed functional data. In [12], the authors focus on functional time series, examining temporal dependencies and proposing an m-dependence construct for both linear and nonlinear series. They assess how this dependence influences statistical techniques like FPCA, covariance estimation, particularly in terms of resilience to weak dependence. In [13], a functional mixed effects model stands out for hierarchical functional data, delineating variations at different levels through distinct principal components, employing penalized splines and an EM algorithm for regularization, optimizing large matrix management via the model's covariance insights. [14] introduce a hierarchical framework to assess treatment impacts on functional data, organized by subjects, weeks, and days. They utilize penalized splines and principal components within a mixed effects context, developing an iterative ECME algorithm for principal component selection and addressing data voids. The work of [15] explores the frequent hospitalizations among end-stage renal disease patients on dialysis, considering temporal and spatial differences across U.S. dialysis centers. They propose a multilevel spatiotemporal functional model employing FPCA and Markov Chain Monte Carlo methods to estimate hospitalization rates based on regional variances. Meanwhile, [16] offered an FPCA approach tailored for functional data indexed spatially on a regular grid, establishing foundations for spectral analysis and linear spatial filters' spectral theory. [17] examined spatial autocorrelation in stock exchange returns during the 2015–2016 global market downturn through functional areal spatial principal component analysis (FASPCA).

When analyzing multiples functions within one or more dimensions, it is crucial to examine the relationships and interactions among the functional datasets. The literature offers multiple methods for conducting FPCA on multivariate functions, which are recorded over the same one-dimensional interval [3,18,19] or across distinct domains [20], in an i.i.d context. These studies are based on a multivariate functional Karhunen-Loève framework. Recent developments have started to extend these multivariate FPCA methods to spatial contexts, although such advancements are still nascent. For instance, [21] proposed a spatio-temporal principal component analysis (STPCA) utilizing a partially functional areal spatial data framework that incorporates spatial autocorrelation through the application of a weight matrix. This methodology comprises a two-step interdependent process: initially, the original multivariate time series are transformed into raw coefficients via a basis function expansion; subsequently, principal spatio-temporal components are derived using the classical Moran's I statistic applicable to areal spatial data for the matrix of basis coefficients. Furthermore, [22] developed a spatial principal component analysis technique utilizing a multivariate functional Moran's I statistic to analyze multivariate functional spatial areal data. Nonetheless, these techniques focus on aggregated datasets, contrary to the geostatistical data that is the main subject of our study (e.g., [23–26], among others within the geostatistical functional context).

This present work aims to present a spectral multivariate FPCA, termed SMFPCA, designed for dimension reduction of multivariate functional spatial data arranged on a regular grid, applicable to both similar and different domains. The goal is to build upon past research such as MFPCA [20], which tackles the i.i.d premise, and SFPCA [16], which lacks multivariate analysis. As far as we know, there are no prior studies on multivariate spectral FPCA for multiple geospatial functional variables across both similar and different domains. The unique aspect of the SMFPCA method, compared to existing work, is its dual focus: it tackles the multivariate functional dimension with several functions on various domains and also incorporates the geospatial dimension. It introduces spatial dependence along with a dimension reduction technique using the spectral density operator, contrasting with MFPCA in [20], which utilizes the covariance operator. Similarly, while [16] use the spectral density operator in their single-dimension SFPCA method, our approach's novelty lies in managing the multivariate functional variables dimension. A significant challenge is effectively addressing spatial dependence within a multivariate functional setting. This is achieved by extending the framework of [16] to incorporate multivariate dimensions with spectral density and by modifying the multivariate strategy of [20] to adopt the spectral density approach.

SMFPCA leverages the spectral density operator to characterize the spatial dependence structure of the data by decomposing it into distinct frequencies. This methodology is advantageous as it facilitates the identification of predominant frequencies that influence the variation within the spatial dependence structure of the data. These predominant frequencies subsequently define the principal components (eigenfunctions, eigenvalues), thereby offering insights into the underlying stochastic processes contributing to data variability. Consequently, the associated spectral decomposition serves as a tool for gaining a deeper understanding of the intricate spatial dependence structure. A study on the finite sample properties was conducted using both simulated and real environmental data, alongside a comparative analysis against existing methods such as MFPCA [20] and STPCA [21]. Observe that the framework of the previous study aligns more closely with aggregated spatial data, employing a weight matrix to assess spatial autocorrelation.

We focus on FPCA method, which has been extensively examined within FDA. However, it is crucial to recognize that the functional literature encompasses a diverse array of dimension reduction techniques in addition to Principal Component Analysis (PCA). [27] introduces functional canonical analysis for stochastic processes. In a different approach, [28] presents a principal varying coefficient model, modeling regression coefficients as covariate functions with reduction via principal directions. Additionally, [29] develops a Bayesian multivariate functional dynamic linear model, projecting data into a lower space. [30] proposes functional lasso, a penalized regression using $L_1$ penalty for variable selection and dimension reduction. Furthermore, [31] uses wavelets to decompose data into coefficients, simplifying complexity and extracting key components. [32] combines variable selection and smoothing in high-dimensional function-on-scalar regression for dimensionality reduction.

The structure of the paper is as follows: Section 2 outlines the SMFPCA methodology, Section 3 discusses the finite sample properties and examines the results, and Section 4 provides the conclusions and future outlook.

## 2. Spectral principal component analysis of multivariate spatial functional data

This section outlines the proposed approach for performing a multivariate spatial functional principal component analysis. This methodology leverages Karhunen-Loève theory, integrating spectral analysis and Functional Principal Component Analysis (FPCA). Moreover, the core

characteristics and principles are drawn from the studies of [16,3,4,20]. The approach is tailored to address a multivariate spatial-functional phenomenon that appears on a regular grid and is defined across either identical or distinct domains. We examine the mathematical structure utilized and give definitions for various components, including the spectral density operator, covariance operator, eigenfunctions (which represent functions that characterize significant variations or dominant modes within the functional data), eigenvalues (associated with the weights assigned to each eigenfunction and serving to quantify the total variance explained), and scores (denoting the projections of the original data onto the basis functions). Upon the establishment of these elements, the subsequent steps of SMFPCA are outlined in Section 2.3.

### 2.1. Multivariate spatial functional data

We consider that at $n$ spatial units located on a region $\mathbf{D} \subset \mathbb{Z}^N$, $N > 1$, representing a rectangular grid, we observe a multivariate spatial functional process $\{\mathbf{X_s}(.) = (X_{\mathbf{s}}^{(1)}(.), ..., X_{\mathbf{s}}^{(p)}(.))^\top\}$, $p \geq 1$, where $\mathbf{s} \in \mathbf{D}$, $X_{\mathbf{s}}^{(j)} = \{X_{\mathbf{s}}^{(j)}(t_j), t_j \in \mathcal{T}_j\}$. For $1 \leq j \leq p$, let $\mathcal{T}_j$ be a compact set in $\mathbb{R}$, with finite (Lebesgue-) measure and such that $X_{\mathbf{s}}^{(j)} : \mathcal{T}_j \longrightarrow \mathbb{C}$ is assumed to belong to $\mathcal{L}^2(\mathcal{T}_j, \mathbb{C})$, the space of complex square-integrable functions on $\mathcal{T}_j$. In the following let $\mathcal{L}^2(\mathcal{T}_j, \mathbb{C}) = \mathcal{L}^2(\mathcal{T}_j)$. Note that the special case $p = 1$ corresponds to the univariate spatial-functional case [16].

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

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

Then, $\mathcal{H}$ is a Hilbert space with respect to the scalar product $\langle\langle \cdot, \cdot \rangle\rangle$ [20].

Let's define the functions observed on $n$ sites $\mathbf{s}_1, ..., \mathbf{s}_n \in \mathbf{D}$ by:

$$\mathbf{X}^{(j)} = \left( X_{\mathbf{s}_1}^{(j)}, ..., X_{\mathbf{s}_n}^{(j)} \right), \quad j = 1, \ldots, p.$$

### 2.2. Univariate spatial functional pca

We independently consider each of the spatial functional univariate samples $\mathbf{X}^{(j)}$, to compute a univariate SFPCA. To achieve this, we apply the univariate spatial FPCA [16]. Let $X^{(j)} \in \mathcal{L}^2(\mathcal{T}_j)$ possess a covariance operator $C_j := \mathbb{E}[(X^{(j)} - \mu^j) \otimes (X^{(j)} - \mu^j)]$ (where $\mu^j$ is the mean curve defined by $\mu^j(t) = \mathbb{E}X^{(j)}(t)$ with $t \in \mathcal{T}_j$) with kernel $c_j(t, s) = cov(X^{(j)}(t), X^{(j)}(s))$ $(t, s \in \mathcal{T}_j)$. Then, the integral operator $C_j$ is defined by

$$(C_j f)(t) = \int_{\mathcal{T}_j} c_j(s, t) f(s) ds, \quad f \in \mathcal{L}^2(\mathcal{T}_j),\ t \in \mathcal{T}_j.$$

Let's suppose that each $\{X_{\mathbf{s}}^{(j)}\}$ is a weakly stationary functional process. We have:

(i) $\mathbb{E}(X_{\mathbf{s}}^{(j)}(t)) = \mathbb{E}(X_{\mathbf{0}}^{(j)}(t)) = \mu^j(t)$, $t \in \mathcal{T}_j$ with $\mathbf{0}$ the null vector in $\mathbb{Z}^N$

(ii) for all $\mathbf{s}, \mathbf{h} \in \mathbf{D}$, and $t, s \in \mathcal{T}_j$; $c_{j,\mathbf{h}}(t, s) := Cov\left( X_{\mathbf{h}}^j(t), X_{\mathbf{0}}^j(s) \right) = Cov\left( X_{\mathbf{s+h}}^j(t), X_{\mathbf{s}}^j(s) \right)$

The integral operator defined by the autocovariance kernel $c_{j,\mathbf{h}}$ is denoted $C_{j,\mathbf{h}}$ and defined by

$$(C_{j,\mathbf{h}} f)(t) = \int_{\mathcal{T}_j} c_{j,\mathbf{h}}(s, t) f(s) ds, \quad f \in \mathcal{L}^2(\mathcal{T}_j),\ t \in \mathcal{T}_j.$$

The following assumptions have been made about the process $X_{\mathbf{s}}^{(j)}$ i.e. it is weakly stationary and has mean zero. We suppose the autocovariance operators are absolutely summable:

$$\sum_{\mathbf{h} \in D} \|C_{j,\mathbf{h}}\| < \infty. \tag{1}$$

Let us denote the spectral density operator of $X_{\mathbf{s}}^{(j)}$ by $\mathcal{F}_\theta^{X^{(j)}}$ with the following kernel:

$$f_\theta^{X^{(j)}}(t, s) := \frac{1}{(2\pi)^N} \sum_{\mathbf{h} \in \mathbb{Z}^N} c_{j,\mathbf{h}}(t, s) \exp(-i\mathbf{h}^\top \theta) \tag{2}$$

$$t, s \in \mathcal{T}_j, \quad \theta \in [-\pi, \pi]^N, \quad i = \sqrt{-1},$$

where $\theta$ is the spatial frequency. We define $\mathcal{L}_U^2([-\pi, \pi]^N)$ as the space of measurable mappings $x : [-\pi, \pi]^N \to U$ satisfying $\int_{[-\pi,\pi]^N} \|x(\theta)\|^2 d\theta < \infty$, with $U$ the Hilbert space of all Hilbert–Schmidt operators from $\mathcal{L}^2(\mathcal{T}_j)$ to $\mathcal{L}^2(\mathcal{T}_j)$ (see [16] for further explanation). The operator $\mathcal{F}_\theta^{X^{(j)}}$ is understood as element of the space $\mathcal{L}_U^2([-\pi, \pi]^N)$ and is defined by

$$\left( \mathcal{F}_\theta^{X^{(j)}} G_\theta \right)(t) = \int_{\mathcal{T}_j} f_\theta^{X^{(j)}}(s, t) G_\theta(s) ds,$$

$$\text{with} \quad G_\theta \in \mathcal{L}_U^2([-\pi, \pi]^N) \quad \text{and} \quad t \in \mathcal{T}_j.$$

Considering the condition (1) and the assumptions previously defined for the process weakly stationary $X_{\mathbf{s}}^{(j)}$, $\mathcal{F}_\theta^{X^{(j)}}$ is a Hilbert-Schmidt operator (positive, self-adjoint) and admits a decomposition [16]:

$$\mathcal{F}_\theta^{X^{(j)}} = \sum_{m \geq 1} \lambda_{j,m}(\theta) \varphi_{j,m}(\theta) \otimes \varphi_{j,m}(\theta), \tag{3}$$

where $\lambda_{j,m}(\theta) \geq \lambda_{j,m}(\theta) \geq ... \geq 0$ are eigenvalues (continuous functions of $\theta$), and $\varphi_{j,m}(\theta)$ are associated eigenfunctions. Let $\varphi_{j,m}(t|\theta)$ be the value of the eigenfunction $\varphi_{j,m}(\theta)$ at $t \in \mathcal{T}_j$. The Fourier coefficients are

$$\phi_{m,\mathbf{l}}^{(j)}(t) := \frac{1}{(2\pi)^N} \int_{[-\pi,\pi]^N} \varphi_{j,m}(t|\theta) \exp(-i\mathbf{l}^\top \theta) d\theta, \tag{4}$$

$t \in \mathcal{T}_j$ and the corresponding expansion of $\varphi_{j,m}(t|\theta)$ is

$$\varphi_{j,m}(t|\theta) = \sum_{\mathbf{l} \in \mathbb{Z}^N} \phi_{m,\mathbf{l}}^{(j)}(t) \exp(-i\mathbf{l}^\top \theta). \tag{5}$$

With the property (1) and the assumptions previously defined for the process weakly stationary $\{X_{\mathbf{s}}^{(j)}\}$, we define the $m$th spatial functional principal component (SFPC) score by:

$$\xi_{m,\mathbf{s}}^{(j)} := \sum_{\mathbf{l} \in \mathbf{D}} \left\langle X_{\mathbf{s-l}}^{(j)}, \phi_{m,\mathbf{l}}^{(j)} \right\rangle, \quad \text{with } \mathbf{s} - \mathbf{l} \in \mathbf{D}, \tag{6}$$

where $\phi_{m,\mathbf{l}}^{(j)}$ is defined by (4). The corresponding SFPC filters are $(\phi_{m,\mathbf{l}}^{(j)})_{\mathbf{l} \in \mathbb{Z}^N}$.

We deduce that

- $\xi_{m,\mathbf{s}}^{(j)}$ converges in mean square with:

  $$\mathbb{E}[\xi_{m,\mathbf{s}}^{(j)}] = 0, \quad \mathbb{E}[(\xi_{m,\mathbf{s}}^{(j)})^2] = \sum_{\mathbf{l} \in \mathbb{Z}^N} \sum_{\mathbf{k} \in \mathbb{Z}^N} \left\langle C_{\mathbf{l-k}}^{X^{(j)}} \phi_{m,\mathbf{l}}^{(j)}, \phi_{m,\mathbf{k}}^{(j)} \right\rangle.$$

- If $X_{\mathbf{s}}^{(j)}$ is real then $\phi_{m,\mathbf{l}}^{(j)}$ and $\xi_{m,\mathbf{s}}^{(j)}$ are also real.
- if $C_{\mathbf{h}}^{X^{(j)}} = 0$ then $\forall\, \mathbf{h} = \mathbf{0}$, $\xi_{m,\mathbf{s}}^{(j)}$ coincides with the scores of the FPCA.
- $\forall\, m \neq m'$ and $\mathbf{s} \neq \mathbf{s}' \in \mathbf{D}$ the SFPCA scores $\xi_{m,\mathbf{s}}^{(j)}$ and $\xi_{m',\mathbf{s}'}^{(j)}$ are uncorrelated.

The spatial Karhunen–Loève expansion of $X_{\mathbf{s}}^{(j)}$ is given by

$$X_{\mathbf{s}}^{(j)}(t) = \sum_{m=1}^{\infty} X_{m,\mathbf{s}}^{(j)}(t) \quad t \in \mathcal{T}_j, \text{with} \tag{7}$$

$$X^{(j)}_{m,\mathbf{s}}(t) := \sum_{\mathbf{l}\in\mathbb{Z}^N} \xi^{(j)}_{m,\mathbf{s}+\mathbf{l}}\phi^{(j)}_{m,\mathbf{l}}(t).$$

In practice, the spectral density operator is unknown and has to be estimated using the sample $\mathbf{X}^{(j)}$ observed on the grid $\mathbf{D} = \{\mathbf{s} = (s_1, ..., s_N),\ 1 \le s_i \le n_i, i = 1, ..., N\}$, the sample size is then $n = \sum_{i=1}^{N} n_i$, and we use the notation $\mathbf{n} = (n_1, n_2, \dots, n_N)$.

The spectral density operator is estimated by:

$$\widehat{\mathcal{F}}^{X^{(j)}}_{\theta} := \frac{1}{(2\pi)^N} \sum_{|\mathbf{h}|\le\mathbf{q}} w(\mathbf{h}/\mathbf{q})\widehat{C}_{j,\mathbf{h}} e^{-i\mathbf{h}^\top\theta} \tag{8}$$

where $w$ represents a weight function and the vector $\mathbf{q} = (q_1, q_2, ..., q_N)$ consists of positive coordinates, the sample autocovariance operators are estimated as follows:

$$\widehat{C}_{j,\mathbf{h}} := \frac{1}{n} \sum_{\mathbf{s}\in M_{\mathbf{h},\mathbf{n}}} \left(X^{(j)}_{\mathbf{s}+\mathbf{h}} - \bar{X}^{(j)}\right) \otimes \left(X^{(j)}_{\mathbf{s}} - \bar{X}^{(j)}\right) \tag{9}$$

with $M_{\mathbf{h},\mathbf{n}} = \left\{\mathbf{s} \in \mathbb{Z}^N : 1 \le s_i, s_i + h_i \le n_i\ \forall 1 \le i \le N\right\}$. If the set $M_{\mathbf{h},\mathbf{n}}$ is empty, we set $\widehat{C}_{j,\mathbf{h}} = 0$.

### 2.3. SMFPCA methodology

In this subsection, we present the methodology for computing SMFPCA. It is divided into two parts. The first part relies on the univariate SFPCA of [16]. This involves considering a spectral analysis on $\mathbf{X}^{(j)}$, following the subsequent steps:

1. Compute the spectral covariance operator.
2. Decompose the spectral covariance operator to obtain the estimated eigenfunctions $\hat{\varphi}_{j,m}(\theta)$, involving SFPC filters $(\hat{\phi}^{(j)}_{m,\mathbf{s}})_{\mathbf{s}\in\mathbf{D}}$, and $\hat{\lambda}_{j,m}(\theta)$, the estimated eigenvalues associated with the spectral variability, where $m = 1, \dots, M_j$ for suitably chosen truncation lags $M_j$. The estimator of the filter function $\phi^{(j)}_{m,\mathbf{s}}$ is given by

$$\hat{\phi}^{(j)}_{m,\mathbf{s}}(t) := \frac{1}{(2\pi)^N} \int_{[-\pi,\pi]^N} \hat{\varphi}_{j,m}(t|\theta)\exp(-i\mathbf{s}^\top\theta)d\theta,$$

   where the functions $\hat{\varphi}_{j,m}(t|\theta)$ are the eigenfunctions of the spectral density operator estimator $\widehat{\mathcal{F}}^{X^{(j)}}_{\theta}$.
3. Finally, the $\mathbf{X}^{(j)}$ are projected onto the spectral eigenfunctions, yielding the estimated scores $\hat{\xi}^{(j)}_{m,\mathbf{s}}$, defined by:

$$\hat{\xi}^{(j)}_{m,\mathbf{s}} := \sum_{\|\mathbf{l}\|_\infty \le L} \left\langle X^{(j)}_{\mathbf{s}-\mathbf{l}}, \hat{\phi}^{(j)}_{m,\mathbf{l}} \right\rangle, \quad \text{with } \mathbf{s}-\mathbf{l} \in \mathbf{D},$$

   assuming that $1 + L \le s_i \le n_i - L$ for all $1 \le i \le N$, where $L$ is an integer-valued truncation parameter.

In the second part, the scores and SFPC filters of $\mathbf{X}^{(j)}$ are utilized to compute the multivariate eigen elements, following the steps:

4. Define the matrix $\mathbf{E} \in \mathbb{R}^{n\times M_+}$ of rows $(\hat{\xi}^{(1)}_{1,\mathbf{s}}, \dots, \hat{\xi}^{(1)}_{M_1,\mathbf{s}}, \dots, \hat{\xi}^{(p)}_{1,\mathbf{s}}, \dots, \hat{\xi}^{(p)}_{M_p,\mathbf{s}}), \mathbf{s} \in \mathbf{D}$ consists of all the scores estimated from the Spectral PCA on each $\mathbf{X}^{(j)}$, with $M_+ = M_1 + \dots + M_p$.
   Let's consider the matrix $\mathbf{Z} \in \mathbb{R}^{M_+\times M_+}$ (Prop.5., p.7 [33]) consisting of blocks $\mathbf{Z}^{(jk)} \in \mathbb{R}^{M_j\times M_k}$ with entries

$$Z^{(jk)}_{ml} = \mathrm{Cov}(\hat{\xi}^{(j)}_{m,\mathbf{s}}, \hat{\xi}^{(k)}_{l,\mathbf{s}}),$$

$$m = 1, \dots, M_j, \quad l = 1, \dots, M_k, \quad j, k = 1, \dots, p.$$

   An estimate $\hat{\mathbf{Z}} \in \mathbb{R}^{M_+\times M_+}$ of the matrix $\mathbf{Z}$ is given by $\hat{\mathbf{Z}} = (n-1)^{-1}\mathbf{E}^T\mathbf{E}$.
5. Perform a matrix eigen-analysis for $\hat{\mathbf{Z}}$ resulting in eigenvalues $\hat{v}_m$ and orthonormal eigenvectors $\hat{c}_m$.
6. The multivariate eigenfunctions applied for each operator of each variable are obtained as follows:

$$\hat{\psi}^{(j)}_{m,\mathbf{s}}(t_j) \approx \sum_{l=1}^{M_j} \left[\hat{c}_m\right]^{(j)}_l \hat{\phi}^{(j)}_{l,\mathbf{s}}(t_j), \tag{10}$$

   $t_j \in \mathcal{T}_j, \mathbf{s} \in \mathbf{D},\ m = 1, ..., M_+$

   where $\left[\hat{c}_m\right]^{(j)} \in \mathbb{R}^{M_j}$ denotes the $j$-th block of the (orthonormal) eigenvector $\hat{c}_m$ of $\hat{\mathbf{Z}}$.
   Furthermore, multivariate scores are calculated as:

$$\hat{\rho}_{m,\mathbf{s}} = \sum_{j=1}^{p}\sum_{l=1}^{M_j} \left[\hat{c}_m\right]^{(j)}_l \hat{\xi}^{(j)}_{l,\mathbf{s}}. \tag{11}$$

We can deduce a sample version of the spatial Karhunen–Loève expansion for each univariate component:

$$X^{(j)}_{\mathbf{s}}(t_j) \approx \sum_{m=1}^{M_j} \hat{X}^{(j)}_{m,\mathbf{s}}(t_j), \quad t_j \in \mathcal{T}_j,$$

$$\text{with} \quad \hat{X}^{(j)}_{m,\mathbf{s}}(t_j) := \sum_{\|\mathbf{l}\|_\infty \le L} \hat{\xi}^{(j)}_{m,\mathbf{s}+\mathbf{l}}\hat{\phi}^{(j)}_{m,\mathbf{l}},$$

assuming $1 + 2L \le s_i \le n_i - 2L$ for $1 \le i \le N$.

## 3. Numerical experiments

This section evaluates the SMFPCA algorithm's performance through analysis of both simulated and real datasets. Following this, we compare these results with those obtained from STPCA [21] and MFPCA [20], with the latter ignoring spatial context.

### 3.1. Simulation study

In this section, we extend the simulation context of (Section 4.1., P.1452 [16]) to a multivariate case.

We consider $N = 2$ and simulate SFARMA process $\{X^j_{s,t}(.)\}$ defined by:

$$\begin{aligned} X^j_{s,t} =& A_{10}X^j_{s-1,t} + A_{01}X^j_{s,t-1} + \varepsilon_{s,t} + B_{10}\varepsilon_{s-1,t} \\ &+ B_{01}\varepsilon_{s,t-1} + B_{11}\varepsilon_{s-1,t-1},\ t = 1, 2, \dots 50, \end{aligned} \tag{12}$$

where $(\mathrm{A}_{kl})_{k,l\in P}$ and $(\mathrm{B}_{kl})_{k,l\in Q}$ are Hilbert–Schmidt operators, P and Q two finite index sets valued in $\mathbb{Z}^N$. The functions $X^j_{s,t}$ are generated using expansions into $d$ Fourier basis functions, and the errors $\varepsilon_{s,t}$ are generated as i.i.d. variables from a multivariate $d$-dimensional Gaussian distribution with mean zero and a $d \times d$-dimensional diagonal covariance matrix $\Sigma$ with entries $\Sigma_{ii} = \exp(-i/2)$. To assess the effectiveness of integrating the spatial and multivariate aspects through SMFPCA, it is common to reconstruct the original functional data from the already computed scores and filters. For each instance, we employ both the novel SMFPCA and the conventional MFPCA. We assess the effectiveness of dimension reduction using the metric known as the normalized mean squared error (NMSE) for each univariate component, as defined by:

$$\mathrm{NMSE}(M_j) = \frac{\sum_{\mathbf{s}\in\mathbf{D_n}} \left\| X^{(j)}_{\mathbf{s}} - \sum_{m=1}^{M_j} \hat{X}^{(j)}_{m,\mathbf{s}} \right\|^2}{\sum_{\mathbf{s}\in\mathbf{D_n}} \left\| X^{(j)}_{\mathbf{s}} \right\|^2}, \tag{13}$$

$\mathbf{D_n}$ represents a region where the average is calculated ($\mathbf{D_n} = \{\mathbf{s} \in \mathbb{Z}^N : 1 \le s_i \le n_i\ 1 \le i \le n\}$)

Another alternative is to compute the error defined in equation (13) using the component eigenvalues $\hat{\lambda}_{j,m}$ of the spectral density operator $\widehat{\mathcal{F}}^{X}_{\theta}$:

**Table 1**
Results of NMSE and NMSE* for both SMFPCA and MFPCA approaches using the first four multivariate principal components (1 PC, 2 PC, 3 PC, 4 PC) under Setting 1.

| Number of Components (PC) | 1 PC | | 2 PC | | 3 PC | | 4 PC | |
|---|---|---|---|---|---|---|---|---|
| Methods | SMFPCA | MFPCA | SMFPCA | MFPCA | SMFPCA | MFPCA | SMFPCA | MFPCA |
| NMSE $SIM_1$ | 0.7403 | 0.8233 | 0.6343 | 0.7768 | 0.6328 | 0.6703 | 0.4561 | 0.6694 |
| NMSE* $SIM_1$ | 0.6850 | 0.8236 | 0.5205 | 0.6714 | 0.4035 | 0.5635 | 0.3125 | 0.4583 |
| NMSE $SIM_2$ | 0.7484 | 0.9952 | 0.6697 | 0.8584 | 0.5026 | 0.8539 | 0.5024 | 0.7338 |
| NMSE* $SIM_2$ | 0.7359 | 0.8554 | 0.5790 | 0.7347 | 0.4655 | 0.6306 | 0.3748 | 0.5399 |

**Table 2**
Results of NMSE and NMSE* for both SMFPCA and MFPCA approaches using the first four multivariate principal components (1 PC, 2 PC, 3 PC, 4 PC) under Setting 2.

| Number of Components (PC) | 1 PC | | 2 PC | | 3 PC | | 4 PC | |
|---|---|---|---|---|---|---|---|---|
| Methods | SMFPCA | MFPCA | SMFPCA | MFPCA | SMFPCA | MFPCA | SMFPCA | MFPCA |
| NMSE $SIM_3$ | 0.5095 | 0.5751 | 0.4096 | 0.5699 | 0.3213 | 0.3318 | 0.2528 | 0.3297 |
| NMSE* $SIM_3$ | 0.4825 | 0.5776 | 0.2979 | 0.3313 | 0.1849 | 0.2180 | 0.1188 | 0.1421 |
| NMSE $SIM_4$ | 0.5993 | 0.9989 | 0.3832 | 0.6267 | 0.2394 | 0.5867 | 0.2175 | 0.3924 |
| NMSE* $SIM_4$ | 0.4310 | 0.6274 | 0.2494 | 0.3930 | 0.1463 | 0.2480 | 0.0859 | 0.1517 |

$$\mathrm{NMSE}^{*}_{\mathrm{spat}}(M_j) = 1 - \frac{\sum_{m \le M_j} \int_{[-\pi,\pi]^N} \hat{\lambda}_{j,m}(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\sum_{m \ge 1} \int_{[-\pi,\pi]^N} \hat{\lambda}_{j,m}(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{14}$$

This measure assesses the quality of the approximation without being influenced by grid's boundary effects unlike the NMSE, where the approximation of $X_{\mathbf{s}}^{(j)}$ is less accurate at the boundary [16].

The experiments are conducted in 4 settings:

1. The first setting involves generating within the same domain $\mathcal{T}_j = [0,1]$ and $j = 1,2$, two simulated functional variables following the process defined in (12), denoted $SIM_1$ and $SIM_2$ with $d = 10$ and $d = 14$ Fourier basis functions respectively. The variables $SIM_1$ and $SIM_2$ are generated without errors ($\Sigma$ is the null matrix).
2. In Configuration 2, we simulate two variables, $SIM_3$ and $SIM_4$, using the same setup as in Configuration 1 but with added errors $\varepsilon_{s,t}$ as defined above.
3. The third setting involves generating two SFARMA variables, denoted $SIM_5$ and $SIM_6$ in distinct domains, with $\mathcal{T}_1 = [0,1]$ and $\mathcal{T}_2 = [2,4]$. We use a set of $d = 14$ and $d = 10$ Fourier basis functions for $SIM_5$ and $SIM_6$ respectively. These variables are generated without errors.
4. In Configuration 4, we simulate two variables, $SIM_7$ and $SIM_8$, using the same setup as in Configuration 3 but with added errors $\varepsilon_{s,t}$ as defined above.

After obtaining the spatial functional data, a centering step is performed, followed by the application of univariate SFPCA for each variable. In the configuration utilized for each setting, we need to define two tuning parameters $\mathbf{q}$ and $L$, as defined by ([16] and Supplemental material). Specifically, we assign:

1. $\mathbf{q}1$ = c(10,10), $\mathbf{q}2$ = c(12,12) and L = (12,12) for setting 1.
2. $\mathbf{q}1$ = c(3,3), $\mathbf{q}2$ = c(15,15) and L = (12,12) for setting 2.
3. $\mathbf{q}1$ = c(10,10), $\mathbf{q}2$ = c(14,14) and L = (12,12) for setting 3.
4. $\mathbf{q}1$ = c(3,3), $\mathbf{q}2$ = c(15,15) and L = (10,10) for setting 4.

Then we proceed with describing steps of SMFPCA defined in procedure 2.3, and compute 4 multivariate principal components to capture approximately 70% of variability for settings 1 and 3, and 80% for settings 2 and 4.

After performing SMFPCA, we conducted a comparative assessment of the reconstructed functional data. We applied equations (13) and (14), and the obtained results are shown in Tables 1, 2, 3, and 4. We can observe the performance of SMFPCA compared to the conventional MFPCA approach, which does not account for spatial considerations. Using the first configuration as an illustrative example, as shown in Table 2, the results emphasize that the enhancement in NMSE and NMSE* quality is linked to the number of functional principal components. This phenomenon is due to the fact that when the first four FPCA are considered, they collectively account for approximately 80% of the total variance. The measure NMSE*, which describes the quality of the approximation with no boundary effects, shows better result compare to NMSE. Regarding the results in Tables 1, 3, and 4, even when considering data defined in various domains and adding errors measurements, a similar trend can be observed in the results.

### 3.2. Application to real data

Following the application of the SMFPCA method to simulated data, we proceed to assess its performance on real data. The data we use in this paper concerns sea surface temperature (SST) data from the NOAA Optimum Interpolation Sea Surface Temperature dataset (Section 4.2., P.1453 [16]). The dataset exhibiting typical spatial dependence characterized by an exponential decay.

The SST dataset is obtained through the aggregation of satellite observations and in-situ measurements from ships and buoys. It provides daily observations on a global grid with a resolution of 0.25°, covering the entire sea area. The temperature data has been recorded since 1982, spanning a period of 33 years. The dataset exhibits annual quasi-periodicity and represents a spatially indexed functional random field. A subset extracted from the Indian Ocean area (60° to 93°$E$ longitude and 15° to 44°$S$ latitude) is chosen for its homogeneity and lack of significant oceanic currents. To address strong correlation among nearby observations, the grid resolution is reduced to 0.75°, reducing computational load and supporting condition (1) where slight differences at the 0.25° grid lead to slow spatial autocorrelation decay. Fig. 1 illustrates a snapshot of this extensive dataset.

For our analysis, we use two representative sea surface temperature variables, namely $TMP-2000$ and $TMP-2001$, which correspond to the years 2000 and 2001 respectively, and represent a multivariate aspect. To verify the efficiency of our methodology, we aim to conduct a comprehensive evaluation and analysis with three variables of $TMP-1996$, $TMP-1998$ and $TMP-1999$, which represent the sea surface temperature data for the respective years 1996, 1998, and 1999.

We apply the SMFPCA, STPCA and ordinary MFPCA on NOAA data, and illustrate the performance estimation procedure in two settings:

**Table 3**
Results of NMSE and NMSE* for both SMFPCA and MFPCA approaches using the first four multivariate principal components (1 PC, 2 PC, 3 PC, 4 PC) under Setting 3.

| Number of Components (PC) | 1 PC | | 2 PC | | 3 PC | | 4 PC | |
|---|---|---|---|---|---|---|---|---|
| Methods | SMFPCA | MFPCA | SMFPCA | MFPCA | SMFPCA | MFPCA | SMFPCA | MFPCA |
| NMSE $SIM_5$ | 0.6863 | 0.8122 | 0.6257 | 0.8106 | 0.4528 | 0.6716 | 0.4494 | 0.6323 |
| NMSE* $SIM_5$ | 0.6792 | 0.8122 | 0.5187 | 0.6720 | 0.4027 | 0.5472 | 0.3109 | 0.4442 |
| NMSE $SIM_6$ | 0.7923 | 0.9823 | 0.6565 | 0.8497 | 0.6306 | 0.8348 | 0.4961 | 0.7362 |
| NMSE* $SIM_6$ | 0.7320 | 0.8506 | 0.5652 | 0.7330 | 0.4454 | 0.6251 | 0.3532 | 0.5347 |

**Table 4**
Results of NMSE and NMSE* for both SMFPCA and MFPCA approaches using the first four multivariate principal components (1 PC, 2 PC, 3 PC, 4 PC) under Setting 4.

| Number of Components (PC) | 1 PC | | 2 PC | | 3 PC | | 4 PC | |
|---|---|---|---|---|---|---|---|---|
| Methods | SMFPCA | MFPCA | SMFPCA | MFPCA | SMFPCA | MFPCA | SMFPCA | MFPCA |
| NMSE $SIM_7$ | 0.7207 | 0.9815 | 0.4009 | 0.5906 | 0.2306 | 0.3750 | 0.2304 | 0.3722 |
| NMSE* $SIM_7$ | 0.4346 | 0.5919 | 0.2256 | 0.3749 | 0.1293 | 0.2185 | 0.0757 | 0.1435 |
| NMSE $SIM_8$ | 0.4565 | 0.5412 | 0.4043 | 0.5408 | 0.4012 | 0.5206 | 0.2489 | 0.3460 |
| NMSE* $SIM_8$ | 0.4375 | 0.5421 | 0.2389 | 0.3469 | 0.1369 | 0.2304 | 0.0802 | 0.1575 |

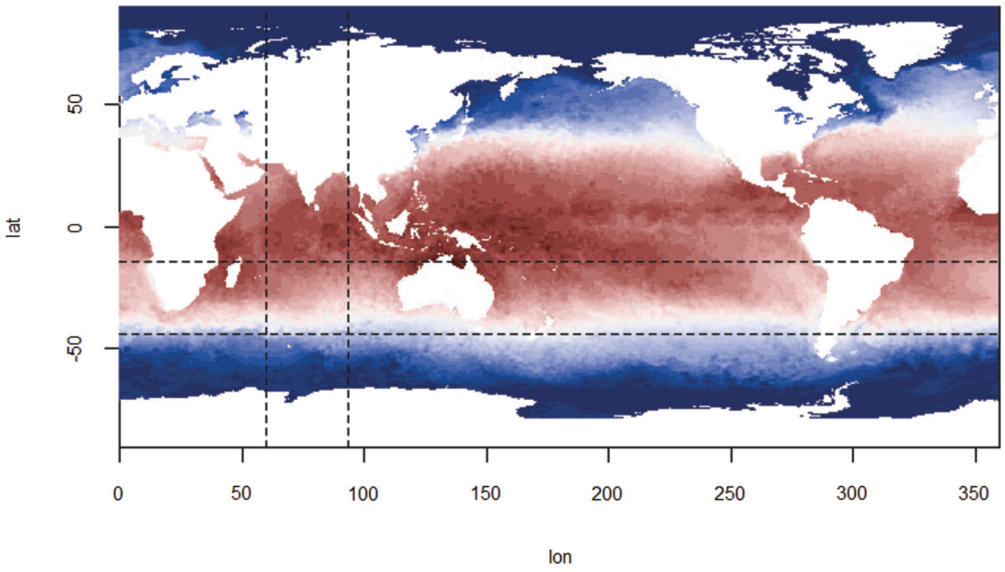


**Fig. 1.** Indian Ocean ranging approximately from 60° to 93° $E$ longitude and 15° to 44° $S$ latitude [16].

1. Application of SMFPCA, STPCA and MFPCA on two NOAA variables $TMP-2000$ and $TMP-2001$.
2. Application of SMFPCA and ordinary MFPCA on three NOAA variables $TMP-1996$, $TMP-1998$ and $TMP-1999$.

To initiate the first setup, the initial step of the SMFPCA algorithm entails calculating the functional data. We convert $TMP-2000$ and $TMP-2001$ into spatial functional data by employing 15 Fourier basis functions, resulting in functions labeled as $X_{s,u}^{(j)}(.)$, where indices $s$, $u$, and $j$ denote longitude, latitude, and year respectively. The time domain $\mathcal{T}_j$, representing intra-year time, is normalized to the interval $[0,1]$. We then execute SFPCA on these variables with a spatial parameter $L=22$ and $q=(18,17)$ for $TMP-2000$, and $L=22$ and $q=(19,19)$ for $TMP-2001$. We approximate the first 15 principal components to encompass the total variance within the functional data. Figs. 2 and 3 present a two-dimensional view of spatial positions using L1 and L2, representing longitude and latitude, showcasing the progression of the three functional spatial filters for $TMP-2000$ and $TMP-2001$ across lags. The first principal component is depicted in black, the second in red, and the third in green. Together, these three components explain over 80% of the total data variability. Notably, results for position 0 embody a standard FPCA, omitting the spatial parameter. This step marks the completion of the initial phase of the methodology, involving independent functional spectral analyses on $TMP-2000$ and $TMP-2001$. Once the functional spatial filters are calculated, the resulting scores and spatial filter operators are derived. Table 5 exhibits the percentage and cumulative percentage of variance accounted for by the principal components obtained using SMFPCA for $TMP-2000$ and $TMP-2001$. Moreover, Figs. 4 and 5 illustrate the filter operators after the first segment of the SMFPCA methodology.

**Table 5**
The cumulative percentage of variance explained by the first three univariate principal components for the variables $TMP-2000$ and $TMP-2001$ utilizing the SFPCA approach.

| Data \ Number of Components | 1 PC | 2 PC | 3 PC |
|---|---|---|---|
| $TMP-2000$ | 56.43 | 74.02 | 83.34 |
| $TMP-2001$ | 49.38 | 72.89 | 83.20 |

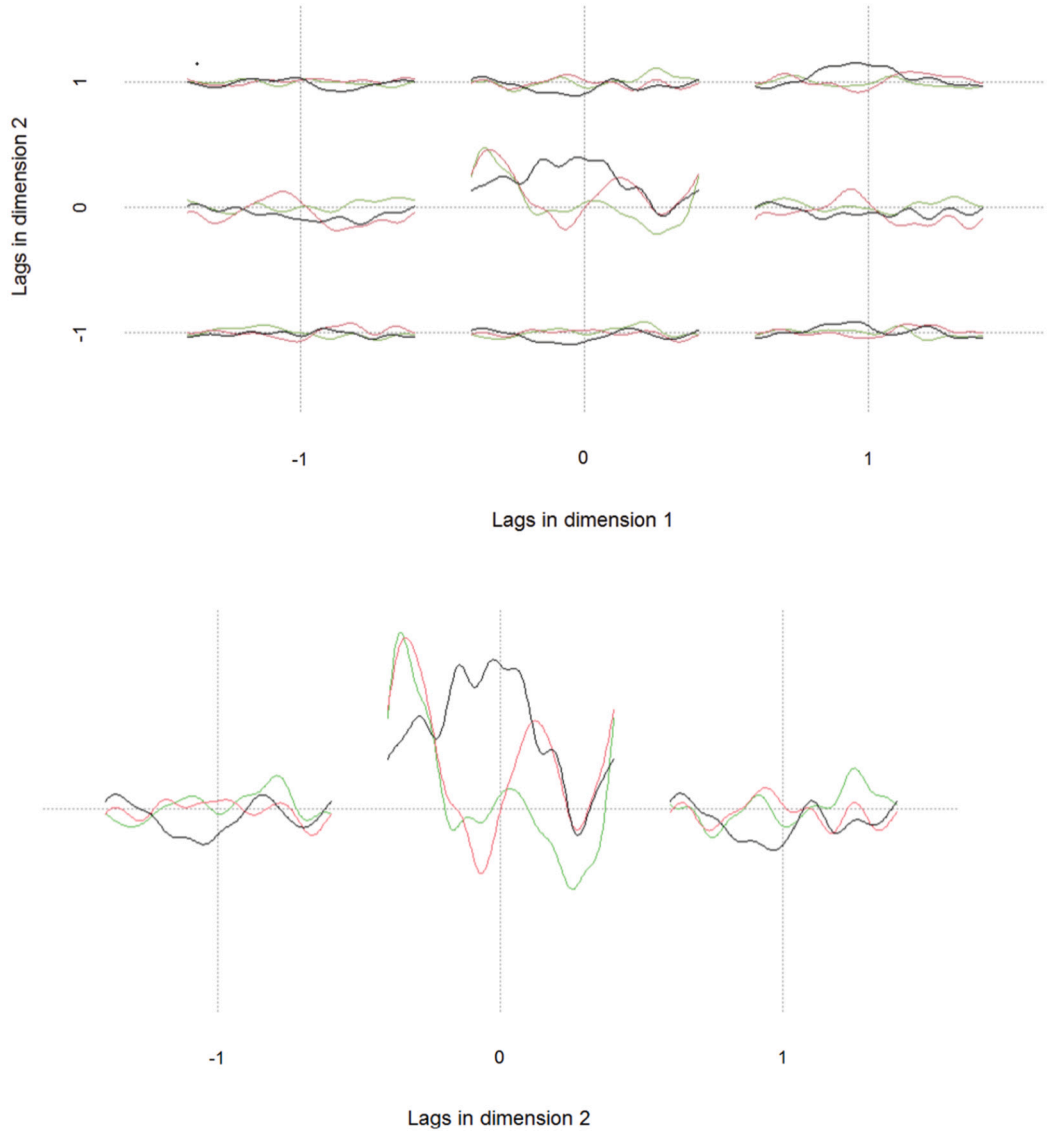


**Fig. 2.** Evolution of three spatial filters of the variable $TMP-2000$ over time following the application of SFPCA. At the top: A representation of two-dimensional spatial positions utilizing L1 and L2. At the bottom: One-dimensional spatial L2, indicating lag positions of (0,-1), (0,0) and (0,1).

Next, we move on to the second segment of the methodology, where the multivariate component is addressed, adhering to the steps outlined in SMFPCA procedure 2.3. We opted to calculate 4 multivariate spatial FPCA to account for over 80% of the total variation.

Similar to the simulated section, we evaluate the effectiveness of dimensionality reduction by calculating the NMSE for each individual component. We utilize equation (13), and the results are presented in Table 7. Moreover, we apply the NMSE$^*$ measure as defined by equation (14) to assess the quality of approximation while considering boundary effects. The findings in Table 7 strongly suggest that incorporating spatial information consistently enhances performance in terms of NMSE and NMSE$^*$. Both metrics, NMSE and NMSE$^*$, show a decreasing pattern as the number of cumulative principal components increases, aligning with the cumulative variance elucidated by the FPCA. Table 6 illustrates the explained cumulative variance of the principal component following the application of SMFPCA to variables $TMP-2000$ and $TMP-2001$. The SFPCA successfully integrates functional, multivariate, and spatial data, capturing disparities between $TMP-2000$ and $TMP-2001$ over marine surfaces and spatial frameworks. Notably, the first FPCA accounts for the majority of data variation, representing 58.80%. Additionally, the first four FPCA capture a substantial portion of the total variation, amounting to 87.33%.

We have included an additional comparison of SMFPCA approach with the STPCA method introduced by [21]. This method is a partially functional areal spatial data framework, which accounts for spatial autocorrelation through the use of a weight matrix. This methodology involves a two-step interdependent process: in the first step, the original multivariate time series are transformed into raw coefficients through a basis function expansion; in the second one, principal spatio-temporal components are constructed using a classical Moran's I statistic applicable to areal spatial data for the matrix of basis coefficients, while, in our approach, we used spatial spectral operator. Additionally, we embedded the data into a multivariate function space, treating them as fully observed functions in a geospatial regular domain. Furthermore, our approach addresses the limitations identified in [21] by incorporating multivariate perspectives within a fully functional geospatial framework, thus extending the univariate functional PCA approach of [16] to a multivariate context. Although the spatial framework employed by [21] pertains mainly to areal spatial data (considering spatial autocorrelation via a weight matrix), we juxtapose it (see Table 6) with our geospa-

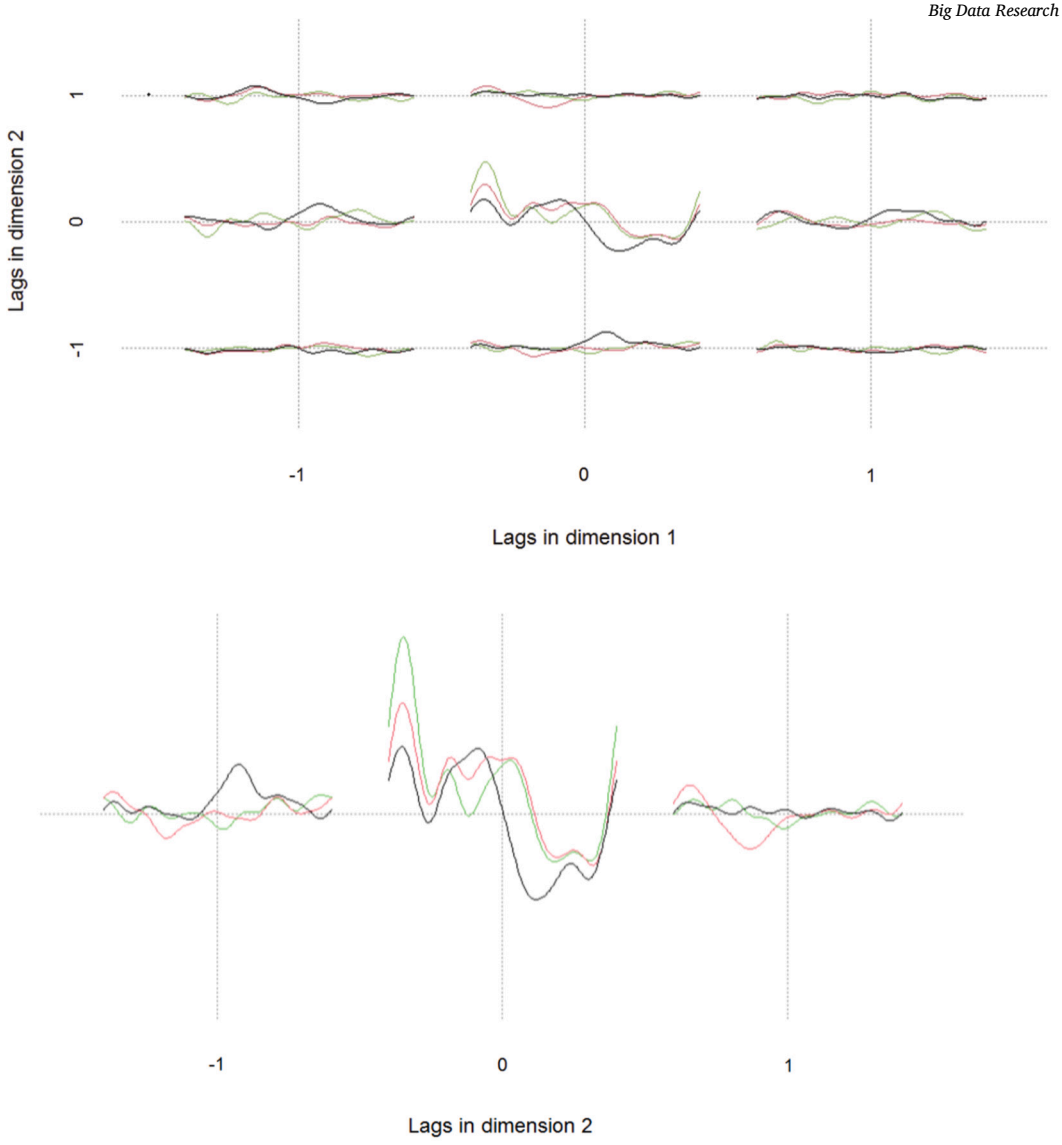


**Fig. 3.** Evolution of the three functional spatial filters of the variable $TMP-2001$ over the lag after the application of SFPCA. At the top: Two-dimensional spatial positions representation using L1 and L2. At the bottom: One dimensional spatial L2, which represent lag position of (0,-1), (0,0) and (0,1).

tial method, which accounts for spatial dependencies through spatial functional spectral operators. The authors of the article [21] tested the STPCA method on nine distinct spatial weight matrices $\boldsymbol{W}$ [21]. For our study, we selected three of these matrices to test the STPCA method on $TMP-2000$ and $TMP-2001$. Specifically, we chose the first matrix, which corresponds to the k-nearest neighbors weights (standard form), the second matrix based on k-nearest neighbors weights (symmetric form), and the third matrix, which uses Power distance-based weights. This selection was motivated by the fact that the eigenvalues of these matrices are all positive, unlike other weight matrices, whose eigenvalues are both positive and negative, rendering comparisons with the results of the SMFPCA method irrelevant, as SMFPCA eigenvalues are exclusively positive. Our proposed SMFPCA method demonstrated superior performance, explaining 58% of the variance with the first principal component and achieving a cumulative variance of approximately 87% with the first four principal components, thus outperforming other methods, as shown in Table 6.

The interpretation of multivariate functional principal components for each spatial location (Figs. 4 and 5) involves analyzing the common variations across different variables over time and space. In the case of multivariate data, the first functional principal components capture these shared patterns. For each spatial location, the analysis reveals how these components reflect the primary sources of variability across the different variables, offering insight into their temporal and spatial dynamics. The first principal component is depicted in black, the second in red, and the third in green. An example of interpreting functional principal components is that the first component represents the primary source of variation common to all studied variables over time. For instance, this could reflect a global trend where temperature of the year 2000 and 2001 increase or decrease synchronously. If the variations captured by the first functional principal component are highly similar across several variables, it suggests a strong temporal correlation between these variables at that particular location. The second component captures secondary variations, such as differences in how certain variables fluctuate independently of others. For example, while temperature of the year 2000 and 2001 may follow similar trends (captured by the two first functional principal component) could capture specific variations in wind speed or climatic anomalies in certain regions. The third functional principal component may capture more complex or subtle phenomena, such as interactions between variables (e.g., a rise in temperature in year 2000 could be accompanied by a decrease in temperature of the year 2001 in specific circumstances).

In setting 2, supplementary experiments were conducted using sea surface temperature data to examine the performance of the SMFPCA al-

**Table 6**
The cumulative percentage of variance explained by the first four principal components applied to the variables $TMP-2000$ and $TMP-2001$ is evaluated using the SMFPCA, MFPCA, and STPCA techniques. Here, $W_1$ denotes the k-Nearest Neighbor weights (standard form), $W_2$ signifies the k-Nearest Neighbor weights (symmetric form), and $W_3$ refers to the Power distance weights (see [21] for more details).

| Methods \ Number of Components | 1 PC | 2 PC | 3 PC | 4 PC |
|---|---|---|---|---|
| SMFPCA | 58.80 | 69.97 | 79.72 | 87.33 |
| MFPCA | 37.59 | 52.11 | 63.78 | 68.63 |
| STPCA ($W_1$) | 41.47 | 57.08 | 69.59 | 74.29 |
| STPCA ($W_2$) | 41.20 | 56.75 | 69.41 | 74.26 |
| STPCA ($W_3$) | 50.33 | 68.15 | 80.92 | 84.22 |

**Table 7**
The NMSE and NMSE* results were calculated for both the SMFPCA and MFPCA approaches, taking into account the variables $TMP-2000$ and $TMP-2001$, as well as the initial four multivariate principal components (1 PC, 2 PC, 3 PC, 4 PC).

| Number of Components (PC) | 1 PC | | 2 PC | | 3 PC | | 4 PC | |
|---|---|---|---|---|---|---|---|---|
| Methods | SMFPCA | MFPCA | SMFPCA | MFPCA | SMFPCA | MFPCA | SMFPCA | MFPCA |
| NMSE 2000 | 0.4796 | 0.5416 | 0.3396 | 0.5147 | 0.2103 | 0.3749 | 0.2090 | 0.3166 |
| NMSE* 2000 | 0.4356 | 0.5156 | 0.2596 | 0.3342 | 0.1664 | 0.2695 | 0.1120 | 0.2086 |
| NMSE 2001 | 0.5178 | 0.6016 | 0.3665 | 0.4121 | 0.3578 | 0.3627 | 0.2493 | 0.2707 |
| NMSE* 2001 | 0.5061 | 0.6021 | 0.2709 | 0.3788 | 0.1678 | 0.2686 | 0.1129 | 0.2135 |

**Table 8**
The NMSE and NMSE* values for both the SMFPCA and MFPCA approaches were determined by analyzing the variables $TMP-1996$, $TMP-1998$, and $TMP-1999$, using the initial three multivariate principal components (1 PC, 2 PC, 3 PC).

| Number of Components (PC) | 1 PC | | 2 PC | | 3 PC | |
|---|---|---|---|---|---|---|
| Methods | SMFPCA | MFPCA | SMFPCA | MFPCA | SMFPCA | MFPCA |
| NMSE 1996 | 0.5090 | 0.6364 | 0.5069 | 0.5215 | 0.3223 | 0.5029 |
| NMSE* 1996 | 0.4523 | 0.5358 | 0.2786 | 0.3772 | 0.1794 | 0.2851 |
| NMSE 1998 | 0.6980 | 0.7111 | 0.3418 | 0.5812 | 0.3026 | 0.5069 |
| NMSE* 1998 | 0.4476 | 0.5791 | 0.2624 | 0.3855 | 0.1640 | 0.2837 |
| NMSE 1999 | 0.4377 | 0.4762 | 0.3053 | 0.3941 | 0.2758 | 0.3237 |
| NMSE* 1999 | 0.4254 | 0.4744 | 0.2739 | 0.3520 | 0.1889 | 0.2778 |

gorithm on three variables. We randomly selected the NOAA variables representing the years 1996, 1998 and 1999, considering both the spatial and non-spatial aspects. We projected the data onto 15 Fourier basis functions, and we performed univariate SFPCA with the spatial parameters $L = 22$ and $q$ defined as:

- 1996: $q = (15, 14)$
- 1998: $q = (22, 22)$
- 1999: $q = (19, 18)$

We then describe the process outlined in procedure 2.3 for SMFPCA across both contexts, and compute the initial 4 multivariate principal components, which account for over 80% of the variability. Upon computing SMFPCA, we carried out a comparative evaluation of the reconstructed functional data against traditional MFPCA, with the results presented in Table 8. The outcomes are consistent with those observed in configuration 1; both the NMSE and NMSE* demonstrate a decline as the number of principal components included increases. This examination attests to positive results, even when considering three variables or more. Dimensionality reduction of multivariate functional spatial data, depicted in Figs. 4 and 5, is accomplished by determining the principal directions of variations among the functional variables. Consequently, the essence of the data is succinctly captured using a limited set of FPCA, enabling more straightforward analysis and interpretation of multiple functional spatial variables. Furthermore, this methodology improves the visualization and understanding of the results by making it easier to discern spatial patterns, structures, and trends within the data.

## 4. Conclusion and perspectives

This study is focused on reducing the dimensionality of multivariate geospatial functional data. Our method termed SMFPCA (Spatial Multivariate Functional Principal Component Analysis) sets itself apart from prior work by tackling the multivariate functional dimension, accurately capturing the data's multivariate nature, SFPCA falls short in when dealing with interactions across different functions if applied separately to each variable. Furthermore, it integrates spatial dependence through spectral density to improve dimension reduction efficacy. We initially demonstrate the finite sample characteristics of SMFPCA using simulated datasets, contrasting our approach with methods that disregard spatial aspects (MFPCA). This allowed us to highlight a significant improvement in performance when spatial components are considered.

In a similar manner, we applied SMFPCA, MFPCA, and another competitor (STPCA) that incorporates spatial elements via a weight matrix and Moran's index to NOAA data, which revealed a very local spatial dependence. We utilized the sea surface temperature variables from the years 2000 and 2001 as illustrative examples of spatially indexed multivariate data. Additionally, we performed further analysis on NOAA data across three different years. SMFPCA notably increases the percentage of variance explained when compared to STPCA and MFPCA methods. These diverse findings highlight SMFPCA's ability to effectively capturing and explaining inter-variable differences as well as the spatial variations inherent in the data. Moreover, SMFPCA can extract latent information from functional multivariate spatial data, revealing previously hidden dimensions and elucidating spatial temporal patterns.

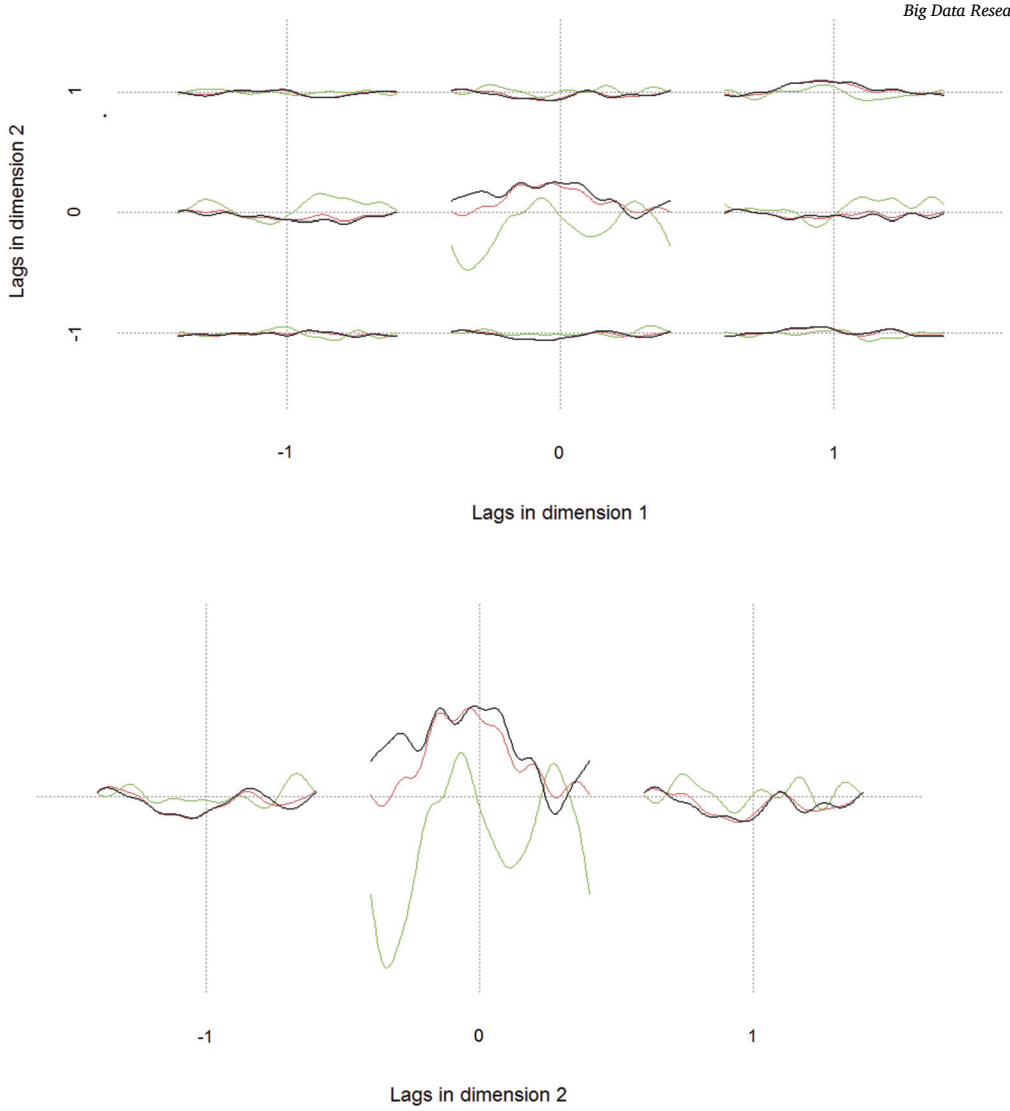


**Fig. 4.** Evolution of the three functional spatial filters of the variable $TMP - 2000$ over the lag after the application of SMFPCA. At the top: Two-dimensional spatial positions representation using L1 and L2. At the bottom: One dimensional spatial L2, which represent lag position of (0,-1), (0,0) and (0,1).

The proposed method paves the way for future research, with some topics outlined below. In the fields of economics and finance, a major hurdle in dimensionality reduction is managing non-stationary time series data. Therefore, a robust methodology is necessary to tackle non-stationary series, facilitating the capturing of crucial fluctuations and uncertainties to improve modeling precision. This necessity has resulted in the creation of factor models or generalized dynamic principal component analysis [34–36], while also prompting the examination of a spatial functional data viewpoint. Moreover, various other methods are actively used, such as the technique proposed by [37], which entails smoothing techniques for transforming irregular data into regular forms. Another approach [38] investigates a non-parametric method to handle irregularly spaced data by reconfiguring a rectangular network into a regular grid. Looking ahead, it is valuable to consider integrating additional functionalities for handling irregular data, which would greatly expand the range of possible analyses over varied data structures.

## CRediT authorship contribution statement

**Idris Si-ahmed:** Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Leila Hamdad:** Visualization, Validation, Supervision, Project administration. **Christelle Judith Agonkoui:** Validation, Methodology, Investigation, Formal analysis. **Yoba Kande:** Methodology, Formal analysis, Data curation. **Sophie Dabo-Niang:** Visualization, Validation, Supervision, Project administration, Methodology, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.bdr.2024.100504.

## Data availability

Data will be made available on request.

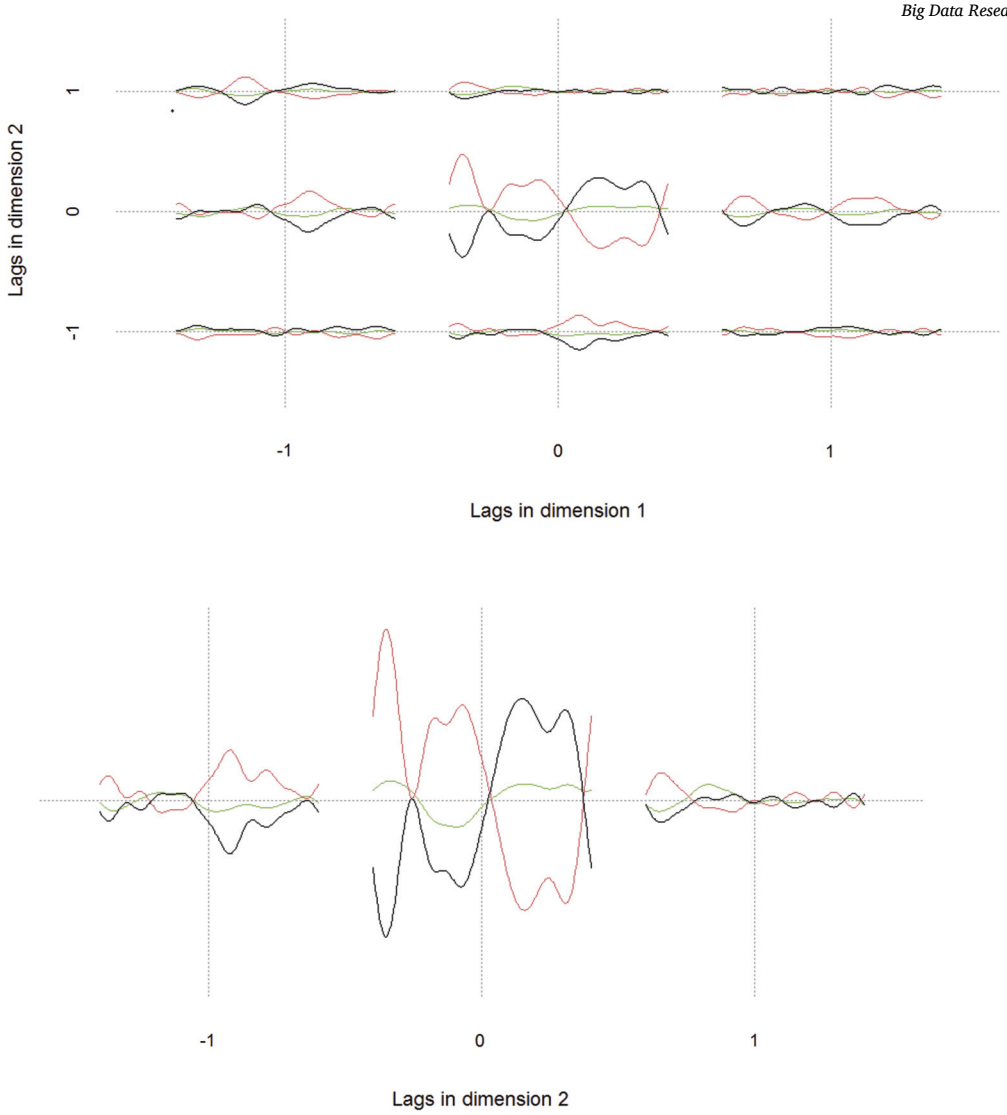


**Fig. 5.** Evolution of the three spatial filters of the variable $TMP - 2001$ over time subsequent to SMFPCA implementation. At the top panel, L1 and L2 depict the two-dimensional spatial positions. In contrast, the lower panel presents a one-dimensional spatial L2 illustrating the lag positions at (0,-1), (0,0), and (0,1).

## References

[1] M.R. Piotr Kokoszka, Introduction to Functional Data Analysis, 1st edition, Chapman and Hall/CRC, New York, 2017.
[2] P.V. Frederic Ferraty, Nonparametric Functional Data Analysis: Theory and Practice, 1st edition, Springer Science & Business Media, 2006, 2006.
[3] J.O. Ramsay, B.W. Silverman, Functional data analysis, in: Springer Series in Statistics, Springer, New York, 2005, pp. 327–348.
[4] J.O. Ramsay, B.W. Silverman (Eds.), Applied Functional Data Analysis: Methods and Case Studies, Springer Series in Statistics, Springer, New York, 2002.
[5] H.L. Shang, A survey of functional principal component analysis, AStA Advances in Statistical Analysis 98 (522) (2014) 121–142.
[6] Y. Li, Y. Qiu, Y. Xu, From multivariate to functional data analysis: fundamentals, recent developments, and emerging areas, Journal of Multivariate Analysis 188 (2022) 104806.
[7] C.K. Klepsch, Johannes, T. Wei, Prediction of functional arma processes with an application to traffic data, Econometrics and Statistics 1 (2017) 119–128.
[8] M.D. Ruiz-Medina, Spatial functional prediction from spatial autoregressive Hilbertian processes, EnvironMetrics 23 (1) (2012) 119–128.
[9] R.W. Reynolds, T.M. Smith, C. Liu, D.B. Chelton, K.S. Casey, M.G. Schlax, Daily high-resolution blended analyses for sea surface temperature, Journal of Climate 20 (22) (2007) 5473–5496.
[10] J. Mateu, E. Romano, Advances in spatial functional statistics, Stochastic Environmental Research and Risk Assessment 31 (2017) 1–6.
[11] S. Koner, A.-M. Staicu, Second-generation functional data, Annual Review of Statistics and Its Application 10 (1) (2023) 547–572.
[12] S. Hörmann, P. Kokoszka, Weakly dependent functional data, The Annals of Statistics 38 (3) (2010) 1845–1884.
[13] L. Zhou, J. Huang, J.G. Martinez, A. Maity, V. Baladandayuthapani, R.J. Carroll, Reduced rank mixed effects models for spatially correlated hierarchical functional data, Journal of the American Statistical Association 105 (490) (2010) 390–400.
[14] H. Li, S. Kozey Keadle, J. Staudenmayer, H. Assaad, J.Z. Huang, R.J. Carroll, Methods to assess an exercise intervention trial based on 3-level functional data, Biostatistics 16 (4) (2015) 754–771.
[15] Y. Li, D.V. Nguyen, S. Banerjee, C.M. Rhee, K. Kalantar-Zadeh, E. Kurum, D. Senturk, Multilevel modeling of spatially nested functional data: spatiotemporal patterns of hospitalization rates in the US dialysis population, Statistics in Medicine 40 (13) (2021) 3112–3132.
[16] T. Kuenzer, S. Hörmann, P. Kokoszka, Principal component analysis of spatially indexed functions, Journal of the American Statistical Association 116 (535) (2021) 1444–1456.
[17] T.H. Khoo, D. Pathmanathan, S. Dabo-Niang, Spatial autocorrelation of global stock exchanges using functional areal spatial principal component analysis, Mathematics 11 (3) (2023) 674.
[18] J.M. Chiou, Y.T. Chen, Y.F. Yang, Multivariate functional principal component analysis: a normalization approach, Statistica Sinica 24 (2014) 1571–1596.
[19] J.R. Berrendero, A. Justel, M. Svarc, Principal components for multivariate functional data, Computational Statistics & Data Analysis 55 (9) (2011) 2619–2634.
[20] C.M. Happ, Statistical methods for data with different dimensions, 2017.
[21] M. Krzyśko, W. Wołyński, A. Zienkiewicz, M. Misztal, A. Krzyżak, Spatio-temporal principal component analysis, Spatial Economic Analysis 19 (1) (2023) 8–29.
[22] D. Pathmanathan, P.S. Kokoszka, M. Reimherr, J. Aston, Spatial principal component analysis and Moran statistics for multivariate functional areal data, preprint, arXiv: 2408.08630, 2024.
[23] P. Delicado, R. Giraldo, C. Comas, J. Mateu, Statistics for spatial functional data: some recent contributions, EnvironMetrics 21 (3–4) (2010) 224–239.

[24] Y. Li, Y. Guan, Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance, Journal of the American Statistical Association 109 (507) (2014) 1205–1215.
[25] C. Liu, S. Ray, G. Hooker, Functional principal component analysis of spatially correlated data, Statistics and Computing 27 (2017) 1639–1654.
[26] W.K.H. Yingxing Li, Chen Huang, Spatial functional principal component analysis with applications to brain image data, Journal of Multivariate Analysis 170 (2019) 263–274.
[27] G. He, H.-G. Müller, J.-L. Wang, Functional canonical analysis for square integrable stochastic processes, Journal of Multivariate Analysis 85 (2003) 54–77.
[28] Q. Jiang, H. Wang, Y. Xia, G. Jiang, On a principal varying coefficient model, Journal of the American Statistical Association 108 (2013) 228–236.
[29] D.R. Kowal, D.S. Matteson, D. Ruppert, A Bayesian multivariate functional dynamic linear model, Journal of the American Statistical Association 112 (2017) 733–744.
[30] R.F. Barber, M. Reimherr, T. Schill, The function-on-scalar lasso with applications to longitudinal gwas, Electronic Journal of Statistics 11 (2017) 1351–1389.
[31] J.S. Morris, R.J. Carroll, Wavelet-based functional mixed models, Journal of the Royal Statistical Society, Series B, Statistical Methodology 68 (2006) 179–199.
[32] A. Parodi, M. Reimherr, Simultaneous variable selection and smoothing for high-dimensional function-on-scalar regression, Electronic Journal of Statistics 12 (2018) 4602–4639.
[33] C. Happ, S. Greven, Multivariate functional principal component analysis for data observed on different (dimensional) domains, Journal of the American Statistical Association 113 (522) (2018) 649–659.
[34] D. Peña, P. Poncela, Nonstationary dynamic factor analysis, Journal of Statistical Planning and Inference 136 (4) (2006) 1237–1257.
[35] D. Peña, V.J. Yohai, Generalized dynamic principal components, Journal of the American Statistical Association 111 (515) (2016) 1121–1131.
[36] T.H. Khoo, I.-M. Dabo, D. Pathmanathan, S. Dabo-Niang, Generalized functional dynamic principal component analysis, preprint, arXiv:2407.16024, 2024.
[37] J.P. French, P.S. Kokoszka, A sandwich smoother for spatio-temporal functional data, Spatial Statistics 42 (2021) 100413.
[38] P.M. Robinson, Asymptotic theory for nonparametric regression with spatial data, Journal of Econometrics 165 (1) (2011) 5–19.