Principal Component Analysis of Multivariate Spatial Functional Data

Idris Si-Ahmed, Leila Hamdad, Christelle Judith Agonkoui, Yoba Kande and Sophie Dabo-Niang

August 7, 2024

Principal Component Analysis of Multivariate Spatial Functional Data

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

[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, Algeria DZ*
[b]*Ecole Nationale Supérieure d'Informatique (ESI) laboratoire LCSI, Oued Smar 16309, Alger, 16000, Algeria, Algeria DZ*
[c]*Institute of mathematics and Physics Sciences, University of Abomey Calavi, 42 rue Paul Duez 59000, Abomey-Calavi, Benin, Benin*
[d]*Université de Lille, CNRS, UMR 8524 Laboratoire Paul Painleve, 42 rue Paul Duez 59000, Lille, France, France FR*
[e]*Université de Lille, 42 rue Paul Duez 59000, Lille, France, France FR*

Abstract

This paper is dedicated to dimension reduction techniques for multivariate spatially indexed functional data. We introduce an innovative method named Spatial Multivariate Funtional Principal Component Analysis (SMFPCA), which stands for principal component analysis for multivariate spatial functional data. Unlike the conventional Multivariate Karhunen-Loève approach, SMFPCA excels at effectively capturing spatial dependencies among multiples functions. SMFPCA conducts spectral functional component analysis on multivariate spatial data, encompassing data points located within a regular grid. The methodological framework and algorithm for SMFPCA have been developed to address the challenges posed by the lack of suitable methods for handling such data. The efficiency of the proposed methodology has been substantiated through comprehensive assessments of its performance using and simulated datasets and sea-surface temperature, providing valuable insights into the properties of multivariate spatial functional data within a finite sample.
Keywords: Spectral Analysis, Functional Data Analysis, Functional Principal Component Analysis, Spatial-functional Principal Component Analysis, Multivariate analysis

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 distinct data type in which each observation is represented by a function rather than a conventional vector of values. These data are of infinite dimension, as it is defined across the entire continuum of points.

Functional Principal component analysis (FPCA) is a technique used for exploratory functional data analysis. It is commonly employed to reduce the multidimensional space, where the information is not easily interpretable, into a space of reduced dimension. Furthermore, FPCA allows us to derive more informative

URL: `i_siahmed@esi.dz` (Idris Si-ahmed), `l_hamdad@esi.dz` (Leila Hamdad), `christelle.agonkoui@imsp-uac.org` (Christelle Judith Agonkoui), `kandeyoba@gmail.com` (Yoba Kande), `sophie.dabo@univ-lille.fr` (Sophie Dabo-Niang)

attributes from the data, identify hidden patterns in a dataset, and discover correlations among variables [1], [2] [3], [4], [5].

Within the domain of spatially indexed functional data analysis, the data are measured over space. This could include, for example, temperature measurements taken at different locations in a geographic region. One of the most common uses of principal component analysis on functional spatial data (SFPCA) is in studies where the functional spatial data consists of spatially distributed objects, such as environmental measurement sampling sites, watersheds, administrative districts, etc. The value of spatial analysis is to understand and explore the interrelationship between the spatial positioning of objects, phenomena, and their characteristics [6]. The analysis of spatially indexed functional data serves two purposes. Firstly, it enables the identification of spatial patterns inherent in the functional data, these patterns provide valuable insights into the underlying spatial structure and dynamics of the phenomena being studied. Secondly, the analysis enables the development of models that can be used for making predictions or drawing conclusions about the spatial distribution of the data. In the literature, [7] provided a useful introduction to selected methods for geostatistical functional data; [8] developed extending ANOVA techniques to analyze spatially correlated functional data. [9] and [10] used a spatially indexed functional data framework to solve space physics problems. [11] and [12] developed spatial-temporal separability tests for functional data. [13] considered tests of anisotropy using Karhunen-Loève expansion but with a semiparametric estimation procedure custom-developed for their objectives. The work of [14] is an example of a substantive application of spatial functional modeling to a problem of practical importance. [15] present a method that combines singular value decomposition with penalized smoothing to avoid high-dimensional covariance estimation, with a focus on brain image analysis. [16] proposed a dimension reduction technique suitable for functional data indexed by spatial locations on a regular grid. They developed the mathematical foundations for the spectral analysis of such data, including spectral theory for linear spatial filters. In our work, we focus on scenarios where multiple variables are examined, defined within both the same and different spatial-temporal domains. For instance, to study temperature co-variability in a specific region over a defined time span, we have positioned ourselves within this framework and utilized a spectral analysis approach to seek viable solutions.

Multivariate functional data analysis is a statistical and mathematical approach designed for the analysis of multiple dependent variables or functions within a one or multidimensional context. It involves studying the relationships and interactions between multiple functional data sets, considering various variables or dimensions simultaneously. In the literature, there are several approaches for multivariate functional principal component analysis (MFPCA) that are restricted to functions observed on the same finite one-dimensional interval ([3]; [17]; [18]), all rely on a multivariate functional Karhunen-Loève data representation. [17] also discusses normalized versions of MFPCA based on a normalized covariance operator. A method [19] has been proposed for conducting MFPCA for independent data and image. This approach is specifically designed to explore patterns of joint variation in multivariate functional data. It can be applied to datasets where the functional variables are observed on potentially different dimensional domains, enabling comprehensive analysis of their interdependence. The estimation algorithm is based on univariate FPCA. However, the authors did not take into consideration spectral analysis of indexed spatial data when applying the FPCA, which may yield better results.

The objective of this study is to introduce a novel approach that enables spectral FPCA on multivariate spatial data, defined within both the same and different domains, all structured on a regular grid. We refer to this approach as SMFPCA, with the aim of extending previous work related to MFPCA [19], which does not consider spectral analysis, and SFPCA [16], which does not consider the multivariate aspect. To our knowledge, no previous work has employed multivariate spectral FPCA on spatial and temporal variables within the same and different domain. We apply spectral spatial univariate FPCA, and the results can be used to estimate multivariate functional principal components, eigenvalues, and scores based on their univariate equivalents. Spectral decomposition is a high-performance method for analyzing functional spatial data. It is based on the spectral density operator, which describes the spatial dependence structure of the data by separating it into distinct frequencies. The advantage of this approach is that it allows the identification of dominant frequencies that drive the variation in the spatial dependence structure of the data. These dominant frequencies can then be used to define the FPCA, providing insights into the underlying stochastic processes contributing to the data variation. Thus, spectral decomposition is an excellent tool for a better understanding of the complex spatial dependence structure. We conducted a test of our method on multivariate and spatially indexed environmental data. SMFPCA facilitates the integration of this type of data within a unified analytical framework. This approach effectively captures and elucidates the variations between variables as well as the inherent spatial variations present in the data. The SMFPCA technique also possesses the capability to extract latent information from functional multivariate spatial data, uncovering previously unobserved latent dimensions. This, in turn, explains variations in the data and reveals spatial-temporal patterns. As a result, SMFPCA provides valuable insights into the underlying mechanisms driving the observed data.

The paper is structured as follows: In Section 2, we describe the SMFPCA methodology; in Section 3, we present finite sample properties and analyze the findings. Section 4 give the conclusion and some perspectives.

2. SPECTRAL PRINCIPAL COMPONENT ANALYSIS OF MULTIVARIATE SPATIAL FUNCTIONAL DATA

This section describes the proposed methodology and the mathematical model for conducting a spectral SMFPCA on multivariate spatial-functional data. The methodology is based on the principles of Karhunen-Loève theory, incorporating spectral analysis and Functional Principal Component Analysis (FPCA). Additionally, key properties and concepts are derived from the works of [16], [3], [4], and [19]. The approach specifically targets a multivariate spatial-functional phenomenon observed on a regular grid and defined within both the same and different domains.

In section 2.1, we delve into the mathematical model employed in our approach. We provide definitions for various components, including the spatial multivariate functional data model, covariance operator, spectral density operator, eigenfunctions (representing functions that characterize significant variations or dominant modes within the functional data), eigenvalues (corresponding to the weights assigned to each eigenfunction and quantifying the total variance explained), and scores (representing the projections of the original data onto the basis functions). Once these elements are clearly established, we proceed to outline the subsequent steps of SMFPCA 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 $\langle\langle \cdot, \cdot \rangle\rangle$ [19].

For each component $X_{\mathbf{s}}^{(j)}$, 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 sample $\mathbf{X}^{(j)}$, to compute a univariate SFPCA. To achieve this, we apply the univariate spatial FPCA [16]. Let $X^{(j)} \in \mathcal{L}^2(\mathcal{T}_j)$ possesses a covariance operator $C_j := \mathbb{E}[(X^{(j)} - \mu^j) \otimes (X^{(j)} - \mu^j)]$ (where μ^j is the mean curve define 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 $\mathbb{R}^N$

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

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 θ 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 θ), 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 mth spatial functional principal component (SFPC) score by:

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

where $\phi_{m,\mathbf{l}}^{(j)}$ is defined by (4). The corresponding SFPC filter 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 :

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

.

- 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

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

$$X_{m,\mathbf{s}}^{(j)}(t) := \sum_{\mathbf{l}\in\mathbb{Z}^N} \xi_{m,\mathbf{s}+\mathbf{l}}^{(j)} \phi_{m,\mathbf{l}}^{(j)}(t).$$

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

The spectral density operator is estimated by:

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

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:

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

with $M_{\mathbf{h},\mathbf{n}} = \left\{ \mathbf{s} \in \mathbb{Z}^N : 1 \leq s_i, s_i + h_i \leq n_i\ \forall 1 \leq i \leq N \right\}$. 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. The methodology 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 $m = 1, \ldots, M_j$ 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 $\widehat{\mathcal{F}}^{X^{(j)}}_{\boldsymbol{\theta}}$.
3. Finally, the $\mathbf{X}^{(j)}$ are projected onto the spectral eigenfunctions, yielding the estimated scores $\hat{\xi}^{(j)}_{m,\mathbf{s}}$, defined by:

$$\hat{\xi}^{(j)}_{m,\mathbf{s}} := \sum_{\|\mathbf{l}\|_\infty \leq L} \langle X^{(j)}_{\mathbf{s}-\mathbf{l}}, \hat{\phi}^{(j)}_{m,\mathbf{l}} \rangle,$$

   assuming that $1 + L \leq s_i \leq n_i - L$ for all $1 \leq i \leq N$, 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:

5. Define the matrix $\mathbf{E} \in \mathbb{R}^{n \times M_+}$ of rows $(\hat{\xi}^{(1)}_{1,\mathbf{s}}, \ldots, \hat{\xi}^{(1)}_{M_1,\mathbf{s}}, \ldots, \hat{\xi}^{(p)}_{1,\mathbf{s}}, \ldots, \hat{\xi}^{(p)}_{M_p,\mathbf{s}}), \mathbf{s} \in \mathbf{D}$ consists of all the scores estimated from the Spectral PCA on each $\mathbf{X}^{(j)}$, with $M_+ = M_1 + \ldots + M_p$.

   Let's consider the matrix $\mathbf{Z} \in \mathbb{R}^{M_+ \times M_+}$ (Prop.5., p.7 [20]) consisting of blocks $\mathbf{Z}^{(jk)} \in \mathbb{R}^{M_j \times M_k}$ with entries

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

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

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

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

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

   where $[\hat{c}_m]^{(j)} \in \mathbb{R}^{M_j}$ denotes the j-th block of the (orthonormal) eigenvector $\hat{c}_m$ of $\hat{\mathbf{Z}}$.

   Furthermore, multivariate scores are calculated as:

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

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

$$X_{\mathbf{s}}^{(j)}(t_j) \approx \sum_{m=1}^{M_j} \hat{X}_{m,\mathbf{s}}^{(j)}(t_j), \quad t_j \in \mathcal{T}_j,$$
$$\text{with} \quad \hat{X}_{m,\mathbf{s}}^{(j)}(t_j) := \sum_{\|\mathbf{l}\|_\infty \le L} \hat{\xi}_{m,\mathbf{s+l}}^{(j)} \hat{\phi}_{m,\mathbf{l}}^{(j)}.$$

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

3. NUMERICAL EXPERIMENTS

In this section, we applied the SMFPCA procedure to simulated and real data. We compare the results of the proposed SMFPCA methodology with MFPCA results, the latter of which does not account for spatial considerations.

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 $\{X_{s,t}^j(.)\}$ defined by:

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

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 errors $\epsilon_{t,s}$ are i.i.d gaussian variables. 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:

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

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

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

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

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 setting:

1. The first setting involves generating within the same domain $\mathcal{T}_j \in [0,1]$ and $j = 1,2$, two simulated functional variables following (12), denoted SIM_1 and SIM_2 with 10 and 14 Fourier basis functions respectively.
2. In setting two, we simulate variables defined in the first configuration and introduce errors following a normal distribution.
3. The third setting involves generating two SFARMA variables, denoted SIM_1 and SIM_2 in distinct domains, with $\mathcal{T}_1 \in [0,1]$ and $\mathcal{T}_2 \in [2,4]$. We use a set of 14 and 10 Fourier basis functions for SIM_1 and SIM_2 respectively.
4. In the fourth setting, we simulate two functional variables as in the third configuration and introduce errors following a normal distribution.

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 q and L. These two parameters value are determined with the automatic routine function established in the FSD package, as defined by ([16] and Supplemental material), specifically, we assign $q = (18,17)$ across the four settings, while the parameter L assumes distinct values of 12, 10, 12, and 10 for each respective setting. Then we proceed with describing steps in procedure 2.3, and compute 4 multivariate principal components to capture more than 80% of variability.

After performing SMFPCA, we conducted a comparative assessment of the reconstructed functional data. We applied equations (13) and (14), and the corresponding results are shown in Tables 1, 2, 3, and 4. We can observe the performance of SMFPCA compare to the conventional MFPCA approach, which does not account for spatial considerations. Using the first configuration as an illustrative example, as depicted in Table 1, the results emphasize that the enhancement in NMSE and NMSE* quality is linked to the number of functional principal components. This phenomenon is attributed to the fact that when the first three 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

Table 1: NMSE and NMSE* results obtained by SMFPCA and MFPCA considering the variables SIM_1 and SIM_2 in different domain [0,1] and [2,4].

Cumulative PCA	PC1		PC2		PC3	
Spatial consideration	Spatial	Ordinary	Spatial	Ordinary	Spatial	Ordinary
NMSE SIM_1	0.5473	0.5877	0.3882	0.5807	0.2214	0.5659
NMSE* SIM_1	0.4366	0.5811	0.2350	0.3924	0.1369	0.2425
NMSE SIM_2	0.5448	0.9078	0.3798	0.6036	0.3621	0.3712
NMSE* SIM_2	0.4253	0.6039	0.2336	0.3710	0.1303	0.2419

Table 2: NMSE and NMSE* results obtained by SMFPCA and MFPCA considering the variables SIM_1 and SIM_2 in different domain [0,1] and [2,4] with introduced errors.

Cumulative PCA	PC1		PC2		PC3	
Spatial consideration	Spatial	Ordinary	Spatial	Ordinary	Spatial	Ordinary
NMSE SIM_1	0.4500	0.4591	0.3359	0.4589	0.2006	0.4503
NMSE* SIM_1	0.3617	0.4590	0.2007	0.2927	0.1193	0.1859
NMSE SIM_2	0.5764	0.9882	0.3305	0.5005	0.3123	0.3137
NMSE* SIM_2	0.3527	0.5005	0.1929	0.3136	0.1099	0.1941

Table 3: NMSE and NMSE* results obtained by SMFPCA and MFPCA considering the variables SIM_1 and SIM_2 in domain [0,1].

Cumulative PCA	PC1		PC2		PC3		PC4	
Spatial consideration	Spatial	Ordinary	Spatial	Ordinary	Spatial	Ordinary	Spatial	Ordinary
NMSE SIM_1	0.6347	0.7171	0.4298	0.6040	0.4294	0.6034	0.2593	0.3840
NMSE* SIM_1	0.4729	0.6040	0.2705	0.3844	0.1542	0.2481	0.0904	0.1631
NMSE SIM_2	0.5566	0.7071	0.4465	0.6257	0.2645	0.4002	0.2643	0.3993
NMSE* SIM_2	0.4702	0.6258	0.2340	0.4002	0.1330	0.2532	0.0788	0.1644

Table 4: NMSE and NMSE* results obtained by SMFPCA and MFPCA considering the variables SIM_1 and SIM_2 in domain [0,1] with introduced errors.

Cumulative PCA	PC1		PC2		PC3		PC4	
Spatial consideration	Spatial	ordinary	Spatial	ordinary	Spatial	ordinary	Spatial	ordinary
NMSE SIM_1	0.5753	0.8565	0.4199	0.6079	0.4174	0.5540	0.2526	0.3707
NMSE* SIM_1	0.4612	0.6080	0.2618	0.3710	0.1517	0.2433	0.0896	0.1577
NMSE SIM_2	0.5922	0.6404	0.4383	0.6234	0.2674	0.3932	0.2671	0.3897
NMSE* SIM_2	0.4655	0.6247	0.2375	0.3903	0.1348	0.2456	0.0793	0.1629

NMSE. Regarding the outcomes depicted in Tables 2, 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 93E longitude and 15 to 44S 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. Figure 1 illustrates a snapshot of this extensive dataset.

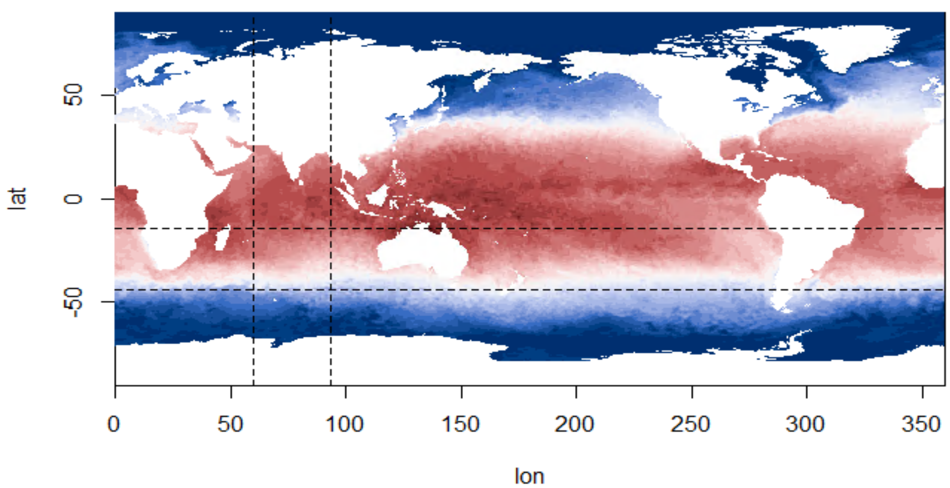

Figure 1: Indian Ocean ranging approximately from 60 to 93E longitude and 15 to 44S latitude.

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 and ordinary MFPCA on NOAA data, and illustrate the performance estimation procedure in two setting:

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

We begin with the first configuration. The starting step of the SMFPCA algorithm involves computing the functional data. We transform the $TMP-2000$ and $TMP-2001$ into spatial function data using 15 Fourier basis functions, this projection yields functions denoted as $X_{s,u}^{(j)}(.)$, where the indices s, u, and

j represent longitude, latitude, and year, respectively. The time domain is $\mathcal{T}_j$, it represents the intra-year time and is scaled to the unit interval $[0, 1]$. Then we perform univariate SFPCA on these variables with a spatial parameter $L = 22$ and $q = (18, 17)$ for the $TMP - 2000$ variable, and $L = 22$ and $q = (19, 19)$ for the $TMP - 2001$ variable. Similar to the approach applied in the simulated data, the values of these two parameters are chosen using the automatic routine function. We estimate the first 15 principal components to capture the total variation present in the functional data. Figures 2 and 3 show the first three principal components, which account for more than 80% of the total data variation. Notice that the results corresponding to position 0 represent an ordinary FPCA without considering the spatial parameter. At this step, we have applied the initial part of the methodology, which involves conducting separate functional spectral analyses on the $TMP - 2000$ and $TMP - 2001$. Once the functional spatial filters have been computed, we obtain the resulting scores and spatial filter operators.

Subsequently, we proceed to the second part of the methodology, wherein the multivariate aspect is considered, following the guidelines specified in procedure 2.3. We have chosen to compute 4 multivariate spatial FPCA to capture more than 80% of the total variation. Table 5 displays the percentage of variance explained by the principal components obtained through the application of SMFPCA to the variables $TMP-2000$ and $TMP - 2001$. Additionally, Figures 4 and 5 show the filter operators presented in Table 5.

Table 5: Cumulative percentage (CP) of explained variance of principal components obtained by SFPCA considering the variables $TMP - 2000$ and $TMP - 2001$.

CP	PC1	PC2	PC3
$TMP - 2000$	56.43	74.02	83.34
$TMP - 2001$	49.38	72.89	83.20

Equivalently to the simulated section, we assess the efficacy of dimensionality reduction through the evaluation of the NMSE for each univariate component. We apply equation (13), and the outcomes are detailed in Table 7. Additionally, we employ the NMSE* metric as defined in equation (14), which characterizes the quality of approximation while accounting for boundary effects. Table 6 displays the explained cumulative variance of the FPCA after applying SMFPCA to variables $TMP - 2000$ and $TMP - 2001$. The SFPCA integrate functional, multivariate, and spatial information, effectively capturing variations between $TMP - 2000$ and $TMP - 2001$ variables on marine surfaces and spatial models. We can observe that the first FPCA accounts for the largest variation in the data, which represents 58.80%. Furthermore, the 4 first FPCA capture a significant amount of the overall variation, which represents 87.33% of the total variation.

The results presented in Table 7 provide strong evidence that the incorporation of the spatial aspect consistently leads to improved performance in terms of NMSE and NMSE*. Both NMSE and NMSE* exhibit a decreasing trend with an increasing number of cumulative principal components, which is consistent with the cumulative variance explained by the FPCA. In other words, as more variance is accounted for by the FPCA, the NMSE and NMSE* values decrease.

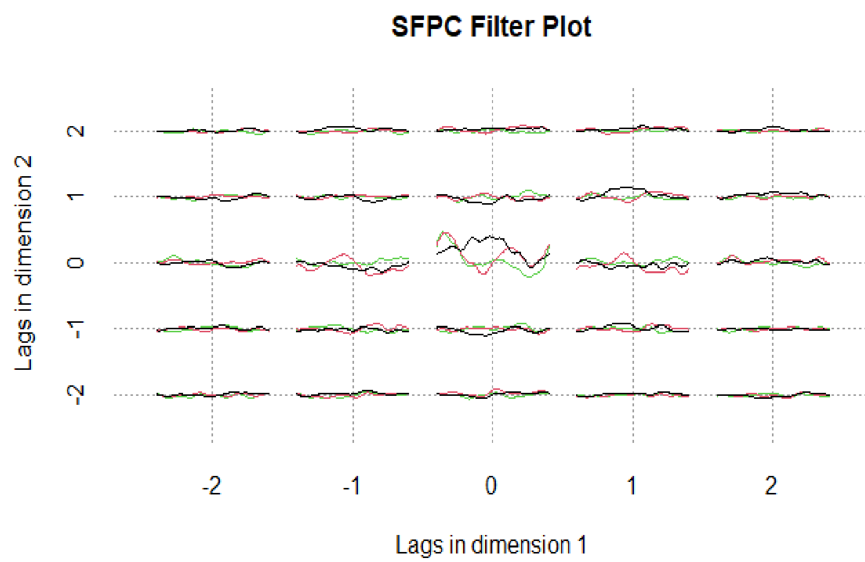

Figure 2: Evolution of three functional spatial filters of $TMP-2000$ variable over the lag.

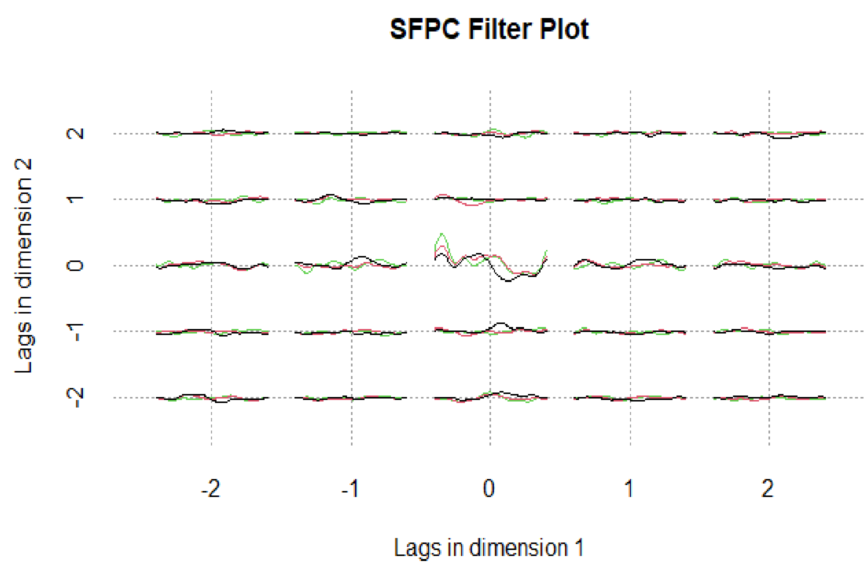

Figure 3: Evolution of three functional spatial filters of $TMP-2001$ variable over the lag.

Table 6: Cumulative percentage (CP) of explained variance of functional principal components obtained by SMFPCA and MFPCA considering the variables $TMP-2000$ and $TMP-2001$.

CP	PC1	PC2	PC3	PC4
SMFPCA	58.80	69.97	79.72	87.33
MFPCA	37.59	52.11	63.78	68.63

Table 7: NMSE and NMSE* results obtained by SMFPCA and MFPCA considering the variables $TMP-2000$ and $TMP-2001$.

Cumulative PC	PC1		PC2		PC3		PC4	
Spatial consideration	Spatial	Ordinary	Spatial	Ordinary	Spatial	Ordinary	Spatial	Ordinary
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

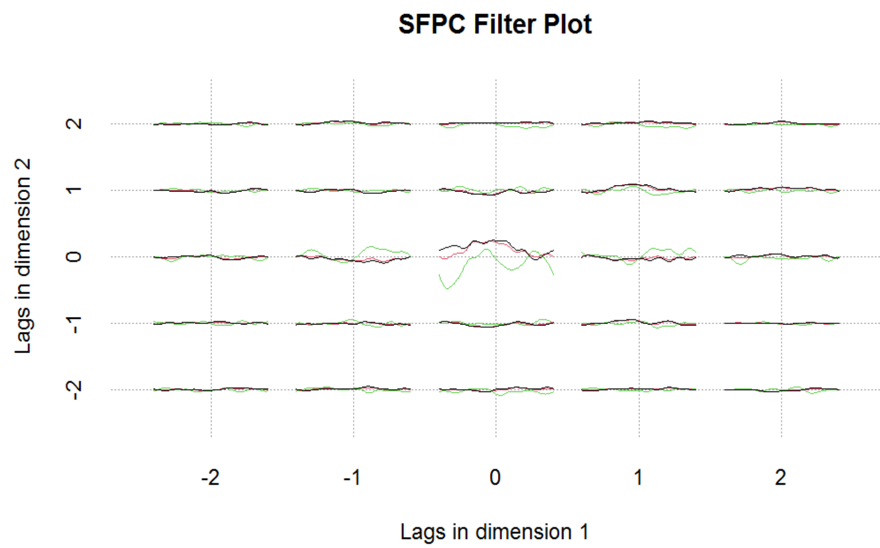

Figure 4: Evolution of three functional spatial filters of $TMP-2000$ variable over the lag after SMFPCA Application.

In setting 2, supplementary tests were conducted using sea surface temperature data to examine the performance of the SMFPCA algorithm 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 : $q = (19, 18)$

Then we proceed with describing steps in procedure 2.3 for both settings, and compute the first 4 multivariate principal components to capture more than 80% of variability.

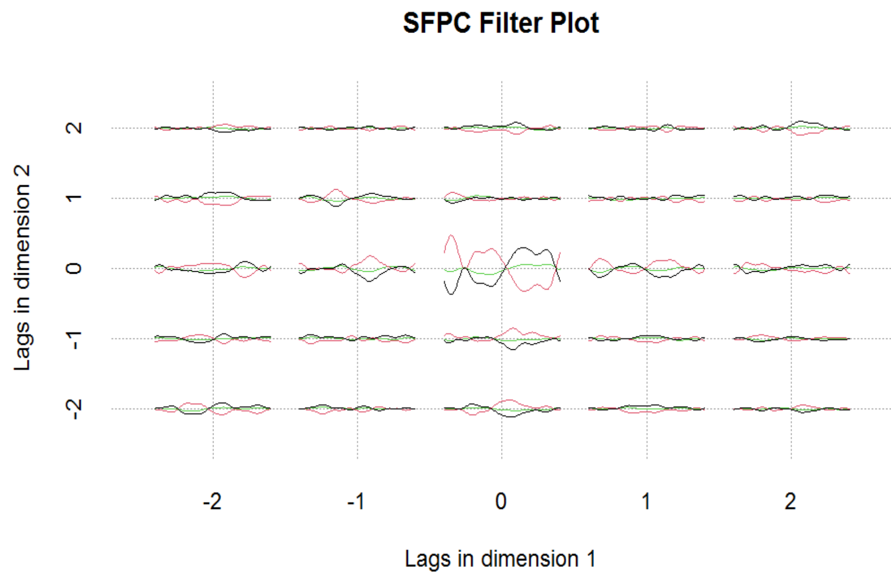

Figure 5: Evolution of three functional spatial filters of $TMP - 2001$ variable over the lag after SMFPCA Application.

After computing SMFPCA, we performed a comparative assessment of the reconstructed functional data with ordinary MFPCA, and the corresponding results are shown in Table 8. We find the same trend in results as in configuration 1, the NMSE and NMSE* results exhibit a decreasing pattern as the cumulative number of principal components increases. This test confirms favorable outcomes, even when considering three or more variables.

Dimensionality reduction of multivariate functional spatial data shown in Figures 4 and 5, is achieved by identifying the principal directions of variation among the functional variables. Consequently, the data is succinctly represented through a restricted number of FPCA, facilitating simplified analysis and interpretation of two or more functional spatial variables. Additionally, this approach enhances the visualization and comprehension of results by enabling the identification of spatial patterns, structures, and trends within the data.

Table 8: NMSE and NMSE* results obtained by SMFPCA and MFPCA considering the variables $TMP-1996$, $TMP-1998$, and $TMP-1999$.

Cumulative PC	PC1		PC2		PC3	
Spatial consideration	Spatial	Ordinary	Spatial	Ordinary	Spatial	Ordinary
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

4. CONCLUSION AND PERSPECTIVES

In this paper, we are interested to dimension reduction for multivariate spatial functional data. We have introduced a novel method known as SMFPCA (Spatial Multivariate Functional Principal Component Analysis). In the literature, the majority of existing methods are primarily based on Karhunen–Loève methodologies, however, there is an absence of methods for processing multivariate spatial data using spectral functional analysis. In this context, we have developed methodologies and mathematical properties for conducting spectral FPCA on multivariate functional spatial data sampled on a regular grid, both within the same domain or different domains.

We initially conducted tests on simulated variables, implementing our methodology without considering the spatial dimension. This allowed us to make a comparison with the results obtained from the data that had spatial indexing. We evaluated our approach by using the Normalized Mean Squared Error (NMSE), which involves reconstructing data based on the calculated scores and filters. Furthermore, we assessed the quality of the approximation without any boundary effects using the NMSE* measure. The results for NMSE and NMSE* presented in Tables 1, 2, 3 and 4, clearly demonstrating the substantial performance improvement achieved through the inclusion of the spatial aspect.

Similarly, we conducted supplementary tests on NOAA data, this data showed a low spatial dependency. We selected sea surface temperature variables for the years 2000 and 2001 as representative examples of the spatially indexed multivariate aspect. Furthermore, we performed supplementary tests on data from three NOAA variables taken from different years. Our findings in tables 7 and 8 mirror those obtained from simulated data, reinforcing that the incorporation of spatial considerations leads to notably improved NMSE and NMSE* outcomes. SMFPCA also enhances data visualization by revealing spatial patterns and trends. This approach unveils spatial-temporal patterns, and provides valuable insights.

As we look forward, it's worth considering the incorporation of additional functionality to handle irregular data. Such an enhancement would significantly broaden the scope of feasible analyses across diverse data structures.

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] T. Feuillet, V. Loonis, M. P. Bellefon, Manuel d'analyse spatiale. Théorie et mise en œuvre pratique avec R, no. 131 in Insee Méthodes, Insee, Eurostat, 2018.

[7] P. Delicado, R. Giraldo, C. Comas, J. Mateu, Statistics for spatial functional data: some recent contributions, Environmetrics 21 (3-4) (2010) 224–239.

[8] J. Aristizabal, R. Giraldo, J. Mateu, Analysis of variance for spatially correlated functional data: application to brain data, Spatial Statistics 32 (2019) 100381.

[9] O. Gromenko, P. Kokoszka, L. Zhu, J. Sojka, Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends, The Annals of Applied Statistics 6 (2) (2012) 669–696.

[10] O. Gromenko, P. Kokoszka, L. Zhu, J. Sojka, Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves, The Annals of Applied Statistics 11 (2017) 898–918.

[11] J. A. Aston, D. Pigoli, S. Tavakoli, Tests for separability in nonparametric covariance operators of random surfaces, The Annals of Statistics 45 (4) (2017) 1431–1461.

[12] P. Constantinou, P. Kokoszka, M. Reimherr, Testing separability of space-time functional processes, Biometrika 104 (2) (2017) 425–437.

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

[14] M. S. Bernardi, G. Mazza, J. O. Ramsay, L. M. Sangalli, A penalized regression model for spatial functional data with application to the analysis of the production of waste in venice province, Stochastic Environmental Research and Risk Assessment 31 (2017) 23–38.

[15] 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.

[16] T. Kuenzer, S. Hörmann, P. Kokoszka, Principal component analysis of spatially indexed functions, Journal of the American Statistical Association 116 (535) (2020) 1444–1456.

[17] J. M. Chiou, Y. T. Chen, Y. F. Yang, Multivariate functional principal component analysis: A normalization approach, Statistica Sinica XX (XX) (2014) 1571–1596.

[18] J. R. Berrendero, A. Justel, M. Svarc, Principal components for multivariate functional data, Computational Statistics Data Analysis 55 (9) (2011) 2619–2634.

[19] C. M. Happ, Statistical methods for data with different dimensions, 2017.

[20] 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.