**UNIVERSIDADE FEDERAL DE MINAS GERAIS**
**Instituto de Ciências Exatas**
**Programa de Pós-Graduação em Estatística**

Alvaro Alexander Burbano Moreno

# Functional data analysis: spatial association of curves and irregular spacing.

Belo Horizonte
2023

Alvaro Alexander Burbano Moreno

# Functional data analysis: spatial association of curves and irregular spacing.

Thesis presented to the Graduate Program in Statistics of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Statistics.

Advisor: Vinícius Diniz Mayrink

Belo Horizonte
2023

Burbano Moreno, Alvaro Alexander.

B946f Functional data analysis: spatial association of curves and irregular spacing [manuscrito] / Alvaro Alexander Burbano Moreno – 2023.
100 f. il.

Orientador: Vinícius Diniz Mayrink.

Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 77-81

1. Estatística – Teses. 2. Análise Espacial (Estatística) – Teses. 3. Inferência Bayesiana – Teses. 4. Processos Gaussianos – Teses. 5. Spline – Teses. 6. Polinômios de Bernstein. I. Mayrink, Vinícius Diniz. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

# UNIVERSIDADE FEDERAL DE MINAS GERAIS

**PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA**

UFMG

ATA DA DEFESA DE TESE DE DOUTORADO DO ALUNO ALVARO ALEXANDER BURBANO MORENO, MATRICULADO SOB O Nº 2019669581, NO PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA, DO INSTITUTO DE CIÊNCIAS EXATAS, DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, REALIZADA NO DIA 12 DE SETEMBRO DE 2023.

Aos 12 dias do mês de setembro de 2023, às 13h30, em reunião pública de número 83 (conforme orientações para a atividade de defesa de tese durante a vigência da Portaria PRPG nº 1819), do Programa de Pós-Graduação em Estatística, do Instituto de Ciências Exatas da UFMG, reuniram-se, de forma híbrida (presencial e virtual), na sala de seminários do LCC, os professores abaixo relacionados, formando a Comissão Examinadora homologada pelo Colegiado do Programa de Pós-Graduação em Estatística, para julgar a defesa de tese do aluno ALVARO ALEXANDER BURBANO MORENO, nº de matrícula 2019669581, intitulada: "*Functional data analysis: spatial association of curves and irregular spacing*", requisito final para obtenção do Grau de doutor em Estatística. Abrindo a sessão, o Senhor Presidente da Comissão, Prof. Vinícius Diniz Mayrink, passou a palavra ao aluno para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do aluno. Após a defesa, os membros da banca examinadora reuniram-se reservadamente sem a presença do aluno e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação:

(X) Aprovado.
( ) Reprovado com resubmissão do texto em ____ dias.
( ) Reprovado com resubmissão do texto e nova defesa em ____ dias.
( ) Reprovado.

Prof. Vinícius Diniz Mayrink
Orientador (EST/UFMG)

Prof. Flávio Bambirra Gonçalves
(EST/ UFMG)

Prof. Marcos Oliveira Prates
(EST/ UFMG)

Profa. Airlane Pereira Alencar
(IME/USP)

Prof. Ronaldo Dias
(IMECC/UNICAMP)

O resultado final foi comunicado publicamente ao aluno pelo Senhor Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 12 de setembro de 2023.

# UNIVERSIDADE FEDERAL DE MINAS GERAIS

## PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

UFMG

Observações:

1. No caso de aprovação da tese, a banca pode solicitar modificações a serem feitas na versão final do texto. Neste caso, o texto final deve ser aprovado pelo orientador da tese. O pedido de expedição do diploma do candidato fica condicionado à submissão e aprovação, pelo orientador, da versão final do texto.
2. No caso de reprovação da tese com resubmissão do texto, o candidato deve submeter o novo texto dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca que então decidirão pela aprovação ou reprovação da tese.
3. No caso de reprovação da tese com resubmissão do texto e nova defesa, o candidato deve submeter o novo texto com a antecedência à nova defesa que o orientador julgar adequada. A nova defesa, mediante todos os membros da banca, deve ser realizada dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca. Baseada no novo texto e na nova defesa, a banca decidirá pela aprovação ou reprovação da tese.

*Dedico este trabalho à minha mãe e em memória ao meu irmão.*

# Acknowledgments

Gostaria de externar meus sinceros agradecimentos a todos aqueles que, direta ou indiretamente, contribuíram para a conclusão do meu doutoramento. Em primeiro lugar, agradeço à minha mãe, Carmen Elisa, por seu amor e por ter me apoiado com palavras de força e alento durante todo o percurso do doutorado. Ao lado dela, agradeço ao meu irmão Jhonny Maurício (in memoriam) por ser um exemplo de motivação e perseverança, e pelo amor dedicado a mim.

Ao meu orientador Vinícius Mayrink, por ter aceitado solicitamente me orientar neste doutorado, me ajudando a formar minhas bases de pesquisa e a ter amor por minha nova área de trabalho, a Estatística. Agradeço pela riqueza de conhecimentos que me proporcionou tanto em Estatística bayesiana como Estatística espacial, pela compreensão em todos os momentos de dificuldade e pela dedicação em sua orientação.

Aos professores Flávio Gonçalves, Marcos Prates, Airlane Alencar e Ronaldo Dias, por terem aceitado participar da minha banca de avaliação, contribuindo com seus conhecimentos e sugestões.

Aos meus colegas Gisele Maia, Cássius Henrique, Alisson Silva e Caio Gabriel pelo companheirismo, pelas conversas agradáveis e pelo incentivo que me dedicaram durante o período do curso de doutorado.

Agradeço à CAPES pela bolsa de estudos, que me permitiu custear minha estadia no Brasil durante todo o período do doutorado. Ademais, agradeço à FAPEMIG e ao CNPq por contribuírem com verbas para equipar os laboratórios de estatística da UFMG, sem os quais não seria possível desenvolver minha tese.

A Universidade Federal de Minas Gerais, por me proporcionar a oportunidade de aprender uma gama de conhecimentos mediados pelos competentes professores do Programa de Pos-Graduacao em Estatística durante o curso de doutoramento e, principalmente, por me conceder o título de doutor em Estatística.

"*Make your life a dream and your dream a reality.*"
(The Little Prince)

# Resumo

A análise de dados funcionais espaciais (SFD) é um área da estatística emergente que combina a análise de dados funcionais (FDA) e a modelagem de dependência espacial. Diferentemente dos métodos estatísticos tradicionais que tratam os dados como valores escalares ou vetores, a SFD considera os dados como funções contínuas, permitindo uma compreensão mais completa de seu comportamento e variabilidade. Essa abordagem é adequada para analisar dados coletados ao longo do tempo, do espaço ou de qualquer outro domínio contínuo. A SFD é aplicada em vários campos, incluindo economia, finanças, medicina, ciências ambientais e engenharia. Esta tese propõe novos modelos funcionais Gaussianos que incorporam estruturas de dependência espacial, com foco em dados tendo espaçamento irregular e que refletem curvas espacialmente correlacionadas. Os modelos são baseados em expansões de base B-spline e Polinômios de Bernstein (BP) e utilizam uma abordagem Bayesiana para estimar quantidades e parâmetros desconhecidos. A tese explora as vantagens e limitações dos modelos baseados em B-spline e BP na captura de formas e padrões complexos, garantindo a estabilidade numérica. As principais contribuições deste trabalho incluem o desenvolvimento de um modelo inovador voltado para SFD usando estruturas B-spline ou BP, incluindo um efeito aleatório para tratar de associações entre observações com espaçamento irregular, e um estudo de simulação abrangente para avaliar o desempenho dos modelos em vários cenários. A tese também apresenta duas aplicações reais relacionadas aos níveis de PM10 e Temperatura na Cidade do México, demonstrando ilustrações práticas dos modelos propostos.

**Palavras-chave:** B-spline, Polinômios de Bernstein, Inferência Bayesiana, Processo Gaussiano, MCMC.

# Abstract

Spatial Functional Data (SFD) analysis is an emerging statistical framework that combines Functional Data Analysis (FDA) and spatial dependency modeling. Unlike traditional statistical methods, which treat data as scalar values or vectors, SFD considers data as continuous functions, allowing for a more comprehensive understanding of their behavior and variability. This approach is well-suited for analyzing data collected over time, space, or any other continuous domain. SFD has found applications in various fields, including economics, finance, medicine, environmental science, and engineering. This thesis proposes new functional Gaussian models incorporating spatial dependence structures, focusing on irregularly spaced data and reflecting spatially correlated curves. The models are based on B-spline basis expansions and Bernstein Polynomials (BP) and utilize a Bayesian approach for estimating unknown quantities and parameters. The thesis explores the advantages and limitations of B-spline-based and BP-based models in capturing complex shapes and patterns while ensuring numerical stability. The main contributions of this work include the development of an innovative model designed for SFD using B-spline or BP structures, including a random effect to address associations between irregularly spaced observations, and a comprehensive simulation study to evaluate models' performance under various scenarios. The thesis also presents two real applications related to levels of PM10 and Temperature in Mexico City, showcasing practical illustrations of the proposed models.

**Keywords:** B-spline, Bernstein polynomials, Bayesian inference, Gaussian Process, MCMC.

# List of Figures

# List of Tables

# Contents

# Chapter 1

# Introduction

Functional Data Analysis (FDA) (Ramsay and Silverman, 2002) is a robust statistical framework that has gained significant attention recently for its ability to analyze and interpret data that vary continuously over a given interval. Traditional statistical methods often assume that observations are represented by scalar values or vectors, which may not adequately capture many real-world datasets with complex and intricate nature. In contrast, FDA treats data as functions, allowing for a more comprehensive understanding of their behavior and variability. The fundamental idea behind FDA is to consider each observation as a curve or function rather than a single data point. This approach is well-suited for analyzing data collected over time, spatial dimensions, or any other continuous domain. Examples of such data are in various fields, including economics, finance, medicine, environmental science, and engineering. For a general introduction to FDA, the reader is referred to Ramsay and Silverman (2005), Ferraty and Vieu (2006), and Kokoszka and Reimherr (2017).

On the other hand, neighborhood dependency is a fundamental concept in analyzing spatial data. It recognizes that observations within a spatial context are often interrelated, and the values observed at neighboring locations tend to be more similar than those farther apart (Cressie, 1993; Banerjee et al., 2014). Incorporating spatial dependency into statistical models is crucial for accurate inference and prediction, particularly in environmental science, geostatistics (Diggle and Ribeiro, 2007), and epidemiology (Lawson, 2018, 2021).

The combination of FDA and spatial dependency modeling has given birth to an intriguing and potent approach called Spatial Functional Data (SFD) analysis. Unlike traditional FDA techniques, which treat observations solely as functions without considering their spatial locations, SFD considers spatial location an additional dimension. This distinction is crucial as it recognizes the significance of spatial relationships in comprehending and modeling underlying processes. By integrating spatial information, SFD enables researchers to explore spatial dependence and heterogeneity in functional data, leading to a more profound understanding of spatial processes.

Numerous notable works demonstrate the extensive and diverse literature on SDF analysis. For example, in their study on Oceanography, Nerini et al. (2010) proposed a spatial functional linear model. Zhou et al. (2010) introduced mixed effects models

for hierarchical functional data with spatial correlation. Giraldo et al. (2012) developed a methodology for clustering spatially correlated functional data; see also, Jiang and Serban (2012); Romano et al. (2017) for additional information. Staicu et al. (2010) proposed a methodology for functional models with a hierarchical structure where the functions at the lowest hierarchy level are spatially correlated. Delicado et al. (2010) and Mateu and Romano (2017) have provided surveys of SFD, while Martínez-Hernández and Genton (2020) presented a review of complex and spatially dependent data. Moreover, Mateu and Giraldo (2021) explored the intersection between geostatistics and functional data analysis, featuring contributions from leading experts in the field.

Some references to SFD with a Bayesian perspective are Baladandayuthapani et al. (2008) employed a Bayesian semi-parametric method using regression splines to handle general between-curve covariance structures. Another study by Zhang et al. (2016) introduced a functional conditional autoregressive (CAR) model for spatially correlated data. Furthermore, Song and Mallick (2019) proposed novel models based on wavelets for spatially correlated functional data. These models enable the regularization of curves observed over space and the prediction of curves at unobserved sites. Finally, Rekabdarkolaee et al. (2019) introduced a novel multivariate space-time functional model with spatially varying coefficients.

In both the classical and Bayesian approaches, it is common to use nonparametric methods to smooth functional data (Hollander et al., 2013). The term "nonparametric" means that specific forms for the underlying functions describing the data are not assumed. This approach allows for greater flexibility and the ability to capture complex patterns. Some commonly used techniques in research for this purpose include Kernels (Wand and Jones, 1994), Splines (De Boor, 2001), and Wavelets (Vidakovic, 2009).

The Kernels approach relies on kernel functions (probability density functions). The shape of these functions determines the influence or weight assigned to each point neighboring the point of interest during the smoothing process. However, Ramsay and Silverman (2002) and Wand and Jones (1994) pointed out some limitations of Kernels, such as the inability to characterize local features adequately. That is, they are not sensitive enough to detect abrupt changes or specific details that may be present in certain parts of the data. On the other hand, splines are polynomial functions defined in different intervals (segments) that combine smoothly at junction points called knots; see details in De Boor (2001). Splines offer a flexible approximation of observed functions and allow smoothness control by choosing the degree of the polynomial and the number of segments. However, it is essential to be careful when selecting the number and position of knots, as this can be subjective and affect the quality of the fit. Finally, the Wavelet smoothing method takes advantage of wavelet transforms to represent the original function at different scales and frequencies, which facilitates noise filtering and function approximation at different levels of detail (Vidakovic, 2009). Wavelet smoothing methods are beneficial

for capturing local and global features of the function. However, it is critical to choose the wavelet and the level of decomposition correctly. Additionally, it is worth noting that the number of observations must be a power of 2, and the spacing between them must be equidistant for this method to work correctly.

The main objective of this thesis is to propose new functional Gaussian models that incorporate a spatial dependence structure. In this context, the study is focused on irregularly spaced data reflecting spatially correlated curves. The developed models are based on fundamental mathematical tools, such as B-spline basis expansions (as described De Boor, 2001) and Bernstein Polynomials (BP) (as discussed in Lorentz, 2012; Farouki and Rajan, 1987, 1988). These methods offer flexibility to capture complex shapes and patterns while ensuring numerical stability. As for the process of estimating the unknown parameters of the models, a Bayesian approach is adopted. This decision is based on its ability to incorporate prior information and consider uncertainty about unknown quantities in the analysis. One key feature of the modeling structures is their capability to integrate the association generated by the irregularly spaced discretely observed measurements in each function.

B-spline-based models offer a more straightforward and robust approach (Aguilera and Aguilera-Morillo, 2013) than other complex smoothing methods (Morris et al., 2003). B-splines are represented by combining local basis functions, allowing each polynomial segment to have a limited scope and only influence a local portion of the curve. The control points, known as knots, do not directly manipulate the curves. Instead, they affect the shape of the curve through their impact on the polynomial segments within their support (Piegl and Tiller, 1996). While B-spline models have been extensively researched in the context of SFD, for example, Giraldo et al. (2010), Giraldo et al. (2011), Giraldo et al. (2012), Giraldo et al. (2012), Cortés-D et al. (2016), Aguilera-Morillo et al. (2017) and Aristizabal et al. (2019), no previous studies have focused explicitly on non-equidistant sample designs. On the other hand, BP is appealing for modeling and analyzing functional data due to its flexibility in approximating continuous functions. These polynomials are applied in various applications, such as density estimation (Tenbusch, 1994; Petrone, 1999; Wang and Guan, 2019) and curve approximation using linear combinations of Bernstein functions (Liang et al., 2022). The BP basis function is determined by its degree and bounded domain, making it easier to apply and compute than other methods. It can also capture complex shapes while ensuring good numerical behavior. However, their application in SFD for equidistant or irregular sample designs still needs to be explored.

Imposing spatial dependence through an association between coefficients defining a linear combination of variables is common in the literature. This idea will be applied in this work but contextualized for basis expansions that structure the functional modeling. In the context of spatial regression, it is possible to mention some studies as examples of the application of a modeling approach that spatially correlates coefficients. Take a look

at the following studies, Gelfand et al. (2003) build spatial modeling with spatially varying coefficient processes. Reich et al. (2011) developed a Bayesian spatial model to predict ozone under different meteorological conditions, and Fan and Huang (2022) presented spatially varying coefficient models using reduced-rank thin-plate splines.

The main contributions of the present work are as follows:

- This proposal presents an innovative data management model explicitly designed for SFD. The model integrates the B-spline or BP structure to model the mean; while establishing a spatial dependence structure in the coefficients of the bases. Notably, the use of the BP represents a significant and previously unexplored contribution in the literature concerning the analysis of SFD.

- Inclusion of a random effect to address the associations between irregularly spaced observations from the target functions. This effect exhibits an autoregressive structure, presenting a novel approach to analyzing SFD.

- Development of a comprehensive simulation study to evaluate the performance of the models under different scenarios. The analysis is based on artificial data from replicas in a Monte Carlo (MC) scheme. The presence of missing data is another topic deserving attention in the study. Notably, the ability of the models to predict missing values showcases another noteworthy contribution. In the literature review conducted for this thesis, no studies were found that dealt with missing data using the structure of the proposed model itself. The widely used strategy is imputing missing observations before fitting the model.

- Exploration of two real applications related to levels of PM10 and Temperature in Mexico City, using a data set covering the period between 2021 and 2022. Unlike other studies in the literature that focus on different periods or conduct spatial analysis in a functional regression setting with PM10 as a covariate explaining Temperature, this work examines PM10 and Temperature separately within a spatial functional model, accounting for irregularly spaced observations and utilizing B-spline or BP structures. This represents another significant contribution to the field of Statistics.

This work is organized as follows: In **Chapter** 2, some basic concepts fundamental to the thesis development are discussed and explained. It will introduce the FDA, a powerful and versatile technique for studying data presented as continuous functions instead of point observations. In addition, it will delve into two essential mathematical tools for functional analysis: B-spline basis functions and BP. These functions play a crucial role in the representation and approximation of curves and surfaces, and their understanding is vital to building adequate models in the study. Finally, discrepancy measures will be

addressed to evaluate the similarity or distance between functions (estimated and target curves). **Chapter** 3 presents the proposed models that represent the central core of the thesis. These models are specifically designed to address the presence of spatial association in functional data, implying that spatial dependence patterns may influence the observed functions. Incorporating this association is crucial for better understanding functional data in contexts with irregular sample designs. **Chapter** 4, a rigorous simulation study will be carried out to evaluate the performance of the proposed models under different scenarios and conditions. Simulation is an essential tool to validate and calibrate the models, allowing us to clearly understand how they behave against simulated data with controlled characteristics.

Also, the procedures and criteria used to generate the simulated data will be described, and the evaluation metrics used to compare and contrast the results obtained. The findings of this simulation study will be crucial to support the appropriate choice of models in real situations. **Chapter** 5, the proposed models are applied to two real data sets. The data are carefully selected to represent different contexts and real-world issues where FDA with spatial association and irregular samples is highly relevant. **Chapter** 6 serves as the final section, presenting the general conclusions drawn from the thesis work. This chapter summarizes the most significant findings and emphasizes the contributions made to SFD Analysis. Moreover, it outlines potential avenues for future research and methodology development, creating exciting prospects for enhancing and applying the proposed models to diverse types of functional data and specific contexts.

# Chapter 2

# Background Functional Data Analysis

One of the purposes of this work is based on the proper treatment of functional data. Therefore, this chapter aims to introduce the related basic concepts. Section 2.1 presents the general definition of functional data and how they can be represented mathematically, for example, through linear combinations of known basis functions. Section 2.2 defines the B-splines via the fundamental recurrence relation discovered by De Boor (1972) and Cox (1972). Section 2.3 introduces Bernstein polynomials as a mathematical tool for representing functional curves. Finally, Section 2.4 indicates discrepancy measurements considered in the analysis.

## 2.1 Functional Data Representation

According to Ramsay and Silverman (2005), the functional data are usually discreetly observed, although its inherent structure is functional. The observation of a function is determined by a set of $N_j$ pairs $(t_{jn}, x_{jn})$, for $n = 1, \ldots, N_j$, where $t_{jn}$ denotes the argument, and $x_{jn}$ the observed functional value. In general, the construction of the original functions starting from the observed data can occur separately or independently for each function with the identifier $j$. The smoothness of a function is related to the number of continuous derivatives it has; this assumption is usually necessary for applying some multivariate analysis techniques.

Typically, the first step in working with functional data is to express them through a basis expansion as follows:

$$x_j(t) = \sum_{r=1}^{\infty} \theta_{jr} f_r(t) \approx \sum_{r=1}^{\boldsymbol{K}} \theta_{jr} f_r(t), \quad j = 1, \ldots, m, \tag{2.1}$$

where $\theta_{jr}$, $r = 1, \ldots, \boldsymbol{K}$ are the coefficients of expansion, and $\{f_r(t)\}_{r=1}^{\infty}$ is a set of known basis functions that are linearly independent. An approximate representation is usually obtained by truncating this basis expansion in terms of a number $\boldsymbol{K}$ (positive integers) of basis functions large enough to represent each curve accurately. This method allows enough flexibility while providing computational efficiency. Furthermore, the dimension

of the data only depends on the number of curves and the order $k$ of the expansion. The choice of the number $\boldsymbol{K}$ basis functions is fundamental and decisive for all subsequent calculations. Small numbers of basis functions mean little flexibility and large numbers result in flexibility but may induce overfitting.

Since the curves are observed with error, some smoothing technique is usually performed to estimate the coefficients. Concerning the choice of a suitable basis, there are multiple options depending on the characteristics of the sample curves, but the most common are Fourier, B-spline, or Wavelet basis (Ramsay and Silverman, 2002; Vidakovic, 2009). In this Thesis, the study is focused on B-spline and BP. These choices were motivated by the fact that the B-spline is widely used in the literature, and the BP is a central novelty in the analysis.

## 2.2 B-splines Basis

A B-spline function can be defined as a piecewise polynomial that is joined by interior knots $\xi_2, \xi_3, \ldots, \xi_l$ so that the function is continuous, and has a certain number of continuous derivatives. An essential feature of B-splines is to simultaneously preserve the flexibility of piecewise polynomials and achieve a certain degree of overall smoothness, with its order of approximation not depending on the degree of the polynomial (Schumaker, 2007).

**Definition 2.2.1** (Piecewise Polynomials)**.** *Let $a = \xi_1 < \xi_2 <, \ldots, < \xi_l < \xi_{l+1} = b$ and $\Delta = \{\xi_1, \xi_2, \ldots, \xi_l, \xi_{l+1}\}$. The set $\Delta$ partitions the interval $[a, b]$ into $l$ subintervals, $I_i = [\,\xi_i, \xi_{i+1}\,)$, for $i = 1, \ldots, l-1$, and $I_l = [\xi_l, \xi_{l+1}]$. Then the corresponding piecewise polynomial function $g(t)$ of order $k$ is defined by*

$$g(t) = p_i(t), \;\; \textit{for}\, t \in I_i, \;\; i =, 1, \ldots, l, \tag{2.2}$$

*where, for $i = 1, \cdots, l$, each function $p_i(t)$ is a polynomial defined in the interval $I_i$.*

The concept of divided differences presented in De Boor (2001) is introduced to define B-splines.

**Definition 2.2.2** (Divided Differences)**.** *The $n$-th divided difference of a function $g$ at the points $\xi_i, \ldots, \xi_{i+n}$ is the leading coefficient (i.e. the coefficient of $t^n$) of the unique polynomial $p_n$ of order $n+1$ which satisfies $p_n(\xi_{i^*}) = g(\xi_{i^*})$, $i^* = i, \ldots, i+n$. It is denoted by $[\xi_i, \ldots, \xi_{i+n}]g$.*

This definition has the following immediate consequences. If $p_i$ agrees with $g$ at $\xi_1, \ldots, \xi_i$ for $i = k$ and $i = k+1$, then

$$p_{k+1}(t) = p_k(t) + (t - \xi_1) \cdots (t - \xi_k)[\xi_1, \ldots, \xi_{k+1}]g,$$

therefore $p_{k+1}(t) - p_k(t)$ is a polynomial of order $k+1$.

**Definition 2.2.3** (Normalized B-spline)**.** *Let $\boldsymbol{\xi} = \{\xi_i\}$ be a nondecreasing sequence, which may be finite or infinite. The $i$-th normalized B-spline of order $k$ for the knot sequence $\boldsymbol{\xi}$ is defined by*

$$B_{i,k}(t) = (\xi_{i+k} - \xi_i)[\xi_i, \ldots, \xi_{i+k}](\bullet - t)_+^{k-1}, \quad \forall\, t \in \mathbb{R}, \tag{2.3}$$

*such that $[\xi_i, \ldots, \xi_{i+k}](\bullet - t)_+^{k-1}$ represents the $k$-th divided difference of the function $(\bullet - t)_+^{k-1}$.*

In Equation (2.3), The expression $(\xi_{i+k} - \xi_i)$ is a normalization factor designed to produce the identity $\sum_i B_{i,k}(t) = 1$, which asserts that the B-splines form a partition of unity. Initially, the B-splines have been defined as a divided difference of the truncated power base (Schoenberg, 1946), and later by the fundamental recurrence relation discovered by De Boor (1972). The recurrence formula is used as it is the most useful for computational implementation.

**Definition 2.2.4** (Recurrence Relation)**.** *Let $\Delta = \{\xi_1, \ldots, \xi_{l+1}\}$ be a nondecreasing sequence of real numbers. The $i$-th B-spline basis function of order $k$ (degree $k-1$), is defined by*

$$B_{i,k}(t) = \frac{t - \xi_i}{\xi_{i+k-1} - \xi_i} B_{i,k-1}(t) + \frac{\xi_{i+k} - t}{\xi_{i+k} - \xi_{i+1}} B_{i+1,k-1}(t), \tag{2.4}$$

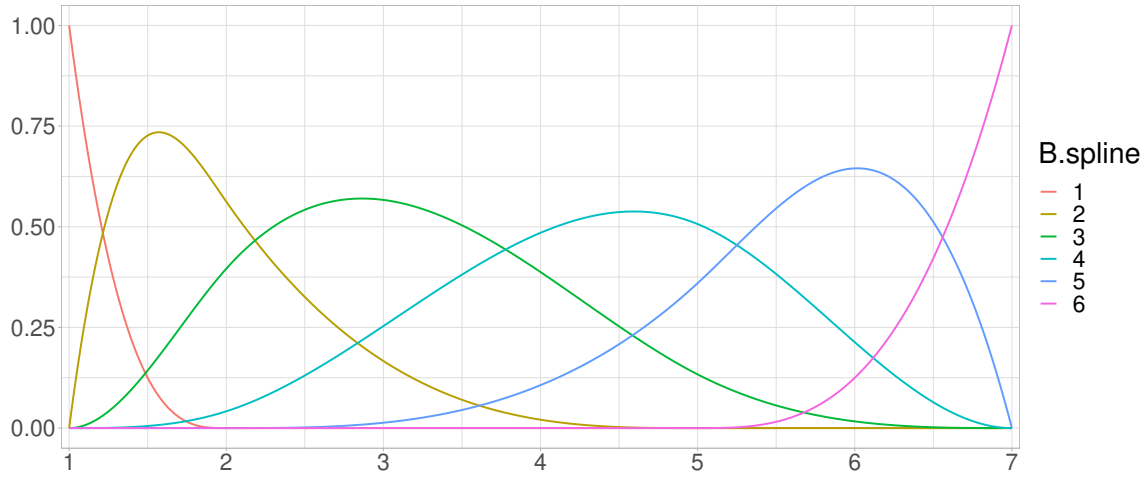
*where*

$$B_{i,1}(t) = \begin{cases} 1 & \text{if } \xi_i \leq t < \xi_{i+1}, \\ 0 & \text{otherwise.} \end{cases}$$

In Equation (2.4), quotients of the form $0/0$ are defined as zero. The $B_{i,k}(t)$ are piecewise polynomials, defined on the entire real line, generally only the interval $[\xi_l, \xi_{l+1}]$ is of interest. Computation of a set of basis functions requires the specification of a knot vector $\Delta$ and the order $k$. The choice of the knot vector influences the shape of the functions defined by the recurrence relation.

The B-spline basis system has a property that is often useful: the sum of the B-spline basis function values at any point $t$ is equal to one. For example, in Figure 2.1, the first and last basis functions are exactly one at the boundaries of the graph domain. This is because all the other basis functions go to zero at these endpoints; for more information, see De Boor (2001) and Ramsay et al. (2009).

Figure 2.1: Cubic B-spline basis functions defined by the knot vector $\Delta = \{1, 2, 5, 7\}$




Source: Prepared by the author

The $B_{i,k}(t)$ functions have the following important properties:

- Non-negativity, $B_{i,k}(t) \geq 0$ for all values of $i$, $k$ and $t$.
- Compact support, that is, it is non-null in a small interval and zero outside this interval.
- Differentiability, meaning that all derivatives of $B_{i,k}(t)$ exist within $[\xi_i, \xi_{i+1})$, for $i = 1, \ldots, l$.
- Partition of unity, $\sum_i B_{i,k}(t) = 1$ for $t \in [a, b]$.

Two aspects determine the approximation of a function by a B-spline curve: The order of the polynomial segments and the knot sequence $\Delta$. The number of parameters required to define a B-spline function is the order plus the number of interior knots, $k + l - 1$. Thus, any function can be expressed as

$$g(t) \approx g_{k,\Delta}(t) = \sum_{r=1}^{\boldsymbol{K}=k+l-1} \theta_r B_{r,k}(t), \tag{2.5}$$

where $B_{r,k}$ is the $r$-th B-spline basis function of order $k$ and $\theta_r$ is the corresponding coefficient.

## 2.3 Bernstein Polynomial Basis

Polynomials are an attractive class of functions for various scientific and engineering computations. They are concisely represented by coefficients on a suitable basis and

are amenable to efficient evaluation by simple algorithms. The set of polynomials is closed under the arithmetic operations of addition, subtraction, multiplication, differentiation, integration, and composition.

The approximative capabilities of polynomials are also of great practical interest in applications. Perhaps the most fundamental result in this context is the theorem of Weierstrass, which is stated below (Davis, 1975):

**Theorem 1.** *Let $g$ be a real and continuous function defined on a compact interval $[a, b]$. Then for each $\varepsilon > 0$ there exists a polynomial $p$ (which depends on $\varepsilon$) such that*

$$|g(t) - p(t)| < \varepsilon \quad \textit{for each } t \textit{ of } [a, b]. \tag{2.6}$$

In other words, it is possible to uniformly approximate any continuous function $g$, defined on a polynomial's closed interval $[a, b]$. An elegant constructive proof of this theorem was published in 1912, in which Bernstein's polynomial basis was first introduced; for more details, see Bernstein (1912) and Lorentz (2012).

**Definition 2.3.1** (Bernstein basis functions)**.** *Let $p$ denote any non-negative integer, and suppose $[a, b]$ is a bounded interval in $\mathbb{R}$. The polynomials*

$$b_{r,p}(t) = \binom{p}{r} \left(\frac{t-a}{b-a}\right)^r \left(1 - \frac{t-a}{b-a}\right)^{p-r}, \quad \textit{for} \quad r = 0, \ldots, p, \tag{2.7}$$

*are called the Bernstein polynomials of degree $p$ ($order\, p + 1$) with respect to the interval $[a, b]$.*

**Remark 2.3.1.** *The domain of the Bernstein basis polynomials can be defined on the interval $[0, 1]$ without loss of generality, replacing*

$$x = \frac{t-a}{b-a}, \quad a \le t \le b, \tag{2.8}$$

*or equivalently,*

$$t = (b-a)x + a, \quad 0 \le x \le 1. \tag{2.9}$$

*By using (2.8), and (2.9), it is observed from (2.7) that*

$$b_{r,p}(t) = \binom{p}{r} x^r (1-x)^{p-r}, \quad \textit{for} \quad r = 0, \ldots, p. \tag{2.10}$$

**Remark 2.3.2.** *For any non-negative integer $p$, and bounded interval $[a, b] \subset \mathbb{R}$, the corresponding Bernstein polynomials, as defined by (2.7), satisfy:*

*a. Recursive generation. The basis of degree p may be generated from the basis of degree* $p-1$

$$\begin{aligned} b_{r,p}(t) &= \binom{p}{r}\left(\frac{t-a}{b-a}\right)^r\left(1-\frac{t-a}{b-a}\right)^{p-r} \\ &= \left[\binom{p-1}{r}+\binom{p-1}{r-1}\right]\left(\frac{t-a}{b-a}\right)^r\left(1-\frac{t-a}{b-a}\right)^{p-r} \\ &= \binom{p-1}{r}\left(\frac{t-a}{b-a}\right)^r\left(1-\frac{t-a}{b-a}\right)^{p-r}+\binom{p-1}{r-1}\left(\frac{t-a}{b-a}\right)^r\left(1-\frac{t-a}{b-a}\right)^{p-r} \\ &= \left(1-\frac{t-a}{b-a}\right)\left[\binom{p-1}{r}\left(\frac{t-a}{b-a}\right)^r\left(1-\frac{t-a}{b-a}\right)^{p-r-1}\right]+ \\ &\qquad \frac{t-a}{b-a}\left[\binom{p-1}{r-1}\left(\frac{t-a}{b-a}\right)^{r-1}\left(1-\frac{t-a}{b-a}\right)^{p-r}\right] \\ &= \left(1-\frac{t-a}{b-a}\right)b_{r,p-1}(t)+\left(\frac{t-a}{b-a}\right)b_{r-1,p-1}(t). \end{aligned} \tag{2.11}$$

*b.*

$$b_{r,p}(a)=\begin{cases}1 & \textit{if } r=0,\\ 0 & \textit{otherwise},\end{cases} \qquad \textit{and} \qquad b_{r,p}(b)=\begin{cases}1 & \textit{if } r=p,\\ 0 & \textit{otherwise.}\end{cases} \tag{2.12}$$

*c. The positivity and partition of unity properties on* $[a,b]$

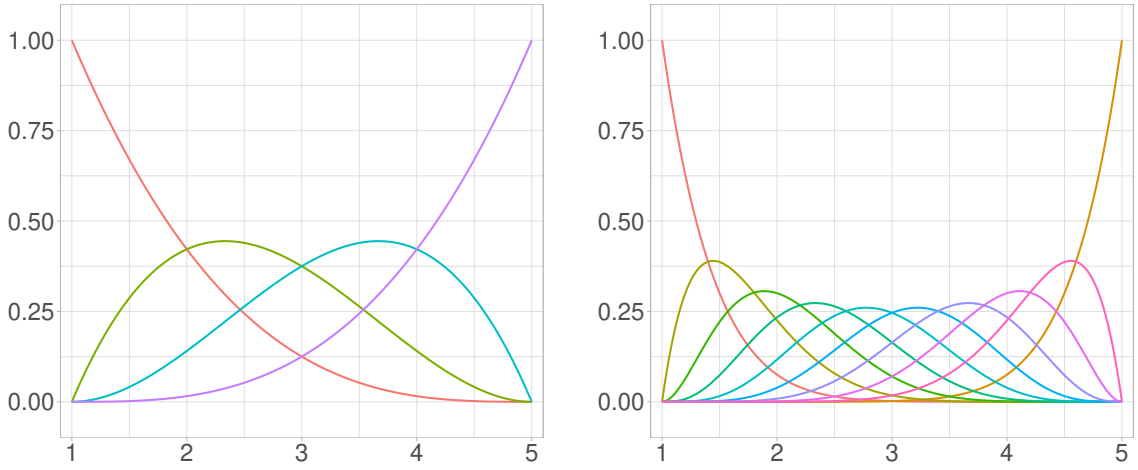
$$b_{r,p}(t)\geq 0, \quad r=0,\ldots,p \quad \textit{and} \quad \sum_{r=0}^{p} b_{r,p}(t)=1. \tag{2.13}$$

*d. Let* $\pi_p$ *be a finite-dimensional linear space, such that* $dim(\pi_p)=p+1$*. Then, the polynomial sequence* $\{b_{r,p}(t),\ r=0,\ldots,p\}$ *is a basis for* $\pi_p$*.*

An illustration of the Bernstein basis can be seen in Figure 2.2. The vector of basis $\mathbf{b}_p(t)=(b_{0,p}(t),\ldots,b_{p,p}(t))$ has a weight role, which varies with $t$, as shown in both panels of Figure 2.2. Thus, the approximation of the target function $g(\cdot)$ is weighted by $p+1$ values coming from the basis vector. It is clear that when $p=9$, more information is available to weigh the function values, resulting in a more accurate approximation.

Figure 2.2: Illustration of Bernstein basis for $p = 3$ and $p = 9$.

(a) $p = 3$ (order 4) (b) $p = 9$ (order 10)




Source: Prepared by the author

The Bernstein polynomial approximation of degree $p$ to any continuous function $g(t)$ on the interval $[a, b]$ is defined by sampling that function at the $p + 1$ equidistant positions $a + \frac{r}{p}(b - a)$, for $r = 0, 1, \ldots, p$, and blending the sampled values with the Bernstein basis functions

$$BP_p(g(t)) = \sum_{r=0}^{p} g\left(a + \frac{r}{p}(b - a)\right) b_{r,p}(t), \tag{2.14}$$

where the $\{b_{r,p}(t)\}$ are the basis functions defined in the Equation (2.7). Generally, in applications, the function that models a data set is not known explicitly, so it is necessary to replace the values of $g\left(a + \frac{r}{p}(b - a)\right)$ with freely specified coefficients $\theta_r$, for $r = 0, \ldots, p$, that can be used to manipulate the behavior of the polynomial intuitively

$$BP_p(t) = \sum_{r=0}^{p} \theta_r b_{r,p}(t), \quad t \in [a, b]. \tag{2.15}$$

The expression (2.14) is called a Bernstein polynomial, while (2.15) is called a polynomial in Bernstein form. Whereas the former refers to a polynomial approximation of a given function $g(t)$, the latter denotes a polynomial with arbitrary coefficients in the Bernstein basis (Farouki and Rajan, 1987, 1988).

## 2.4 Measures of Discrepancy

The statistical characteristics of an estimator for a function at a specific point share similarities with the conventional statistical attributes of an estimator for a single scalar

parameter. The statistical properties involve expectations and other aspects of random variables. In the following, when reference is made to an expectation (denoted as $\mathbb{E}(\cdot)$) or a variance ($\mathbb{V}(\cdot)$), it is important to note that these values are computed concerning the (unknown) distribution of the underlying random variable. In practice, the empirical distribution is considered to calculate these quantities.

Analyzing the performance of the function estimator requires the specification of appropriate criteria to measure the error. When considering estimation at a single point $t$, a natural measure is the mean square error (MSE), defined by

$$\mathrm{MSE}\left(\widehat{g}(t)\right) = \mathbb{E}\left(\widehat{g}(t) - g(t)\right)^2. \tag{2.16}$$

By elementary properties of mean and variance,

$$\mathrm{MSE}\left(\widehat{g}(t)\right) = \left(\mathbb{E}\left(\widehat{g}(t)\right) - g(t)\right)^2 + \mathbb{V}\left(\widehat{g}(t)\right). \tag{2.17}$$

This error criterion is often preferred to other criteria, such as mean absolute error (MAE), defined by

$$\mathrm{MAE}\left(\widehat{g}(t)\right) = \mathbb{E}\left|\widehat{g}(t) - g(t)\right|. \tag{2.18}$$

It is more difficult to do a mathematical analysis of the MAE than for the MSE. In addition, it does not have a simple decomposition into other meaningful quantities like the MSE, see Equation (2.17).

Instead of simply estimating $g(t)$ at a fixed point, estimating the function over the entire real line is often desirable, especially from a data analysis viewpoint. In this case, the estimate is the function $\widehat{g}(t)$, so it is necessary to consider an error criterion that globally measures the distance between the functions $\widehat{g}(t)$ and $g(t)$. Generally, these criteria can be defined as a norm of the function.

The $L^p$ norm (Gentle, 2009; Wand and Jones, 1994) of the error is

$$\left(\int_T \left|\widehat{g}(t) - g(t)\right|^p dt\right)^{1/p}, \tag{2.19}$$

where $T$ is the domain of $Y$ (true function). The estimator $\widehat{g}(t)$ must also be defined over the same domain. The integral may not exist.

Two useful measures are the $L^1$ norm, also called the integrated absolute error (IAE),

$$\mathrm{IAE}\left(\widehat{g}(t)\right) = \int_T \left|\widehat{g}(t) - g(t)\right| dt, \tag{2.20}$$

and the $L^2$, also called the integrated squared error (ISE),

$$\mathrm{ISE}\left(\widehat{g}(t)\right) = \int_T \left(\widehat{g}(t) - g(t)\right)^2 dt. \tag{2.21}$$

The $L^1$ measure is invariant under monotone transformations of the coordinate axes. However, the measure based on the $L^2$ norm is not[1]. Another natural way to compare estimators of functions is to use the expected value of previous measurements. The mean integrated absolute error (MIAE) is

$$\begin{aligned}\mathrm{MIAE}\left(\widehat{g}(t)\right) &= \mathbb{E}\left(\mathrm{IAE}\left(\widehat{g}(t)\right)\right)\\ &= \mathbb{E}\left(\int_T \left|\widehat{g}(t) - g(t)\right| dt\right).\end{aligned} \tag{2.22}$$

The mean integrated squared error (MISE) is defined by

$$\begin{aligned}\mathrm{MISE}\left(\widehat{g}(t)\right) &= \mathbb{E}\left(\mathrm{ISE}\left(\widehat{g}(t)\right)\right)\\ &= \mathbb{E}\left(\int_T \left(\widehat{g}(t) - g(t)\right)^2 dt\right).\end{aligned} \tag{2.23}$$

Since the integrand is nonnegative, the order of integration and expectation in (2.22) and (2.23) can be inverted to obtain the alternative forms.

$$\begin{aligned}\mathrm{MIAE}\left(\widehat{g}(t)\right) &= \int_T \mathbb{E}\left|\widehat{g}(t) - g(t)\right| dt\\ &= \int_T \mathrm{MAE}\left(\widehat{g}(t)\right) dt,\end{aligned} \tag{2.24}$$

and

$$\begin{aligned}\mathrm{MISE}\left(\widehat{g}(t)\right) &= \int_T \mathbb{E}\left(\widehat{g}(t) - g(t)\right)^2 dt\\ &= \int_T \mathrm{MSE}\left(\widehat{g}(t)\right) dt.\end{aligned} \tag{2.25}$$

By using Equation (2.17) in Expression (2.25), it is obtained that

$$\begin{aligned}\mathrm{MISE}\left(\widehat{g}(t)\right) &= \int_T \left[\left(\mathbb{E}\left(\widehat{g}(t)\right) - g(t)\right)^2 + \mathbb{V}\left(\widehat{g}(t)\right)\right] dt\\ &= \int_T \left(\mathbb{E}\left(\widehat{g}(t)\right) - g(t)\right)^2 dt + \int_T \mathbb{V}\left(\widehat{g}(t)\right) dt.\end{aligned} \tag{2.26}$$

The first and second integrals represent the integrated squared bias ISB $\left(\widehat{g}(t)\right)$ and the integrated variance $\mathrm{IV}(\widehat{g}(t))$, respectively.

The measures presented in this section will be applied to evaluate the performance of the proposed modeling against artificial data and in real applications developed later in this thesis. Reference will be made to the nomenclatures (" acronyms") in the chapters where the data are explored.

Chapter 2 ends here with the basic definitions of B-spline, BP, and discrepancy measures. The next chapter is the most important of the study, as it presents the proposed models that will be studied in the thesis.

[1]In the simulation studies of this thesis (Chapter 4), transformations are not applied to the distances $d_i$, which will be generated in the interval (0,1). A transformation is necessary for the real application (Chapter 5), and then the ISE is not considered to evaluate results.

# Chapter 3

# Spatially Dependent Functional Data Model

In recent years, the explosion of data and the growing complexity of spatially related information have given rise to innovative techniques that effectively capture the underlying structures within datasets. One such powerful approach is the functional data model with spatial dependence, which offers a unique framework for analyzing and interpreting data exhibiting functional and spatial characteristics.

The traditional FDA methodology is primarily designed to handle data represented as smooth curves or functions, typically observed over a continuous interval. It has proven successful in diverse fields, including economics, biology, environmental sciences, and more (Ramsay and Silverman, 2002; Ferraty and Vieu, 2006). However, as our understanding of the interconnectedness of spatial data has deepened, the need to integrate spatial dependencies into functional models has become increasingly evident. The functional data model with spatial dependence addresses this requirement by accommodating the inherent spatial relationships present in the data. Unlike standard FDA techniques that often treat observations as independent, this model considers the spatial context of each functional curve. By doing so, it leverages functional information and accounts for the spatial structure, enabling the analysis of complex datasets that combine functional and spatial aspects.

This chapter introduces a statistical functional model that utilizes a Bayesian hierarchical structure to model spatially correlated curves effectively. These curves are observations collected at irregularly spaced discrete points within their respective domains. A crucial aspect of this model is incorporating an autoregressive random effect component, which accounts for the dependence arising from non-equidistant distances between the data points. Two smoothing techniques have been incorporated into the structure of the proposed model. First, there is the B-spline basis that is used generally in most of the SFD research, e.g., Baladandayuthapani et al. (2008); Giraldo et al. (2012); Kokoszka and Reimherr (2017); Aguilera-Morillo et al. (2017) and Aristizabal et al. (2019). Second, there is the basis of BP, which is a little-explored tool when the functional forms of the discrete data are unknown from different spatial locations.

The organization of this chapter is as follows: Section 3.1 describes the functional,

statistical model, while Section 3.2 presents the Bayesian hierarchical structure. It explains how the Bayesian approach is utilized to model the data hierarchically. Bayesian methods incorporate prior knowledge or beliefs about the parameters into the analysis, allowing uncertainty quantification and more robust inference.

## 3.1 Statistical Model

A spatial functional process is described as $\left\{X_s : s \in \mathcal{D} \subset \mathbb{R}^d\right\}$, where $X_s$ are the functional random variables, located at location $s$ in the $d$-dimensional Euclidean space, usually $d = 2$ or 3. Each $X_s$ is defined on the interval $T = [a, b] \subseteq \mathbb{R}$ and is assumed to belong to a Hilbert space of square-integrable functions, that is, in

$$L^2(T) = \left\{ X_s : T \longrightarrow \mathbb{R}, \text{ such that } \int_T X_s(t)^2 dt < \infty \right\},$$

with the inner product $\langle X_s, X_{s^*} \rangle = \int_T X_s(t) X_{s^*}(t) dt$.

Let $\{X_{s_1}(t), X_{s_2}(t), \ldots, X_{s_m}(t)\}$ be a sample of $m$ non-independent curves that are indexed by a common domain $t \in T$ at each spatial location $s_j$, $j = 1, \ldots, m$. In practice, these functions are observed over a finite set of discrete points $\{t_i,\, i = 1, \ldots, n\}$, and the measurements are often contaminated with noise. For a fixed site $s_j$, it is assumed that the following model generates the observed functions:

$$Y_{s_j}(t_i) = X_{s_j}(t_i) + \delta_i + \epsilon_{s_j}(t_i), \tag{3.1}$$

where $\delta_i$ represents a random effect and $\epsilon_{s_j}(t_i)$ are independent random errors for each fixed $t_i$, with mean zero $E(\epsilon_j(t_i)) = 0$ and common variance, unknown, $Var(\epsilon_{s_j}(t)) = \tau$. In general, $\epsilon_{s_j}(t_i)$ is assumed to follow a Gaussian distribution. From now on, it will be assumed that each realization $X_{s_j}(t_i)$ of an underlying random function $X_s(t)$; can be expressed by a finite number of basis functions $\{f_1(t), \ldots, f_k(t)\}$ to get a reasonable approximation. Then, each curve admits an expansion into this basis as follows

$$X_{s_j}(t_i) \approx \sum_{\substack{r=r^* \\ r^* \in \{0,1\}}}^{\boldsymbol{K}} \theta_{rj} f_r(t_i), \tag{3.2}$$

for $\boldsymbol{K} \in \mathbb{N}$, with $\theta_{r,j}$'s, which are real random variables representing the basis coefficients. Based on the works of Liu et al. (2017) and Aguilera-Morillo et al. (2017), to construct the correlation between curves, $\theta_{r,j}$'s are assumed to be associated across $j$ for each $r$ but not across $r$, i.e., $Cor(\theta_{r,j}, \theta_{r,j^*}) \neq 0$ for different locations $\{j, j^*\}$ and for any given basis function index $r$. One can specify a full spatial correlation structure between $\theta_{r,j}$'s by allowing for a non-zero covariance, e.g., $Cov(\theta_{r,j}, \theta_{r,j^*}) = C(s_j, s_{j^*}) = C(h)$, for $j \neq j^*$. Note that the covariance function depends on the location $s_j$ and $s_{j^*}$ only through the Euclidean geographic distance $h = \parallel s_j - s_{j^*} \parallel \in \mathbb{R}$.

If Equation (3.2) is taken into account, then model (3.1) can be expressed as follows

$$Y_{s_j}(t_i) = \sum_{\substack{r=r^* \\ r^* \in \{0,1\}}}^{\boldsymbol{K}} \theta_{rj} f_r(t_i) + \delta_i + \epsilon_{s_j}(t_i). \tag{3.3}$$

A commonly adopted framework in SFD is to consider that the discrete realizations of the functional data at each location are collected at equally spaced points in the domain. However, an important and relevant aspect of the methodological proposal of this thesis is to work with the distances between non-equidistant measurement points, that is, $d_i = (t_i - t_{i-1}) \neq (t_{i^*} - t_{i^*-1}) = d_{i^*}$ for some $i \neq i^* \in \{2, \ldots, n\}$. For modeling convenience, a transformation is applied to the scale of the functional domain $T$ so that $d_i \in (0, 1)$.

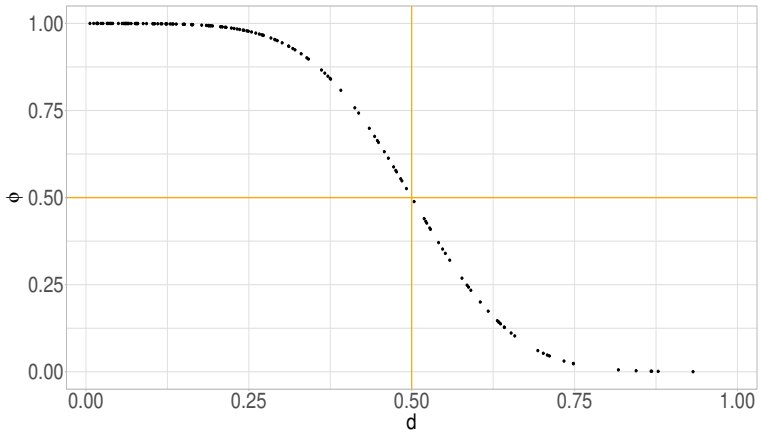
In the statistical model (3.3), the geographical positioning of the locations $s_j$ is used to impose an association between the functional trajectories observed at nearby sites. In addition to this spatial dependence, the modeling takes advantage of the distances $d_i$ to establish a similarity between the close discrete measurements of each function $Y_{s_j}(t)$ and thereby insert the dependence structure motivated by the irregular spacing of the functional domain. This information is specified through a random effect $\delta = (\delta_1, \ldots, \delta_n)^\top$, which has the following structure

$$\delta_i = \phi_i \delta_{i-1} + \epsilon_{\delta_i} \quad \text{with} \quad \phi_i = \Phi(4 - 8d_{i-1}). \tag{3.4}$$

In the above specifications, $\epsilon_{\delta_i}$, for $i = 1, \ldots, n$, is a sequence of uncorrelated identically distributed Gaussian random variables with zero mean and variance $\nu$, and $\Phi(\cdot)$ is the Cumulative Distribution Function (CDF) of the Standard Normal whose values fixed in its argument (4 and 8) allow controlling the impact of distance on the probability of $\phi_i$, then $\lim_{di \to 0} \phi_i \approx 1$, $\lim_{di \to 1} \phi_i \approx 0$ and $\phi_i = 0.5$ for $d_i = 0.5$. This formulation was motivated by the work of Mayrink and Gonçalves (2017), who used the CDF to explain (probit link) the probability of having a Markov dependence on a Bayesian mixture model. In addition, the authors use the probit structure $\phi$ due to the Gibbs sampling construction, as it facilitates the calculations necessary to obtain the full conditional distributions.

Note that an autoregressive structure is defined to associate the random effects in $\delta$, i.e., that the level of relationship between $\delta_i$ and $\delta_{i-1}$, concerning the positions $t_i$ and $t_{i-1}$ of the functional domain, is controlled by the coefficient $\phi_i \in (0, 1)$, see Figure 3.1. Assume $d_0 = 0$ (so $\phi_1 = 1$) and $\delta_0 = 0$, it implies that $\delta_1 \sim Normal(0, \nu)$.

Figure 3.1: Graph of the relationship between distances $d_i$ and the weight $\phi_i$, with 150 $d_i$'s generated from the Beta(1,2).




Source: Prepared by the author

The configuration presented in Figure 3.1 considers several distances $d_i$ distributed over the entire interval $(0, 1)$. This means a significant number of observed values will be below and above the center point 0.5. The reader must know that when the practical study focuses on distances less than or greater than 0.5, it is necessary to adjust the curve in Figure 3.1 to represent a faster or slower decay. If this adjustment is not made, having only large or small distances will result in strong or weak associations between the $\delta_i$ and $\delta_{i-1}$ effects, as determined by Equation (3.4). In the context of this thesis, one assumes that the study covers distances spanning the entire interval $(0, 1)$, and therefore any adaptation to Equation (3.4) is left for future work. As for the $\phi_i$-structure argument, it was decided not to estimate the values of 4 and 8. This decision is based on the consideration that if the values of $d_i$ do not present a high variation, i.e., if the distances between the observations of the series are concentrated in a small subinterval of $(0, 1)$, the model will lack the necessary information to estimate these values. This could result in an identification problem and would not reflect the desired behavior, as illustrated in Figure 3.1. In addition, it is decided not to establish a prior distribution for $\phi$ because not only will the variability of $\epsilon_{\delta_i}$ be taken into account, but also the variability of $\phi$ in order to distort the relationship between $\delta_i$ and $\delta_{i-1}$.

It is important to note that the modeling proposed in Equation (3.3) is not well-explored in the FDA literature. Therefore, the main objective of this thesis is to study, from a Bayesian approach, the smoothing and prediction performance of the model in different scenarios, taking into account the spatial correlation and the dependence that may exist between discrete observations, motivated by the irregular spacing over a small sample of the functional domain.

## 3.2 Bayesian Hierarchical Models

Consider that $Y_{s_j}(t_i)$ denotes the $i$-th discrete observation for the $j$-th curve, together with a collection $\{f_r(t_i)\}$ of basis functions (B-spline or BP). Let $\theta_{rj}$, $\delta_i$ and $\tau$ be the list of model parameters indicated in Section 3.1. Then, the hierarchical structure of the model can be expressed as follows:

- Observations: for $i = 1, \ldots, n$ and $j = 1, \ldots, m$

$$Y_{s_j}(t_i) | \, \theta_{rj}, \delta_i, \tau \sim Normal \left( \sum_{\substack{r=r^* \\ r^* \in \{0,1\}}}^{K} \theta_{rj} f_r(t_i) + \delta_i, \, \tau \right). \tag{3.5}$$

For this work, the Gaussian covariance function (Banerjee et al., 2014), $C(h) = \kappa \exp(-(\varphi h)^2)$, is used to build an isotropic spatial process in the $m$-order covariance matrix, $\mathbf{\Sigma}_m$, where $\kappa$ represents the spatial variation, $\varphi$ is the spatial decay parameter fixed by the researcher. At the same time, $h$ is the Euclidean distance between locations $s_j$ and $s_{j^*}$. Note that if these locations are very close in the space $\mathbb{R}^d$, the basis coefficients are similar, i.e., curves with the same shapes and $C(h) \approx \kappa$. Conversely, the larger the distance between these locations, the higher the curve's dissimilarity, and the closer to zero is $C(h)$.

The spatial association between locations is established through the adoption of the following specifications:

- The multivariate prior of the coefficients

$$\boldsymbol{\theta}_r | \, \mu_{\theta_r}, \kappa \sim Normal_m \left( \mu_{\theta_r} \mathbf{1}_m, \mathbf{\Sigma}_m \right), \quad \text{with} \quad \mathbf{1}_m = (1, \ldots, 1)^\top. \tag{3.6}$$

In this case, $\boldsymbol{\theta}_r = (\theta_{r1}, \ldots, \theta_{rm})^\top$ is a vector containing the values of the $r$-th coefficient at the $m$ locations, and $\mu_{\theta_r}$ represents the means of the $r$ coefficient of the basis functions. Furthermore, the same set of means is used for each location $j$.

- Some hyper-prior distributions for the hyper-parameters $\mu_{\theta_r}$ and $\kappa > 0$ are

$$\mu_{\theta_r} \sim Normal \left( o, v \right), \quad \text{for} \quad \kappa \sim Gamma(a_\kappa, \, b_\kappa). \tag{3.7}$$

The terms $o \in \mathbb{R}$, $v > 0$, $a_\kappa > 0$, and $b_\kappa > 0$ are scalars to be defined by the analyst. The full specification of the Bayesian model is completed with the priors:

- For $\delta_i$, $i = 2, \ldots, n$,

$$\delta_i | \, \delta_{i-1}, \nu \sim Normal \left( \phi_i \delta_{i-1}, \, \nu \right), \tag{3.8}$$

with

$$\delta_1 | \, \nu \sim Normal(0, \nu), \quad \phi_i = \Phi(4 - 8 d_{i-1}) \quad \text{and} \quad \nu \sim Gamma(a_\nu, b_\nu). \tag{3.9}$$

- For $\tau > 0$,

$$\tau \sim Gamma(a_\tau, b_\tau). \tag{3.10}$$

The elements $a_\nu > 0$, $b_\nu > 0$, $a_\tau > 0$, and $b_\tau > 0$ are specified by the researcher.

The present chapter is complete with the full description of the proposed models based on B-spline or BP. In the following chapter, a simulation study is developed to understand how the proposed models behave comprehensively. For this purpose, an artificial environment is established to reproduce various scenarios, which allows us to evaluate the performance of the models under different conditions.

# Chapter 4

# Simulation Studies

The previous Chapter 3 presents methodologies for smoothing discretely observed measurement data that exhibit spatial correlation. These techniques rely on the concept that curves nearby often show similar behavior, allowing for incorporating spatial dependence structure in the smoothing process. Moreover, these methods also consider the influence or association between irregularly spaced observations within each curve when fitting models to the data. In order to evaluate the performance of these methodologies in real scenarios, it is necessary to test them on data where the correct answer is known. The main objective is to determine the effectiveness of using the B-spline basis functions and BP together with the autoregressive random effect component. Specifically, assessing how well these techniques perform with varying levels of between-curve variability and different types of spatial correlation (low, moderate, and high) is essential.

The proposed statistical model specification uses the `Stan` programming language (Stan Development Team, 2023). `Stan` enables full Bayesian inference for continuous variable models using Markov Chain Monte Carlo (MCMC) methods, specifically the No-U-Turn sampler (NUTS), which is an adaptive form of Hamiltonian Monte Carlo (HMC) (Hoffman and Gelman, 2014). In certain situations, NUTS is presented as a more efficient and robust sampling method compared to the Gibbs or Metropolis-Hastings sampling techniques (Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984; Gelfand and Smith, 1990). However, it is essential to note that `Stan` has a limitation: it does not support inference for discrete parameters. While `Stan` can handle discrete data and discrete data models like logistic regressions, it cannot perform inference for discrete unknowns. Various interfaces are available to interact with `Stan`. The `cmdstan` interface is used for the command line shell, `pystan` for `Python` (van Rossum, 2023), and `rstan` or `cmdstanr` for `R` (R Core Team, 2023). The last-mentioned interface is considered to implement the models proposed in this thesis.

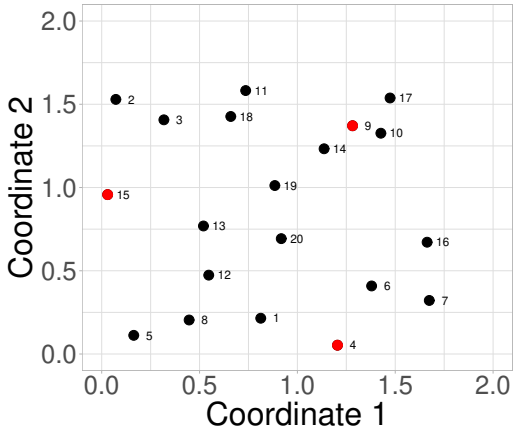
This chapter is structured as follows. Section 4.1 explains the different configurations for generating artificial datasets. The simulation results are then presented, focusing on spatial dependence, random effects, prediction, and handling missing data in scenarios with 150 observations in the series. Section 4.2, the simulation study considers functional domains with 300 and 500 discretely observed measurements. Finally, additional scenarios are introduced in Section 4.3, involving datasets generated for three different

configurations of geographic location (varying the number of sites).

## 4.1 Simulation Study Part I

Suppose that 20 locations are established in the map, as shown in Figure 4.1, and 150 discretely observed data are obtained for each site. The measure of each observation is related to a point $t_i$, $i = 1, \ldots, 150$, of the functional domain that is the same for each location. Remember that the set of $t_i$'s is considered irregularly spaced.

Figure 4.1: Grid of simulated locations.




Source: Prepared by the author

For the present simulation study, the distances between measurements are considered to be in the interval $(0, 1)$ and are generated from a Uniform distribution, $U(0, 1)$, and from a Beta distribution, $Beta(1, 2)$. The first option will induce a similar number of small and large distances. In contrast, the second option will generate a more significant number of small distances.

The inference results obtained with irregular spacing generated via the $U(0, 1)$ are generally similar to the conclusions obtained using the $Beta(1, 2)$. Estimating the coefficients and variances was similar and showed no significant disadvantage when considering distances from the Uniform or Beta. However, it was noticed that obtaining spacing from $U(0, 1)$ caused a loss of importance for the random effect $\delta$, which was introduced to incorporate associations between neighboring observations in the data series. This result is reasonable since the Uniform generates fewer small distances than the $Beta(1, 2)$; the lower the number of small distances, the lower the dependency level. Considering this mentioned behavior and aiming for a more concise and non-repetitive presentation, this thesis will focus on the results obtained with spacing generated via the $Beta(1, 2)$. Some results obtained using $U(0, 1)$ are shown in Appendix C.1 and C.2.

**Generating Artificial Data**

To generate the artificial observations $Y_{s_j}(t_i)$, $j = 1, \ldots, 20$, of the functional datasets with different levels of variability and spatial dependence for each model, consider the structures defined in Table 4.1.

Table 4.1: Description and notation of the models to be considered in the simulation study. Assume that $\theta^B_{r,j}$ and $\theta^{BP}_{r,j}$ are coefficients related to the B-spline and BP, respectively.

| Notation | Description | Functional structure of the model |
|---|---|---|
| $\mathcal{M}_{B_{k,l};\delta}$ | B-spline of order $k$, subinterval number $l$, and random effect $\delta$. | $Y_{s_j}(t_i) = \sum_{r=1}^{k+l-1} \theta^B_{r,j} B_{r,k}(t_i) + \delta_i + \epsilon_{s_j}(t_i).$ |
| $\mathcal{M}_{BP_p;\delta}$ | BP of degree $p$, order $p+1$ and random effect $\delta$. | $Y_{s_j}(t_i) = \sum_{r=0}^{p} \theta^{BP}_{r,j} b_{r,p}(t_i) + \delta_i + \epsilon_{s_j}(t_i).$ |

The appropriate number of B-spline or BP basis to utilize in a given dataset depends on various factors, including the data's complexity and the desired modeling accuracy. While there is no fixed rule for determining the precise number of basis, following some general guidelines can be helpful.

When using the B-spline, one can start with a small number of basis and gradually increase until a satisfactory approximation is achieved for the specific purpose. It is possible to use techniques such as cross-validation to evaluate the model's performance with different numbers of basis and select the optimal number that minimizes error or maximizes accuracy according to the evaluation criteria. For further information, please refer to Ramsay and Silverman (2002) and Kokoszka and Reimherr (2017). On the other hand, for the BP basis, the higher the value of $p$, the more flexibility one has to adjust the function to be approximated. However, as $p$ increases, the complexity of the resulting polynomials also increases, which can lead to numerical instability and require more computational resources (Lorentz, 2012; De Villiers, 2012).

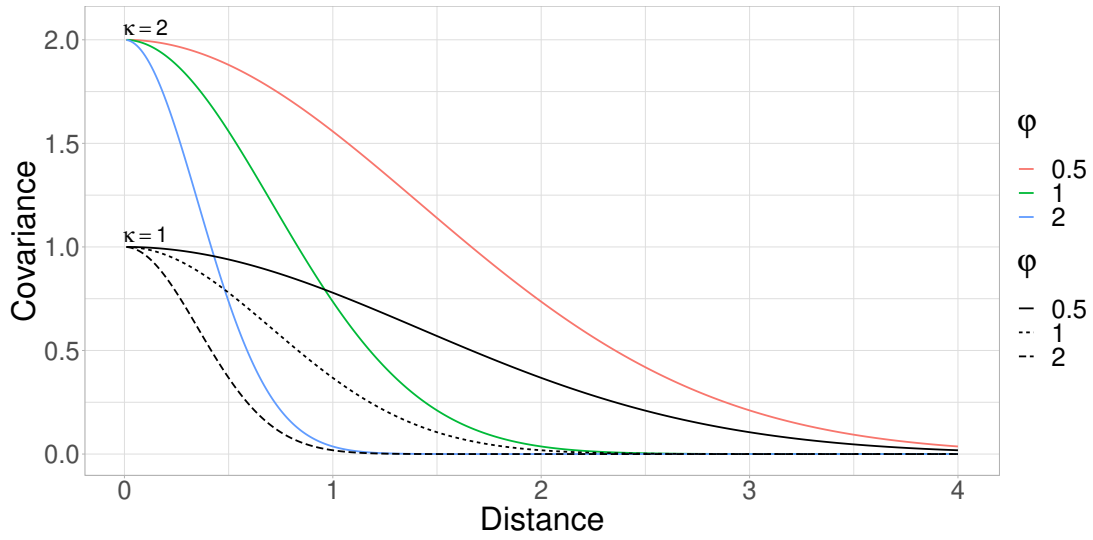
This section uses the models $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$ to generate artificial functional datasets. The first model uses an order $k = 4$ and $l = 4$, which divides the interval $[a, b]$ (functional domain) into four sub-intervals. This model comprises 7 B-spline basis functions along with the random effect component. In contrast, the second model utilizes a degree of $p = 3$, involving four BP as the basis and the random effect. In the present study, it is essential to note that the model considered to generate the data is the same one used to fit the data. Some variations include fitting the data with/without the effect $\delta_i$ and with/without the spatial dependence. A central objective here is to evaluate performance and to show that implementation is correct. Concerning the coefficients $\theta^B$ and $\theta^{BP}$ of the basis expansions, these have been obtained from

a multivariate Normal distribution $\boldsymbol{\theta}_r^B = (\theta_{r1}^B, \cdots, \theta_{r20}^B)^\top \sim Normal_{20}\left(\mu_{\theta_r^B}\mathbf{1}_{20}, \boldsymbol{\Sigma}\right)$ and $\boldsymbol{\theta}_r^{BP} = (\theta_{r1}^{BP}, \cdots, \theta_{r20}^{BP})^\top \sim Normal_{20}\left(\mu_{\theta_r^{BP}}\mathbf{1}_{20}, \boldsymbol{\Sigma}\right)$. The values used for their corresponding means are as follows:

- $\mu_{\theta_1^B} = 5$, $\mu_{\theta_2^B} = 11$, $\mu_{\theta_3^B} = 6$, $\mu_{\theta_4^B} = 4$, $\mu_{\theta_5^B} = 6$, $\mu_{\theta_6^B} = 8$, $\mu_{\theta_7^B} = 10$ and
- $\mu_{\theta_0^{BP}} = 5$, $\mu_{\theta_1^{BP}} = 11$, $\mu_{\theta_2^{BP}} = 6$, $\mu_{\theta_3^{BP}} = 4$.

The spatial structure of the data is determined by the matrix $\boldsymbol{\Sigma}$ and is defined using the Gaussian covariance function $C(h) = \kappa \exp\{-(\varphi h)^2\}$, where $h = \| s_j - s_{j^*} \|$ represents the Euclidean distance between points $s_j$ and $s_{j^*}$, with $j, j^* \in \{1, \ldots, 20\}$. To illustrate different scenarios, consider two levels of variability (moderate with $\kappa = 1$, high with $\kappa = 2$) and three levels of spatial correlation (low with $\varphi = 2$, moderate with $\varphi = 1$, and high with $\varphi = 0.5$). The behavior of the covariance is shown in Figure 4.2.

Figure 4.2: Behavior of the Gaussian covariance function concerning the distance. Each curve represents a combination $(\kappa, \varphi)$.




Source: Prepared by the author

To explain the scaling established in this work for the three levels of correlation controlled by the decay parameter $\varphi$, a constant distance of $h = 1$ with $\kappa = \{1, 2\}$ is first maintained. Then, the value of the Gaussian covariance function is calculated using $\varphi = 0.5$. The result is compared with values obtained from the other two configurations. A strong correlation can be identified when the association is twice the value obtained with $\varphi = 1$ and 43 times greater than the value obtained with $\varphi = 2$ (lowest level). When the association ($\varphi = 1$) is 20 times higher than the lowest level, it is considered a moderate correlation.

An MC scheme will be explored in this study with 250 replications of the datasets. It is important to highlight that the generated sets of coefficients for the B-spline basis

expansions and BP are kept the same across all replications for each combination ($\kappa$, $\varphi$). The component $\delta_i$ is specified according to the Equations (3.8) and (3.9), while $\epsilon_j(t_i) \sim Normal_{20}(\mathbf{0}_{20}, \boldsymbol{I}_{20})$ is a random error generated independently for each $j = 1, \ldots, 20$ and each fixed $t_i$, $i = 1, \ldots, 150$. The term $\mathbf{0}_{20}$ is a $20 \times 1$ null vector and $\boldsymbol{I}_{20}$ is a $20 \times 20$ identity matrix suggesting that $\tau = 1$; see Equation (3.5).

**Prior Specifications**

Table 4.2 shows the values for the prior distribution for each unknown hyper-parameter of the model. The choice of the value 0.1 for the first and third Gamma distributions reflects the high uncertainty related to $\tau$ and $\kappa$. These values result in Gamma distributions with a mean of 1 and a variance of 10. The second Gamma has arguments $a_\nu$ and $b_\nu$ equal to 1, defining a mean and variance of 1. This more informative specification is considered to avoid a situation where $\delta_i|\delta_{i-1}$ has a large variability, which is problematic since the mean $\phi_i\delta_{i-1}$ becomes less important; see Equation (3.8). As for the hyper-parameter $\mu_{\theta_r}$, a Normal prior with a mean of 0 and variance of 10 is proposed to represent low certainty about the real values of the coefficient means.

Table 4.2: Prior specifications considered in the simulation study.

| Hyper-parameter | Prior | Values for posterior inference |
|---|---|---|
| $\tau$ | $Gamma(a_\tau, b_\tau)$ | $a_\tau = b_\tau = 0.1$ |
| $\nu$ | $Gamma(a_\nu, b_\nu)$ | $a_\nu = b_\nu = 1$ |
| $\kappa$ | $Gamma(a_\kappa, b_\kappa)$ | $a_\kappa = b_\kappa = 0.1$ |
| $\mu_{\theta_r}$ | $Normal(o, v)$ | $o = 0$, $v = 10$ |

Again, for each of the proposed models and each of the $\kappa$ with $\varphi$ configurations within each simulation scenario, the MC scheme includes 250 replicates. As estimators, the posterior means are considered to summarize the inference results. The MCMC algorithm with the HMC dynamics implemented in `Stan` is used to obtain the samples of the posterior distributions. A total of 5,000 iterations are performed, discarding the first 2,500 observations (burn-in period) and taking the rest as samples. Only one chain is obtained for each parameter. Convergence is achieved for all the chains visually inspected during the study. In terms of mixing, the inspected MCMC chains suggested low autocorrelation.

### 4.1.1 Results of the Simulation Study Part I

**Spatial Dependence**

This simulation aims to prove the relevance of considering the spatial dependence between curves corresponding to a fixed configuration of geographical locations, see Figure 4.1. As for the generation of the artificial data, the steps described in Section 4.1 are followed

(generate data with spatial dependence) so that they subsequently fit the proposed model of the Equation (3.5), as well as the same model without the spatial structure (in this case, the matrix built by the covariance function is replaced by the identity matrix) for each MC dataset.

Once the smoothed curves are obtained from the proposed model with and without the spatial dependence, proceed to calculate the IAE and ISE measurements that analyze the goodness of fit of the trajectories in each of the 20 spatial locations for each of the 250 replicates considered. After collecting this information, the following ratios are determined:

$$\text{Ratio}^{(\text{IAE})} = \frac{\text{IAE}\left(\widehat{Y}_{s_j}(t)_{(\text{Ind})}\right)}{\text{IAE}\left(\widehat{Y}_{s_j}(t)_{(\text{Dep})}\right)}, \quad \text{Ratio}^{(\text{ISE})} = \frac{\text{ISE}\left(\widehat{Y}_{s_j}(t)_{(\text{Ind})}\right)}{\text{ISE}\left(\widehat{Y}_{s_j}(t)_{(\text{Dep})}\right)}, \tag{4.1}$$

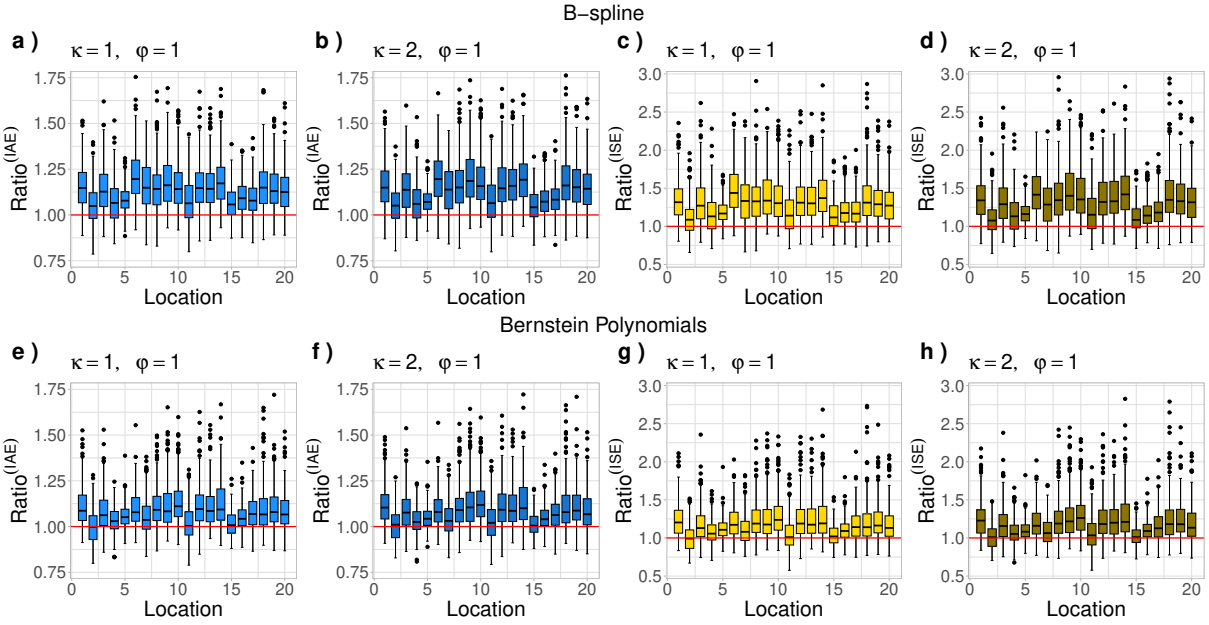
where $\widehat{Y}_{s_j}(t)_{(\text{Dep})}$ and $\widehat{Y}_{s_j}(t)_{(\text{Ind})}$ represent the smoothed curves when spatial dependence and independence are taken into account. When the expressions in (4.1) are analyzed, it is observed that if the value of the numerator is less than that of the denominator (proper fraction), the result is less than 1. On the contrary, if the numerator is larger than its denominator (improper fraction), the resulting value is greater than 1, being is the desired case for this simulation study since it means that there is a better approximation between the smoothed curves and the target curves when the modeling takes into account the spatial structure than without it on SFD sets.

In this subsection, recall that only the results of the datasets that come from the irregular spacing of the measures of the functional domain of a $Beta(1, 2)$ distribution are presented, together with the fixed values of the parameters $\kappa = \{1, 2\}$ and $\varphi = 1$ (spatial variation and spatial decay, respectively), because this is the most appropriate configuration when modeling the different datasets for the simulation study. Additional results involving $\varphi = \{0.5, 2\}$ are detailed in Appendix A.1.

Figure 4.3 illustrates the results of the quotients in (4.1) corresponding to the smoothing methods of the functional data (B-spline and BP). At first glance, it can be seen that there is little difference between the two levels of spatial variability $\kappa$ considered for both the IAE and ISE ratios. In addition, no significant difference is detected between the base functions (B-spline and BP). In Panels $(a)$, $(b)$, $(c)$, and $(d)$, it is observed that, for all geographic locations corresponding to each curve, when B-splines are used, most of the 250 values composing each boxplot are above 1; this indicates a better performance of the smoothed trajectories when the spatial structure is considered in the proposed model compared to the model without it. This suggests that spatial dependence is essential to ensure a better fit. For the Panels $(e)$, $(f)$, $(g)$, and $(h)$ corresponding to the BP, it is observed that in locations 2, 11, and 15, which are the curves that have few or no near neighbors, the proposed model with the spatial structure presents about half of the MC replicates higher values of the discrepancy measures concerning the model without the

spatial dependence. However, for trajectories with several nearest neighbors, the model's performance with spatial dependence is better in most samples when compared to the model without it.

Figure 4.3: Comparison of the IAE and ISE ratios of the B-spline and BP models with two spatial variation parameter settings: $\kappa = 1$ and $\kappa = 2$, together with a decay parameter $\varphi = 1$.




Source: Prepared by the author

### Autoregressive Random Effect

In this simulation scenario, the objective is to test the importance of the random effect $\delta$, which is a component of the model proposed in Equation (3.5), and which allows the insertion of the dependency motivated by the irregular distancing of the observations in the functional domain.

The results presented here correspond to the artificial datasets generated from the model shown in Chapter 3, considering the two smoothing methodologies (B-spline, BP), with the following adjustments: the spatial decay parameter $\varphi = 1$, the variability levels $\kappa = \{1, 2\}$ and the irregular spacing of the observations given by a $Beta(1, 2)$ distribution. Additional results (not shown here) providing valuable information for understanding the context and scope of the study are presented in Appendix A.2.

Tables 4.3 and 4.4 present the MIAE and MISE discrepancy measures of the curves fitted using the models described in Table 4.1. The tables compare the performance of these models with and without the random effect component. The comparison is made for two variability options and the same spatial correlation value. The analysis reveals that the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$ perform better when the formats of the artificial curves

are similar and close to each other (specifically, when $\kappa = 1$). In addition, functions located close to each other on a spatial scale lend strength to each other for inference purposes; for example, the smoothed curves in places 9 and 14 are close and have other neighbors, the reason for which they have the lowest values for the discrepancy measures mentioned above for the two models. In contrast, curves at distant locations without nearby companions, such as 2 and 5, demonstrate higher MIAE and MISE metrics values.

In summary, both metrics used to assess the accuracy of the smoothed curves show lower values when the random effect component is incorporated in the models (see columns 2-3 and 6-7 in both tables), in contrast to the models that exclude this component. In the latter scenario, the values are significantly higher (see columns 4-5 and 8-9 in both tables), suggesting that the curves move away from the target functions. This pattern of behavior is repeated in both proposed models (B-spline and BP). It is important to remember that a lower value indicates a more accurate fit in the context of discrepancy metrics.

Table 4.3: Discrepancy measures for the smoothed curves using the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{B_{4,4};\bullet}$, with and without the random effect component, respectively.

| | $\mathcal{M}_{B_{4,4};\delta}$ | | $\mathcal{M}_{B_{4,4};\bullet}$ | | $\mathcal{M}_{B_{4,4};\delta}$ | | $\mathcal{M}_{B_{4,4};\bullet}$ | |
|---|---|---|---|---|---|---|---|---|
| | MIAE | MISE | MIAE | MISE | MIAE | MISE | MIAE | MISE |
| | $\kappa = 1, \varphi = 1$ | | | | $\kappa = 2, \varphi = 1$ | | | |
| $S_1$ | 0.187 | 0.055 | 0.718 | 0.789 | 0.192 | 0.057 | 0.719 | 0.792 |
| $S_2$ | 0.209 | 0.070 | 0.730 | 0.806 | 0.215 | 0.074 | 0.732 | 0.812 |
| $S_3$ | 0.191 | 0.057 | 0.724 | 0.795 | 0.195 | 0.059 | 0.726 | 0.798 |
| $S_4$ | 0.206 | 0.067 | 0.728 | 0.806 | 0.210 | 0.070 | 0.729 | 0.808 |
| $S_5$ | 0.216 | 0.074 | 0.733 | 0.816 | 0.220 | 0.077 | 0.734 | 0.819 |
| $S_6$ | 0.189 | 0.056 | 0.724 | 0.792 | 0.193 | 0.059 | 0.726 | 0.794 |
| $S_7$ | 0.197 | 0.061 | 0.726 | 0.796 | 0.203 | 0.065 | 0.729 | 0.798 |
| $S_8$ | 0.187 | 0.055 | 0.718 | 0.791 | 0.189 | 0.056 | 0.718 | 0.792 |
| $S_9$ | 0.180 | 0.051 | 0.717 | 0.787 | 0.182 | 0.052 | 0.718 | 0.787 |
| $S_{10}$ | 0.185 | 0.054 | 0.719 | 0.791 | 0.188 | 0.056 | 0.720 | 0.792 |
| $S_{11}$ | 0.200 | 0.063 | 0.727 | 0.803 | 0.204 | 0.066 | 0.729 | 0.806 |
| $S_{12}$ | 0.183 | 0.053 | 0.721 | 0.788 | 0.187 | 0.055 | 0.722 | 0.791 |
| $S_{13}$ | 0.185 | 0.054 | 0.722 | 0.790 | 0.188 | 0.056 | 0.724 | 0.791 |
| $S_{14}$ | 0.182 | 0.052 | 0.719 | 0.787 | 0.185 | 0.054 | 0.719 | 0.789 |
| $S_{15}$ | 0.204 | 0.066 | 0.726 | 0.797 | 0.213 | 0.072 | 0.729 | 0.804 |
| $S_{16}$ | 0.208 | 0.069 | 0.723 | 0.812 | 0.212 | 0.073 | 0.724 | 0.816 |
| $S_{17}$ | 0.195 | 0.060 | 0.723 | 0.793 | 0.201 | 0.064 | 0.725 | 0.797 |
| $S_{18}$ | 0.185 | 0.054 | 0.720 | 0.791 | 0.187 | 0.055 | 0.720 | 0.792 |
| $S_{19}$ | 0.187 | 0.055 | 0.723 | 0.792 | 0.191 | 0.058 | 0.724 | 0.794 |
| $S_{20}$ | 0.189 | 0.056 | 0.726 | 0.793 | 0.192 | 0.059 | 0.728 | 0.796 |

Table 4.4: Discrepancy measures for the smoothed curves using the $\mathcal{M}_{BP_3;\delta}$ and $\mathcal{M}_{BP_3;\bullet}$, with and without the random effect component, respectively.

| | $\mathcal{M}_{BP_3;\delta}$ | | $\mathcal{M}_{BP_3;\bullet}$ | | $\mathcal{M}_{BP_3;\delta}$ | | $\mathcal{M}_{BP_3;\bullet}$ | |
|---|---|---|---|---|---|---|---|---|
| | MIAE | MISE | MIAE | MISE | MIAE | MISE | MIAE | MISE |
| | $\kappa = 1, \varphi = 1$ | | | | $\kappa = 2, \varphi = 1$ | | | |
| $S_1$ | 0.178 | 0.050 | 0.756 | 0.886 | 0.181 | 0.051 | 0.757 | 0.888 |
| $S_2$ | 0.198 | 0.063 | 0.769 | 0.905 | 0.200 | 0.064 | 0.769 | 0.907 |
| $S_3$ | 0.179 | 0.050 | 0.757 | 0.888 | 0.182 | 0.052 | 0.757 | 0.890 |
| $S_4$ | 0.191 | 0.057 | 0.758 | 0.895 | 0.194 | 0.060 | 0.759 | 0.898 |
| $S_5$ | 0.192 | 0.058 | 0.758 | 0.896 | 0.195 | 0.060 | 0.759 | 0.898 |
| $S_6$ | 0.181 | 0.052 | 0.755 | 0.889 | 0.185 | 0.054 | 0.755 | 0.891 |
| $S_7$ | 0.186 | 0.054 | 0.759 | 0.890 | 0.189 | 0.057 | 0.761 | 0.892 |
| $S_8$ | 0.178 | 0.049 | 0.756 | 0.887 | 0.180 | 0.051 | 0.756 | 0.888 |
| $S_9$ | 0.175 | 0.048 | 0.758 | 0.886 | 0.175 | 0.048 | 0.758 | 0.886 |
| $S_{10}$ | 0.178 | 0.050 | 0.760 | 0.889 | 0.180 | 0.051 | 0.760 | 0.889 |
| $S_{11}$ | 0.190 | 0.057 | 0.755 | 0.901 | 0.192 | 0.058 | 0.756 | 0.901 |
| $S_{12}$ | 0.177 | 0.049 | 0.761 | 0.888 | 0.180 | 0.051 | 0.762 | 0.889 |
| $S_{13}$ | 0.177 | 0.049 | 0.759 | 0.886 | 0.179 | 0.051 | 0.760 | 0.887 |
| $S_{14}$ | 0.176 | 0.049 | 0.760 | 0.886 | 0.178 | 0.050 | 0.760 | 0.887 |
| $S_{15}$ | 0.192 | 0.058 | 0.759 | 0.894 | 0.197 | 0.061 | 0.761 | 0.897 |
| $S_{16}$ | 0.187 | 0.055 | 0.759 | 0.892 | 0.190 | 0.057 | 0.759 | 0.894 |
| $S_{17}$ | 0.184 | 0.053 | 0.759 | 0.890 | 0.187 | 0.055 | 0.759 | 0.891 |
| $S_{18}$ | 0.179 | 0.050 | 0.755 | 0.890 | 0.179 | 0.050 | 0.755 | 0.889 |
| $S_{19}$ | 0.179 | 0.050 | 0.762 | 0.888 | 0.181 | 0.051 | 0.762 | 0.889 |
| $S_{20}$ | 0.180 | 0.051 | 0.762 | 0.890 | 0.183 | 0.053 | 0.761 | 0.891 |

### Prediction

This simulation study aims to predict functional data in areas where no observations are available. For this purpose, a series of data observed at different geographic points in a particular region is used. These datasets are then employed to build a model that provides information about the behavior of these series in unsampled locations. See Figure 4.1 for the complete map. The datasets explored here are those with spatial dependence evaluated in the first study presented in Subsection 4.1.1. In this particular study, the observations from locations $S_4$, $S_9$, and $S_{15}$ are assumed as missing. In other words, the whole series related to these locations are ignored when fitting the models, but they are the target for prediction. The same models and configurations used to generate the artificial observations (as detailed in Section 4.1) are used to predict the values of the unobserved curves. Therefore, it is sufficient to specify only the geographic coordinates of the locations

where the series are not observed. This procedure is executed within the `Stan`, treating the unobserved values as unknown quantities with defined prior specifications that align with the model's structure. In other words, when $Y_{s_j}(t_i)$ is missing, the expression in (3.5) serves as a prior specification rather than a term forming the likelihood.

The results presented in Table 4.5 can be analyzed to determine the impact of the proximity of observed curves on the trajectory of the predictions at unobserved locations. By examining the MIAE and MISE from MC replications, it becomes evident that unobserved sites surrounded by nearby neighbors exhibit a better fit to the target functions, regardless of the level of spatial variability. This trend is demonstrated in the case of location $S_9$ displaying lower values. On the other hand, the somewhat isolated locations $S_4$ and $S_{15}$ show inferior performance in their predictions, especially when there is high variability ($\kappa = 2$) between the curves.

In summary, the prediction of curves at unsampled locations shows better performance in both models (B-spline and BP) when nearby neighbors are present in the surrounding areas, as they significantly influence the estimation process. This strategy is based on the idea that trends observed in adjacent locations provide valuable information about underlying patterns and behaviors, allowing for more accurate and reliable estimates.

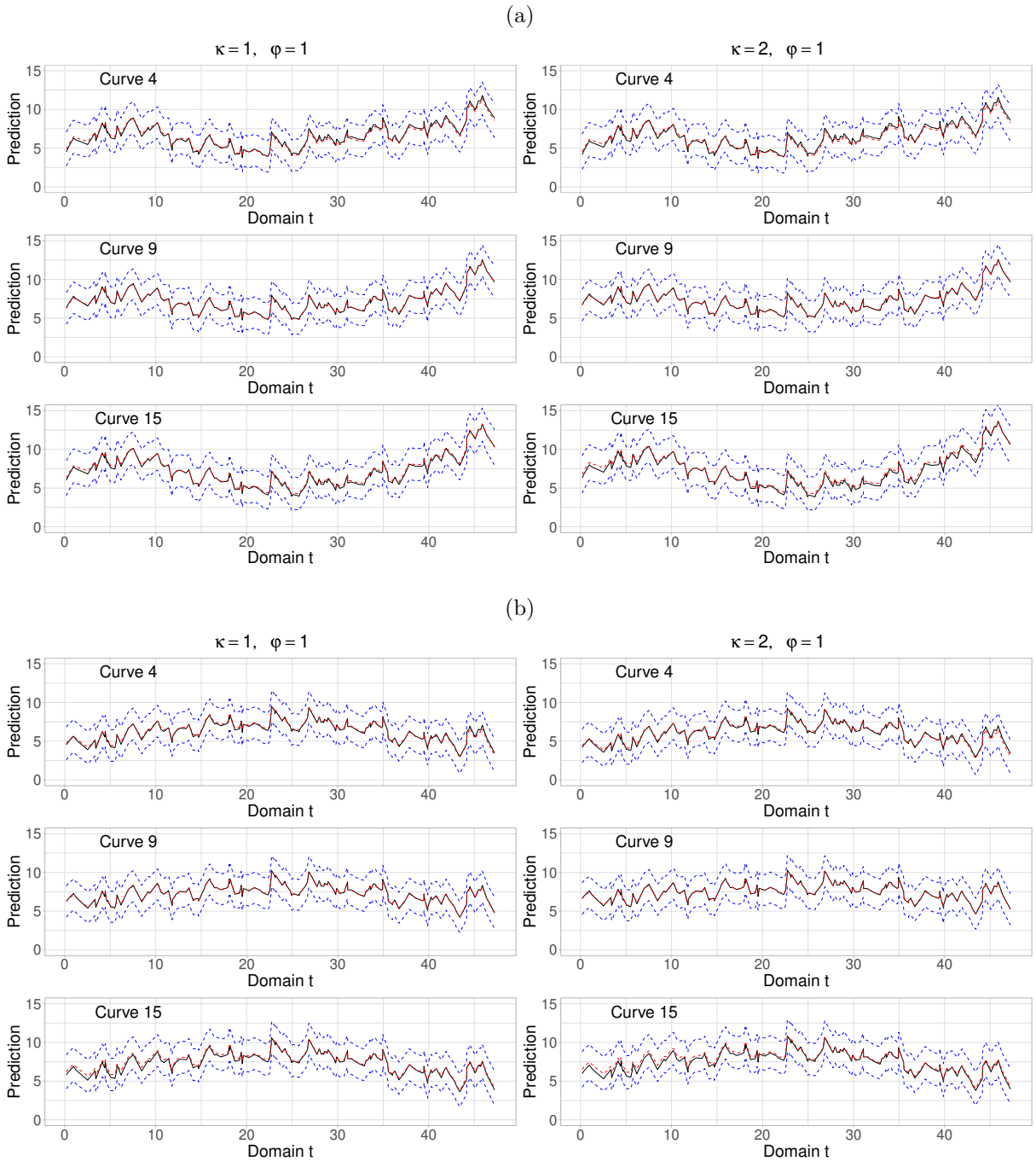
Table 4.5: Comparison of MIAE and MISE discrepancy measures for the prediction of unobserved curves using the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$, with two configurations of the spatial variation parameters: $\kappa = 1$ and $\kappa = 2$, together with a decay parameter $\varphi = 1$.

| Artificial functional data structure | Prediction model | Location | Measure of discrepancy | | | |
|---|---|---|---|---|---|---|
| | | | MIAE | MISE | MIAE | MISE |
| | | | $\kappa = 1, \varphi = 1$ | | $\kappa = 2, \varphi = 1$ | |
| $\mathcal{M}_{B_{4,4};\delta}$ | $\mathcal{M}_{B_{4,4};\delta}$ | $S_4$ | 0.277 | 0.119 | 0.306 | 0.144 |
| | | $S_9$ | 0.197 | 0.061 | 0.199 | 0.063 |
| | | $S_{15}$ | 0.258 | 0.104 | 0.319 | 0.156 |
| $\mathcal{M}_{BP_3;\delta}$ | $\mathcal{M}_{BP_3;\delta}$ | $S_4$ | 0.237 | 0.088 | 0.255 | 0.103 |
| | | $S_9$ | 0.191 | 0.057 | 0.192 | 0.058 |
| | | $S_{15}$ | 0.267 | 0.113 | 0.318 | 0.159 |

Figure 4.4 shows results summarized from the MC scheme. The average of the estimates obtained from each sample is considered to build the 95% Highest Posterior Density (HPD) intervals (blue lines) and the estimated curve (red line). Note that average intervals (region between blue lines) manage to capture the real trajectories (black line). Furthermore, the average predicted curves for the locations $S_4$, $S_9$, and $S_{15}$ are close to their respective target functions (red and black lines are too close). This indicates that good predictions are obtained even for location $S_{15}$ having distant neighbors. When comparing Panels $(a)$ and $(b)$, no strong distinction in terms of performance is detected

between the B-spline and BP. For other cases ($\varphi = \{0.5, 2\}$), the results are presented and explained in Appendix A.3

Figure 4.4: Target function (black solid lines), point-wise mean prediction curves (red dashed line), and 95% point-wise mean HPD intervals (blue dashed line) of the estimated curves for the $\mathcal{M}_{B_{4,4};\delta}$ (Panel a) and $\mathcal{M}_{BP_3;\delta}$ (Panel b). Each model is evaluated with two settings of the spatial variation parameters, specifically $\kappa = 1$ and $\kappa = 2$, and a decay parameter $\varphi = 1$.




Source: Prepared by the author

### Missing Data

This simulation study evaluates the proposed models' performance in dealing with missing data in the discretely observed measurement samples at different locations. In the previous analysis, the whole series related to a location $S_j$ is unknown. In contrast, the present study assumes that a few observations are missing within one or more series. For this purpose, the data from the previous scenario (prediction) are used, and some observations are deleted randomly to create samples with missing data. It is essential to mention that the configuration of the missing data positions is the same for all MC replications. The study is designed with 450 missing values that should be estimated, representing 15% of the 3,000 observations in each MC sample.

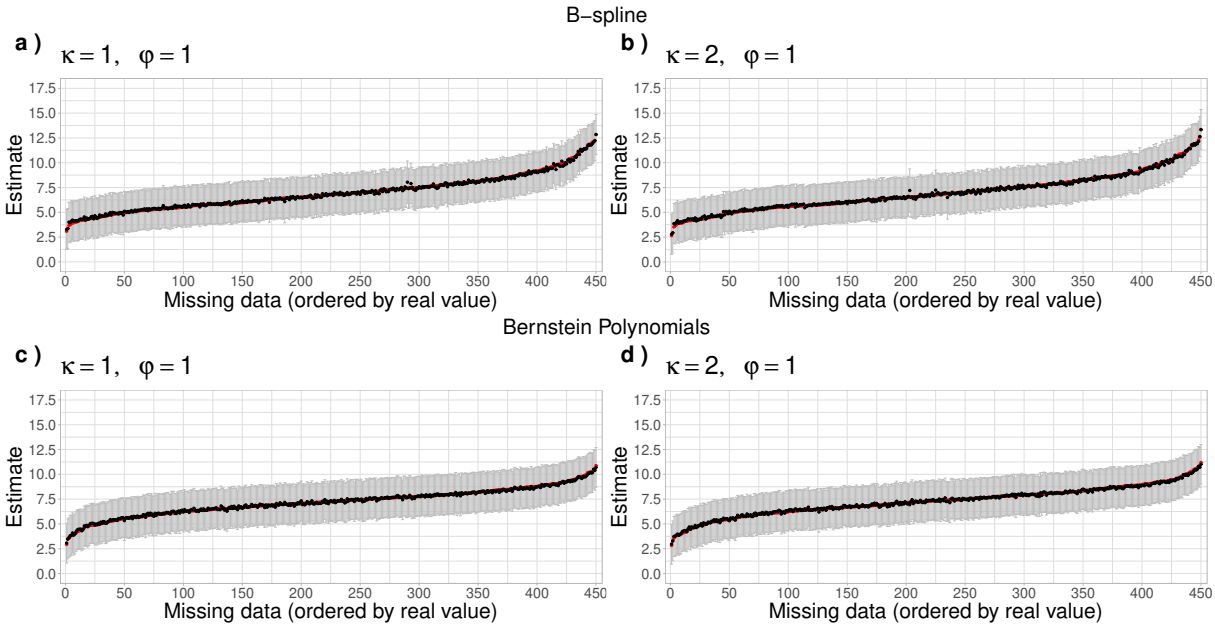
Figure 4.5 shows the mean 95% HPD intervals for the 450 missing data from the MC datasets. The black points represent the average posterior estimate, and the true observations are the red points. The graphs are ordered concerning the true values to improve the visual analysis. Please note that these graphs do not represent a curve in the functional domain. The full study covers 6 different scenarios (Beta distances), each characterized by two levels of variability between the curves and three correlation values. The proposed $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$ models are employed to fit the data in all cases. Only four combinations involving B-spline, BP, and $\kappa = \{1, 2\}$ are presented here, while the other analyses ($\varphi = \{0.5, 2\}$) performed can be found in Appendix A.4.

Figures 4.5 ($a$) and 4.5 ($b$) show results from datasets generated and fitted using the $\mathcal{M}_{B_{4,4};\delta}$. It can be observed that the HPD intervals successfully capture the actual values of the missing data. Although the intervals have some degree of uncertainty, most averages of the mean posterior estimates from each MC replicate closely align with the desired targets. However, two and three estimates, in particular, stand out in both graphs. These values correspond to the curves at geographical positions 2 and 16 at functional domain points 23 (curve 2), 141, and 150 (curve 16). A possible explanation for this behavior is that these curves are relatively far away from their closest neighbors, indicating that the spatial dependence is not strong enough to share strength during the estimation process, especially when there is higher variability between the curves.

Figures 4.5 ($c$) and 4.5 ($d$) display the analyses for the data generated and fitted with the $\mathcal{M}_{BP_3;\delta}$ model. These graphs clearly illustrate that the average HPD intervals effectively capture all the real values, and the averaged posterior mean estimates from the MC scheme are close to the true magnitudes of the data. The posterior uncertainty is similar across all panels, suggesting no significant difference between $\kappa = 1$ or 2 and B-spline or BP.

In conclusion, the data imputation strategy of utilizing the B-spline and BP models has effectively yielded reliable results for point estimates. Furthermore, across all evaluated scenarios, there's a consistent pattern of similarity in the lengths of the mean

Figure 4.5: Mean of the 95% HPD Intervals for the 450 missing data using the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$. The actual values are the red dots, and the averages of the mean posterior estimates from each MC replicate are the black dots. Each model is evaluated with two settings of the spatial variation parameters ($\kappa = 1$ and $\kappa = 2$) and a decay parameter $\varphi = 1$.




Source: Prepared by the author

intervals for each imputed value. This indicates a similar level of uncertainty in each case, with no noticeable significant differences between them.

## 4.2 Simulation Study Part II

In this second part of the study, the configurations and structures used in Part I to generate artificial data and estimate the parameters and hyper-parameters of the proposed models are maintained. However, this time, the functional domain is expanded to 300 and 500 discrete points, which gives us more insight into the behavior of the underlying functions in the data. In addition, the analysis focuses on only two variability levels: $\kappa = 1$ and 2. Likewise, a decay parameter value of $\varphi = 1$ is used, which allows us to attribute moderate correlations.

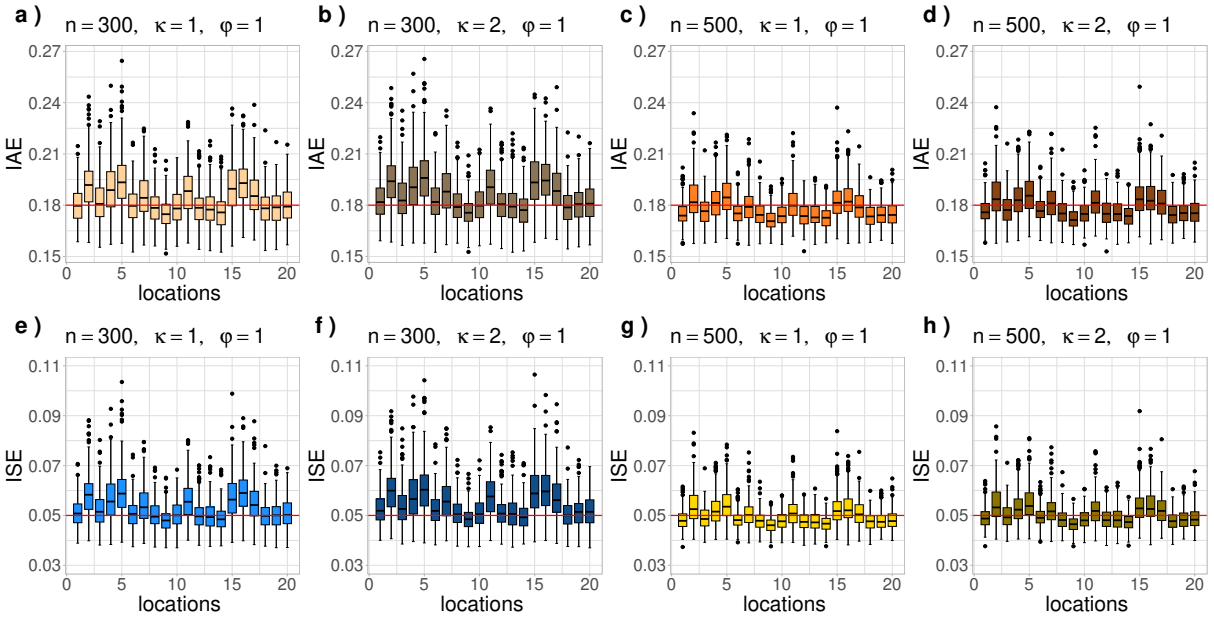
### 4.2.1 Results of the Simulation Study Part II

Figure 4.6 presents the IAE and ISE discrepancy measurement results for the 250 MC replicates. These measurements allow us to evaluate how close the smoothed (or estimated) curves of $\mathcal{M}_{B_{4,4};\delta}$ are to the target curves. From the analysis of the results,

some conclusions can be drawn. First, no significant differences are observed between the results for the two levels of variation for each size of the functional domain. For example, Panels $(a)$ and $(b)$, as well as $(c)$ and $(d)$, show similar behavior. Second, when analyzing the sets of curves with a domain composed of 300 discrete observations, it is observed that the interquartile range (IQR) in the boxplots is higher compared to the results obtained for functions with a domain of 500 observations. This indicates higher measurement variability when working with smaller datasets, which is an expected result.

Regarding the position of the medians, the majority of them are below the reference values of 0.18 and 0.5 (represented by the red line arbitrarily defined to aid the visual analysis) when the smoothed curves are based on a broader domain, as seen in Panels $(c)$, $(d)$, $(g)$, and $(h)$. On the other hand, when the domain has fewer points (300), most of the medians lie above the reference levels. This indicates that the estimated curves show improved performance when a substantial amount of information is accessible.

Figure 4.6: Comparison of IAE and ISE discrepancy measures of the 20 estimated curves. The $\mathcal{M}_{B_{4,4};\delta}$ is used with spatial variation parameter $\kappa = 1$ and $\kappa = 2$. In addition, a constant value for the decay parameter $\varphi = 1$ and two sizes of measurements discretely observed in the functional domain are used: $n = 300$ and $n = 500$.

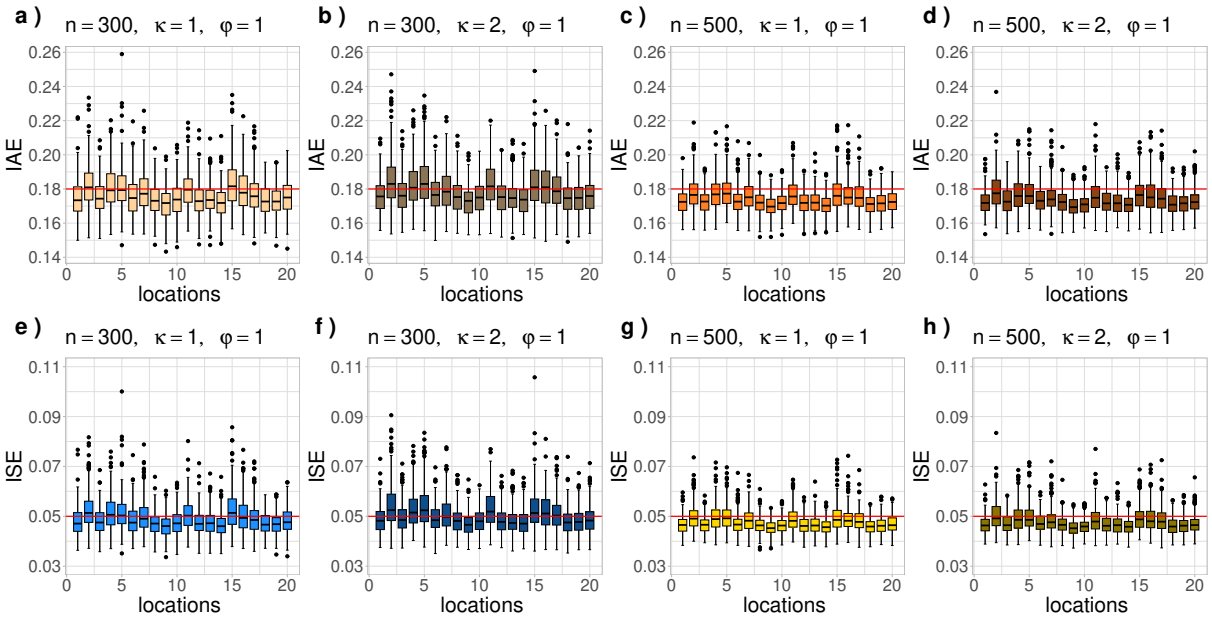



Source: Prepared by the author

Figure 4.7 presents the results for the datasets generated and estimated using the $\mathcal{M}_{BP_3;\delta}$ model. The graphs show the variability differences (size of the boxplots) in the discrepancy measures for the two levels of variation corresponding to each sample size. To easily observe these differences, one can compare the box lengths on the panels, with a slightly higher variability for $\kappa = 2$. This visualization provides valuable insights into the model's performance and how it responds to different levels of variation and moderate spatial correlation for the given sample sizes. When analyzing the curves consisting of

300 discrete observations, it is evident that the IQR is greater than the results obtained for functions with a domain of 500 data. This indicates higher MC dispersion in the measurements when working with smaller sizes, which is an expected result. In most cases, the medians are below the arbitrary reference values of 0.18 and 0.5 (red line) included to facilitate interpretation. However, it is essential to note that the medians in Panels $(c)$, $(d)$, $(g)$, and $(h)$ have a lower value than the other graphs, indicating an improvement in curve fitting. In summary, the behavior observed for the BP case is similar to those observed for the B-spline.

Figure 4.7: Comparison of IAE and ISE discrepancy measures of the 20 estimated curves. The $\mathcal{M}_{BP_3;\delta}$ is used with spatial variation parameter $\kappa = 1$ and $\kappa = 2$. In addition, a constant value for the decay parameter $\varphi = 1$ and two sizes of measurements discretely observed in the functional domain are used: $n = 300$ and $n = 500$.




Source: Prepared by the author

Table 4.6 presents the MISE of the estimators for the 20 functions evaluated over the entire common domain shared by these functions. This discrepancy measure can be effectively decomposed into Integrated Variance (IV) and Integrated Squared Bias (ISB); see Section 2.4. Such decomposition allows for a comprehensive and detailed analysis of the variance and bias inherent in the functions smoothed by the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$ models. When analyzing the performance of the two fitting models with a sample size of $n = 300$ in both cases of variation, it is highlighted that the smoothed curve of site 9 shows the lowest MISE value. Moreover, this curve exhibits a more accurate fit when considering the variation of $\kappa = 1$. In contrast, the curves of sites 2 and 5 present the highest values of both MISE and ISB in any of the scenarios studied. On the other hand, by increasing the size of the discretely observed measurements to $n = 500$, it is observed that the MISE and IV indices decrease even more in curves $S_9$, $S_2$, and $S_5$ compared

to the 300 discrete points of the previous functional domain. However, it is essential to mention a case in which the smoothed curve of site 9, modeled by the $\mathcal{M}_{B_{4,4};\delta}$, presented an increase in the ISB value with the larger sample. In summary, the results show that all the smoothed curves present a good fitting performance in both cases of variation. The closeness between locations $S_j$ is also beneficial in the estimation process. In addition, it is observed that increasing the sample size to $n = 500$ achieves higher accuracy in the models. These findings support the importance of considering an adequate sample size to obtain more reliable estimates in model fit analysis.

Table 4.6: Results of the MISE, IV, and ISB discrepancy measures for the curves estimated with the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$, with two levels of variation $\kappa = 1$ and $\kappa = 2$ and a fixed value for the decay parameter $\varphi = 1$, in two sample sizes $n = 300$ and $n = 500$.

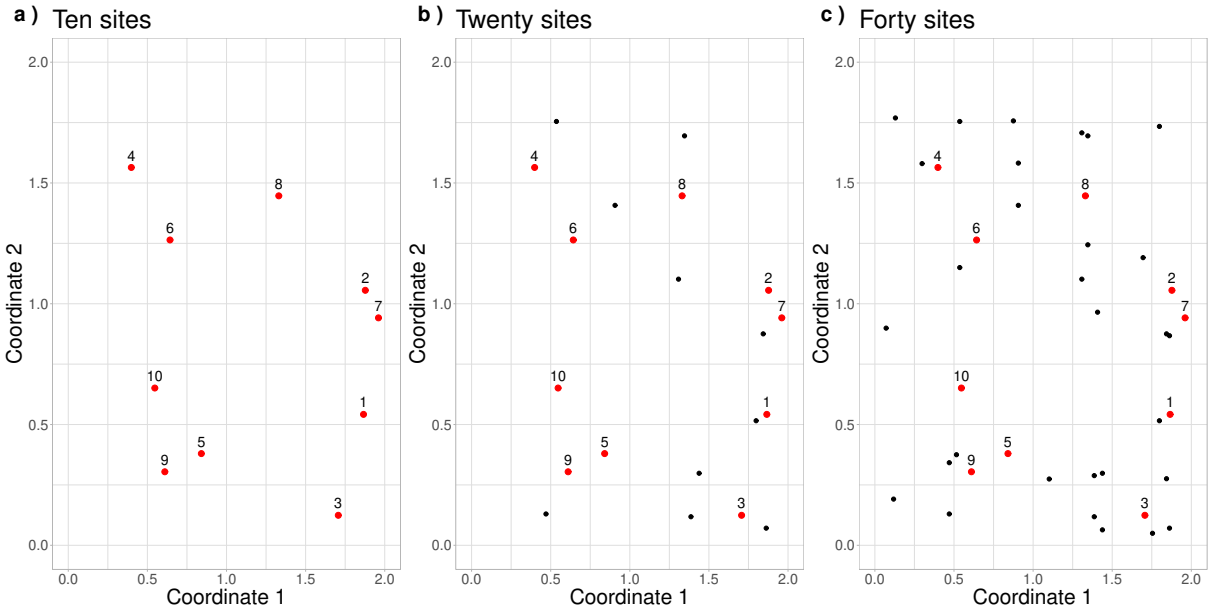
| | | $\mathcal{M}_{B_{4,4};\delta}$ | | | | | | $\mathcal{M}_{BP_3;\delta}$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | Site | MISE | IV | ISB | MISE | IV | ISB | MISE | IV | ISB | MISE | IV | ISB |
| | | $\kappa = 1, \varphi = 1$ | | | $\kappa = 2, \varphi = 1$ | | | $\kappa = 1, \varphi = 1$ | | | $\kappa = 2, \varphi = 1$ | | |
| 300 | $S_1$ | 0.05113 | 0.04494 | 0.00619 | 0.05265 | 0.04619 | 0.00646 | 0.04786 | 0.04198 | 0.00588 | 0.04863 | 0.04324 | 0.00538 |
| | $S_2$ | 0.05861 | 0.05036 | 0.00825 | 0.06021 | 0.05230 | 0.00792 | **0.05264** | **0.04486** | **0.00778** | **0.05406** | **0.04709** | **0.00696** |
| | $S_3$ | 0.05230 | 0.04444 | 0.00785 | 0.05371 | 0.04559 | 0.00812 | 0.04815 | 0.04132 | 0.00683 | 0.04915 | 0.04313 | 0.00602 |
| | $S_4$ | 0.05689 | 0.04903 | 0.00786 | 0.05828 | 0.05088 | 0.00739 | 0.05178 | 0.04467 | 0.00711 | 0.05275 | 0.04589 | 0.00686 |
| | $S_5$ | **0.05997** | **0.05156** | **0.00841** | **0.06126** | **0.05344** | **0.00781** | 0.05157 | 0.04500 | 0.00658 | 0.05392 | 0.04778 | 0.00615 |
| | $S_6$ | 0.05083 | 0.04415 | 0.00667 | 0.05231 | 0.04558 | 0.00673 | 0.04824 | 0.04150 | 0.00674 | 0.04939 | 0.04309 | 0.00630 |
| | $S_7$ | 0.05446 | 0.04846 | 0.00600 | 0.05647 | 0.05029 | 0.00617 | 0.05006 | 0.04419 | 0.00587 | 0.05120 | 0.04575 | 0.00545 |
| | $S_8$ | 0.05036 | 0.04454 | 0.00581 | 0.05112 | 0.04530 | 0.00582 | 0.04736 | 0.04122 | 0.00614 | 0.04845 | 0.04280 | 0.00566 |
| | $S_9$ | **0.04818** | **0.04281** | **0.00537** | **0.04865** | **0.04329** | **0.00537** | **0.04618** | **0.04031** | **0.00587** | **0.04679** | **0.04140** | **0.00539** |
| | $S_{10}$ | 0.05050 | 0.04400 | 0.00651 | 0.05128 | 0.04464 | 0.00664 | 0.04744 | 0.04098 | 0.00645 | 0.04823 | 0.04222 | 0.00601 |
| | $S_{11}$ | 0.05587 | 0.04744 | 0.00843 | 0.05765 | 0.04937 | 0.00828 | 0.05074 | 0.04274 | 0.00800 | 0.05289 | 0.04532 | 0.00757 |
| | $S_{12}$ | 0.04992 | 0.04343 | 0.00650 | 0.05124 | 0.04433 | 0.00690 | 0.04753 | 0.04076 | 0.00677 | 0.04833 | 0.04218 | 0.00615 |
| | $S_{13}$ | 0.05005 | 0.04439 | 0.00566 | 0.05116 | 0.04558 | 0.00557 | 0.04760 | 0.04169 | 0.00591 | 0.04839 | 0.04280 | 0.00559 |
| | $S_{14}$ | 0.04877 | 0.04297 | 0.00581 | 0.04972 | 0.04383 | 0.00589 | 0.04644 | 0.04043 | 0.00601 | 0.04753 | 0.04184 | 0.00568 |
| | $S_{15}$ | 0.05743 | 0.05067 | 0.00675 | 0.06001 | 0.05301 | 0.00700 | 0.05257 | 0.04602 | 0.00655 | 0.05305 | 0.04718 | 0.00587 |
| | $S_{16}$ | 0.05945 | 0.04889 | 0.01056 | 0.06061 | 0.05095 | 0.00965 | 0.05056 | 0.04435 | 0.00621 | 0.05192 | 0.04601 | 0.00592 |
| | $S_{17}$ | 0.05515 | 0.04839 | 0.00676 | 0.05743 | 0.05017 | 0.00726 | 0.04943 | 0.04358 | 0.00586 | 0.05102 | 0.04576 | 0.00527 |
| | $S_{18}$ | 0.05005 | 0.04423 | 0.00581 | 0.05056 | 0.04498 | 0.00558 | 0.04709 | 0.04087 | 0.00623 | 0.04780 | 0.04243 | 0.00537 |
| | $S_{19}$ | 0.05038 | 0.04373 | 0.00665 | 0.05149 | 0.04470 | 0.00679 | 0.04705 | 0.04096 | 0.00609 | 0.04814 | 0.04239 | 0.00575 |
| | $S_{20}$ | 0.05112 | 0.04396 | 0.00715 | 0.05219 | 0.04506 | 0.00714 | 0.04822 | 0.04141 | 0.00681 | 0.04887 | 0.04254 | 0.00633 |
| 500 | $S_1$ | 0.04834 | 0.04227 | 0.00607 | 0.04924 | 0.04304 | 0.00620 | 0.04670 | 0.04064 | 0.00605 | 0.04671 | 0.04126 | 0.00545 |
| | $S_2$ | 0.05365 | 0.04679 | 0.00686 | 0.05462 | 0.04802 | 0.00659 | **0.04971** | **0.04272** | **0.00700** | **0.05038** | **0.04443** | **0.00595** |
| | $S_3$ | 0.04911 | 0.04204 | 0.00707 | 0.04986 | 0.04279 | 0.00707 | 0.04662 | 0.04030 | 0.00632 | 0.04716 | 0.04133 | 0.00582 |
| | $S_4$ | 0.05216 | 0.04547 | 0.00669 | 0.05312 | 0.04658 | 0.00653 | 0.04950 | 0.04296 | 0.00654 | 0.04928 | 0.04320 | 0.00609 |
| | $S_5$ | **0.05429** | **0.04680** | **0.00749** | **0.05493** | **0.04793** | **0.00700** | 0.04996 | 0.04351 | 0.00645 | 0.04954 | 0.04391 | 0.00563 |
| | $S_6$ | 0.04833 | 0.04212 | 0.00621 | 0.04941 | 0.04316 | 0.00625 | 0.04708 | 0.04063 | 0.00645 | 0.04732 | 0.04145 | 0.00586 |
| | $S_7$ | 0.05086 | 0.04502 | 0.00584 | 0.05217 | 0.04620 | 0.00597 | 0.04864 | 0.04252 | 0.00612 | 0.04847 | 0.04291 | 0.00556 |
| | $S_8$ | 0.04798 | 0.04208 | 0.00590 | 0.04857 | 0.04259 | 0.00598 | 0.04654 | 0.04036 | 0.00618 | 0.04686 | 0.04103 | 0.00583 |
| | $S_9$ | **0.04623** | **0.04079** | **0.00545** | **0.04659** | **0.04112** | **0.00546** | **0.04543** | **0.03951** | **0.00592** | **0.04537** | **0.04001** | **0.00536** |
| | $S_{10}$ | 0.04791 | 0.04155 | 0.00636 | 0.04845 | 0.04199 | 0.00645 | 0.04628 | 0.04008 | 0.00620 | 0.04608 | 0.04049 | 0.00559 |
| | $S_{11}$ | 0.05154 | 0.04404 | 0.00751 | 0.05247 | 0.04519 | 0.00727 | 0.04911 | 0.04191 | 0.00720 | 0.04878 | 0.04240 | 0.00637 |
| | $S_{12}$ | 0.04770 | 0.04154 | 0.00617 | 0.04855 | 0.04213 | 0.00643 | 0.04656 | 0.04009 | 0.00647 | 0.04681 | 0.04082 | 0.00599 |
| | $S_{13}$ | 0.04779 | 0.04215 | 0.00564 | 0.04857 | 0.04288 | 0.00569 | 0.04657 | 0.04039 | 0.00618 | 0.04672 | 0.04101 | 0.00571 |
| | $S_{14}$ | 0.04695 | 0.04124 | 0.00571 | 0.04758 | 0.04183 | 0.00575 | 0.04579 | 0.03971 | 0.00608 | 0.04597 | 0.04045 | 0.00553 |
| | $S_{15}$ | 0.05244 | 0.04609 | 0.00635 | 0.05377 | 0.04737 | 0.00640 | 0.04949 | 0.04301 | 0.00648 | 0.04939 | 0.04359 | 0.00580 |
| | $S_{16}$ | 0.05330 | 0.04537 | 0.00793 | 0.05399 | 0.04667 | 0.00731 | 0.04886 | 0.04267 | 0.00619 | 0.04895 | 0.04333 | 0.00561 |
| | $S_{17}$ | 0.05102 | 0.04474 | 0.00628 | 0.05239 | 0.04581 | 0.00658 | 0.04841 | 0.04246 | 0.00595 | 0.04862 | 0.04315 | 0.00547 |
| | $S_{18}$ | 0.04745 | 0.04174 | 0.00571 | 0.04775 | 0.04217 | 0.00558 | 0.04590 | 0.03991 | 0.00599 | 0.04603 | 0.04060 | 0.00542 |
| | $S_{19}$ | 0.04798 | 0.04182 | 0.00616 | 0.04867 | 0.04244 | 0.00623 | 0.04637 | 0.04023 | 0.00615 | 0.04636 | 0.04077 | 0.00559 |
| | $S_{20}$ | 0.04811 | 0.04184 | 0.00627 | 0.04880 | 0.04253 | 0.00628 | 0.04686 | 0.04039 | 0.00648 | 0.04710 | 0.04096 | 0.00614 |

## 4.3 Simulation Study Part III

Figure 4.8 illustrates three different scenarios of locations of observations for the discrete functional data. For the first scenario, there are ten fixed spatial locations (Panel $a$). The second scenario (Panel $b$) keeps the initial ten locations and adds ten new ones. Finally, in the third scenario (Panel $c$), the number of sites increases to 40, of which 20 are the exact locations as in Panel ($b$).

The observations of each curve are related to a specific point $t$ in the functional domain, which is the same for all trajectories. In this case, the number of measurements considered is $n = 150$, and the spacing between them has been generated from the $Beta(1, 2)$. The artificial data generation and fitting process are based on the same configurations used in the first phase of the simulation study (Spatial dependence study in Subsection 4.1.1). However, in this simulation phase, a single value of spatial variation $\kappa = 1$ is considered, along with the three levels of correlation, which are controlled by the decay parameter $\varphi = \{0.5, 1, 2\}$.

Figure 4.8: Grid of simulated locations.



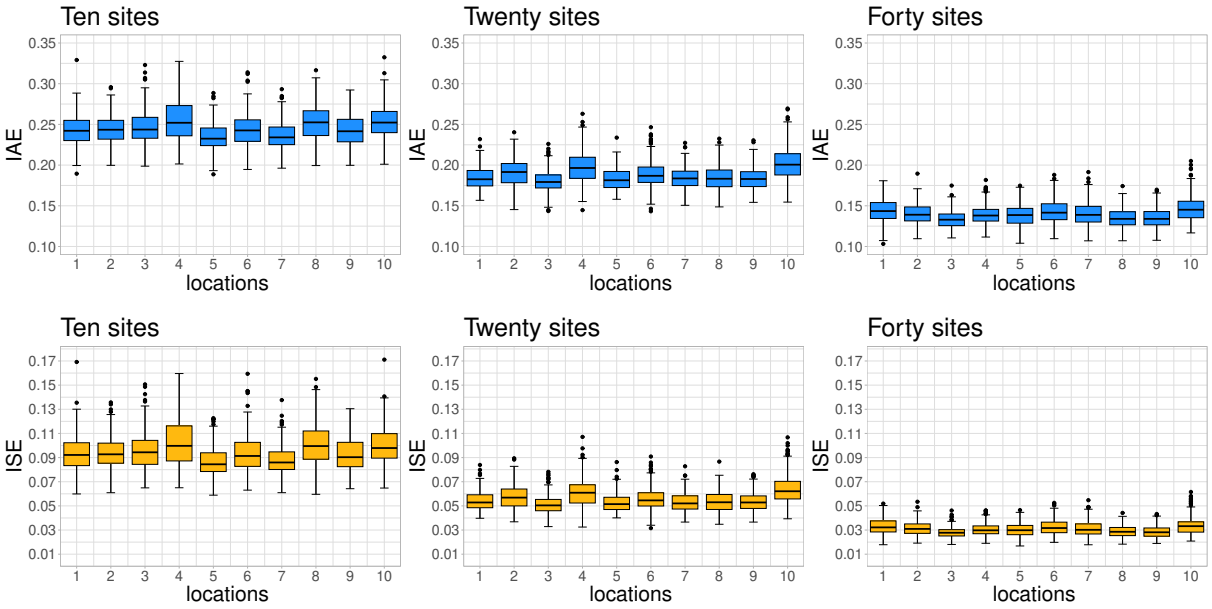

Source: Prepared by the author

### 4.3.1 Results of the Simulation Study Part III

This subsection focuses on the ($\kappa = 1$, $\varphi = 1$) scenario. The results of the other scenarios are presented in Appendix B. Figure 4.9 shows boxplots for the two discrepancy measures, IAE and ISE, summarizing the 250 MC replicates. In this case, functional data were generated and fitted using the $\mathcal{M}_{B_{4,4};\delta}$ model. By observing the left panels

(10 sites), it can be noted that most of the boxplots show a moderate dispersion, except for site 4, where a higher variability in the discrepancy indices is observed. Please note that location $S_4$ is somewhat isolated in the spatial configuration. The medians show a tendency around the values 0.25 and 0.09 (scenario with 10 sites), which indicates that these boxplots are at a higher level than those from the scenarios with 20 or 40 sites. Outliers are present in some boxplots, except those related to sites 4 and 9. As the number of locations increases, the number of near neighbors also increases. As a result, the indices IAE and ISE decrease, suggesting an improved fit of the curves. Moreover, the boxplots exhibit smaller dispersion, especially in the case of 40 sites.

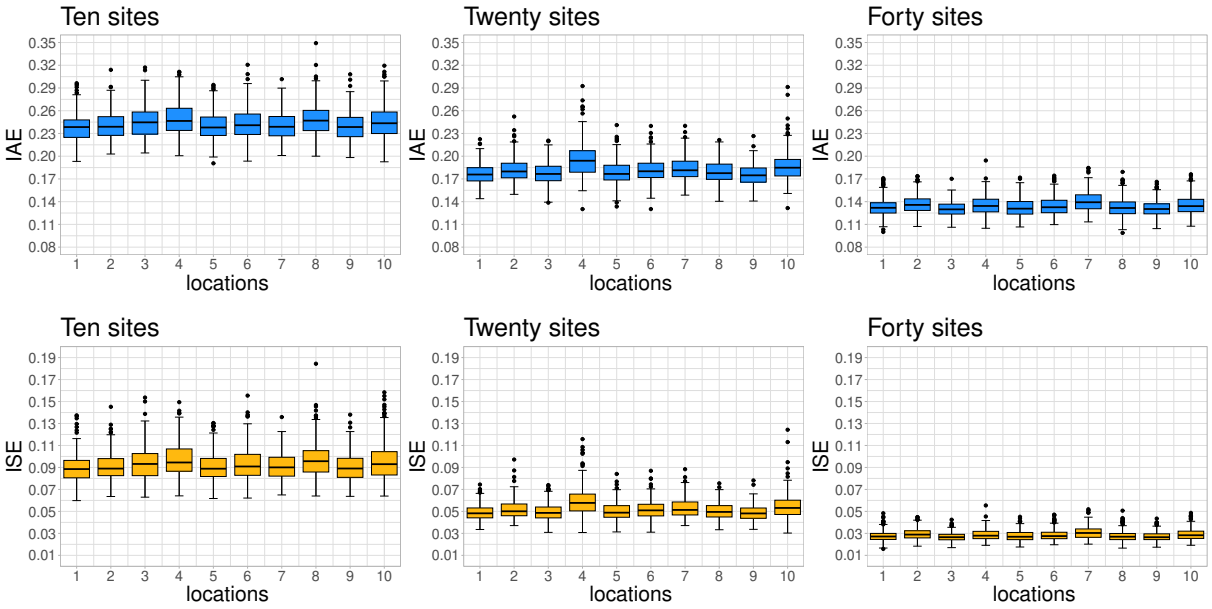
Figure 4.9: Comparison of IAE and ISE discrepancy measures for functions estimated using the $\mathcal{M}_{B_{4,4};\delta}$. This analysis is conducted in different spatial configurations where the number of sites is 10, 20, or 40.




Source: Prepared by the author

Figure 4.10 illustrates the results obtained from the $\mathcal{M}_{BP_3;\delta}$; data is generated and fitted assuming this model. Initially, the measurements of the ten reference sites displayed moderate variability, with median values of approximately 0.23 (IAE) and 0.09 (ISE). Additionally, a few outliers are observed. However, as the number of neighbors for the reference locations increases, the IAE and ISE values in each replicate tend to decrease. In addition, the MC dispersions become considerably lower. This trend is particularly pronounced when there are 40 sites, resulting in medians around 0.14 for IAE and 0.03 for ISE. These findings indicate a better performance of the smoothed curves. In conclusion, when comparing the results from B-spline and BP, one can see similar behaviors; therefore, no significant distinction is detected between these model versions.

Figure 4.10: Comparison of IAE and ISE discrepancy measures for functions estimated using the $\mathcal{M}_{BP_3;\delta}$ model. This analysis is conducted in different spatial configurations where the number of sites is 10, 20, or 40. The analysis is focused on the 10 red locations.




Source: Prepared by the author

To perform a detailed and general analysis of the 250 MC replicas, the MISE, IV, and ISB measures are used, which can be found in Table 4.7. In these measures, it can be observed that in the scenario of 10 sites (reference case) using model $\mathcal{M}_{B_{4,4};\delta}$, the curve of location 5 presents smaller measurements (MISE, IV, and ISB) compared to the others. This indicates that the mean of the smoothed functions is closer to the target function, and its variance is small, suggesting greater consistency and stability in the results of each sample. Furthermore, the bias is low, meaning it tends to approximate the true function without deviating. On the other hand, the curve of location 4 has the highest MISE value along with the other indices, indicating that the smoothed functions have a higher error and, therefore, are less accurate.

As the number of close neighbors increases for the 10 reference sites, a reduction is observed in all measures (MISE, IV, and ISB) for all curve means. This suggests that having multiple nearby trajectories improves the fit of the estimated functions. A clear illustration of this is found in Figure 4.8 ($c$) for the scenario involving 40 sites. Here, it's evident that site 3 exhibits the lowest values for MISE, IV, and ISB, while site 10 displays the highest. This distinction arises from the fact that site 3 has a more significant number of nearby trajectories, whereas site 10 has fewer such trajectories.

In the case of model $\mathcal{M}_{BP_3;\delta}$, a similar analysis to the three scenarios considered for model $\mathcal{M}_{B_{4,4};\delta}$ is conducted. In the first scenario, Figure 4.8 ($a$), by examining the discrepancy measures in Table 4.7, it is observed that site 1 has the lowest value of the indices while site 4 has the highest. In the second scenario, Figure 4.8 ($b$), performance

changes are observed when ten additional geographic locations are included. The curves for sites 9 and 4 are now identified as the best and worst performers, respectively. Lastly, in the third scenario, Figure 4.8 $(c)$, twenty new locations are added along with the previous ones. In this case, the smoothed functions of site 3 indicate a better approximation to the observed data, while site 7 presents the worst fit. This indicates that including nearby neighbors in the estimation process leads to improvement.

Table 4.7: Comparison of the MISE, IV, and ISB discrepancy measures for curves estimated using the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$. These models have a fixed decay parameter value of $\varphi = 1$ and a level of variation $\kappa = 1$. The evaluation was done for three scenarios with 10, 20, and 40. geographic sites.

| | | $m = 10$ | | | $m = 20$ | | | $m = 40$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Site | MISE | IV | ISB | MISE | IV | ISB | MISE | IV | ISB |
| | | | | | $\kappa = 1, \varphi = 1$ | | | | | |
| $\mathcal{M}_{B_{4,4};\delta}$ | $S_1$ | 0.09371 | 0.06874 | 0.02497 | 0.05404 | 0.04267 | 0.01138 | 0.03303 | 0.02684 | 0.00619 |
| | $S_2$ | 0.09397 | 0.06854 | 0.02543 | 0.05739 | 0.04511 | 0.01228 | 0.03138 | 0.02710 | 0.00428 |
| | $S_3$ | 0.09553 | 0.07655 | 0.01898 | **0.05117** | **0.04416** | **0.00701** | **0.02796** | **0.02568** | **0.00228** |
| | $S_4$ | **0.10281** | **0.07845** | **0.02436** | 0.06159 | 0.04912 | 0.01247 | 0.03055 | 0.02725 | 0.00330 |
| | $S_5$ | **0.08695** | **0.06802** | **0.01893** | 0.05270 | 0.04492 | 0.00777 | 0.03043 | 0.02673 | 0.00370 |
| | $S_6$ | 0.09335 | 0.07068 | 0.02267 | 0.05598 | 0.04609 | 0.00989 | 0.03229 | 0.02774 | 0.00455 |
| | $S_7$ | 0.08784 | 0.06878 | 0.01906 | 0.05308 | 0.04487 | 0.00822 | 0.03119 | 0.02800 | 0.00319 |
| | $S_8$ | 0.10129 | 0.07683 | 0.02446 | 0.05353 | 0.04459 | 0.00893 | 0.02897 | 0.02580 | 0.00316 |
| | $S_9$ | 0.09271 | 0.07013 | 0.02258 | 0.05310 | 0.04494 | 0.00816 | 0.02870 | 0.02612 | 0.00258 |
| | $S_{10}$ | 0.09994 | 0.07004 | 0.02989 | **0.06398** | **0.04944** | **0.01454** | **0.03377** | **0.02810** | **0.00567** |
| $\mathcal{M}_{BP_3;\delta}$ | $S_1$ | **0.08895** | **0.07392** | **0.01504** | 0.04903 | 0.04239 | 0.00663 | 0.02774 | 0.02530 | 0.00245 |
| | $S_2$ | 0.09100 | 0.07353 | 0.01747 | 0.05208 | 0.04341 | 0.00867 | 0.02934 | 0.02557 | 0.00377 |
| | $S_3$ | 0.09450 | 0.07708 | 0.01741 | 0.04960 | 0.04204 | 0.00756 | **0.02684** | **0.02437** | **0.00247** |
| | $S_4$ | **0.09745** | **0.07854** | **0.01892** | **0.05978** | **0.04636** | **0.01341** | 0.02882 | 0.02508 | 0.00374 |
| | $S_5$ | 0.09088 | 0.07346 | 0.01742 | 0.05032 | 0.04342 | 0.00690 | 0.02780 | 0.02526 | 0.00255 |
| | $S_6$ | 0.09312 | 0.07414 | 0.01899 | 0.05174 | 0.04425 | 0.00749 | 0.02851 | 0.02573 | 0.00278 |
| | $S_7$ | 0.09076 | 0.07309 | 0.01767 | 0.05300 | 0.04315 | 0.00985 | **0.03087** | **0.02626** | **0.00461** |
| | $S_8$ | 0.09715 | 0.07983 | 0.01732 | 0.05068 | 0.04294 | 0.00774 | 0.02759 | 0.02468 | 0.00291 |
| | $S_9$ | 0.09000 | 0.07380 | 0.01620 | **0.04878** | **0.04223** | **0.00654** | 0.02722 | 0.02487 | 0.00235 |
| | $S_{10}$ | 0.09521 | 0.07459 | 0.02062 | 0.05524 | 0.04661 | 0.00863 | 0.02925 | 0.02636 | 0.00289 |

Finally, the current chapter containing results based on artificial data is complete. In the studies developed here, it was possible to verify the quality of the proposed models to adjust functional datasets with spatial dependence and irregular spacing between observations. The next chapter shows a real analysis involving two datasets related to the environmental variable.

# Chapter 5

# Real Data Application

This chapter focuses on the presentation and exploration of two real applications that illustrate the modeling approach proposed in this thesis. Temperature and particulate matter (PM10) are two interconnected aspects of environmental conditions that significantly impact our daily lives. Temperature refers to the measure of heat in the atmosphere, while PM10 is a component of atmospheric pollution with a crucial physical characteristic: its diameter. Understanding the implications of these factors is vital as they can profoundly affect our health and well-being.

Particulate matter, commonly called PM10, consists of tiny particles suspended in the air. These particles have a diameter of 10 micrometers or less, making them small enough to be inhaled. This characteristic poses a significant risk to human health as it can penetrate deep into our respiratory system, causing damage to tissues and organs. Moreover, PM10 can serve as a carrier for bacteria and viruses, potentially exacerbating the spread of diseases.

Extensive research has established a positive relationship between exposure to PM10 and various adverse health outcomes. Studies by Greenbaum et al. (2001) and Paldy et al. (2006) have highlighted the association between PM10 exposure and an increased risk of respiratory and cardiovascular diseases, cancer, influenza, and asthma. These findings underscore the importance of monitoring and mitigating the levels of particulate matter in the air we breathe.

Temperature, on the other hand, is a fundamental aspect of weather and climate. It refers to the measure of hotness or coldness of the atmosphere, influenced by factors such as solar radiation, air pressure, and wind patterns. Temperature variations have far-reaching effects on human activities, ecosystems, and the overall functioning of the planet. Changes in Temperature patterns, particularly global warming, have raised concerns worldwide. Rising Temperatures can lead to heatwaves, droughts, and altered precipitation patterns, impacting agricultural productivity, water resources, and human health. Heatwaves, in particular, pose a significant risk to vulnerable populations, including the elderly and those with underlying health conditions.

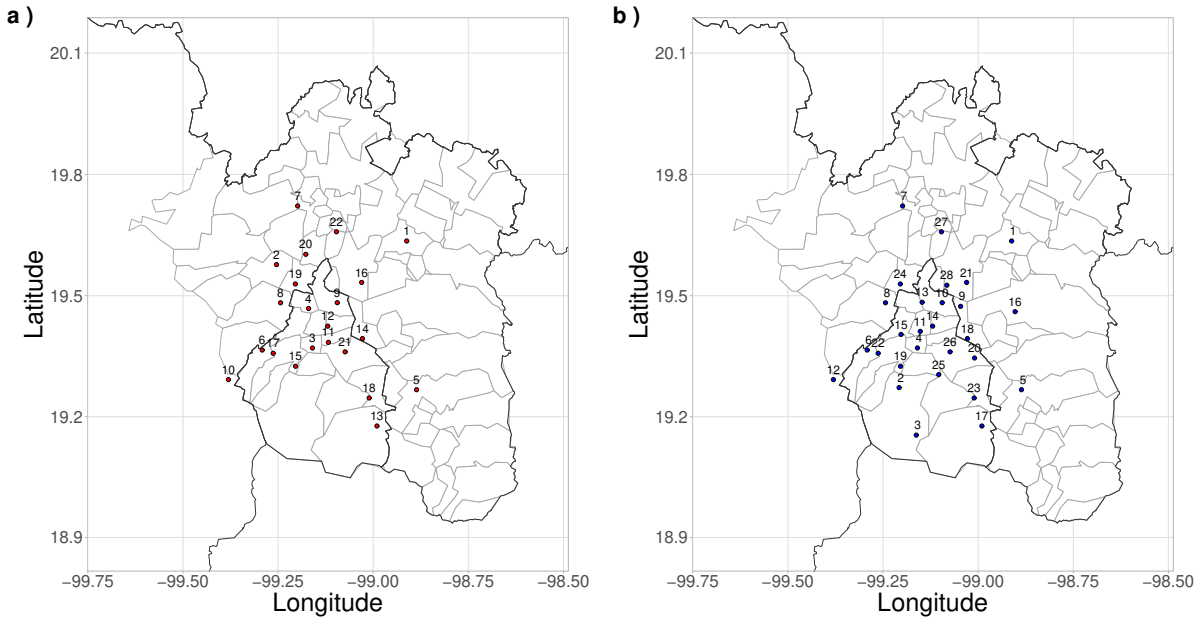
The main focus of this study does not involve evaluating the association between temperature and PM10. These two variables will be explored separately in the analyses using the proposed models for spatial functional data. The study of these data sets is

divided into three parts. Section 5.1 explains the origin of the samples and the strategy used to reorganize the functional domain. Section 5.2 provides a detailed description of the data, which is crucial for statistical analysis and understanding the problem. In Section 5.3, several models (B-spline or BP versions) are fitted and compared with each other to determine a final result. All the fitted models employ four chains, with a burn-in value of 2,500 and storage of 2,500 posterior values for each chain.

## 5.1 Data Origin

For two years, data on air quality in Mexico City was collected through a monitoring network. This data corresponds to consecutive hours, from 1:00 a.m. on January 1, 2021, to midnight, on December 31, 2022. The measurement was carried out at 22 environmental stations in different parts of the city, as shown in Figure 5.1 ($a$). These stations are part of the air quality network known as RAMA (*Red Automática de Monitoreo Atmosférico*) and monitor particles up to 10 micrometers ($\mu$m) in size every hour, among other things. The data can be accessed freely through the internet on a webpage[1] maintained by the Mexico City government. The mentioned dataset, which includes the period $2021 - 2022$ selected for this study, is just a tiny portion of the complete dataset available at the provided address.

Figure 5.1: Map of Mexico City with the monitoring stations: ($a$) sampled sites of PM10 and ($b$) sampled sites of Temperature.




Source: Prepared by the author

[1] `http://www.aire.cdmx.gob.mx`

Mexican Official Standard NOM-025-SSA1-2014 (*Norma Oficial Mexicana*) establishes the concentration limits for suspended particulate matter PM10 in ambient air to protect the population's health. It also provides the criteria for assessing such concentration. Specifically, it establishes a 24-hour average limit (acute exposure) of 75 $g/m^3$ and an average annual limit for chronic exposure of 40 $g/m^3$. These values are considerably higher than the World Health Organization (WHO) air quality guidelines, which establish a limit of 50 $g/m^3$.

The Temperature data in Mexico City were collected during the same period as the PM10 data. These measurements were made at 28 stations belonging to the Meteorological Monitoring Network REDMET (*Red de Meteorología y Radiación Solar*), located in different areas of the city and its surroundings; see Figure 5.1 (*b*). Some of these stations are common with RAMA.

**Functional Domain Reorganization**

The PM10 and Temperature sets consist of 17,520 observations collected at discrete equidistant time points (hours) for each station. However, both data sets have missing values. Since this work focuses on samples with irregular distances between measurement points in the domain, it is necessary to incorporate this feature naturally into the mentioned data. To achieve this, the following steps are required:

- The total number of missing data per month for 2021 and 2022, which comprise the study period for each sample, is analyzed separately. It is important to note that the PM10 and Temperature datasets comprise 22 and 28 respective stations, each with 8,760 hours recorded.
- Next, the quartiles are calculated for each year, which provides information on the dispersion of the months with the least and most missing data. Based on this, the pattern of distancing between the observations recorded for each month is established, as shown in Table 5.1. Suppose a particular month has a percentage of missing data between 25.1% and 50% of its hours. In that case, the analysis will consider a spacing that summarizes the information for 48 hours (2 days).

Table 5.1: Ranking of the months with the most missing data

| Classification of the months | Recorded in intervals Hours | Domain in days |
|---|---|---|
| $0\% - 25\%$ | 24 | 1 |
| $25.1\% - 50\%$ | 48 | 2 |
| $50.1\% - 75\%$ | 92 | 4 |
| $75.1\% - 100\%$ | 192 | 8 |

- Once the above configuration is established, the median of the data recorded in intervals of 24, 48, 96, and 192 hours is calculated to work with a functional domain of days. The last day of the interval for which the summarization via median is performed will be identified as the moment $t_i$ in the functional domain where the observation is recorded. If only missing observations can be found within the interval to be summarized, then the location $t_i$ will be considered a point with a missing value. The presence o missing values in the series is not a problem for the proposed models since they can handle this situation, as shown in Subsection 4.1.1.

After following these steps, the resulting samples contain curves with the same irregular spacing configuration between discrete points in the domain. In addition, each station in the monitoring network for PM10 and Temperature has 342 and 344 observations, which include both observed and missing data. Note that these sizes of series in the functional domain are higher than the value of 150 explored in the simulated study of Chapter 4. Remember that the results indicated that better approximations are obtained by increasing the amount of data in the functional domain; see Section 4.2.

## 5.2 Descriptive Analysis

### Particulate matter - PM10

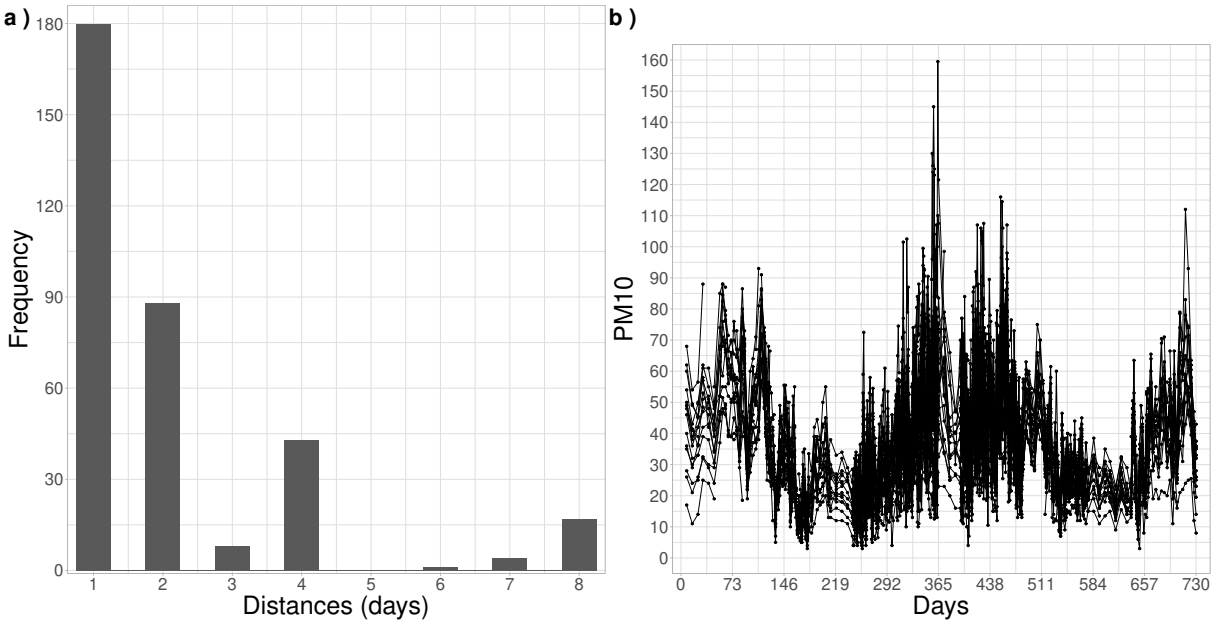
Here, the dataset consisting of 7,528 measurements of PM10 collected from 22 monitoring stations is analyzed through descriptive statistics. Of these measurements, 18.5%(1,389) are considered missing data, distributed among all stations, as shown in Table 5.2. Station 9 has the highest number of missing data points, 291, while Station 12 has the fewest, with only 10 missing observations. Each station has 342 irregularly spaced measurements. Figure 5.2 $(a)$ shows that a spacing of 1 day is the most common among consecutive observations in the functional domain. The configuration of 2 days is the second most frequent, followed by a spacing of 4 days. As can be seen, the largest distance between observations in this study involves 8 days. It is important to note that this spacing pattern reflects the behavior of a $Beta(1, 2)$, mentioned previously in Section 4.1.

Table 5.2: Number of missing data for each station.

| | Automatic Atmospheric Monitoring Network | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| Missing data | 42 | 36 | 18 | 71 | 62 | 117 | 50 | 31 | 291 | 40 | 86 | 10 | 77 | 44 | 36 | 44 | 73 | 39 | 32 | 83 | 38 | 69 |

Throughout the year, Mexico City undergoes significant fluctuations in PM10 concentration. Generally, the highest levels of particles are observed during the dry season, which lasts from November to April. This corresponds to days 1 to 120, 305 to 485, and 670 to 730 of the two years considered in this application, as shown in Figure 5.2 $(b)$.

Figure 5.2: Descriptive analysis: ($a$) bar-plot of measurement spacing and ($b$) PM10 Curves concerning non-missing observations for each station.




Source: Prepared by the author

During this time, the absence of rainfall and stable atmospheric conditions contribute to the buildup of pollutants in the city. On the contrary, the summer months, such as June, July, and August (days 152 to 243 and 517 to 608), typically exhibit lower PM10 values. Throughout this season, the city benefits from enhanced dispersion of pollutants due to favorable weather conditions, including stronger winds and increased solar radiation. Moreover, rainfall also aids in purifying the air and reducing the concentration of suspended particulate pollutants.

Table 5.3 presents descriptive statistics regarding the behavior of the PM10 variable in Mexico City and its surrounding areas. Upon analyzing these statistics, various ranges and levels of dispersion can be observed at each station. For instance, Station 22 exhibits the highest standard deviation (26.9), indicating high variability. On the other hand, Stations 10 and 17 demonstrate low standard deviations (8.6 and 10.6, respectively), suggesting lesser variability. Furthermore, the medians and means at most stations are similar, implying that the distributions of the discrete data are likely symmetric or unaffected by significant outliers. Regarding the minimum and maximum values, they provide information about the range of the data. The stations with the lowest minimum value are 5 and 10 (3.0), followed by 2 (4.0). In terms of the maximum value, Station 22 has the highest maximum (159.5), followed by 5 (124.0) and 18 (123.0). Note that the PM10 values are always positive, and the smallest recorded value is 3.0, which is not very close to zero.

Table 5.3: Descriptive statistics of non-missing observations at each station.

| Station | Median | Mean | Standard Deviation | Minimum | Maximum | Range |
|---|---|---|---|---|---|---|
| 1 | 38.0 | 40.7 | 19.2 | 6.5 | 105.0 | 98.5 |
| 2 | 27.8 | 28.8 | 11.5 | 4.0 | 63.0 | 59 |
| 3 | 30.5 | 31.7 | 12.2 | 7.0 | 68.5 | 61.5 |
| 4 | 41.0 | 42.7 | 14.5 | 10.0 | 98.5 | 88.5 |
| 5 | 43.2 | 45.3 | 22.7 | 3.0 | 124.0 | 121 |
| 6 | 30.5 | 30.9 | 11.0 | 9.0 | 79.0 | 70 |
| 7 | 40.0 | 41.1 | 16.0 | 10.0 | 100.0 | 90 |
| 8 | 36.0 | 36.9 | 14.5 | 6.0 | 84.5 | 78.5 |
| 9 | 48.5 | 48.0 | 14.6 | 12.0 | 76.0 | 64 |
| 10 | 16.5 | 18.1 | 8.6 | 3.0 | 59.5 | 56.5 |
| 11 | 34.0 | 35.8 | 14.9 | 7.0 | 79.5 | 72.5 |
| 12 | 36.8 | 37.5 | 15.6 | 5.0 | 77.5 | 72.5 |
| 13 | 29.5 | 31.0 | 13.7 | 4.0 | 81.0 | 77 |
| 14 | 41.8 | 42.2 | 17.9 | 5.0 | 92.5 | 87.5 |
| 15 | 26.5 | 28.2 | 11.7 | 6.0 | 74.5 | 68.5 |
| 16 | 43.5 | 44.0 | 16.9 | 7.0 | 97.0 | 90 |
| 17 | 25.0 | 26.6 | 10.6 | 5.0 | 86.5 | 81.5 |
| 18 | 38.0 | 38.7 | 18.7 | 4.0 | 123.0 | 119 |
| 19 | 42.5 | 44.2 | 14.7 | 14.0 | 98.0 | 84 |
| 20 | 45.0 | 47.2 | 20.6 | 9.0 | 116.0 | 107 |
| 21 | 39.0 | 40.4 | 15.0 | 6.0 | 83.0 | 77 |
| 22 | 48.0 | 52.2 | 26.9 | 9.0 | 159.5 | 150.5 |

**Temperature**

For this application, 9,632 Temperature measurements are available from 28 stations in Mexico City and its surroundings. Of these measurements, 1,564 data points are missing (16.23% of the total), distributed among all stations, as shown in Table 5.4. Station 2 has the highest amount of missing data, with a total of 314, while Stations 24, 12, and 9 have 0, 1, and 1 missing data, respectively.

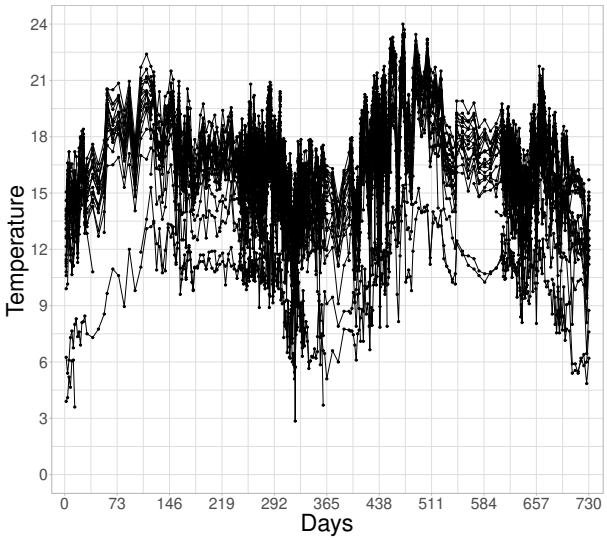
Each station provides 344 individual measurements, which are irregularly spaced in time. The spacing pattern is similar to the previous application, where half of the data is collected at one-day intervals, followed by two-day periods, until reaching a few measurements taken in a maximum interval of 8 days. It is essential to mention that the frequency varies slightly concerning the PM10 data. The spacing pattern of the observations can be considered almost the same between PM10 and Temperature; therefore, the bar-plot is omitted (see Figure 5.2 ($a$)) again.

Table 5.4: Number of missing data for each Temperature monitoring station.

| | Automatic Atmospheric Monitoring Network | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| Missing data | 22 | 314 | 60 | 29 | 83 | 129 | 47 | 120 | 1 | 11 | 6 | 1 | 11 | 53 | 34 | 55 | 48 | 20 | 31 | 110 | 24 | 128 | 14 | 0 | 6 | 102 | 35 | 70 |

Figure 5.3 shows the Temperature behavior in Mexico City. The months with the highest values are April, May, and June, corresponding to days 91 to 181 and 456 to 546, during spring and early summer. During this period, the Temperature ranges between 20 and 24 degrees Celsius. It is important to note that the city's altitude, approximately 2,240 meters above sea level, moderates the climate and prevents Temperatures from reaching extreme levels. On the other hand, the coldest months in the Mexican capital are December, January, and February. Specifically, this period includes the dates from 1 to 59, 335 to 424, and 700 to 730. During these months, the minimum Temperatures can drop below 10 degrees Celsius and occasionally even approach 0 degrees Celsius.

Figure 5.3: Temperature curves concerning non-missing observations for each station.




Source: Prepared by the author

Table 5.5 displays the statistical measures for the Temperature variable at each station. Upon analyzing the data, one can observe that, in general, the stations exhibit similar values for both the median and mean, indicating a relatively symmetrical distribution. Regarding the standard deviation, Stations 7 (2.8), 20 (2.5), and 27 (2.4) have the highest values, indicating more variability in the data. Conversely, Stations 6 (1.6) and 5 (1.8) have the lowest standard deviations, indicating less dispersion. In addition, the range provides information about the difference between the maximum and minimum values in the data. Based on this, Station 7 (18.6) exhibits the highest range, indicating a significant variation in the data. In contrast, Stations 6 (9.0) and 2 (9.2) have the lowest ranges, suggesting less variability. It is worth noting that the last station has lim-

ited information, with only 30 recorded observations. Please note that the Temperature observations are positive, but a negative value is not unrealistic in this application.

Table 5.5: Descriptive statistics of non-missing observations at each Temperature monitoring station.

| Station | Median | Mean | Standard Deviation | Minimum | Maximum | Range |
|---|---|---|---|---|---|---|
| 1 | 15.2 | 15.1 | 2.0 | 7.9 | 20.0 | 12.2 |
| 2 | 11.6 | 11.6 | 2.1 | 6.4 | 15.7 | 9.2 |
| 3 | 10.9 | 10.5 | 2.2 | 3.6 | 15.3 | 11.7 |
| 4 | 17.0 | 17.2 | 2.0 | 11.4 | 23.1 | 11.7 |
| 5 | 16.3 | 16.3 | 1.8 | 10.2 | 21.2 | 11.0 |
| 6 | 13.4 | 13.5 | 1.6 | 8.6 | 17.5 | 9.0 |
| 7 | 15.4 | 15.1 | 2.8 | 2.9 | 21.4 | 18.6 |
| 8 | 15.5 | 15.5 | 2.0 | 9.6 | 20.8 | 11.2 |
| 9 | 18.2 | 18.3 | 1.9 | 13.1 | 24.0 | 10.9 |
| 10 | 17.6 | 17.6 | 2.0 | 11.5 | 23.3 | 11.8 |
| 11 | 17.1 | 17.3 | 1.9 | 12.2 | 22.6 | 10.5 |
| 12 | 10.5 | 10.0 | 2.1 | 5.1 | 15.4 | 10.3 |
| 13 | 17.0 | 17.2 | 2.1 | 12.2 | 22.3 | 10.1 |
| 14 | 17.3 | 17.5 | 2.1 | 12.2 | 23.5 | 11.2 |
| 15 | 16.8 | 17.0 | 1.9 | 11.8 | 22.8 | 10.9 |
| 16 | 16.8 | 16.9 | 2.1 | 10.6 | 22.5 | 11.9 |
| 17 | 14.7 | 14.8 | 2.0 | 10.2 | 21.0 | 10.8 |
| 18 | 17.2 | 17.3 | 2.0 | 12.2 | 23.4 | 11.2 |
| 19 | 15.9 | 16.1 | 2.1 | 10.6 | 22.0 | 11.4 |
| 20 | 13.2 | 13.6 | 2.5 | 7.1 | 20.6 | 13.5 |
| 21 | 17.2 | 17.4 | 2.1 | 11.9 | 23.4 | 11.5 |
| 22 | 14.4 | 14.8 | 2.0 | 10.4 | 21.3 | 10.9 |
| 23 | 15.7 | 15.6 | 2.0 | 8.0 | 21.5 | 13.5 |
| 24 | 16.3 | 16.5 | 2.1 | 11.2 | 22.5 | 11.3 |
| 25 | 15.9 | 15.9 | 1.9 | 9.8 | 22.0 | 12.2 |
| 26 | 16.9 | 17.0 | 2.0 | 12.0 | 23.2 | 11.2 |
| 27 | 15.6 | 15.7 | 2.4 | 8.1 | 22.5 | 14.4 |
| 28 | 17.0 | 17.1 | 1.9 | 12.3 | 22.4 | 10.1 |

## 5.3 Modeling Approaches

In order to perform the modeling of the PM10 and Temperature data, it is necessary to reorganize the function's domain because the original spacing is equidistant. Hence, the initial step involves applying the strategy introduced in Section 5.1 to obtain irregularly

spaced samples. Once the domain has been restructured, it is essential to note that the proposed models assume $d_i \in (0, 1)$. For this requirement, the distances between adjacent time points are divided by the maximum separation distance between discrete measurements, which in this case is 8 days. This new scaling is relevant for the random effects component, $\delta_i$, which exhibits an autoregressive structure.

In this section, the models mentioned in Table 4.1 are used to fit the two applications, PM10 and Temperature. The models' structure comprises seven basis B-spline functions and four basis BP, similar to the previous simulation study. Remember that these choices configure models with the smallest number of basis explored in Chapter 4. In addition, the modeling allows treating the missing values of the samples as a vector of unknown parameters, which are estimated through the Bayesian approach with priors defined by the distribution attributed to the observations. The imputation of missing values is a crucial analysis step as it accounts for the uncertainty related to missing data and improves the accuracy of model parameter estimates. The configuration of the parameters and hyperparameters and their prior distributions are identical to that used in the simulation study. The imputation of missing values is a crucial analysis step as it accounts for the uncertainty related to missing data and improves the accuracy of model parameter estimates. The configuration of the parameters and hyperparameters and their prior distributions are identical to that used in the simulation study.

A sensitivity analysis was conducted to find the best possible values for the Gaussian covariance function's decay parameter $\varphi$. This study closely examined the estimated trajectories to determine the magnitudes and variabilities of the curves. The findings revealed that values of 1 for PM10 and 1.2 for Temperature data were the most suitable for fitting the observed data. For further information, please refer to Appendix D. The spatial variation parameter $\kappa$ is treated as unknown and estimated in the inference.

In order to evaluate the models using a higher number of basis functions, one set of 13 is selected. This decision is based on maintaining a reasonable computational cost, which is adequate for the execution of the MCMC on a standard computer. At the same time, this number of bases makes it possible to maintain a complexity in the models that do not hinder subsequent analysis. It is important to emphasize that the limitations of not using a higher set of bases are related to computational complexity and numerical accuracy.

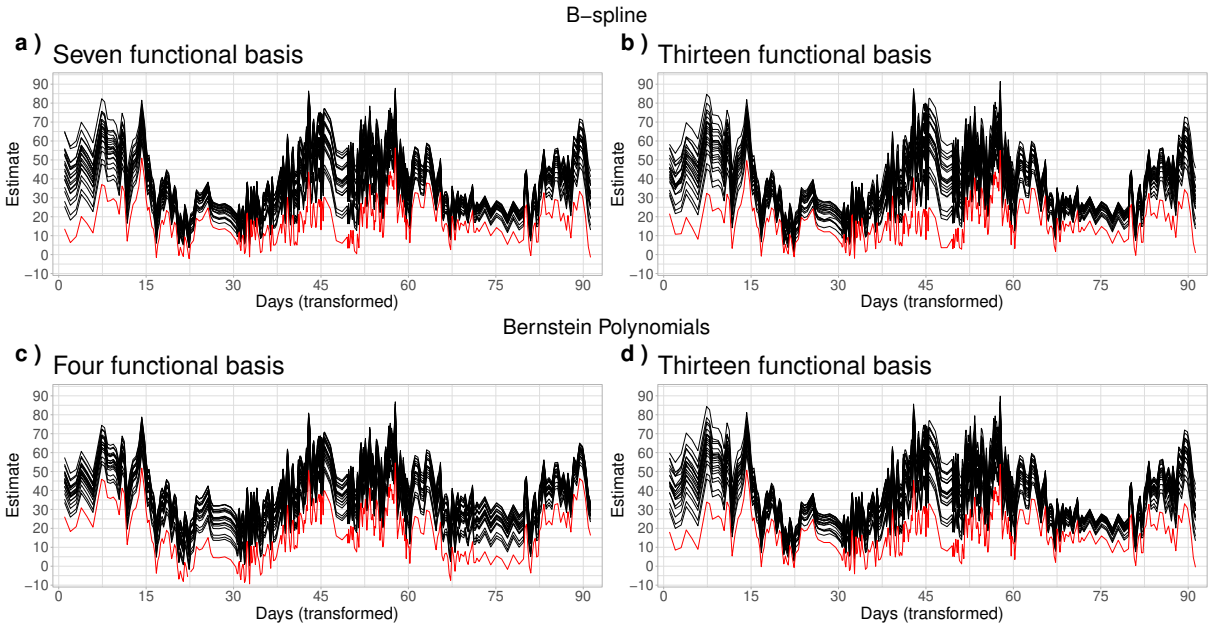
To evaluate the performance of the models, the discrepancy index IAE is used for the measurements discretely observed in each curve that make up the real datasets. Its purpose is to measure the difference between the observed values and the values estimated by the models over time. A lower IAE value indicates a better fit of the model to the observed data. In addition, three measures are used in this work to compare and evaluate the fitted models. These are the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), the Logarithm of Pseudo-Marginal Likelihood (LPML) (Gelman et al., 2013),

and the Watanabe-Akaike Information Criterion (WAIC) (Watanabe and Opper, 2010). The last two metrics are multiplied by -2 to standardize the measurements alongside the DIC. These measures help us identify the best-fitted model among the options. The lower the values of these metrics, the better the model performs in fit and overall performance against the observed data.

### Analysis of PM10 Data

Figure 5.4 displays the smoothed curves of each fitted model. Upon examining the panels, it becomes apparent that most of the estimated functions follow the patterns of the target function sample in Figure 5.2 ($b$). However, it is crucial to note that all models slightly underestimate some observations with high PM10 levels, specifically in the range of 100 to 160. This discrepancy can be seen by comparing Figure 5.2 ($b$) with the panels in Figure 5.4. In addition, in each panel of Figure 5.4, it has been identified that some estimates of the PM10 index are negative. This fact occurs only for Station 10 (red path). For example, see Panels ($a$) and ($b$), which displays $\mathcal{M}_{B_{4,l};\delta}$ with $l = 4$ and 10 subintervals (7 and 13 bases). In this case, 7 and 6 negative values are estimated in ($a$) and ($b$), respectively. On the other hand, model $\mathcal{M}_{BP_p;\delta}$ (with degree $p$ =4 and 13) shows 36 and 7 negative values with 4 and 13 BP as basis functions, respectively.

Figure 5.4: Smoothed PM10 curves (estimated). In Panels ($a$) and ($b$), $\mathcal{M}_{B_{4,l};\delta}$ was applied using 4 and 10 subintervals, respectively. In Panels ($c$) and ($d$), consider the $\mathcal{M}_{BP_p;\delta}$ with BP of degrees 3 and 12.



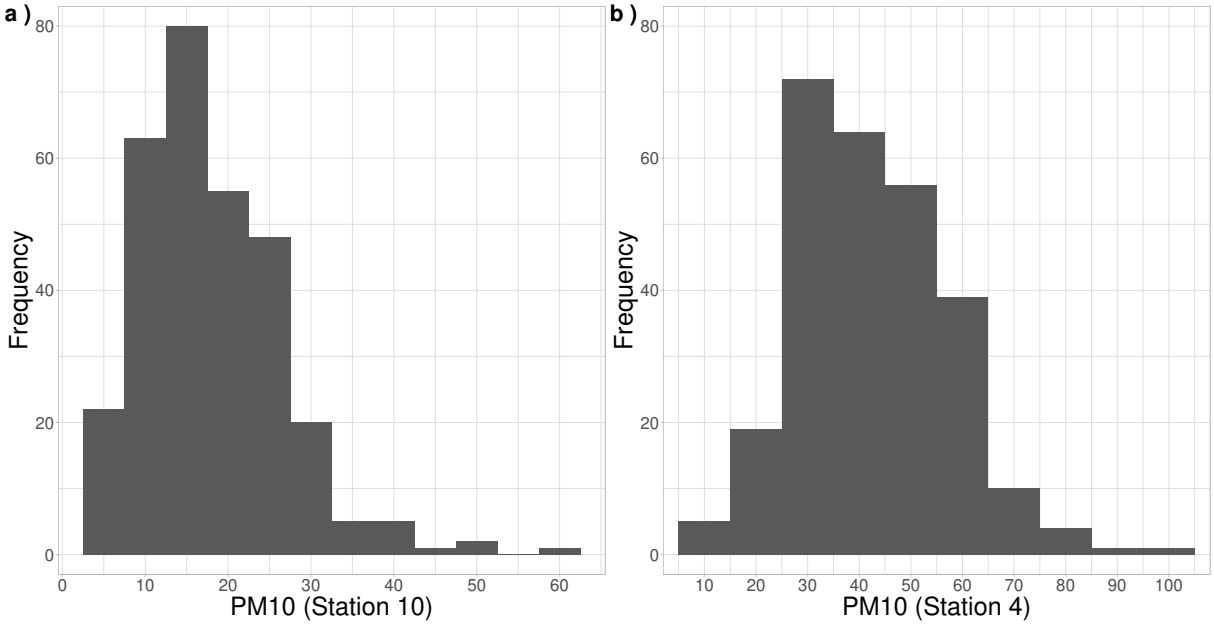

Source: Prepared by the author

The estimated negative PM10 values can be attributed to the observations being close to zero. Several factors contribute to this proximity. Firstly, Station 10 is located

on the outskirts of Mexico City in a rural area characterized by abundant vegetation and a lack of public transportation flow or industrial activities. These conditions generally result in lower levels of PM10 recorded at this particular station. Consequently, when applying a Gaussian model to the data, it is highly probable to obtain negative values under these circumstances. It is important to note that this does not imply that the model is inappropriate. It is worth mentioning that most stations have PM10 values far from zero, and this estimation issue does not occur for them. However, when using the Gaussian model, the analyst should be aware of this problem when evaluating a series with observations close to zero.

Figure 5.5 represents the distributions of the observed data from Stations 10 and 4. In contrast with Station 10, Station 4 is approximately located in the center of the map with several neighbors around; see Figure 5.1 ($a$). As can be seen, the two histograms have slightly different shapes, with Panel ($a$) indicating a distribution whose left tail is closer to the threshold of 0 compared to the graph in Panel ($b$). Therefore, there will be no issue of negative response estimation for most stations that exhibit a left-tailed behavior similar to that of Station 4.

Figure 5.5: Comparison of PM10 levels at Stations 10 and 4.




Source: Prepared by the author

Table 5.6 displays the IAE discrepancy values for each fitted model. It is important to note that this index is calculated based solely on the measurements observed at each station. Upon comparing the performance of the curve sets, the following results emerge for models $\mathcal{M}_{B_{4,\mathbf{4};\delta}}$ and $\mathcal{M}_{B_{4,\mathbf{10};\delta}}$. Among the 22 curves estimated by $\mathcal{M}_{B_{4,\mathbf{10};\delta}}$, 13 of them exhibit a lower IAE index than the $\mathcal{M}_{B_{4,\mathbf{4};\delta}}$. This indicates that these curves are closer to the target functions. Conversely, for the $\mathcal{M}_{BP_{\mathbf{12};\delta}}$, the results demonstrate that

17 curves in the sample outperform the $\mathcal{M}_{BP_4;\delta}$. Furthermore, when comparing these two models, which exhibit the highest number of well-fitted curves, it becomes evident that the $\mathcal{M}_{B_{4,\mathbf{10}};\delta}$ outperforms the $\mathcal{M}_{BP_{\mathbf{12}};\delta}$ in 17 instances with lower IAE measures. This suggests a higher accuracy in the functional curves. In summary, the best fit is obtained assuming a higher number of bases and, comparing the B-spline against BP; the B-spline shows a better performance.

Table 5.6: IAE discrepancy measurement of the smoothed PM10 curves obtained from the proposed models.

| | IAE | | IAE | |
|---|---|---|---|---|
| Site | $\mathcal{M}_{B_{4,\mathbf{4}};\delta}$ | $\mathcal{M}_{B_{4,\mathbf{10}};\delta}$ | $\mathcal{M}_{BP_{\mathbf{3}};\delta}$ | $\mathcal{M}_{BP_{\mathbf{12}};\delta}$ |
| 1 | 6.25 | 6.77 | 7.93 | 6.70 |
| 2 | 8.74 | 9.01 | 8.91 | 9.10 |
| 3 | 4.07 | 4.05 | 4.39 | 4.08 |
| 4 | 5.57 | 5.61 | 5.69 | 5.66 |
| 5 | 7.40 | 6.93 | 9.49 | 7.60 |
| 6 | 4.23 | 4.07 | 4.98 | 4.23 |
| 7 | 6.06 | 6.12 | 5.91 | 6.04 |
| 8 | 4.26 | 4.26 | 4.45 | 4.31 |
| 9 | 4.44 | 4.52 | 4.47 | 4.51 |
| 10 | 4.84 | 4.70 | 7.45 | 4.97 |
| 11 | 3.69 | 3.66 | 3.75 | 3.67 |
| 12 | 4.40 | 4.31 | 4.29 | 4.33 |
| 13 | 6.56 | 6.69 | 6.91 | 6.85 |
| 14 | 4.55 | 4.47 | 4.92 | 4.49 |
| 15 | 4.18 | 4.23 | 5.13 | 4.26 |
| 16 | 6.23 | 6.18 | 6.06 | 6.11 |
| 17 | 3.95 | 3.91 | 5.73 | 4.02 |
| 18 | 5.61 | 5.50 | 5.71 | 5.46 |
| 19 | 7.03 | 7.00 | 7.22 | 7.06 |
| 20 | 7.11 | 7.17 | 7.47 | 7.27 |
| 21 | 3.83 | 3.79 | 3.88 | 3.87 |
| 22 | 11.05 | 11.03 | 12.32 | 11.35 |

Table 5.7 displays each fitted model's comparison measures (DIC, -2WAIC, and -2LPML). The results from these three metrics indicate that the $\mathcal{M}_{B_{4,\mathbf{10}};\delta}$ better fits the observed data compared to the $\mathcal{M}_{B_{4,\mathbf{4}};\delta}$. Furthermore, when comparing $\mathcal{M}_{BP_{\mathbf{3}};\delta}$ and $\mathcal{M}_{BP_{\mathbf{12}};\delta}$, it is evident that $\mathcal{M}_{BP_{\mathbf{3}};\delta}$ exhibits the lowest values for WAIC and LPML, suggesting a better performance of the simpler configuration. Now comparing B-spline vs. BP with fewer basis functions (7 and 4, respectively), two of the three measures

suggest that the model most adapted to the data uses BP of degree 3. On the other hand, in the models with more basis functions (both with 13), all three measures indicate the choice that employs B-splines.

Table 5.7: Comparison goodness-of-fit measurements concerning the PM10 data.

| | $\mathcal{M}_{B_{4,4};\delta}$ | $\mathcal{M}_{B_{4,10};\delta}$ | $\mathcal{M}_{BP_3;\delta}$ | $\mathcal{M}_{BP_{12};\delta}$ |
|---|---|---|---|---|
| DIC | 56520.27 | **56074.51** | 57217.93 | 56200.36 |
| WAIC | 106020.59 | **104811.61** | **104810.23** | 107055.22 |
| LPML | 943216.38 | **877605.81** | **887727.91** | 954893.07 |

### Analysis of Temperature Data

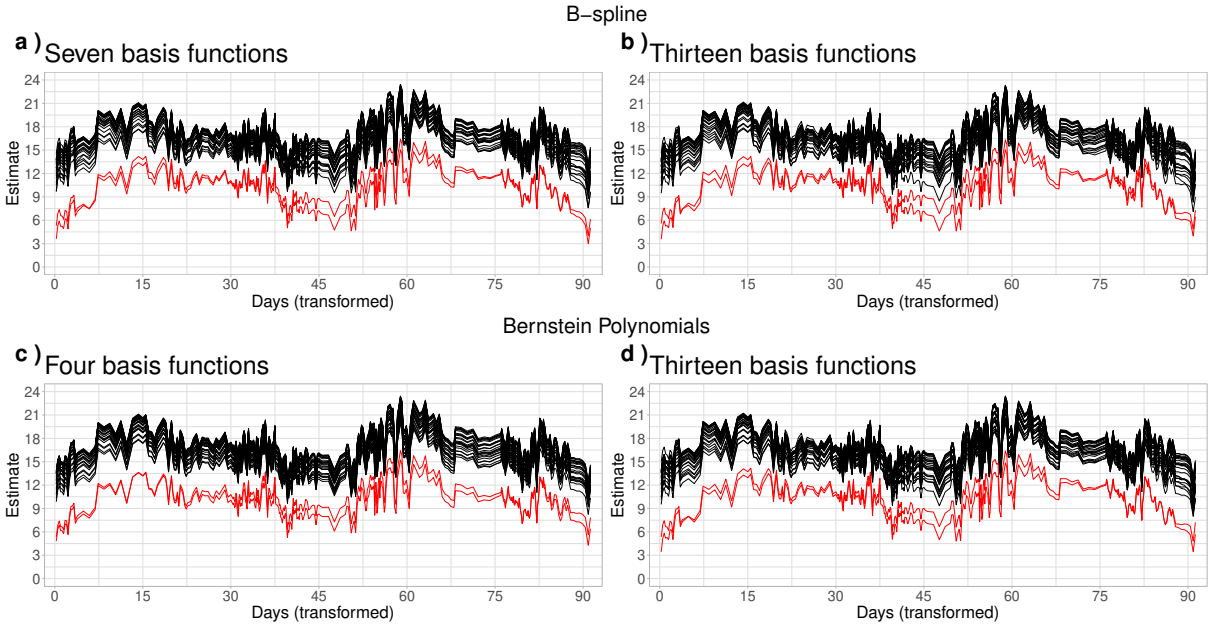
Figure 5.6 displays the smoothed curves for each fitted model. Each panel in the figure illustrates how the trajectories align with the behavior of the target functions presented in Figure 5.3. It is important to note that, according to Table 5.4, some functions are not complete with 344 observed Temperature measurements. In other words, missing values are present.

Analyzing the curves estimated by the models at Stations 3 and 12, represented by the red trajectories, provides interesting insights. These stations are known for having the lowest Temperature values during the considered periods. At Station 3, the Temperature ranges between 3.6°C and 15.3°C, while at Station 12, it fluctuates between 5.1°C and 15.4°C, according to the observed data. This temperature pattern is unique to these stations and not observed at other spatially distant locations. At those remote stations, the Temperature varies between 11°C and 24°C, and the estimated curves exhibit similar patterns, indicating a high level of spatial dependence due to their proximity. It is worth noting that despite the significant distance between Stations 3 and 12, the estimated curves at both locations share similar shapes.

In order to clearly understand the behavior of the curves in black in Figure 5.6, it is essential to consider the specific locations of the monitoring stations in Mexico City. Most of these stations are in urban areas characterized by heavy vehicular traffic, factories, and a scarcity of green spaces. However, Stations 3 and 12 are in rural areas with ample vegetation, devoid of vehicular traffic, industrial activities, and significant population density. Moreover, based on the latest report from SEDEMA (*Secretaria del Medio Ambiente de la Ciudad de México*) in 2019, Station 3 recorded an average annual temperature of 11.2°C, while Station 12 had an average of 10 °C. The suggested Gaussian models consider spatial dependency and include a random effect with an autoregressive structure. This helps to effectively deal with associations arising from geographic locations and uneven spacing of observed data. This idea accurately represents the Temperature data, enabling the

estimation process to yield curves that closely align with the target functions. A comprehensive analysis uses the IAE measure to evaluate the curve approximation performance.

Figure 5.6: Smoothed Temperature curves (estimated). In the case of Panels $(a)$ and $(b)$, the $\mathcal{M}_{B_{4,l};\delta}$ model was applied using 4 and 10 subintervals, respectively. On the other hand, for Panels $(c)$ and $(d)$, the $\mathcal{M}_{BP_p;\delta}$ model with BP of degrees 3 and 12 was used.




Source: Prepared by the author

Table 5.8 displays each fitted model's IAE discrepancy measure values. This measure is calculated based on the observed data from each Temperature monitoring station. The following results are observed when comparing different options for the number of basis functions used in each model. In the Gaussian model with B-spline smoothing methodology and $l = 10$ (where $l$ is the number of subintervals dividing the function domain), 19 out of 28 estimated curves have lower IAE values than the same model with $l = 4$. On the other hand, in the Gaussian model using BP, degree 12 (thirteen basis functions) has 20 IAE values lower than the model of degree 3. In summary, the higher number of basis approximates the observed values better. Likewise, in comparing the $\mathcal{M}_{B_{4,\mathbf{4}};\delta}$ and $\mathcal{M}_{BP_{\mathbf{3}};\delta}$, the first model showed lower IAE index values in 20 curves than the second model (with BP). Finally, when analyzing the models with more basis, it is noted that the $\mathcal{M}_{B_{4,\mathbf{10}};\delta}$ outperforms the $\mathcal{M}_{BP_{\mathbf{12}};\delta}$ since the values of 17 IAE were lower than the first model. In conclusion, these results indicate that a better approximation is obtained using the model based on the B-spline structure.

When analyzing the smoothed curves in the peripheral (with very few neighbors) areas of Mexico City, specifically in Stations 1, 12, 16, and 27, Table 5.8 shows that their IAE values are not very high (compared to more centralized stations), despite being in more isolated locations and without strong communication to seek information from the

curves in the neighborhood. These values indicate that even though the nearest neighbors are distant, each model accurately captures the information in the data, resulting in well-behaved curves. Additionally, in areas where the functions are closer to each other in spatial scale, such as Stations 4, 10, 11, 14, and 15, which have multiple nearest neighbors, discrepancy measures of less than 0.5 are observed. This suggests that these curves share similar characteristics with their neighbors and mutually benefit from each other during the inference process. In other words, the proximity between functions at these locations promotes consistency and accuracy in the obtained results.

Table 5.8: IAE discrepancy measurement of the smoothed Temperature curves obtained from the proposed models.

| | IAE | | IAE | |
|---|---|---|---|---|
| Site | $\mathcal{M}_{B_{4,\mathbf{4}};\delta}$ | $\mathcal{M}_{B_{4,\mathbf{10}};\delta}$ | $\mathcal{M}_{BP_{\mathbf{3}};\delta}$ | $\mathcal{M}_{BP_{\mathbf{12}};\delta}$ |
| 1 | 0.52 | 0.51 | 0.52 | 0.51 |
| 2 | 1.37 | 1.21 | 1.53 | 1.23 |
| 3 | 0.87 | 0.84 | 0.95 | 0.86 |
| 4 | 0.43 | 0.45 | 0.45 | 0.45 |
| 5 | 0.58 | 0.54 | 0.66 | 0.58 |
| 6 | 0.73 | 0.74 | 0.70 | 0.78 |
| 7 | 0.92 | 0.90 | 1.13 | 0.89 |
| 8 | 0.77 | 0.74 | 0.73 | 0.73 |
| 9 | 0.85 | 0.86 | 0.83 | 0.86 |
| 10 | 0.32 | 0.31 | 0.33 | 0.32 |
| 11 | 0.35 | 0.35 | 0.40 | 0.35 |
| 12 | 0.75 | 0.73 | 1.04 | 0.75 |
| 13 | 0.60 | 0.58 | 0.66 | 0.59 |
| 14 | 0.33 | 0.33 | 0.37 | 0.34 |
| 15 | 0.38 | 0.42 | 0.40 | 0.40 |
| 16 | 0.61 | 0.62 | 0.62 | 0.63 |
| 17 | 0.49 | 0.50 | 0.48 | 0.50 |
| 18 | 0.91 | 0.89 | 0.87 | 0.87 |
| 19 | 0.56 | 0.58 | 0.56 | 0.59 |
| 20 | 2.52 | 2.64 | 2.57 | 2.68 |
| 21 | 0.49 | 0.47 | 0.50 | 0.50 |
| 22 | 0.51 | 0.48 | 0.51 | 0.49 |
| 23 | 0.78 | 0.73 | 0.80 | 0.73 |
| 24 | 0.44 | 0.44 | 0.47 | 0.44 |
| 25 | 0.50 | 0.50 | 0.50 | 0.49 |
| 26 | 0.62 | 0.60 | 0.63 | 0.59 |
| 27 | 0.64 | 0.62 | 0.64 | 0.63 |
| 28 | 0.71 | 0.69 | 0.76 | 0.68 |

Note that the magnitudes of the IAE metric are smaller in the fitting for Temperature data (Table 5.8) than for PM10 data (Table 5.6). Comparing Figures 5.2 (*b*) and 5.3, notice that the scale on the vertical axis differs. The PM10 data oscillates more strongly, reaching levels between 0 and 160. It is possible to observe a smoother behavior in the Temperature series. This characteristic of lighter oscillation seems to have favored the model in determining smaller IAE measures, which indicates a better approximation between the observed and estimated values.

Table 5.9 shows the DIC, WAIC, and LPML measurements to evaluate the model's goodness of fit. Among the versions utilizing B-splines, the $\mathcal{M}_{B_{4,\mathbf{10};\delta}}$ stands out as the top performer based on the WAIC and LPML criteria. For the models employing BP, the WAIC and the LPML suggest that the $\mathcal{M}_{BP_{\mathbf{12};\delta}}$ provides the best fit. In conclusion, the analyzed metrics suggest that the models that best fit the data are those with more bases. However, the fits with fewer bases do not seem bad and have the advantage of providing a more parsimonious structure and better computational ease to run the MCMC. Now, considering the models that performed well in both cases mentioned earlier (B-spline and BP), based explicitly on the DIC and WAIC metrics, it is concluded that the preferred model, which accurately explains the observed Temperature data, is $\mathcal{M}_{B_{4,\mathbf{10};\delta}}$. It is essential to highlight that a slight difference is detected in this comparison, so the BP cannot be judged as a less advantageous option. The reader should know that the BP has a more straightforward structure, which only requires defining the degree. In contrast, the B-spline requires specifying the order and the number of knots.

Table 5.9: Comparison measurements of fitted models concerning Temperature data.

| | $\mathcal{M}_{B_{4,\mathbf{4};\delta}}$ | $\mathcal{M}_{B_{4,\mathbf{10};\delta}}$ | $\mathcal{M}_{BP_{\mathbf{3};\delta}}$ | $\mathcal{M}_{BP_{\mathbf{12};\delta}}$ |
|---|---|---|---|---|
| DIC | 29363.50 | 30046.58 | 30109.26 | 30222.95 |
| WAIC | 55433.22 | **54257.99** | 57493.22 | **55645.21** |
| LPML | 719167.38 | **684983.10** | 839835.17 | **684203.48** |

# Chapter 6

# Conclusions and Future Works

The main objective of this thesis was to propose new Gaussian functional models that include a spatial dependence structure to handle functional data observed as irregularly-spaced series in different geographical locations. The models developed were based on mathematical tools, such as B-spline basis expansions, defined using the recurrence relation discovered by De Boor (2001), and BP, mathematically specified in a similar way as described by Farouki and Rajan (1987). These methods offered flexibility to capture complex shapes and patterns while ensuring numerical stability to allow estimation without computational difficulties. A Bayesian approach was adopted to estimate the unknown parameters of the models. This decision was based on its ability to incorporate prior information and account for uncertainty in the analysis. One crucial aspect of the model structures was their capacity to incorporate the association motivated by the irregular spacing of the observed measurements for each function through a random effect $\delta$ with an autoregressive structure. This means that each observation is influenced by its neighbors. Including this effect represents one of the main contributions of this thesis, as it is a novel, robust and promising approach to SFD analysis.

The proposed modeling was fitted via MCMC using the `Stan` platform. A thorough simulation study assessed and compared the models' performance under various scenarios and configurations. The procedures and criteria for generating the simulated data were described, along with the evaluation metrics used to compare and contrast the results. The findings of this simulation study were essential in guiding the selection of appropriate models for real-world situations. A highlight of the simulation study focuses on the prediction and handling of missing data, achieving satisfactory performance for the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$ models. The results revealed that, at three levels of spatial correlation considered (low $\varphi = 2$, moderate $\varphi = 1$, and high $\varphi = 0.5$), along with two levels of variation (moderate $\kappa = 1$ and high $\kappa = 2$), prediction curves close to the target functions were obtained, with low MISE values. This was especially evident when the functions were located close to each other on a spatial scale, allowing them to share strength for inference or prediction purposes. In addition, the posterior means were close to the true values and showed little variability, suggesting increased confidence in the accuracy of the estimates. These results in curve prediction and complete information management (with Bayesian imputation) represent another significant contribution of this thesis for FDA

analysis.

The proposed model with BP emerges as another valuable contribution from this study, offering an innovative and efficient alternative compared to other approaches, especially in the face of B-splines. By introducing this new option in the modeling, the need to worry about the number and location of the B-splines knots is eliminated, which considerably simplifies both the implementation and the analysis of the process.

Concerning the applications discussed in this work lead to the following conclusions. Thoroughly analyzing the fitted models, PM10 data has revealed several important insights. The smoothed curves of each model, displayed in Figure 5.4, demonstrate that most estimated functions closely follow the patterns of the set of target curves in Figure 5.2 ($b$). However, all models underestimate observations with high PM10 levels, specifically from 100 to 160. Additionally, it has been identified that some estimates of the PM10 index are negative in each panel of Figure 5.4, particularly for Station 10. This phenomenon is attributed to the station's location on the outskirts of a rural area with minimal industrial activity and abundant vegetation. While this issue does not affect most stations with PM10 values far from zero, it does underscore the importance of considering the proximity to zero when applying Gaussian models to data with positive domains. Finally, the model comparison results presented in Tables 5.6 and 5.7 shed light on the performance of different fitted models. Notably, the $\mathcal{M}_{B,4,\mathbf{10};\delta}$ model demonstrated superior performance based on various metrics, indicating its capability to better fit the observed data compared to other configurations, including the $\mathcal{M}_{B,4,\mathbf{4};\delta}$ and $\mathcal{M}_{BP,\mathbf{12};\delta}$. Additionally, the comparison between B-spline and BP models revealed that, depending on the number of bases, the choice of model varied, with B-splines excelling when a higher number of basis were used and simpler BP models demonstrating better performance with fewer bases.

On the other hand, by analyzing the smoothed curves obtained from the proposed models fitted to Mexico City Temperature data, a clearer understanding of how temperature patterns behave at different monitoring stations has been received. Particularly Stations 3 and 12, in rural areas with unique environmental characteristics, exhibit distinct temperature patterns compared to other locations. The Gaussian models (B-spline and BP) effectively capture this behavior under moderate spatial dependency ($\varphi = 1.2$) and with the association motivated by irregular spacing. The comparison of different model configurations highlights the significance of the number of functional bases in achieving better target curve approximation. Models with more bases generally exhibit improved performance, accurately capturing trajectory changes. Notably, the B-spline-based model $\mathcal{M}_{B,4,\mathbf{10};\delta}$ emerges as a preferred choice based on various goodness-of-fit metrics (IAE, DIC, WAIC, and LPML), closely followed by the simpler structure of the BP model $\mathcal{M}_{BP,\mathbf{12};\delta}$.

**Future Works**

As for future research opportunities, this thesis proposes a modeling approach that can be extended in the following aspects:

- Evaluate other functional basis options in terms of inference and quality of fit. For example, consider second-generation Wavelets (Sweldens, 1996), which allow working with irregular sample designs, while retaining the characteristics of traditional Wavelets to adequately capture local behaviors that may occur in certain parts of the data.

- Explore alternative covariance functions, such as the Matérn function. This function is handy for modeling the spatial correlation between two measurements taken at different locations. Assume that the distance between observations $i$ and $j$ is represented as $d$. The Matérn correlation can be defined as follows:

$$\rho_{ij} = \rho(d; \zeta, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{d}{\zeta}\right)^{\nu} K_{\nu}\left(\frac{d}{\zeta}\right),$$

  where $K_{\nu}(\cdot)$ denotes the modified Bessel function of the second kind of order $\nu$. The parameter $\zeta > 0$ determines the rate at which the correlation decays to zero with increasing $d$. The parameter $\nu > 0$ controls the degree of smoothness. The $\nu$ is an exciting element to allow the researcher to modify the smoothness level of the covariance function, which can be helpful in some applications.

- Instead of using the discretized position $t_i$ in the functional domain to record the measurement and $d_i = t_{i+1} - t_i$ (distance between the two positions $t_{i+1}$ and $t_i$), it would be more convenient to consider $t_{ij}$ (the $i$-th measurement position performed at site $j$) and $d_{ij}$. In this variant, irregular spacing is allowed, which may vary between locations. This extension would be exciting to improve the modeling developed in this thesis.

- Develop a sensitivity analysis to investigate different priors specifications for the parameters in the hierarchical model. In this case, one can consider distinct levels of information, including informative, vague, or non-informative specifications.

- Exploring values other than 4 and 8 in the Expression (3.4) is an interesting aspect for future studies.

# Bibliography

Aguilera, A. M. and M. Aguilera-Morillo (2013). Comparative study of different b-spline approaches for functional data. *Mathematical and Computer Modelling 58*(7-8), 1568–1579.

Aguilera-Morillo, M. C., M. Durbán, and A. M. Aguilera (2017). Prediction of functional data with spatial dependence: a penalized approach. *Stochastic Environmental Research and Risk Assessment 31*(1), 7–22.

Aristizabal, J.-P., R. Giraldo, and J. Mateu (2019). Analysis of variance for spatially correlated functional data: Application to brain data. *Spatial Statistics 32*, 100381.

Baladandayuthapani, V., B. K. Mallick, M. Young Hong, J. R. Lupton, N. D. Turner, and R. J. Carroll (2008). Bayesian Hierarchical Spatially Correlated Functional data Analysis with Application to Colon Carcinogenesis. *Biometrics 64*(1), 64–73.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. London: Chapman and Hall/CRC.

Bernstein, S. (1912). Démonstration du Théoreme de Weierstrass Fondée Sur le Calcul des Probabilities. *Communications of the Kharkov Mathematical 13*, 1–2.

Cortés-D, D. L., J. H. Camacho-Tamayo, and R. Giraldo (2016). Spatial prediction of soil penetration resistance using functional geostatistics. *Scientia Agricola 73*, 455–461.

Cox, M. G. (1972). The numerical evaluation of b-splines. *IMA Journal of Applied Mathematics 10*(2), 134–149.

Cressie, N. (1993). *Statistics for Spatial Data, Revised Edition*. New York: John Wiley Sons, Inc.

Davis, P. (1975). *Interpolation and Approximation*. New York: Dover Publications.

De Boor, C. (1972). On Calculating with B-splines. *Journal of Approximation Theory 6*(1), 50–62.

De Boor, C. (2001). *A Practical Guide to Splines*. New York: Springer.

De Villiers, J. (2012). *Mathematics of Approximation*, Volume 1. Springer Science & Business Media.

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

Diggle, P. and P. Ribeiro (2007). *Model-Based Geostatistics*. Springer New York.

Fan, Y.-T. and H.-C. Huang (2022). Spatially varying coefficient models using reduced-rank thin-plate splines. *Spatial Statistics 51*, 100654.

Farouki, R. T. and V. Rajan (1987). On the Numerical Condition of Polynomials in Bernstein Form. *Computer Aided Geometric Design 4*(3), 191–216.

Farouki, R. T. and V. Rajan (1988). Algorithms for Polynomials in Bernstein Form. *Computer Aided Geometric Design 5*(1), 1–26.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer.

Gelfand, A. E., H.-J. Kim, C. Sirmans, and S. Banerjee (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association 98*(462), 387–396.

Gelfand, A. E. and A. F. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), 398–409.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian Data Analysis, Third Edition*. London: Chapman and Hall/CRC.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (6), 721–741.

Gentle, J. (2009). *Computational Statistics*. Springer New York.

Giraldo, R., P. Delicado, and J. Mateu (2010). Continuous time-varying kriging for spatial prediction of functional data: An environmental application. *Journal of Agricultural, Biological, and Environmental Statistics 15*, 66–82.

Giraldo, R., P. Delicado, and J. Mateu (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics 18*, 411–426.

Giraldo, R., P. Delicado, and J. Mateu (2012). Hierarchical Clustering of Spatially Correlated Functional Data. *Statistica Neerlandica 66*(4), 403–421.

Giraldo, R., J. Mateu, and P. Delicado (2012). geofd: an r package for function-valued geostatistical prediction. *Revista Colombiana de Estadística 35*(3), 385–407.

Greenbaum, D. S., J. D. Bachmann, D. Krewski, J. M. Samet, R. White, and R. E. Wyzga (2001). Particulate air pollution standards and morbidity and mortality: case study. *American Journal of Epidemiology 154*(12), S78–S90.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika 57*(1), 97–109.

Hoffman, M. D. and A. Gelman (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

Hollander, M., D. Wolfe, and E. Chicken (2013). *Nonparametric Statistical Methods*. New Jersey: Wiley.

Jiang, H. and N. Serban (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics 54*(2), 108–119.

Kokoszka, P. and M. Reimherr (2017). *Introduction to Functional Data Analysis*. London: Chapman and Hall/CRC.

Lawson, A. B. (2018). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. CRC press.

Lawson, A. B. (2021). *Using R for Bayesian Spatial and Spatio-Temporal Health Modeling*. CRC Press.

Liang, Z., F. Weng, Y. Ma, Y. Xu, M. Zhu, and C. Yang (2022). Measurement and analysis of high frequency assert volatility based on functional data analysis. *Mathematics 10*(7), 1140.

Liu, C., S. Ray, and G. Hooker (2017). Functional principal component analysis of spatially correlated data. *Statistics and Computing 27*(6), 1639–1654.

Lorentz, G. G. (2012). *Bernstein Polynomials, Second Edition*. New York: American Mathematical Society.

Martínez-Hernández, I. and M. G. Genton (2020). Recent developments in complex and spatially correlated functional data. *Brazilian Journal of Probability and Statistics 34*(2), 204–229.

Mateu, J. and R. Giraldo (2021). *Geostatistical Functional Data Analysis*. John Wiley & Sons.

Mateu, J. and E. Romano (2017). Advances in spatial functional statistics. *Stochastic Environmental Research and Risk Assessment 31*, 1–6.

Mayrink, V. D. and F. B. Gonçalves (2017). A Bayesian Hidden Markov Mixture Model to Detect Overexpressed Chromosome Regions. *Journal of the Royal Statistical Society, Series C 66*(2), 387–412.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087–1092.

Morris, J. S., M. Vannucci, P. J. Brown, and R. J. Carroll (2003). Wavelet-Based Nonparametric Modeling of Hierarchical Functions in Colon Carcinogenesis. *Journal of the American Statistical Association 98*(463), 573–583.

Nerini, D., P. Monestiez, and C. Manté (2010). Cokriging for Spatial Functional Data. *Journal of Multivariate Analysis 101*(2), 409–418.

Paldy, A., J. Bobvos, M. Lustigova, H. Moshammer, E. M. Niciu, P. Otorepec, V. Puklova, K. Szafraniec, T. Zagargale, M. Neuberger, et al. (2006). Health impact assessment of pm10 on mortality and morbidity in children in central-eastern european cities. *Epidemiology 17*(6), S131.

Petrone, S. (1999). Bayesian density estimation using bernstein polynomials. *Canadian Journal of Statistics 27*(1), 105–126.

Piegl, L. and W. Tiller (1996). *The NURBS Book*. Springer Science & Business Media.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and MATLAB*. New York: Springer.

Ramsay, J. and B. Silverman (2002). *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis*. New York: Springer.

Reich, B. J., M. Fuentes, and D. B. Dunson (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association 106*(493), 6–20.

Rekabdarkolaee, H. M., C. Krut, M. Fuentes, and B. J. Reich (2019). A Bayesian Multivariate Functional Model with Spatially Varying Coefficient Approach for Modeling Hurricane Track Data. *Spatial Statistics 29*, 351–365.

Romano, E., A. Balzanella, and R. Verde (2017). Spatial variability clustering for spatially dependent functional data. *Statistics and Computing 27*, 645–658.

Schoenberg, I. J. (1946). Contributions to the Problem of Approximation of Equidistant Data by Analytic Functions. *Quarterly of Applied Mathematics 4*(2), 45–99.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. New York: Cambridge University Press.

Song, J. J. and B. Mallick (2019). Hierarchical Bayesian Models for Predicting Spatially Correlated Curves. *Statistics 53*(1), 196–209.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology 64*(4), 583–639.

Staicu, A.-M., C. M. Crainiceanu, and R. J. Carroll (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics 11*(2), 177–194.

Stan Development Team (2023). *Stan Modeling Language Users Guide and Reference Manual*. version 2.18.0.

Sweldens, W. (1996). The Lifting Scheme: A Custom-Design Construction of Biorthogonal Wavelets. *Applied and Computational Harmonic Analysis 3*(2), 186–200.

Tenbusch, A. (1994). Two-dimensional bernstein polynomial density estimators. *Metrika 41*(1), 233–253.

van Rossum, G. (2023). Python(programming language).

Vidakovic, B. (2009). *Statistical Modeling by Wavelets*. John Wiley & Sons.

Wand, M. P. and M. C. Jones (1994). *Kernel smoothing*. CRC press.

Wang, T. and Z. Guan (2019). Bernstein polynomial model for nonparametric multivariate density. *Statistics 53*(2), 321–338.

Watanabe, S. and M. Opper (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research 11*(12).

Zhang, L., V. Baladandayuthapani, H. Zhu, K. A. Baggerly, T. Majewski, B. A. Czerniak, and J. S. Morris (2016). Functional CAR Models for Large Spatially Correlated Functional Datasets. *Journal of the American Statistical Association 111*(514), 772–786.

Zhou, L., J. Z. Huang, J. G. Martinez, A. Maity, V. Baladandayuthapani, and R. J. Carroll (2010). Reduced Rank Mixed Effects Models for Spatially Correlated Hierarchical Functional Data. *Journal of the American Statistical Association 105*(489), 390–400.

# Appendix A

# Simulation Study Part I: Extra Results

Subsection 4.1.1 presents the results of the artificial data sets generated and fitted using the models proposed in Table 4.1. These data sets were created with two levels of variability, denoted by $\kappa = \{1, 2\}$ while maintaining a fixed correlation value of $\varphi = 1$. Additionally, the irregular spacing of the functional domain was established using the $Beta(1, 2)$.

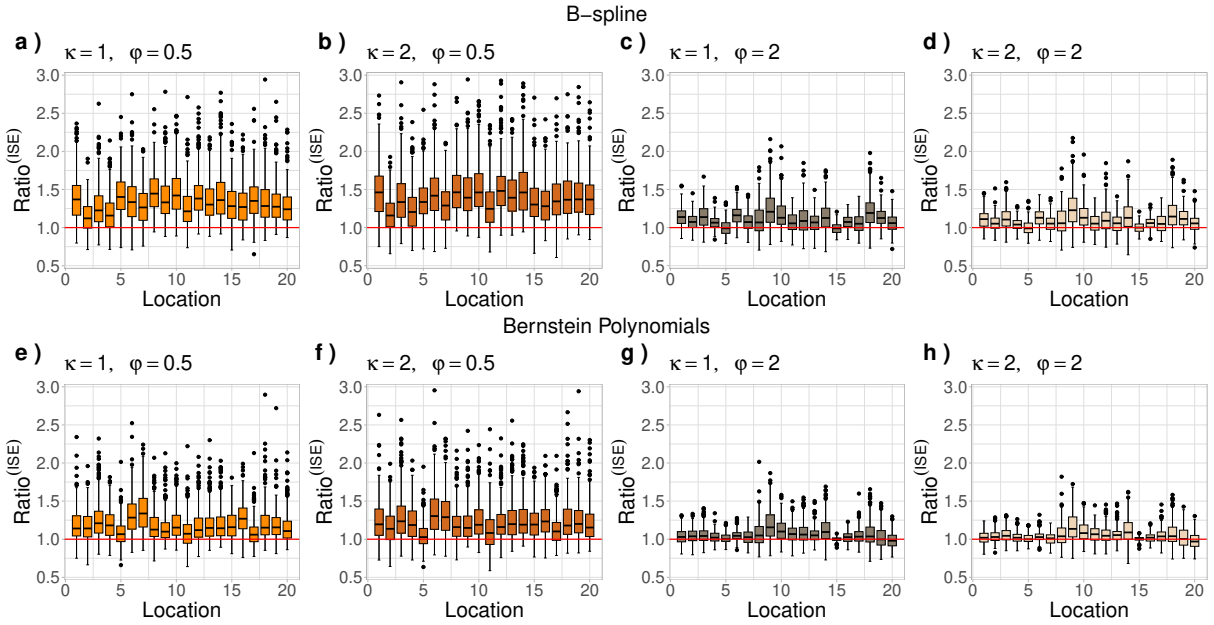
It is essential to highlight that our complete study encompasses various scenarios, each having different levels of correlation, namely $\varphi = 0.5$ and $\varphi = 2$ representing high and low spatial correlation, respectively. A detailed analysis is provided in this Appendix to gain a comprehensive understanding of these additional scenarios.

## A.1 Spatial Dependence

Figure A.1 shows that the results obtained are similar for the two levels of variability chosen with each spatial correlation value and the two smoothing techniques. Note also that by having a strongly associated set of curves regardless of whether the distances of the geographic locations are close or distant, the correct proposed model fit results in smoothed trajectories that are closer to the true ones than the model fit without the spatial structure, see panels $(a)$, $(b)$, $(e)$, and $(f)$.

For SFD with a low level of spatial dependence, the distances of the curve locations are an essential aspect since having several close neighbors results in better performance of smooth trajectories in MC replicates when fitting the exposed model with the spatial association than the model without it, see panels $(c)$, $(d)$, $(g)$, and $(f)$.

Figure A.1: Comparison of the IAE and ISE ratios of the B-spline and BP models with two spatial variation parameter settings: $\kappa = \{1, 2\}$, together with a decay parameter $\varphi = \{0.5, 2\}$.




Source: Prepared by the author

## A.2 Random Effect

Table A.1 shows the results of the MISE discrepancy measure for each fitting model. First, it is observed that, both for variability values and for a high and low level of spatial correlation, the performance of the functions is superior when the $\mathcal{M}_{B_{4,4};\delta}$ fit model is considered. On the other hand, when the curves have a high spatial dependence and similar formats, regardless of the distance of the locations, they present close or equal values in the MISE measure, as occurs with the curves of the locations $S_{12}$, $S_{13}$, $S_{14}$, $S_{19}$ and $S_{20}$. In the case of $\varphi = 2$, curves surrounding several close neighbors show better performance in the two variability values considered.

Table A.2 shows the results of the MISE measure for the BP models with and without the random effect component $\delta$. For each scenario of variability and spatial correlation, a behavior analogous to the previously analyzed results of the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{B_{4,4};\bullet}$ fit models are observed.

Table A.1: Discrepancy measures for the smoothed curves using the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{B_{4,4};\bullet}$ models, with and without the random effect component, respectively. Each model is evaluated with two settings of the spatial variation parameters, specifically $\kappa = \{1, 2\}$, together with a decay parameter $\varphi = \{0.5, 2\}$.

| Local | MISE | | MISE | | MISE | | MISE | |
|---|---|---|---|---|---|---|---|---|
| | $\mathcal{M}_{B_{4,4};\delta}$ | $\mathcal{M}_{B_{4,4};\bullet}$ | $\mathcal{M}_{B_{4,4};\delta}$ | $\mathcal{M}_{B_{4,4};\bullet}$ | $\mathcal{M}_{B_{4,4};\delta}$ | $\mathcal{M}_{B_{4,4};\bullet}$ | $\mathcal{M}_{B_{4,4};\delta}$ | $\mathcal{M}_{B_{4,4};\bullet}$ |
| | $\kappa = 1,\ \varphi = 0.5$ | | $\kappa = 2,\ \varphi = 0.5$ | | $\kappa = 1,\ \varphi = 2$ | | $\kappa = 2,\ \varphi = 2$ | |
| $S_1$ | 0.049 | 0.785 | 0.050 | 0.786 | 0.073 | 0.813 | 0.077 | 0.815 |
| $S_2$ | 0.059 | 0.793 | 0.062 | 0.797 | 0.077 | 0.815 | 0.082 | 0.819 |
| $S_3$ | 0.050 | 0.786 | 0.051 | 0.787 | 0.068 | 0.804 | 0.073 | 0.808 |
| $S_4$ | 0.055 | 0.790 | 0.059 | 0.794 | 0.074 | 0.808 | 0.079 | 0.813 |
| $S_5$ | 0.066 | 0.810 | 0.068 | 0.812 | 0.078 | 0.816 | 0.082 | 0.820 |
| $S_6$ | 0.049 | 0.785 | 0.049 | 0.785 | 0.069 | 0.802 | 0.074 | 0.806 |
| $S_7$ | 0.058 | 0.795 | 0.060 | 0.797 | 0.074 | 0.810 | 0.079 | 0.816 |
| $S_8$ | 0.050 | 0.788 | 0.051 | 0.788 | 0.066 | 0.805 | 0.070 | 0.809 |
| $S_9$ | 0.049 | 0.785 | 0.049 | 0.785 | 0.059 | 0.794 | 0.062 | 0.798 |
| $S_{10}$ | 0.049 | 0.785 | 0.050 | 0.785 | 0.066 | 0.801 | 0.071 | 0.807 |
| $S_{11}$ | 0.054 | 0.791 | 0.056 | 0.795 | 0.071 | 0.812 | 0.075 | 0.814 |
| $S_{12}$ | 0.047 | 0.783 | 0.048 | 0.784 | 0.070 | 0.812 | 0.074 | 0.816 |
| $S_{13}$ | 0.047 | 0.784 | 0.048 | 0.784 | 0.072 | 0.811 | 0.077 | 0.816 |
| $S_{14}$ | 0.047 | 0.783 | 0.047 | 0.783 | 0.066 | 0.804 | 0.070 | 0.808 |
| $S_{15}$ | 0.053 | 0.789 | 0.055 | 0.790 | 0.076 | 0.809 | 0.080 | 0.814 |
| $S_{16}$ | 0.057 | 0.797 | 0.059 | 0.799 | 0.072 | 0.807 | 0.077 | 0.811 |
| $S_{17}$ | 0.054 | 0.791 | 0.056 | 0.792 | 0.070 | 0.808 | 0.075 | 0.812 |
| $S_{18}$ | 0.048 | 0.784 | 0.049 | 0.785 | 0.064 | 0.798 | 0.069 | 0.803 |
| $S_{19}$ | 0.047 | 0.783 | 0.048 | 0.784 | 0.078 | 0.823 | 0.081 | 0.827 |
| $S_{20}$ | 0.047 | 0.783 | 0.049 | 0.785 | 0.079 | 0.823 | 0.082 | 0.828 |

Table A.2: Discrepancy measures for the smoothed curves using the $\mathcal{M}_{BP_3;\delta}$ and $\mathcal{M}_{BP_3;\bullet}$ models, with and without the random effect component, respectively. Each model is evaluated with two settings of the spatial variation parameters, specifically $\kappa = \{1, 2\}$, together with a decay parameter $\varphi = \{0.5, 2\}$.

| Local | MISE | | MISE | | MISE | | MISE | |
|---|---|---|---|---|---|---|---|---|
| | $\mathcal{M}_{BP_3;\delta}$ | $\mathcal{M}_{BP_3;\bullet}$ | $\mathcal{M}_{BP_3;\delta}$ | $\mathcal{M}_{BP_3;\bullet}$ | $\mathcal{M}_{BP_3;\delta}$ | $\mathcal{M}_{BP_3;\bullet}$ | $\mathcal{M}_{BP_3;\delta}$ | $\mathcal{M}_{BP_3;\bullet}$ |
| | $\kappa = 1,\ \varphi = 0.5$ | | $\kappa = 2,\ \varphi = 0.5$ | | $\kappa = 1,\ \varphi = 2$ | | $\kappa = 2,\ \varphi = 2$ | |
| $S_1$ | 0.047 | 0.884 | 0.048 | 0.885 | 0.064 | 0.906 | 0.066 | 0.908 |
| $S_2$ | 0.055 | 0.894 | 0.057 | 0.896 | 0.061 | 0.896 | 0.063 | 0.898 |
| $S_3$ | 0.047 | 0.885 | 0.048 | 0.885 | 0.057 | 0.895 | 0.059 | 0.897 |
| $S_4$ | 0.050 | 0.888 | 0.052 | 0.889 | 0.061 | 0.897 | 0.063 | 0.899 |
| $S_5$ | 0.059 | 0.904 | 0.059 | 0.904 | 0.064 | 0.904 | 0.065 | 0.905 |
| $S_6$ | 0.046 | 0.885 | 0.047 | 0.884 | 0.060 | 0.895 | 0.063 | 0.898 |
| $S_7$ | 0.051 | 0.890 | 0.053 | 0.891 | 0.063 | 0.903 | 0.066 | 0.906 |
| $S_8$ | 0.047 | 0.886 | 0.048 | 0.885 | 0.055 | 0.892 | 0.057 | 0.895 |
| $S_9$ | 0.046 | 0.884 | 0.047 | 0.884 | 0.052 | 0.888 | 0.054 | 0.891 |
| $S_{10}$ | 0.047 | 0.884 | 0.048 | 0.885 | 0.056 | 0.893 | 0.059 | 0.896 |
| $S_{11}$ | 0.052 | 0.893 | 0.054 | 0.896 | 0.057 | 0.893 | 0.059 | 0.895 |
| $S_{12}$ | 0.046 | 0.883 | 0.046 | 0.884 | 0.057 | 0.896 | 0.060 | 0.898 |
| $S_{13}$ | 0.046 | 0.884 | 0.047 | 0.885 | 0.059 | 0.896 | 0.061 | 0.898 |
| $S_{14}$ | 0.045 | 0.883 | 0.046 | 0.883 | 0.057 | 0.895 | 0.059 | 0.897 |
| $S_{15}$ | 0.050 | 0.887 | 0.051 | 0.888 | 0.062 | 0.898 | 0.065 | 0.901 |
| $S_{16}$ | 0.051 | 0.892 | 0.051 | 0.892 | 0.059 | 0.896 | 0.062 | 0.898 |
| $S_{17}$ | 0.050 | 0.888 | 0.051 | 0.888 | 0.059 | 0.896 | 0.061 | 0.898 |
| $S_{18}$ | 0.046 | 0.884 | 0.047 | 0.885 | 0.058 | 0.898 | 0.060 | 0.900 |
| $S_{19}$ | 0.046 | 0.884 | 0.046 | 0.884 | 0.061 | 0.900 | 0.063 | 0.903 |
| $S_{20}$ | 0.046 | 0.885 | 0.047 | 0.885 | 0.064 | 0.906 | 0.066 | 0.908 |

## A.3 Prediction

Figures A.2 ($a$) and A.3 ($a$) demonstrate that the trajectories estimated by the proposed $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$ models (represented by the red line) closely align with the actual curves (represented by the black line) when there is a strong correlation between the curves. This consistency is observed irrespective of whether the unobserved locations have close neighbors or if there is a significant level of spatial variability between the functions. Conversely, in scenarios with low correlation, illustrated in Figures A.2 ($b$) and A.3 ($b$), the proposed models provide accurate trajectory estimates for unobserved locations with nearby neighbors. However, the prediction curves deviate from the target functions at somewhat isolated sites. The discrepancy becomes more noticeable in the panels of the curve $S_{15}$, especially when the variability value is 2.

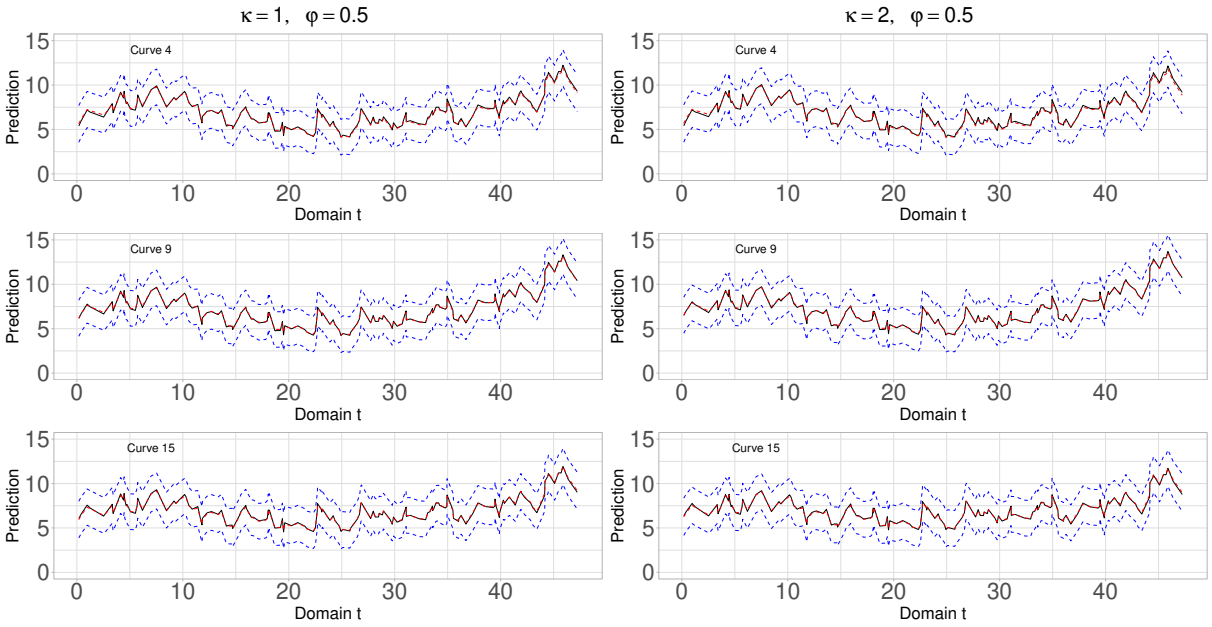
### Misspecification

In Figures A.4 ($a$) and A.5 ($a$), it can be observed that when the sets of MC curves exhibit high spatial dependence, the average predictions of the trajectories of unobserved sites are not affected by an incorrect specification of the level of dependence. In this case, even with weak correlation, the mean of the HPD intervals manages to capture the target functions, and the estimated average curves by the proposed models approach their targets.

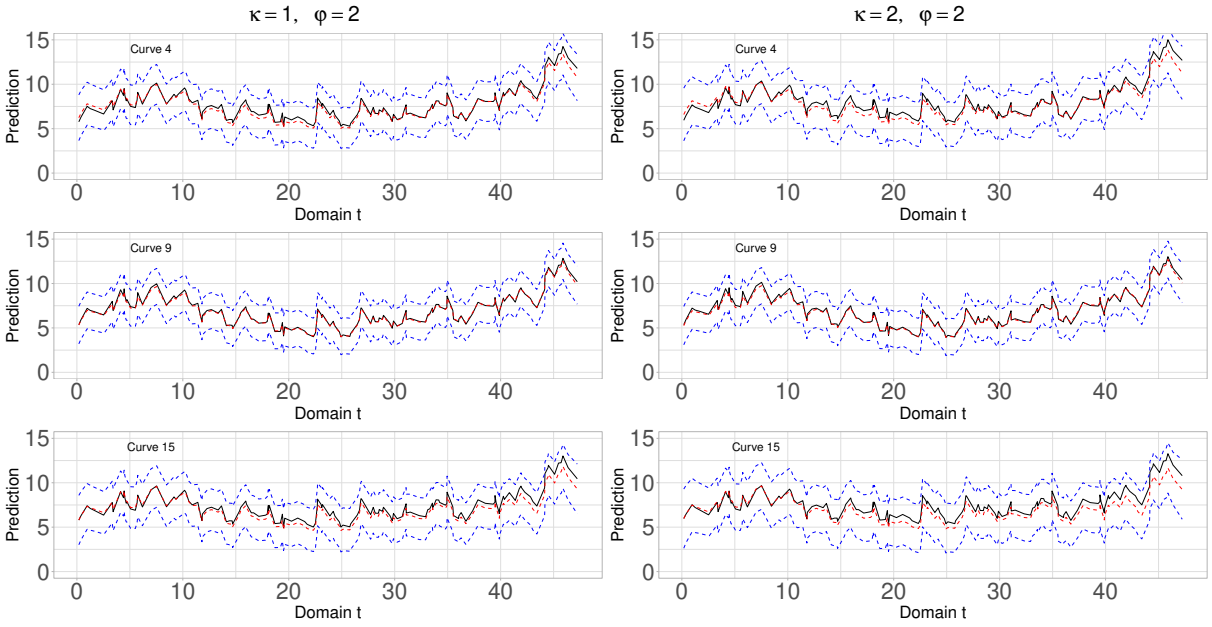
In the panels of Figure A.4 ($b$), different behaviors can be observed when specifying an incorrect level of strong dependence in MC sets with a weak association between their curves. For location $S_9$, which has not been observed but is surrounded by several nearby neighbors, the HPD intervals' mean successfully captures the curve's true shape, and the average prediction approaches its target. However, in the case of the peripheral location $S_{15}$, which has also not been observed, the mean of the HPD intervals fails to capture a portion of the actual curve, and the estimates deviate slightly from the target, especially when considering the variability of value 2. As for the model using BP, the plots in Figure A.5 ($b$) show similar behavior to that explained above for the model using B-splines.

Figure A.2: Target function (black solid lines), point-wise mean prediction curves (red dashed line), and 95% point-wise mean HPD intervals (blue dashed line) of the estimated curves for the $\mathcal{M}_{B_{4,4};\delta}$ model. The model is evaluated with two settings of spatial variation parameters and decay parameters: Panel $(a)$ $\kappa = \{1, 2\}$ and $\varphi = 0.5$. Panel $(b)$ $\kappa = \{1, 2\}$ and $\varphi = 2$.

(a)



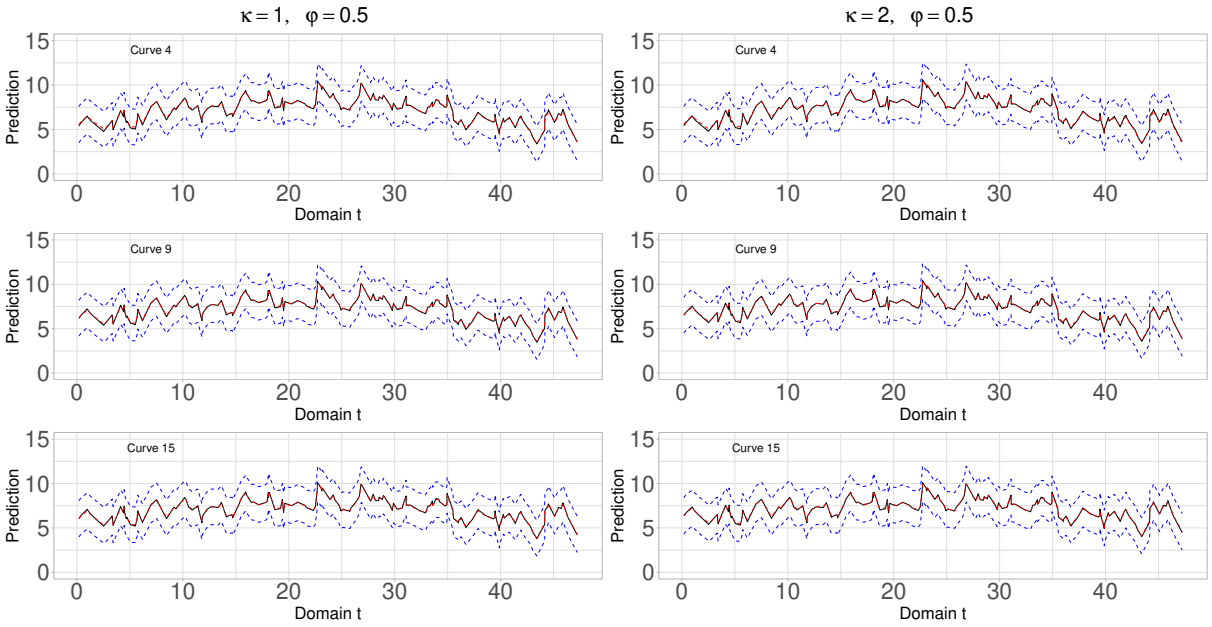

(b)



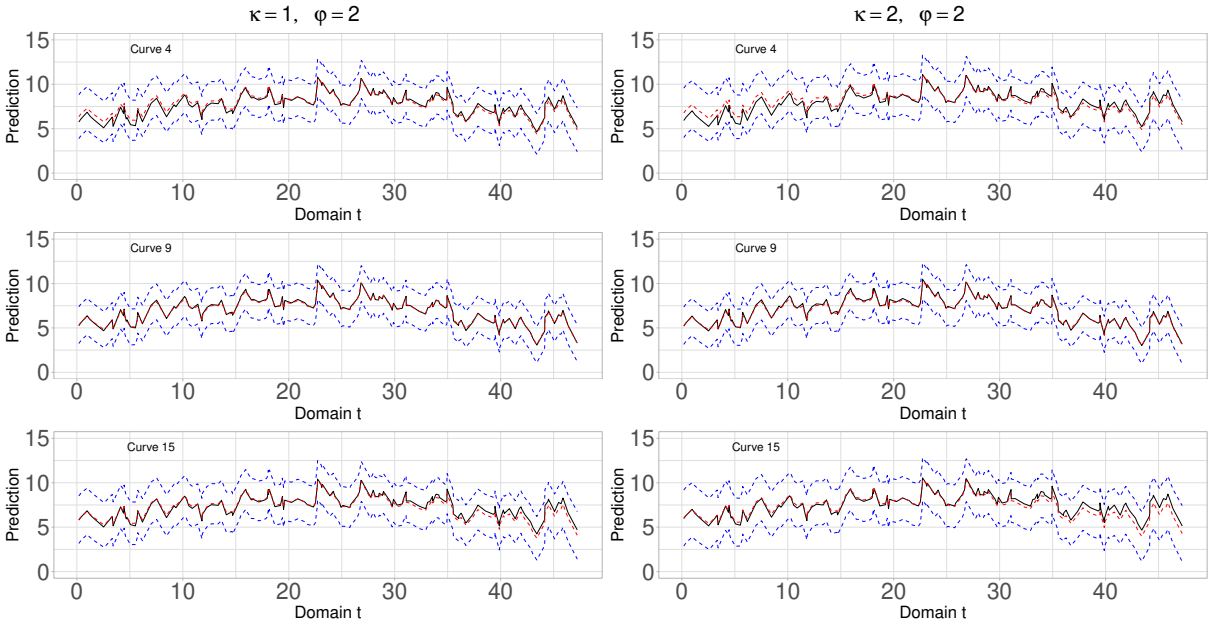

Source: Prepared by the author

Figure A.3: Target function (black solid lines), point-wise mean prediction curves (red dashed line), and 95% point-wise mean HPD intervals (blue dashed line) of the estimated curves for the $\mathcal{M}_{BP_3;\delta}$ model. The model is evaluated with two settings of spatial variation parameters and decay parameters: Panel $(a)$ $\kappa = \{1, 2\}$ and $\varphi = 0.5$. Panel $(b)$ $\kappa = \{1, 2\}$ and $\varphi = 2$.

(a)



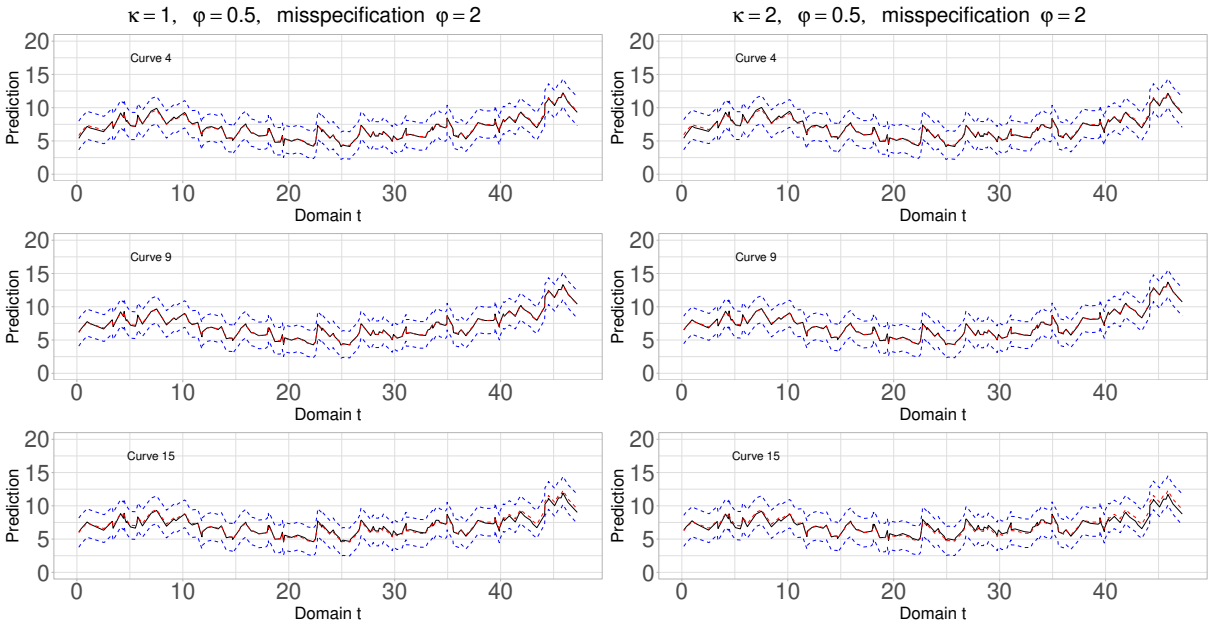

(b)



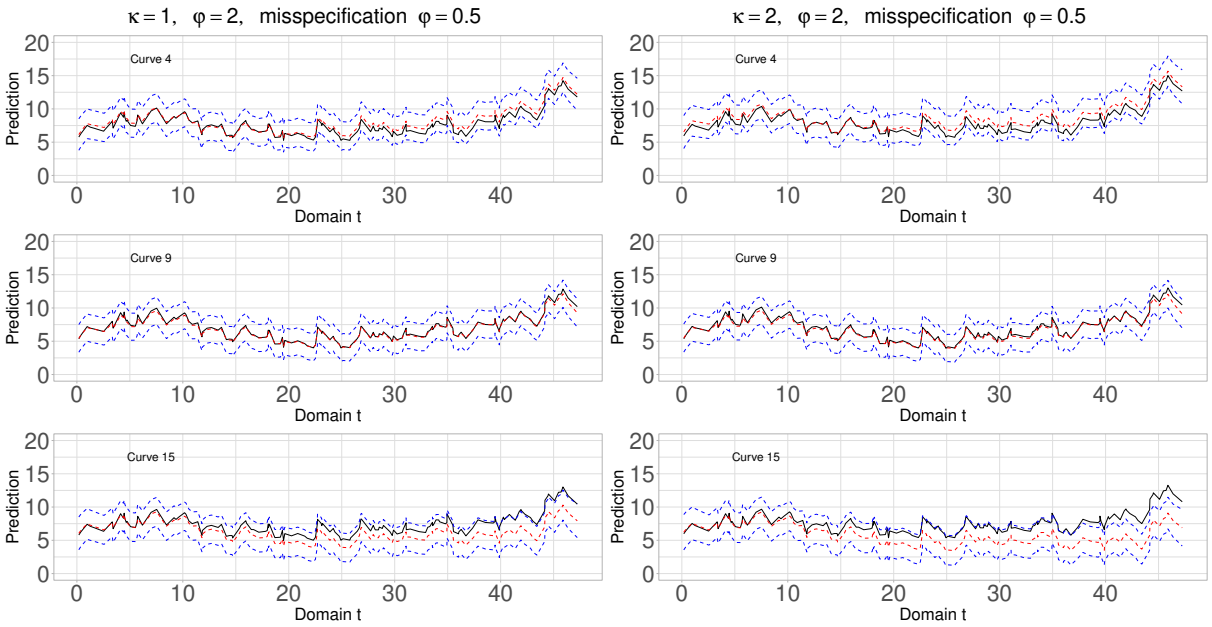

Source: Prepared by the author

Figure A.4: Target function (black solid lines), point-wise mean prediction curves (red dashed line), and 95% point-wise mean HPD intervals (blue dashed line) of the estimated curves for the $\mathcal{M}_{B_{4,4};\delta}$ model. The model is evaluated with two settings of spatial variation parameters and decay parameters misspecification: Panel $(a)$ $\kappa = \{1, 2\}$, and $\varphi = 2$ (real value $\varphi = 0.5$). Panel $(b)$ $\kappa = \{1, 2\}$ and $\varphi = 0.5$ (real value $\varphi = 2$).

(a)



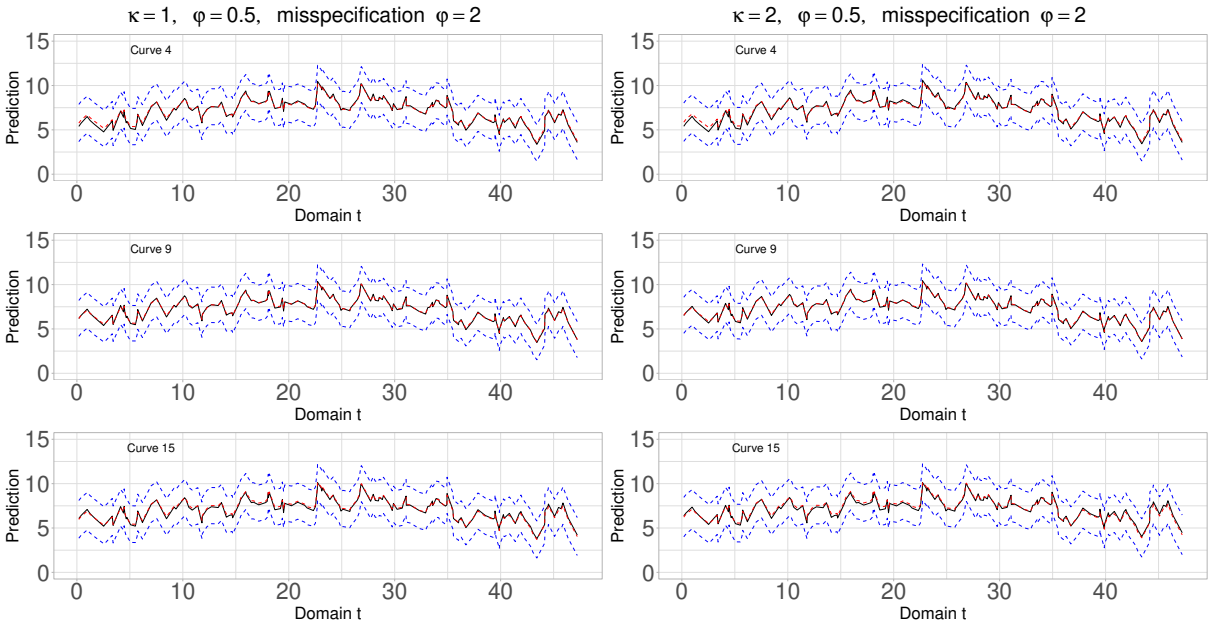

(b)



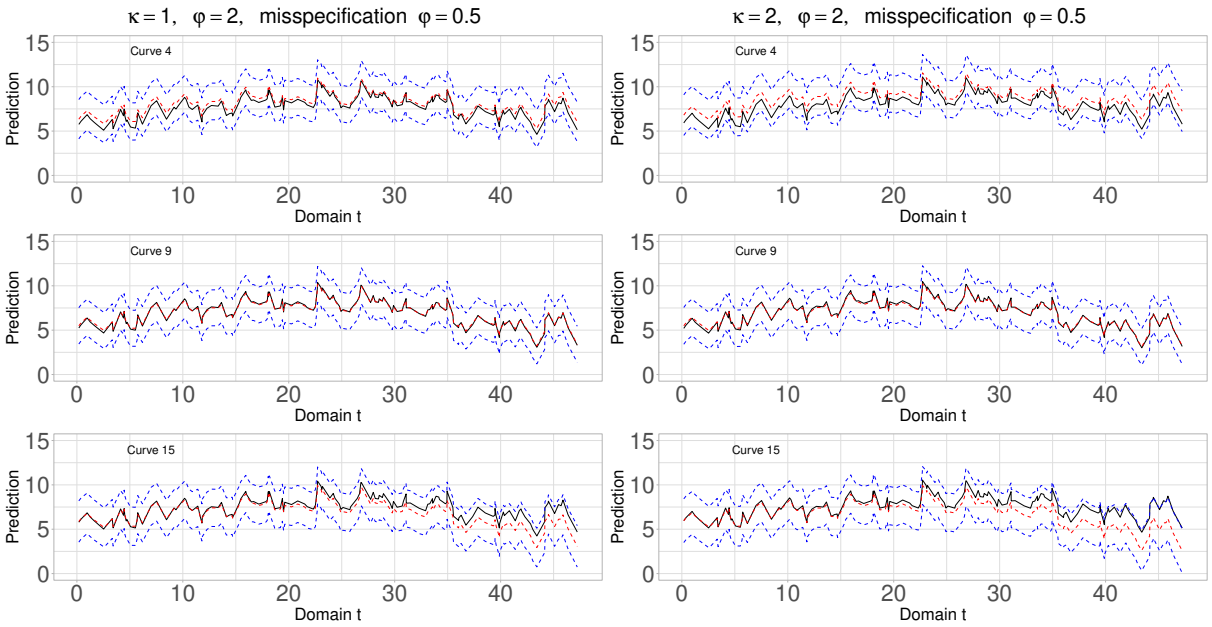

Source: Prepared by the author

Figure A.5: Target function (black solid lines), point-wise mean prediction curves (red dashed line), and 95% point-wise mean HPD intervals (blue dashed line) of the estimated curves for the $\mathcal{M}_{BP_3;\delta}$ model. The model is evaluated with two settings of spatial variation parameters and decay parameters misspecification: Panel $(a)$ $\kappa = \{1, 2\}$, and $\varphi = 2$ (real value $\varphi = 0.5$). Panel $(b)$ $\kappa = \{1, 2\}$ and $\varphi = 0.5$ (real value $\varphi = 2$).

(a)



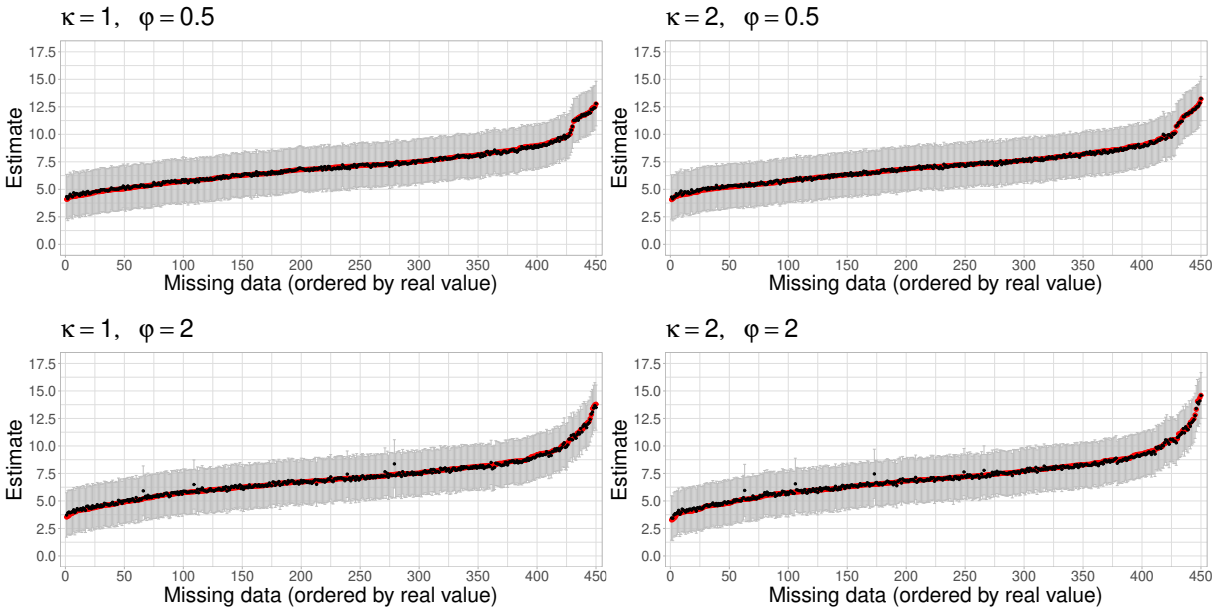

(b)



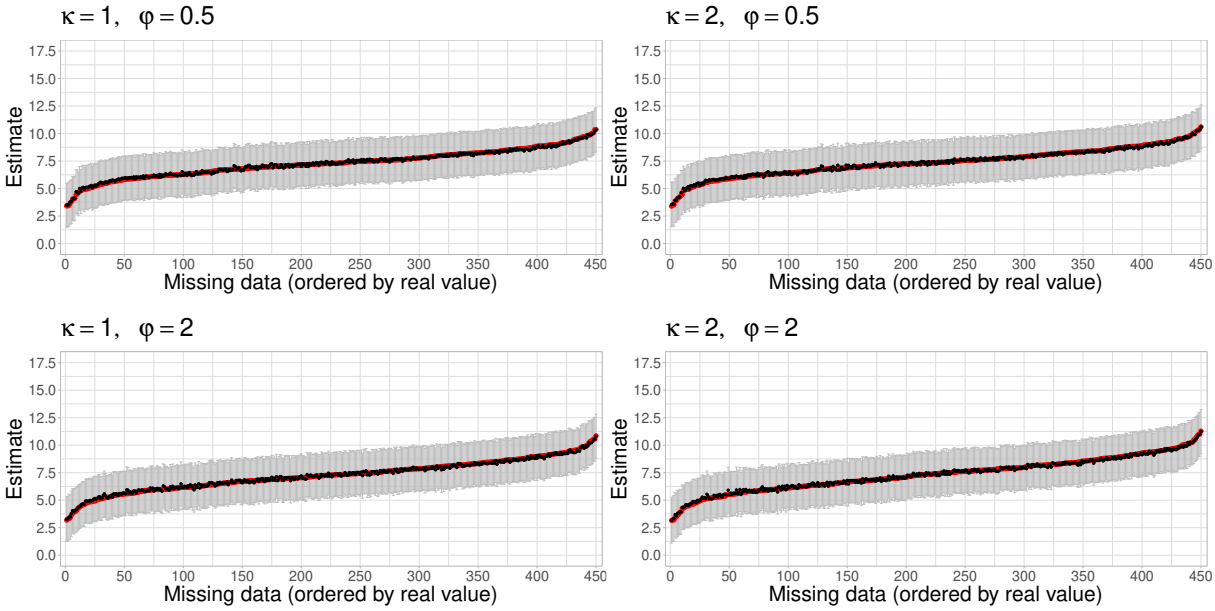

Source: Prepared by the author

Figure A.6: Mean of the 95% HPD Intervals for the 450 missing data using the $\mathcal{M}_{B_{4,4};\delta}$ (Panel $(a)$) and $\mathcal{M}_{BP_3;\delta}$ (Panel $(b)$) models. The actual values are the red dots, and the averages of the mean posterior estimates from each MC replicate are the black dots. Each model is evaluated with two settings of the spatial variation parameters, specifically $\kappa = \{1, 2\}$, and a decay parameter $\varphi = \{0.5, 2\}$.

(a)




(b)




Source: Prepared by the author

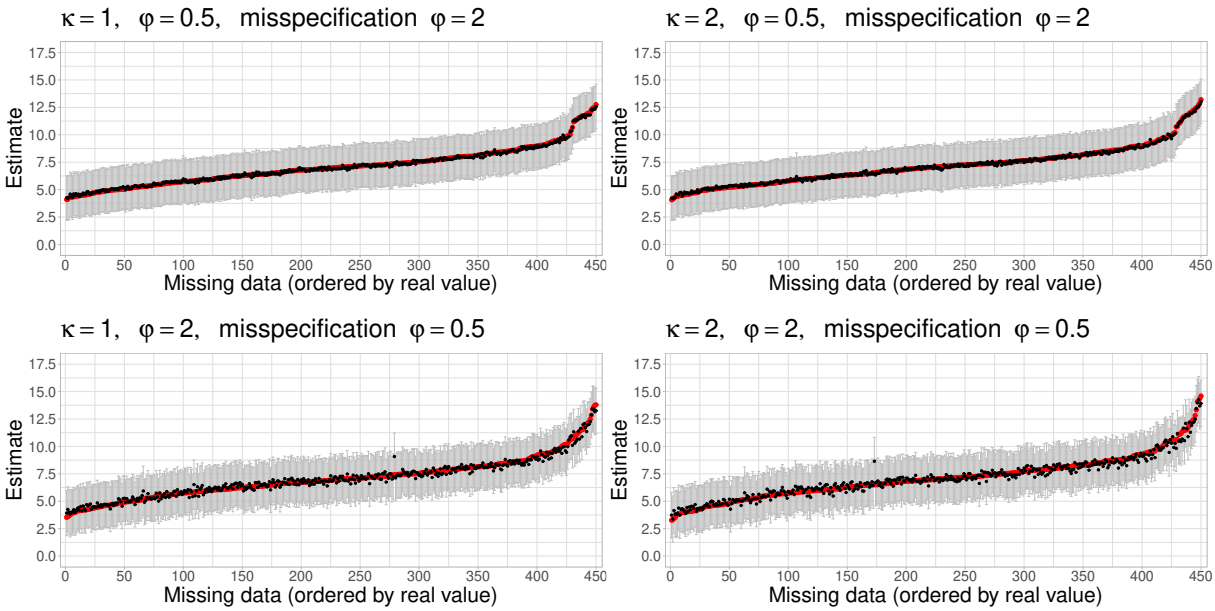
## A.4 Missing Data

In Figures A.6 ($a$) and A.6 ($b$), it is evident that both model $\mathcal{M}_{B_{4,4};\delta}$ and model $\mathcal{M}_{BP_3;\delta}$ can obtain mean estimates that closely match the actual values of the missing data for each curve, given a high correlation between the sets of curves, regardless of spatial variability. However, in cases where samples exhibit weak spatial dependence, model $\mathcal{M}_{B_{4,4};\delta}$ tends to estimate some missing observations further away from the actual values, unlike Model $\mathcal{M}_{BP_3;\delta}$.

### Misspecification

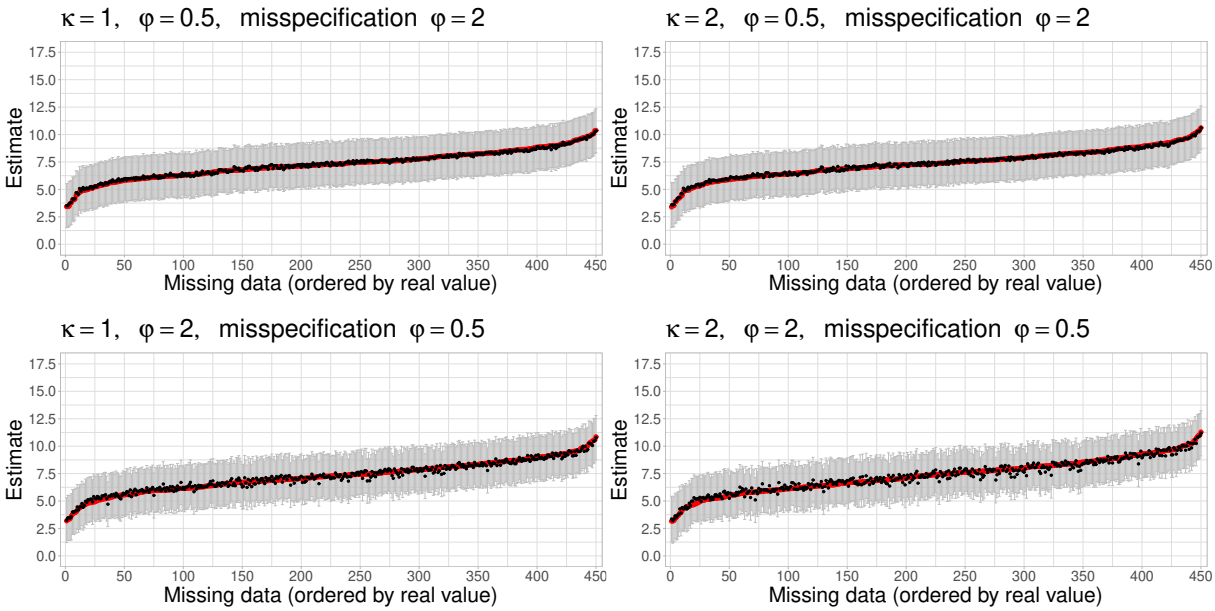
In Figures A.7 and A.8, the following can be observed: when the replicates show strong spatial dependence but correlation effects are assumed to decrease rapidly as the distance between curve sites increases ($\varphi = 2$) in the fitting models, it happens that the average of the posterior mean estimates of the MC ensembles for the missing data of the curves approaches the target values and the posterior uncertainty is not so high. On the other hand, when samples have weak spatial dependence but a small value for the decay parameter is used (more distant data significantly influence the estimate), some average estimates of the posterior means are found to be overestimated and underestimated. This occurs in both fitting models, regardless of the level of spatial variability.

Figure A.7: Performance of the $\mathcal{M}_{B_{4,4};\delta}$-fitting model when the degree of spatial dependence of the smoothed curves is incorrectly specified.




Source: Prepared by the author

Figure A.8: Performance of the $\mathcal{M}_{BP_3;\delta}$-fitting model when the degree of spatial dependence of the smoothed curves is incorrectly specified.



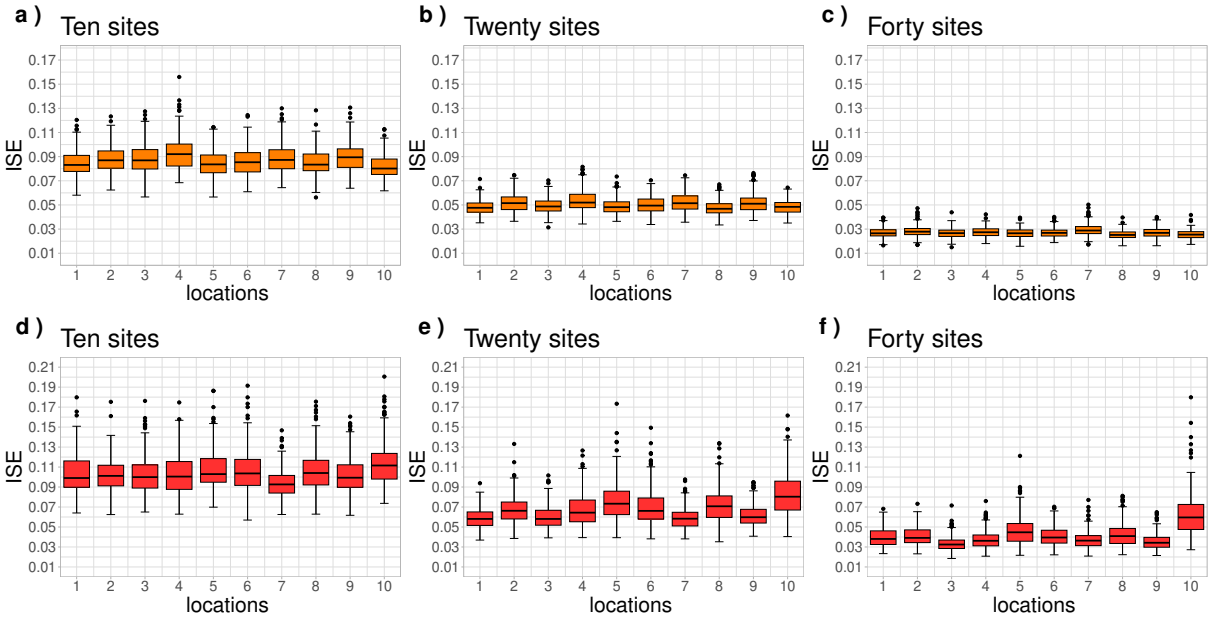

Source: Prepared by the author

# Appendix B

# Simulation Study Part III: Extra Results

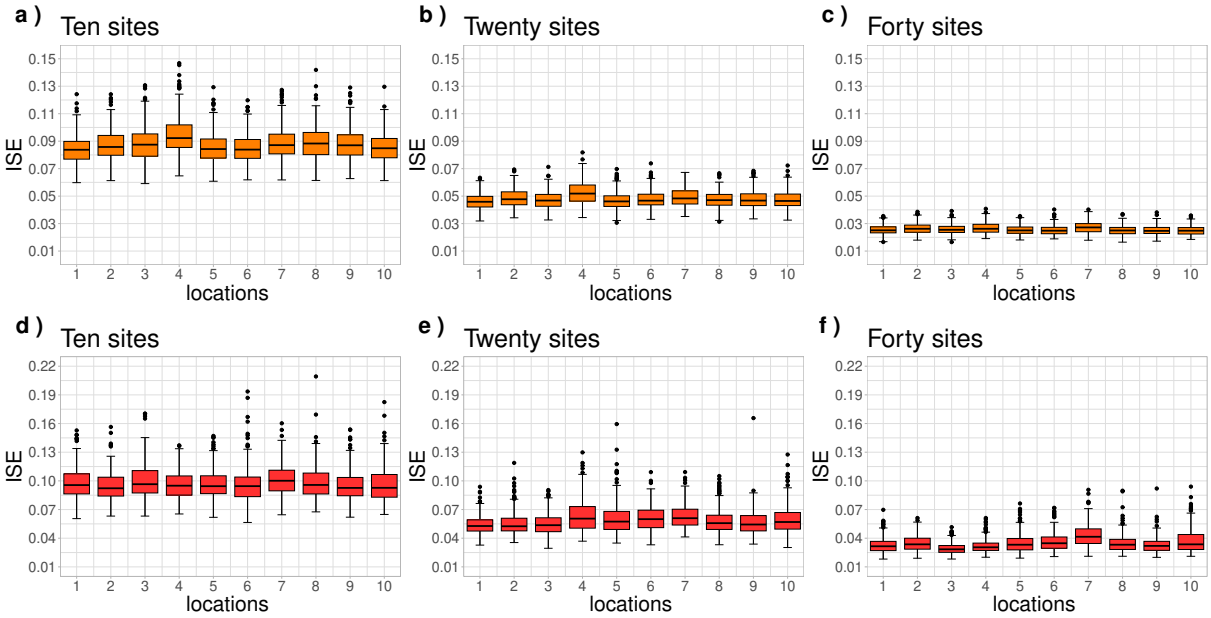
Figures B.1 and B.2 show the graphs representing the analysis of other values considered for the decay parameter ($\varphi = 0.5$ and 2) in the proposed $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$ models. These graphs reveal behavior and performance practically identical to those analyzed and described in Chapter 4 for the value of $\varphi = 1$. As the number of sites increases, the number of close neighbors also rises. Consequently, the ISE indices decrease, indicating a closer fit of the curves. Furthermore, the boxplots exhibit reduced dispersion, particularly noticeable when there are 40 sites.

Figure B.1: Comparison of IAE and ISE discrepancy measures for functions estimated using the $\mathcal{M}_{B_{4,4};\delta}$ model. This analysis is conducted on ten common sites within the three simulated location Grids. The model is evaluated with a configuration of the spatial variation parameter and two values for the decay parameters: Panels $(a)$, $(b)$, and $(c)$ $\kappa = 1$, and $\varphi = 0.5$. Panels $(d)$, $(e)$, and $(f)$ $\kappa = 1$, and $\varphi = 2$.




Source: Prepared by the author

Figure B.2: Comparison of IAE and ISE discrepancy measures for functions estimated using the $\mathcal{M}_{BP_3;\delta}$ model. This analysis is conducted on ten common sites within the three simulated location Grids. The model is evaluated with a configuration of the spatial variation parameter and two values for the decay parameters: Panels $(a)$, $(b)$, and $(c)$ $\kappa = 1$, and $\varphi = 0.5$. Panels $(d)$, $(e)$, and $(f)$ $\kappa = 1$, and $\varphi = 2$.




Source: Prepared by the author

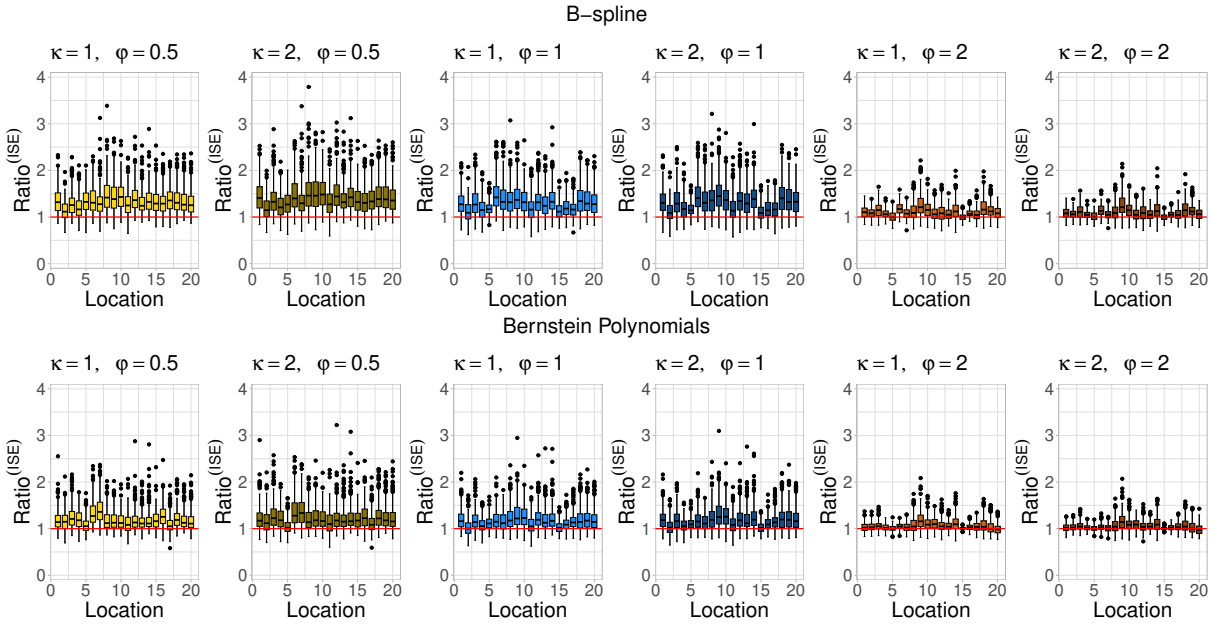
# Appendix C

# Simulation Study: Considering the Distances of the Distribution $U(0,1)$

## C.1 Spatial Dependence

Figure C.1 illustrates the results of the quotients in (4.1) corresponding to the smoothing methods of the functional data (B-spline and BP). At first glance, it can be seen that there is little difference between the two levels of spatial variability $\kappa$ considered for the ISE ratio for both structures. However, it becomes evident that the outcomes differ noticeably between the two basis functions. Notably, the B-spline approach excels in capturing the spatial patterns within the data.

Figure C.1: Comparison of the IAE and ISE ratios of the B-spline and BP models with two spatial variation parameter settings: $\kappa = \{1, 2\}$, together with three options for the decay parameter $\varphi = \{0.5, 1.2\}$.




Source: Prepared by the author

## C.2 Autoregressive Random Effect

Table C.1: Discrepancy measures for the smoothed curves using the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{B_{4,4};\bullet}$ models, with and without the random effect component, respectively.

| | MISE | | | | | | MISE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{M}_{B_{4,4};\delta}$ | | | $\mathcal{M}_{B_{4,4};\bullet}$ | | | $\mathcal{M}_{B_{4,4};\delta}$ | | | $\mathcal{M}_{B_{4,4};\bullet}$ | | |
| Site | $\varphi=0.5$ | $\varphi=1$ | $\varphi=2$ | $\varphi=0.5$ | $\varphi=1$ | $\varphi=2$ | $\varphi=0.5$ | $\varphi=1$ | $\varphi=2$ | $\varphi=0.5$ | $\varphi=1$ | $\varphi=2$ |
| | $\kappa=1$ | | | | | | $\kappa=2$ | | | | | |
| $S_1$ | 0.050 | 0.056 | 0.074 | 1.091 | 1.094 | 1.120 | 0.051 | 0.059 | 0.078 | 1.092 | 1.098 | 1.121 |
| $S_2$ | 0.059 | 0.069 | 0.076 | 1.098 | 1.110 | 1.120 | 0.063 | 0.074 | 0.080 | 1.102 | 1.117 | 1.123 |
| $S_3$ | 0.051 | 0.058 | 0.071 | 1.091 | 1.100 | 1.113 | 0.052 | 0.061 | 0.075 | 1.092 | 1.104 | 1.116 |
| $S_4$ | 0.057 | 0.067 | 0.075 | 1.095 | 1.112 | 1.113 | 0.060 | 0.070 | 0.080 | 1.100 | 1.114 | 1.118 |
| $S_5$ | 0.067 | 0.075 | 0.076 | 1.120 | 1.126 | 1.120 | 0.069 | 0.077 | 0.079 | 1.121 | 1.127 | 1.122 |
| $S_6$ | 0.050 | 0.056 | 0.071 | 1.091 | 1.098 | 1.108 | 0.051 | 0.059 | 0.076 | 1.091 | 1.100 | 1.112 |
| $S_7$ | 0.059 | 0.063 | 0.078 | 1.103 | 1.102 | 1.118 | 0.061 | 0.067 | 0.083 | 1.104 | 1.104 | 1.125 |
| $S_8$ | 0.052 | 0.057 | 0.070 | 1.095 | 1.098 | 1.114 | 0.052 | 0.058 | 0.073 | 1.094 | 1.099 | 1.118 |
| $S_9$ | 0.049 | 0.052 | 0.060 | 1.090 | 1.092 | 1.100 | 0.050 | 0.052 | 0.063 | 1.091 | 1.092 | 1.104 |
| $S_{10}$ | 0.050 | 0.056 | 0.065 | 1.091 | 1.097 | 1.105 | 0.051 | 0.057 | 0.070 | 1.091 | 1.099 | 1.110 |
| $S_{11}$ | 0.055 | 0.065 | 0.075 | 1.097 | 1.112 | 1.123 | 0.057 | 0.069 | 0.079 | 1.101 | 1.115 | 1.125 |
| $S_{12}$ | 0.048 | 0.054 | 0.072 | 1.089 | 1.094 | 1.121 | 0.049 | 0.056 | 0.075 | 1.090 | 1.097 | 1.126 |
| $S_{13}$ | 0.049 | 0.055 | 0.072 | 1.089 | 1.096 | 1.116 | 0.050 | 0.057 | 0.076 | 1.090 | 1.097 | 1.121 |
| $S_{14}$ | 0.048 | 0.052 | 0.065 | 1.088 | 1.092 | 1.110 | 0.048 | 0.054 | 0.069 | 1.088 | 1.093 | 1.112 |
| $S_{15}$ | 0.055 | 0.068 | 0.078 | 1.095 | 1.103 | 1.114 | 0.057 | 0.074 | 0.083 | 1.097 | 1.109 | 1.119 |
| $S_{16}$ | 0.059 | 0.070 | 0.075 | 1.105 | 1.121 | 1.116 | 0.061 | 0.073 | 0.080 | 1.108 | 1.125 | 1.119 |
| $S_{17}$ | 0.056 | 0.062 | 0.073 | 1.098 | 1.101 | 1.117 | 0.058 | 0.066 | 0.076 | 1.099 | 1.104 | 1.120 |
| $S_{18}$ | 0.049 | 0.055 | 0.065 | 1.090 | 1.098 | 1.104 | 0.050 | 0.056 | 0.070 | 1.091 | 1.099 | 1.109 |
| $S_{19}$ | 0.048 | 0.055 | 0.077 | 1.088 | 1.097 | 1.132 | 0.049 | 0.058 | 0.079 | 1.089 | 1.099 | 1.133 |
| $S_{20}$ | 0.048 | 0.057 | 0.081 | 1.089 | 1.099 | 1.134 | 0.050 | 0.059 | 0.085 | 1.090 | 1.101 | 1.139 |

Table C.2: Discrepancy measures for the smoothed curves using the $\mathcal{M}_{BP_3;\delta}$ and $\mathcal{M}_{BP_3;\bullet}$ models, with and without the random effect component, respectively.

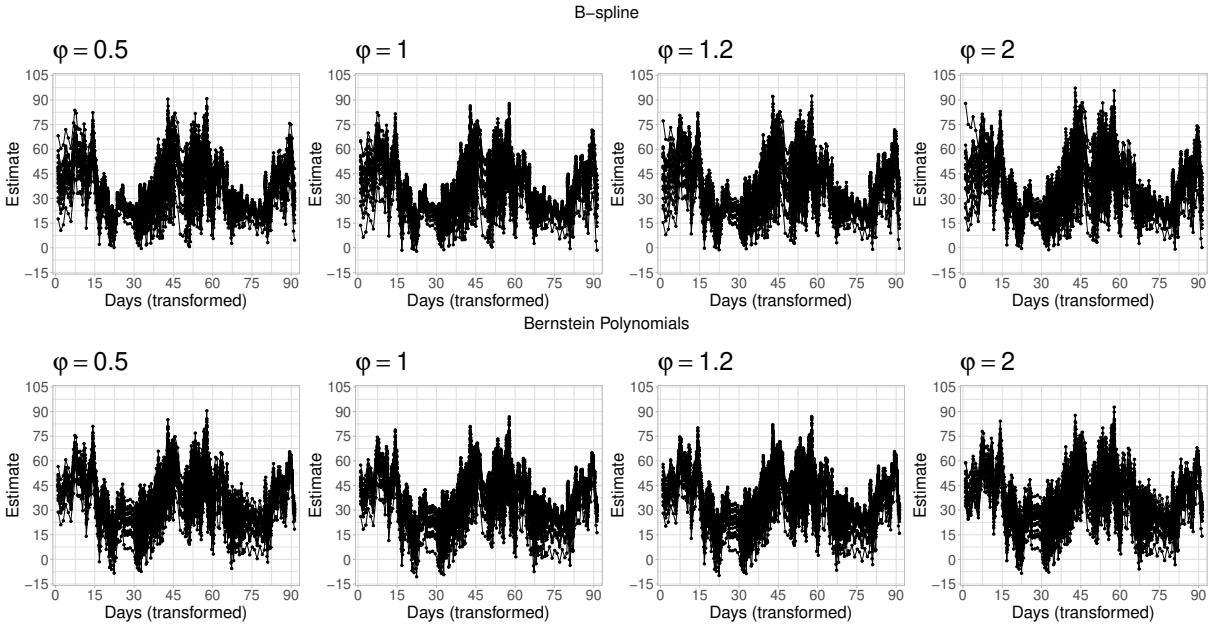
| | MISE | | | | | | MISE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{M}_{BP_3;\delta}$ | | | $\mathcal{M}_{BP_3;\bullet}$ | | | $\mathcal{M}_{BP_3;\delta}$ | | | $\mathcal{M}_{BP_3;\bullet}$ | | |
| Site | $\varphi=0.5$ | $\varphi=1$ | $\varphi=2$ | $\varphi=0.5$ | $\varphi=1$ | $\varphi=2$ | $\varphi=0.5$ | $\varphi=1$ | $\varphi=2$ | $\varphi=0.5$ | $\varphi=1$ | $\varphi=2$ |
| | $\kappa=1$ | | | | | | $\kappa=2$ | | | | | |
| $S_1$ | 0.048 | 0.052 | 0.064 | 1.408 | 1.410 | 1.433 | 0.049 | 0.053 | 0.066 | 1.409 | 1.412 | 1.433 |
| $S_2$ | 0.055 | 0.061 | 0.061 | 1.417 | 1.430 | 1.418 | 0.057 | 0.062 | 0.063 | 1.419 | 1.431 | 1.420 |
| $S_3$ | 0.048 | 0.052 | 0.059 | 1.409 | 1.413 | 1.421 | 0.049 | 0.054 | 0.061 | 1.409 | 1.414 | 1.422 |
| $S_4$ | 0.052 | 0.059 | 0.062 | 1.412 | 1.421 | 1.420 | 0.054 | 0.061 | 0.064 | 1.414 | 1.423 | 1.423 |
| $S_5$ | 0.059 | 0.060 | 0.065 | 1.434 | 1.420 | 1.429 | 0.060 | 0.062 | 0.066 | 1.432 | 1.423 | 1.431 |
| $S_6$ | 0.048 | 0.052 | 0.061 | 1.410 | 1.413 | 1.419 | 0.048 | 0.054 | 0.064 | 1.409 | 1.415 | 1.422 |
| $S_7$ | 0.053 | 0.056 | 0.067 | 1.416 | 1.413 | 1.431 | 0.054 | 0.059 | 0.071 | 1.415 | 1.416 | 1.435 |
| $S_8$ | 0.049 | 0.052 | 0.058 | 1.411 | 1.411 | 1.417 | 0.049 | 0.053 | 0.060 | 1.410 | 1.412 | 1.421 |
| $S_9$ | 0.048 | 0.049 | 0.054 | 1.408 | 1.410 | 1.412 | 0.048 | 0.050 | 0.056 | 1.408 | 1.410 | 1.414 |
| $S_{10}$ | 0.049 | 0.052 | 0.057 | 1.408 | 1.415 | 1.416 | 0.049 | 0.053 | 0.060 | 1.409 | 1.415 | 1.419 |
| $S_{11}$ | 0.054 | 0.059 | 0.059 | 1.419 | 1.432 | 1.418 | 0.057 | 0.060 | 0.062 | 1.424 | 1.431 | 1.420 |
| $S_{12}$ | 0.047 | 0.051 | 0.060 | 1.407 | 1.413 | 1.423 | 0.048 | 0.053 | 0.063 | 1.408 | 1.414 | 1.425 |
| $S_{13}$ | 0.048 | 0.051 | 0.059 | 1.409 | 1.411 | 1.420 | 0.048 | 0.053 | 0.061 | 1.409 | 1.412 | 1.422 |
| $S_{14}$ | 0.047 | 0.050 | 0.058 | 1.406 | 1.410 | 1.420 | 0.047 | 0.051 | 0.060 | 1.407 | 1.411 | 1.422 |
| $S_{15}$ | 0.052 | 0.062 | 0.065 | 1.411 | 1.419 | 1.422 | 0.054 | 0.064 | 0.067 | 1.412 | 1.423 | 1.425 |
| $S_{16}$ | 0.053 | 0.056 | 0.061 | 1.421 | 1.415 | 1.422 | 0.054 | 0.058 | 0.064 | 1.421 | 1.417 | 1.423 |
| $S_{17}$ | 0.051 | 0.055 | 0.060 | 1.412 | 1.415 | 1.420 | 0.053 | 0.057 | 0.062 | 1.412 | 1.415 | 1.421 |
| $S_{18}$ | 0.048 | 0.052 | 0.059 | 1.409 | 1.417 | 1.423 | 0.049 | 0.052 | 0.061 | 1.410 | 1.415 | 1.425 |
| $S_{19}$ | 0.047 | 0.051 | 0.062 | 1.408 | 1.413 | 1.426 | 0.048 | 0.052 | 0.064 | 1.408 | 1.413 | 1.428 |
| $S_{20}$ | 0.048 | 0.052 | 0.066 | 1.409 | 1.414 | 1.434 | 0.048 | 0.054 | 0.068 | 1.409 | 1.415 | 1.437 |

# Appendix D

# Sensitivity Study for Parameter $\varphi$: PM10 and Temperature

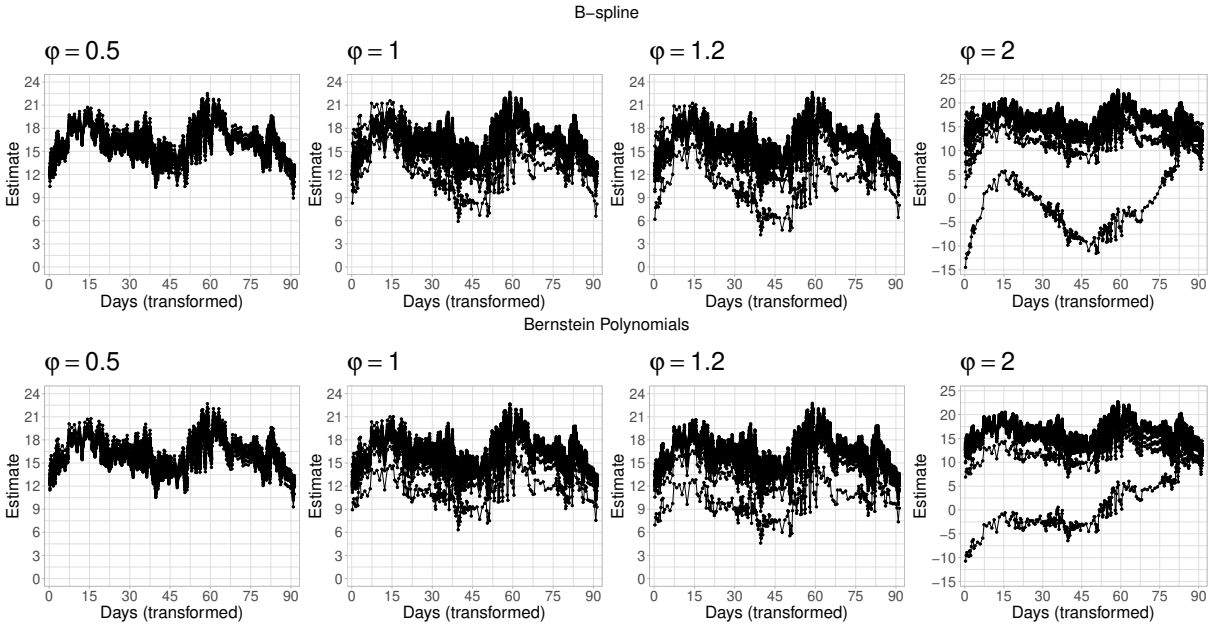
A sensitivity study was conducted to select the appropriate value for the decay parameter in the Gaussian covariance function. This structure is integrated into the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$ models, introducing a dependent relationship between the curves. Four alternatives were evaluated for this study, represented by $\varphi = \{0.5, 1, 1.2, 2\}$. The final choice of $\varphi$ should be based on a balance between model fit and a sound understanding of the underlying spatial processes in your data.

Figure D.1: Smoothed curves with the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$ models for the PM10 data set of Mexico City. Four options are considered for the decay parameter $\varphi = \{0.5, 1, 1.2, 2\}$.




Source: Prepared by the author

Figure D.2: Smoothed curves with the $\mathcal{M}_{B_{4,4};\delta}$ and $\mathcal{M}_{BP_3;\delta}$ models for the Mexico City Temperature dataset. Four options are considered for the decay parameter $\varphi = \{0.5, 1, 1.2, 2\}$.




Source: Prepared by the author

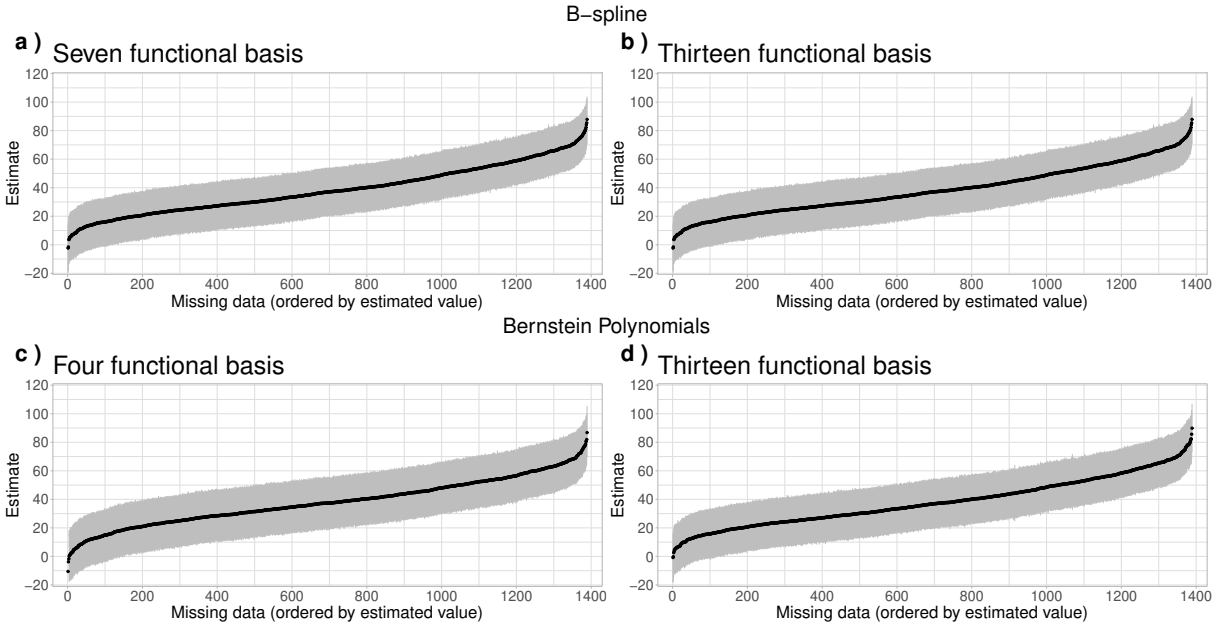
In Figures D.1 and D.2, it is evident that the distances between stations do not significantly influence when $\varphi = 0.5$, indicating a strong correlation. This is evident because all the curves closely cluster together. However, when $\varphi$ takes values of 1 and 1.2, one begins to observe that the curves are affected by the distance between locations. In simpler terms, Curves at closely spaced sites exhibit similar behavior to those at distant locations. Finally, when using a value of $\varphi = 2$ (weak correlation), It is evident that both applications display a higher variability in their measurements, which is not consistent with the characteristics of observed data.

# Appendix E

# Extra Results of the Application

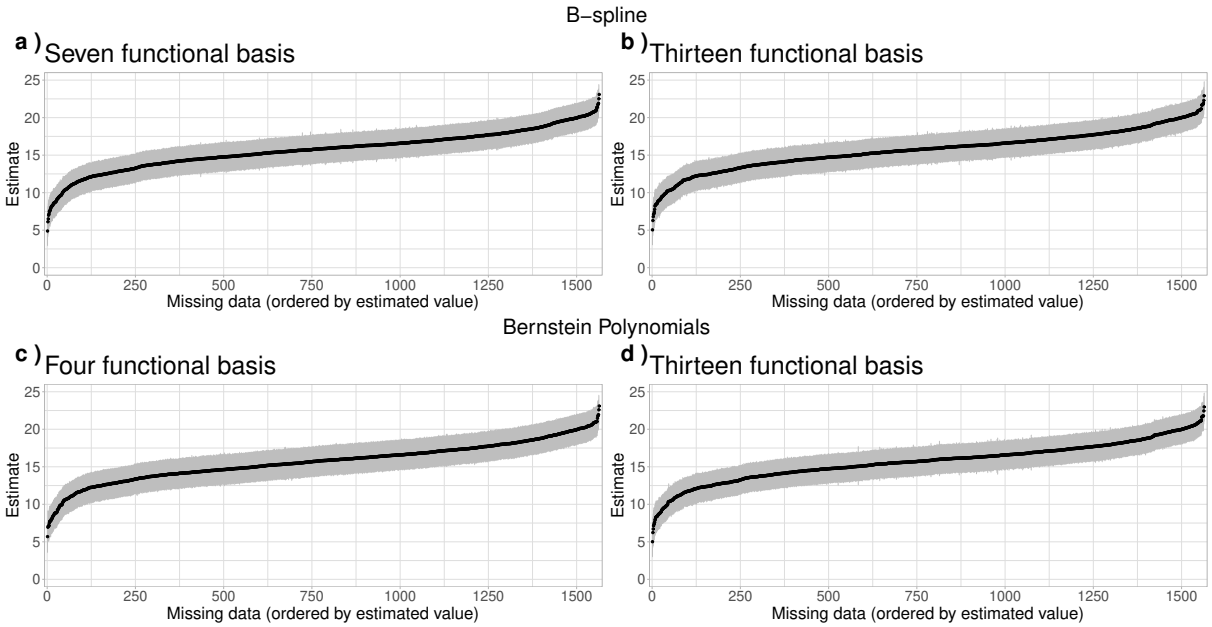
Figure E.1 displays the 95% HPD intervals for each posterior mean estimate (shown as black dots) for the missing data in the 22 PM10 data series. It is observed that the intervals are quite wide in both structures, B-spline or BP, indicating high uncertainty in the estimates. Conversely, Figure E.2 shows that the HPD intervals for the missing data in the 28 temperature series, which have narrower widths indicate a higher level of accuracy in the estimation process. Note that this behavior is similar for both structures (B-spline and BP).

Figure E.1: The 95% HPD Intervals for the 1,389 missing data using the $\mathcal{M}_{B_{4,4};\delta}$ (Panel $a$), $\mathcal{M}_{B_{4,10};\delta}$ (Panel $b$), $\mathcal{M}_{BP_3;\delta}$ (Panel $c$) and $\mathcal{M}_{BP_{12};\delta}$ (Panel $d$) models. The posterior means estimators are the black dots. Each model is fitted to the PM10 set for Mexico City.




Source: Prepared by the author

Figure E.2: The 95% HPD Intervals for the 1,564 missing data using the $\mathcal{M}_{B_{4,4};\delta}$ (Panel $a$), $\mathcal{M}_{B_{4,10};\delta}$ (Panel $b$), $\mathcal{M}_{BP_3;\delta}$ (Panel $c$) and $\mathcal{M}_{BP_{12};\delta}$ (Panel $d$) models. The posterior means estimators are the black dots. Each model is fitted to the Temperature set for Mexico City.




Source: Prepared by the author