Wiley Series in Probability and Statistics

GEOSTATISTICAL FUNCTIONAL DATA ANALYSIS

JORGE MATEU | RAMÓN GIRALDO

WILEY

Geostatistical Functional Data Analysis

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by *Walter A. Shewhart and Samuel S. Wilks*

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at http://www.wiley.com/go/wsps

Geostatistical Functional Data Analysis

Edited by

Jorge Mateu
University Jaume I of Castellon
Castellon, Spain

Ramón Giraldo
National University of Colombia
Bogota, Colombia

WILEY

This edition first published 2022

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Mateu, Jorge, editor. | Giraldo, Ramón, editor.
Title: Geostatistical functional data analysis / edited by Jorge Mateu, Ramón Giraldo.
Description: Hoboken, NJ : Wiley, 2022. | Series: Wiley series in probability and statistics | Includes bibliographical references and index.
Identifiers: LCCN 2021015788 (print) | LCCN 2021015789 (ebook) | ISBN 9781119387848 (hardback) | ISBN 9781119387909 (adobe pdf) | ISBN 9781119387886 (epub)
Subjects: LCSH: Geology–Statistical methods. | Kriging. | Spatial analysis (Statistics) | Functional analysis.
Classification: LCC QE33.2.S82 G434 2022 (print) | LCC QE33.2.S82 (ebook) | DDC 551.072/7–dc23
LC record available at https://lccn.loc.gov/2021015788
LC ebook record available at https://lccn.loc.gov/2021015789

Cover Design: Wiley
Cover Image: © Googee/Shutterstock

Set in 9.5/12.5pt STIXTwoText by Straive, Chennai, India

10 9 8 7 6 5 4 3 2 1

Contents

List of Contributors

Ana M. Aguilera
University of Granada
Department of Statistics and
Operational Research
Spain

Mohamed-Salem Ahmed
University of Lille
France

Mara S. Bernardi
Politecnico di Milano
MOX - Department of Mathematics
Italy

Gregory Bopp
Pennsylvania State University
Department of Statistics
USA

Martha Bohorquez
National University of Colombia
Department of Statistics
Colombia

Laurence Broze
University of Lille
France

María del Carmen Aguilera Morillo
Universitat Politècnica de València
Department of Statistics and
Operational Research and Quality
Spain

Sophie Dabo-Niang
University of Lille
France

Maria Durban
Universidad Carlos III
Department of Statistics
Spain

John Ensley
Pennsylvania State University
Department of Statistics
USA

Maria Franco-Villoria
Università di Modena e Reggio Emilia
Department of Economics
"*Marco Biagi*"
Italy

Zied Gharbi
University of Lille
France

Ramón Giraldo
National University of Colombia
Department of Statistics
Colombia

Alberto Guadagnini
Politecnico di Milano
Department of Civil and Environmental Engineering
Italy

and

The University of Arizona
Department of Hydrology and Atmospheric Sciences
USA

Rosaria Ignaccolo
Università degli Studi di Torino
Dipartimento di Economia e Statistica "Cognetti de Martiis"
Italy

Antonio Irpino
University of Campania "Luigi Vanvitelli"
Department of Mathematics and Physics
Italy

Piotr Kokoszka
Colorado State University
Department of Statistics
USA

Sara Sjöstedt de Luna
Umeå University
Department of Mathematics and Mathematical Statistics
Sweden

Dae-Jin Lee
BCAM–Basque Center for Applied Mathematics
Spain

Claude Manté
Université du Sud Toulon-Var
CNRS/INSU, IRD, MIO, Aix-Marseille Université
France

Jorge Mateu
University Jaume I of Castellon
Department of Mathematics
Spain

Alessandra Menafoglio
Politecnico di Milano
MOX - Department of Mathematics
Italy

Pascal Monestiez
INRAE - Unité BioSP
France

David Nerini
Université du Sud Toulon-Var
CNRS/INSU, IRD, MIO, Aix-Marseille Université
France

Federica Passamonti
Politecnico di Milano
MOX - Department of Mathematics
Italy

Davide Pigoli
King's College London
UK

Alessia Pini
Università Cattolica del Sacro Cuore
Department of Statistical Sciences
Italy

Cristian Preda
Institute of Statistics and Applied Mathematics of the Romanian Academy
Romania

Matthew Reimherr
Pennsylvania State University
Department of Statistics
USA

Elvira Romano
University of Campania "Luigi Vanvitelli"
Department of Mathematics and Physics
Italy

Laura M. Sangalli
Politecnico di Milano
MOX - Department of Mathematics
Italy

Piercesare Secchi
Politecnico di Milano
MOX - Department of Mathematics
Italy

and

CADS - Center for Analysis Decisions and Society
Human Technopole
Italy

Johan Strandberg
Umeå University
Department of Statistics
Sweden

Camille Ternynck
University of Lille
France

Baba Thiam
University of Lille
France

Vincent Vandewalle
University of Lille
France

Simone Vantini
Politecnico di Milano
MOX - Department of Mathematics
Italy

Valeria Vitelli
University of Oslo
Oslo Center for Biostatistics and Epidemiology
Department of Biostatistics
Norway

Anne-Françoise Yao
Université Clermont-Auvergne
France

Foreword

Functional data analysis (FDA) is a branch of statistics that analyses data providing information about curves, surfaces, or anything else varying over a continuum. In its most general form, under an FDA framework each sample element is a function. The continuum over which these functions are defined is often time, but may also be spatial location, wavelength, probability, etc. In the 20 years since the first books and papers on this topic, this field of statistics has received the attention and encouragement of researchers in statistics and many applied disciplines and has become an important and dynamic area of modern statistics. Topics that have been covered include descriptive techniques, statistical inference, multivariate and nonparametric methods, regression, generalized linear models, time series, and spatial statistics.

Modern technology has made it possible to obtain large spatial and spatiotemporal data sets, and poses the challenge of statistical modeling of such data. The combination of spatial statistics with FDA has emerged as a key approach. This book presents new theories and methods to define, describe, characterize, and model functional data indexed in spatial or spatio-temporal domains. The main focus is on functional data obtained under a geostatistical framework, where the domain is fixed and continuous. Specific topics considered include kriging, clustering, regression, and optimal sampling, moving on in the last part of the book to spatiotemporal data. Some chapters also consider the treatment of functional data on lattices.

When we wrote our original book on the subject in the 1990s, James Ramsay and I hoped that we would encourage FDA as a way of thinking, not simply a collection of techniques. It has therefore been very pleasing to see the development of the field since then, and the abundance of research activity in the area has confirmed our hopes. I would urge readers and researchers to raise their sights above any specific methods, obviously important that they are, to ask how considering data as functions changes and broadens our statistical horizons. Particularly in the new era of data science, this concerns both what data can be collected and how they can be analyzed. I am sure this book will make a valuable contribution in helping them to do so.

November 2020

Sir Bernard Silverman
University of Oxford
University of Nottingham
Oxford and Nottingham

1

Introduction to Geostatistical Functional Data Analysis

Jorge Mateu[1] *and Ramón Giraldo*[2]

[1] *Department of Mathematics, University Jaume I of Castellon, Spain*
[2] *Department of Statistics, National University of Colombia, Bogota, Colombia*

1.1 Spatial Statistics

Spatial statistics has developed rapidly during the last 30 years. We have seen an interesting progress both in theoretical developments and in practical studies. Some early applications were in mining, forestry, and hydrology. It seems to be honest to remark that the increasing availability of computer power and skillful computer software has stimulated the ability to solve increasingly complex problems. Clearly, these problems have some common elements: they were all of a spatial nature. Some theory was available, for example the random function theory as developed by Yaglom and others in the 1960s. But that was largely insufficient to find generic solutions for the whole class of problems, and hence, the applications required a new theory. Thereupon some far-reaching theories have been developed: image reconstruction, Markov random fields, point process statistics, geostatistics, and random sets, to mention just a few. As a next stage, these theories were applied successfully to new disciplinary problems leading to modifications and extensions of mathematical and statistical procedures. We therefore notice a general scientific process that has occurred in the field of spatial statistics: well-defined problems with a common character were suddenly on the agenda, and data availability and intensive discussion with practical and disciplinary researchers resulted in new theoretical developments. Often, it is difficult to say which was first, and what followed, but we see different theoretical models developed for different applications.

Spatial statistics has hence emerged as an important new field of science. One of the peculiarities is its power for visualization. A common cold-water fear of many statisticians and mathematicians to analyze images, to communicate their

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

results by maps, and to have to trust information in pictures was overcome. It has led to interesting theories and better and more objective procedures for dealing with spatial variation. Following Wittgenstein, we could state that we needed some geniuses to tackle the obvious. Now, many results of a spatial statistical analysis could be communicated smoothly toward the nonstatistical audience, like a disciplinary scientist, a policy-maker, or an interested student. They, in turn, were able to judge whether a problem was solved, whether a policy measure was relevant or was inspired by the beautiful pictures expressing deep thoughts on relevant issues.

The role in policy-making may be once more stressed. It is known that many policy-makers are inclined to make a decision on the basis of a well developed, well organized, and well understandable figure. They find it (rightly so!) rather boring to use long lists of statistical data. But as political decisions affect us all, it puts another responsibility on the back of statisticians: to make statistically sound maps. It is often hard to say what that should be, but at the very least, we should be able to generate pictures, maps, and graphs that rely on good data and that show important aspects for decision-making.

In this way, spatial statistics has become a refreshing wind in statistics. We do not need to do well much longer on difficult equations, long lists of data, and tables with simulated controlled scenarios. But, to be clear on the back of all these nice pictures a sound science with sometimes difficult and tedious derivations and deep thoughts are still required to make serious progress.

Spatial statistics recognizes and exploits the spatial locations of data when designing for, collecting, managing, analyzing, and displaying such data. Spatial data are typically dependent, for which there are classes of spatial models available that allow process prediction and parameter estimation. Spatially arranged measurements and spatial patterns occur in a surprisingly wide variety of scientific disciplines. The origins of human life link studies of the evolution of galaxies, the structure of biological cells, and settlement patterns in archaeology. Ecologists study the interactions among plants and animals. Foresters and agriculturalists need to investigate plant competition and account for soil variations in their experiments. The estimation of rainfall and of ore and petroleum reserves is of prime economic importance. Rocks, metals, and tissue and blood cells are all studied at a microscopic level. Geology, soil science, image processing, epidemiology, crop science, ecology, forestry, astronomy, atmospheric science, or simply any discipline that works with data collected from different spatial locations, need to develop models that indicate when there is dependence between measurements at different locations. Spatiotemporal variability is a relatively new area within Spatial Statistics, which explains the scarcity of space-time statistical tools 20 years ago. There has been a growing realization in the last decade that knowing where data were observed could help enormously in answering the substantive questions that precipitated their collection. One of the most powerful

tools for spatial data analysis is the map. For example, in military applications, the battlespace is mapped for command and control. The sensors are both *in situ* and remote, and they generate spatially distributed data of many different kinds. Producing a statistically optimal map, together with measures of map uncertainty, which is always up to date, is a complicated task. Once these types of statistical problems are solved, a geographic information system, or GIS, is well suited to forming the decision-making maps.

Spatial statistics can be considered a natural generalization of signal processing to higher dimensions. In traditional signal processing, one has a signal dependent on a scalar variable t, which may belong to a discrete set or which may be continuous. Spatial statistics is concerned with cases in which t is a multidimensional index of dimension $d > 1$. In most practical examples $d = 2$, though much of the basic theory and methodology is the same whatever the dimension. Although the models and methods of spatial statistics have not developed as rapidly as those for one-dimensional signal processing, there have nevertheless been substantial new developments in recent years. Standard and modern references on spatial statistics include the books of [1–4] among others.

Following Cressie [5], spatial data can be thought of as resulting from observations on the stochastic process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$, where D is possibly a random set in $\mathbb{R}^d$. If we believe that the roots of statistical science are in data, we can classify spatial areas according to the type of observations encountered. Thus, (i) if D is a fixed subset of $\mathbb{R}^d$ and $Z(\boldsymbol{s})$ is a random vector at location $\boldsymbol{s} \in D$, we are dealing with *geostatistical data*; (ii) if D is a fixed (regular or irregular) collection of countably many points of $\mathbb{R}^d$ and $Z(\boldsymbol{s})$ is a random vector at location $\boldsymbol{s} \in D$, we are dealing with *lattice data*; (iii) if D is a point process in $\mathbb{R}^d$ and $Z(\boldsymbol{s})$ is a random vector at location $\boldsymbol{s} \in D$, we are dealing with *point patterns*; (iv) if D is a point process in $\mathbb{R}^d$ and $Z(\boldsymbol{s})$ is itself a random set, we are dealing with *spatial objects*. Geostatistical-type problems are distinguished most clearly from lattice-and point-pattern-type problems by the ability of the spatial index s to vary continuously over a subset of $\mathbb{R}^d$. A space-time process can be denoted by $\{Z(\boldsymbol{s}, t) : \boldsymbol{s} \in D(t), t \in T\}$, where each of Z, D, and T is possibly random.

Spatial statistics is one of the major methodologies of *environmental statistics*. Its applications include producing spatially smoothed or interpolated representations of air pollution fields, calculating regional average means or regional average trends based on data at a finite number of monitoring stations, and performing regression analyses with spatially correlated errors to assess the agreement between observed data and the predictions of some numerical model. The notion of proximity in space is implicitly or explicitly present in the environmental sciences. Proximity is a relative notion, relative to the spatial scale of the scientific investigation. When a spatial dimension is present in an environmental study, the statistician's job is to create a statistical framework within which one carries out

defensible inferences on processes and parameters of interest. These modeling and inference strategies are not always easy to do, but are never impossible. If statistics is to continue to be the broker of variability, it must address difficult questions such as those found in the environmental sciences, otherwise, it will become marginalized as a discipline. Problems in the environmental sciences are inherently spatial (and temporal), observational in nature, and have experimental units that are highly variable.

In the last decade, spatial statistics has undergone enormous development in the area of statistical modeling. It started slowly, building from models that were purely descriptive of spatial dependence. Then, it became apparent that the process of interest was usually hidden by measurement error and that the principal goal should be inference on the hidden process from the noisy data. It has only been in the last few years that the full potential for hierarchical spatial statistical modeling has been glimpsed. There is an enormous amount of flexibility in hierarchical statistical models, such as the opportunity to account for nonlinearities. Their attractive feature is that at each level of the hierarchy, the model specification is simple, yet globally, the model can be quite complex. This approach could be summarized as a model locally, analyze globally.

Applications of spatial statistics cover many areas. Much of the original impetus for the subject was driven by geostatistics. It was in this context that the technique of kriging, optimal least squares interpolation over a random spatial field, was originally developed. In recent years, the applications of spatial statistics have increased enormously, with particularly fruitful applications in the environmental and ecological sciences. A typical problem is the sampling of a pollution field, such as ozone in the atmosphere or toxic chemicals in rivers and lakes. Another example is the use of meteorological measurements in studies of global climate change. In these fields, as in geostatistics, the objective may be to interpolate spatially between measurements, but there are also other objectives which may be quite different. Spatial statistics has also found applications in such diverse fields as sociology, for example social networks theory and financial economics.

The usual approach in geostatistics is based on an assumption that the spatial random field is stationary and isotropic. In the original geophysical applications which motivated the development of the field, this assumption was often justified by the fact that with sparse data, there was no reasonable alternative. A further point is that many geostatistical applications involved only one measurement at each site (or equivalently, only one replication of the random field) so there was no way of determining the complete spatial covariance function without some kind of stationarity assumption. In modern environmental applications, however, there are very often enough monitoring stations to go beyond such assumptions, and with multiple observations per site, it is also possible to estimate the covariance between any pair of sites without assuming stationarity across the field. Another

consideration is that very often, simple topography makes a stationary assumption implausible. Therefore, there are by now many reasons to go beyond a stationary model. In spite of this obvious need for nonstationary models; however, there is not, as yet, a wide variety of approaches to the problem.

Environmental issues have brought atmospheric science to the center of science and technology, where it now plays a key role in shaping national and international policy. Weather prediction plays a significant role in the planning of human affairs. Further, a broader appreciation of the role of weather and climate impacts on the environment of the planet has now led to nearly universal concern regarding potential climate change, its causes, impacts, and possible remedying. A large variety of statistical methods are used routinely in the atmospheric sciences. For example, techniques of multivariate time series are especially common. These include multivariate autoregressive, moving average models and Kalman filtering. Statistical methods for spatial data are also standard. A major tool in the analysis of space-time data is empirical orthogonal functions (EOF). Virtually, all atmospheric and oceanographic processes (e.g. wind, temperature, sea surface temperature, moisture) involve variability over space and time. One only needs examine the governing partial differential equations for wind processes, or their selected spatial-temporal averages, to see that mathematical and statistical descriptions of these dynamical processes depend on complicated temporal and spatial relationships. Furthermore, observations of geophysical processes typically include measurement errors and are often temporally and spatially incomplete, which may obscure the signal of interest.

In studies involving spatial data, it is seldom the case that data for only a single process are collected. Typically, there is a great expense associated with establishing spatial monitoring networks or other mechanisms of spatial data collection (e.g. satellites) and so measurements are usually made on two or more variables. Thus, statistical techniques for multivariate spatial data are critical for effective modeling of spatial processes.

Lately, there has been a rich and growing literature on space–time modeling. Fundamentally, it is clear that in the absence of a temporal component, second-order geostatistical models can be used to represent spatial variability. These are descriptive in the sense that, although they model spatial correlation, there is no causative interpretation associated with them. Thus, for space-time modeling, the geostatistical paradigm assumes a descriptive structure for both space and time (i.e. covariance structures are directly specified). For example, one can extend the geostatistical kriging methodology for spatial processes by assuming that time is just another spatial dimension. Alternatively, one can treat time slices of a spatial field as variables and apply a multivariate or cokriging approach. Although these approaches have been successful in many applications, there are fundamental differences between space and time, and it is not likely

that realistic covariance structures can be specified that accurately capture the complicated dynamical processes as found in geophysical applications.

In the absence of a spatial component, there is a large class of time series models that could be used to represent the temporal variability. These are dynamic in the sense that they exploit the fact that time flows in only one direction, and so the state of the process at the current time is related to what happened at previous times. Thus, one might consider the space–time process as a collection of spatially correlated time series in continuous space, or on a spatial lattice. Although these approaches include dynamical structures, without a descriptive spatial component one lacks the ability to perform spatial prediction at locations without observations. If both temporal and spatial components are present, it is natural to combine the temporally dynamic state-space approach and the spatially descriptive approach. These models are referred to as space–time dynamic models.

Spatial interpolation is an essential feature of many GIS. It is a procedure for estimating values of a variable at unsampled locations. A map with isolines is usually the visual output of such a process and plays a crucial role in decision-making. Based on Tobler's law of geography, which stipulates that observations close together in space are more likely to be similar than those farther apart, the development of models attempting to represent the way close observations are related can sometimes be very problematic. The approaches can be divergent and may therefore lead to very different results. As a consequence, an understanding of the initial assumptions and methods used is the key to the spatial interpolation process.

Surprisingly, when spatial interpolation tools are integrated within GIS, they are often implemented in such a way that users have no real choice in selecting the best possible methods, and if they do have a choice, required input parameters are sometimes fixed, without any possible way of modifying them. One reason for the frequent blind use of spatial interpolation methods, and spatial statistics in general, probably has its origins in teaching. Despite the large variety of its applications, the discipline has been confined to those fields where it has seen its major developments. The progress made in spatial statistics is therefore usually presented only in journals dedicated to statistics, mining, and petroleum engineering. As a consequence, GIS users who have a different technical background often do not have an in-depth knowledge of such spatial interpolation techniques. Furthermore, since the conventional tests used in basic statistics usually generate some kind of categorical answer, the prerequisite experience and statistical knowledge necessary for the proper use of spatial interpolation techniques are often discouraging to this type of users. Nevertheless, during the last few years, the diversity of the applications of these methods has encouraged the publication of new books and new case studies and has stimulated a number of conferences on the subject.

1.2 Spatial Geostatistics

This section has been partially taken and summarized in parts from [6], intending to provide a brief overview to spatial geostatistics. The reader is referred to [6] for further and more complete details.

1.2.1 Regionalized Variables

Geostatistics can be defined as the study of regionalized phenomena, that is, phenomena that stretch across space and which have a certain spatial organization or structure. However, geostatistics is not applied to the regionalized phenomenon as such, which is a physical reality, but to a mathematical description of that reality, that is, a numerical function called *regionalized variable* or *regionalization*, defined in a geographical space, which is supposed to correctly represent and measure that phenomenon.

In order to delve deeper into the concept of regionalized variable, let us imagine we are interested in a feature of a given phenomenon that spans across space and that several measurements are taken in a domain D at a given moment in time. If the measurements are taken on objects or similar, the objects sampled can be considered a subset of a larger collection of objects, as many more measurements could have been taken, but were not for many possible reasons. If the observations were made at certain points in the domain, infinite measurements could be taken.

When $\boldsymbol{s}$ spans across the domain under study, D, the set $\{z(\boldsymbol{s}), \boldsymbol{s} \in D\}$, is called a regionalized variable or regionalization, the set $\{z(\boldsymbol{s}_i),\ i = 1, 2, 3, \dots\}$ being a collection of values of the regionalized variable, and each value of that collection being a regionalized value.

It is true that a deterministic approach can be employed to describe or model a regionalized phenomenon and obtain an accurate assessment of the values of the regionalization on the basis of a limited number of observations. However, this requires in-depth knowledge of the origin of the phenomenon and the physical or mathematical laws that govern the evolution of the regionalized variable. Furthermore, many of the regionalized phenomena that are usually studied are so complex that a deterministic approach can only partially portray them. That is why the deterministic approach is discarded and the probabilistic approach, which permits modeling both the knowledge of and also the uncertainty surrounding the regionalized random phenomenon, is adopted.

1.2.2 Random Functions

From a probabilistic perspective, the regionalized value can be seen as the result of a random mechanism, resulting in a *random variable* (r.v.). If the regionalized

values at all the points in the domain D are considered, it can be seen as a reality of an infinitely large set of r.v.s, one at each point in the domain, which is known as spatial *random function* (synonyms: stochastic process, random field).

When $\boldsymbol{s}$ spans across the domain under study, D, we have a family of r.v.s, $\{Z(\boldsymbol{s}),\ \boldsymbol{s} \in D\}$, which constitutes a spatial random field (r.f.).

This methodological decision is one of the cornerstones of geostatistics: the regionalized variable is interpreted as a realization of a spatial r.f. At this point, we must state that the regionalized variable is often highly locally irregular (which makes it impossible to represent using a deterministic mathematical function) and has a certain spatial organization or structure. The probabilistic approach, or probabilistic geostatistics, which interprets the regionalized variable as a realization of a r.f., can take into account all the aspects of regionalization mentioned above, because, as stated in page 55 of [7]:

i) At each location $\boldsymbol{s}$, $Z(\boldsymbol{s})$ is a r.v. (hence, the erratic aspect).
ii) For any given set of points $\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_k$, the r.v.s $Z(\boldsymbol{s}_1), Z(\boldsymbol{s}_2), \dots, Z(\boldsymbol{s}_k)$ are linked by a network of spatial correlations responsible for the similarity of the values they take (hence the structured aspect).

Let $Z(\boldsymbol{s})$ be a r.f. and let us consider the set of points $(\boldsymbol{s}_1, \dots, \boldsymbol{s}_k)$. Then, the r.f. $Z(\boldsymbol{s})$ is characterized by its k-dimensional distribution function. The set of k-dimensional distribution functions for all values of k and all possible choices of $(\boldsymbol{s}_1, \dots, \boldsymbol{s}_k)$ in the domain is called the *spatial law of probability*.

For a given r.f., $Z(\boldsymbol{s})$, the k-dimensional distribution function $F(z(\boldsymbol{s}_1)), \dots, (z(\boldsymbol{s}_k))$: $\mathbb{R}^d \to [0, 1]$ is defined as

$$F_{Z_1,\dots,Z_k}(z(\boldsymbol{s}_1)), \dots, (z(\boldsymbol{s}_k)) = P[Z(\boldsymbol{s}_1) \leq z(\boldsymbol{s}_1), \dots, Z(\boldsymbol{s}_k) \leq z(\boldsymbol{s}_k)]. \tag{1.1}$$

In linear geostatistics, it is enough to know the first two moments of the distribution of $Z(\boldsymbol{s})$. What is more, in most practical applications, the available information does not allow to infer higher-order moments.

The expectation, expected value or first-order moment of a r.f. is defined as a nonrandom function of $\boldsymbol{s}$ that coincides at each point with the expectation of the r.v. at that point $\mu(\boldsymbol{s}) = E(Z(\boldsymbol{s}))$, where $\mu(\boldsymbol{s}_i) = E(Z(\boldsymbol{s}_i))$, $\forall i \in \mathbb{N}$. It is also called the drift of the r.f., especially when it varies with location.

The variance of a r.f. is defined as a nonrandom function of $\boldsymbol{s}$ that coincides at each point with the variance of the r.v. at that point, i.e. $V(\boldsymbol{s}) = V(Z(\boldsymbol{s}))$, where $V(\boldsymbol{s}_i) = V(Z(\boldsymbol{s}_i))$, $\forall i \in \mathbb{N}$.

The covariance function of a r.f. is defined as a nonrandom function of $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, such that for any pair of values $(\boldsymbol{s}_i, \boldsymbol{s}_j)$ coincides with the covariance between the r.v. at those two points

$$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = C(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)) = E\left((Z(\boldsymbol{s}_i) - \mu(\boldsymbol{s}_i))(Z(\boldsymbol{s}_j) - \mu(\boldsymbol{s}_j))\right), \quad \forall \boldsymbol{s}_i, \boldsymbol{s}_j \in D. \tag{1.2}$$

The variogram of the r.f. is defined as the variance of the first differences of the r.f.

$$2\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) = V\left((\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)\right), \quad \forall \boldsymbol{s}_i, \boldsymbol{s}_j \in D. \tag{1.3}$$

The function $\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is called semivariogram.

$Z(\boldsymbol{s})$ is a Gaussian r.f. if for all k and any given set of points $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k$, the joint distribution of $Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_k)$ is a multivariate Gaussian distribution. A multivariate Gaussian distribution is characterized by a mean vector and a variance–covariance matrix, such that the two first moments of a Gaussian r.f. completely determine its probability structure. The Gaussianity of the r.f. is a common assumption in geostatistics.

1.2.3 Stationarity and Intrinsic Hypothesis

regionalized variable in probabilistic terms as a particular realization of a given r.f. $\{Z(\boldsymbol{s}),\ \boldsymbol{s} \in D\}$ makes operational sense when it is possible to infer part or all of the law of probability which defines that r.f. In this sense, stationarity, which indicates a certain degree of homogeneity in the regionalization across space, is a desirable quality.

Indeed, it would be impossible to infer the probability law of a r.f. if there was only one realization of the r.f. In order to make inferences consistently, many realizations are necessary. However, in reality there is only one. The solution to this problem is to adopt the *hypothesis of stationarity or spatial homogeneity*. The idea behind the hypothesis of stationarity is to substitute repetitions of the (inaccessible) realizations of the r.f. with repetitions in space, that is, the values observed at different locations in the domain under study have the same characteristics and can be considered as realizations of the same r.f. in mathematical terms. However, these realizations are not independent, and an additional hypothesis, ergodicity, is normally assumed; see pages 19–22 of [8] for details. The hypothesis of stationarity means that the spatial law of probability of the r.f. or part of it, is translation invariant. That is, the probabilistic properties of a set of observations do not depend on the specific locations where they have been measured, but only on their separations.

Therefore, in mathematical and probabilistic terms, the hypothesis of stationarity refers to the regular behavior in space of the moments of the r.f., or the function itself and, as we will see later, there are different degrees of stationarity. This hypothesis will allow us to act as if all the variables that make up the r.f. had the same probability distribution (or the same moments; we can even relax this assumption) and, as a consequence, to be able to make inferences.

Using the assumed level of spatial homogeneity of the r.f. that (supposedly) generates the observed realization as a basis, we have the following cases: Stationary random function in the strict sense, second-order stationary random function, and

intrinsically stationary random function or random function of stationary increments. Let us briefly introduce these concepts.

The r.f. $\{Z(\boldsymbol{s}),\ \boldsymbol{s} \in D\}$ is said to be *stationary in the strict sense*, or strictly stationary, if the families of r.v.s $Z(\boldsymbol{s}_1), \dots, Z(\boldsymbol{s}_k)$ and $Z(\boldsymbol{s}_1 + \mathbf{h}), \dots, Z(\boldsymbol{s}_k + \mathbf{h})$ have the same joint distribution function for all k, and for any given spatial points and any translation vector $\mathbf{h} \in \mathbb{R}^d$.

In other words, the joint distribution function of $\{Z(\boldsymbol{s}_1), Z(\boldsymbol{s}_2), \dots, Z(\boldsymbol{s}_k)\}$ is unaffected by the translation of an arbitrary quantity $\mathbf{h}$. As a result, density functions with dimension lower than k do not depend on location either. Generally speaking, this is a strongly strict condition, which is why this hypothesis is normally relaxed to the so-called "assumption of second-order stationarity," which limits the stationarity hypothesis to the first two moments of the r.f. (recall that in linear geostatistics, we are only interested in the two first moments of the r.f.).

The r.f. $\{Z(\boldsymbol{s}),\ \boldsymbol{s} \in D\}$ is said to be *second-order stationary, weakly stationary or stationary in the broad sense*, if it has finite second-order moments (that is the covariance exists) and verifies that

- The expectation exists and is constant, and therefore does not depend on the location $\boldsymbol{s}$

$$E\left(Z(\boldsymbol{s})\right) = \mu(\boldsymbol{s}) = \mu. \tag{1.4}$$

- The covariance exists for every pair of r.v.s, $Z(\boldsymbol{s})$ and $Z(\boldsymbol{s} + \mathbf{h})$, and only depends on the vector $\mathbf{h}$ that joins the locations $\boldsymbol{s}$ and $(\boldsymbol{s} + \mathbf{h})$

$$C\left(Z(\boldsymbol{s}), Z(\boldsymbol{s} + \mathbf{h})\right) = C(\mathbf{h}),\ \forall \boldsymbol{s} \in D \text{ and } \mathbf{h}. \tag{1.5}$$

As the covariance function $C(\mathbf{h})$ of a second-order Stationary, r.f. is only a function of $\mathbf{h}$, the variance of the r.f. exists and is finite and constant:

$$V(Z(\boldsymbol{s})) = C(\mathbf{0}) = \sigma^2. \tag{1.6}$$

In light of Eqs. (1.4) and (1.6), the second-order stationarity hypotheses can be interpreted as if the regionalized variable takes values that fluctuate around a constant value (the mean), and the variation of these fluctuations is the same everywhere in the domain.

In some cases, in order to model the spatial dependence of second-order stationary r.f.s, the *correlogram*, or *correlation function*, is used instead of the covariogram, and is defined as

$$\text{Corr}(Z(\boldsymbol{s}), Z(\boldsymbol{s} + \mathbf{h})) = \frac{C(\mathbf{h})}{C(\mathbf{0})} = \rho(\mathbf{h}). \tag{1.7}$$

In case of second-order stationarity, the covariance function and the semivariogram are equivalent when it comes to defining the structure of spatial

dependence displayed by the phenomenon, as they verify the following mutual relationship:

$$\begin{aligned}\gamma(\mathbf{h}) &= \frac{1}{2}V\left(Z(\boldsymbol{s}+\mathbf{h})-Z(\boldsymbol{s})\right)\\ &= \frac{1}{2}\left(V(Z(\boldsymbol{s}+\mathbf{h}))+V(Z(\boldsymbol{s}))+2C(Z(\boldsymbol{s}+\mathbf{h}),Z(\boldsymbol{s}))\right)\\ &= \frac{1}{2}\left(C(\mathbf{0})+C(\mathbf{0})+2C(\mathbf{h})\right)\\ &= C(\mathbf{0})-C(\mathbf{h}).\end{aligned}$$

Notice that if a r.f. is strictly stationary, then it is also stationary in the broad sense. The converse, however, is generally not true. Obviously, for Gaussian r.f.s, second-order stationarity is equivalent to strict stationarity.

A r.f. is said to be quasistationary when the corresponding stationary hypothesis (usually, the hypothesis of second-order stationarity) is valid only for distances $|\mathbf{h}| < d$, where d is a limit distance. That is, in the second-order quasistationary case (usually referred as the quasistationary case) $\mu(\boldsymbol{s}+\mathbf{h}) \approx \mu(\boldsymbol{s})$ if $|\mathbf{h}| < d$ and $C\left(Z(\boldsymbol{s}+\mathbf{h})-Z(\boldsymbol{s})\right) = C(\mathbf{h})$ if $|\mathbf{h}| < d$.

Second-order stationarity can also be considered a strict assumption on many occasions, as it implies the existence of the variance in the r.f. A phenomenon may have infinite variation capacity and be impossible to model using a r.f. with finite variance. However, there are cases in which the increments or differences $Z(\boldsymbol{s}+\mathbf{h})-Z(\boldsymbol{s})$ do have finite variance and, therefore, are second-order stationary. This type of r.f. is described as being intrinsically stationary.

The r.f. $\{Z(\boldsymbol{s}),\ \boldsymbol{s}\in D\}$ is said to be *intrinsically stationary (or simply intrinsic)* if, for any given translation vector $\mathbf{h}$, the first-order increments $Z(\boldsymbol{s}+\mathbf{h})-Z(\boldsymbol{s})$ are second-order stationary, that is,

$$E\left(Z(\boldsymbol{s}+\mathbf{h})-Z(\boldsymbol{s})\right) = \mu(\boldsymbol{s}), \tag{1.8}$$

where $\mu(\boldsymbol{s})$, the drift, is necessarily linear in $\mathbf{h}$, and

$$C\left(\left(Z(\boldsymbol{s}+\mathbf{h})-Z(\boldsymbol{s})\right),\left(Z(\boldsymbol{s}+\mathbf{h}+\mathbf{h}')-Z(\boldsymbol{s}+\mathbf{h}')\right)\right) = C(\mathbf{h},\mathbf{h}'), \tag{1.9}$$

which is equivalent to

$$\frac{1}{2}V\left(Z(\boldsymbol{s}+\mathbf{h})-Z(\boldsymbol{s})\right) = \gamma(\mathbf{h}), \tag{1.10}$$

which is only a function of $\mathbf{h}$.

Obviously, in case that the linear drift is zero

$$E\left(Z(\boldsymbol{s}+\mathbf{h})-Z(\boldsymbol{s})\right) = 0$$

and

$$E(Z(\boldsymbol{s}+\mathbf{h})-Z(\boldsymbol{s}))^2 = \gamma(\mathbf{h}),$$

If a r.f. is second-order stationary, then it is also intrinsically stationary. However, the converse is not necessarily true. Intrinsic r.f.s that are not second-order stationary are called *strictly intrinsic* r.f.s. In particular, a Gaussian intrinsic r.f. is an intrinsic r.f. whose increments follow a multivariate Gaussian distribution.

A r.f. is said to be quasiintrinsic when the intrinsic hypotheses is valid only for distances $|\mathbf{h}| < d$, where d is a limit distance.

A r.f. $\{Z(\boldsymbol{s}),\ \boldsymbol{s} \in D\}$ for which the mean and/or the covariance function depends on the location (are not translation invariant), is said to be a nonstationary r.f.

When a r.f. $\{Z(\boldsymbol{s}),\ \boldsymbol{s} \in D\}$ has a drift, i.e. its mean is nonconstant and varies with location, and its first-order increments $Z(\boldsymbol{s}+\mathbf{h}) - Z(\boldsymbol{s})$ are nonstationary, it is said that the r.f. is a *nonintrinsic* r.f. (some authors call them *intrinsic random functions* of order $k > 0$).

1.3 Spatiotemporal Geostatistics

This section has been partially taken and summarized in parts from [9], intending to provide a brief overview to spatiotemporal geostatistics. The reader is referred to [9] for further and more complete details.

Geostatistical research has typically analyzed r.f.s, in which every spatiotemporal location can be seen as a point on $\mathbb{R}^d \times \mathbb{R}$. While from a mathematical point of view $\mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+1}$, from a physical perspective, it would make no sense to consider spatial and temporal aspects in the same way, due to the significant differences between the two axes of coordinates. Therefore, while the time axis is ordered intrinsically (as it exists in the past, present, and future), the same does not occur with the spatial coordinates.

Assume that observations stem from a r.f. given by $Z(x, \boldsymbol{s}, t) = \eta(x(\boldsymbol{s}, t), \boldsymbol{s}, t, \beta) + \epsilon(x, \boldsymbol{s}, t), \boldsymbol{s} \in D, t \in T$, where $\boldsymbol{s}$ denotes a spatial location, t a time point, x some potentially space and time-dependent regressors, η a parametrized trend model, $D \subset \mathbb{R}^d$ (very often $d = 2$), and $T \subset \mathbb{R}$. For ease of notation, we remove the term in the covariates x, and write $Z(\boldsymbol{s}, t)$, assuming whenever necessary that any trend coming from a set of covariates has already been removed.

1.3.1 Relevant Spatiotemporal Concepts

A spatiotemporal r.f. $Z(\boldsymbol{s}, t)$ is said to be *Gaussian* if the random vector $\mathbf{Z} = (Z(\boldsymbol{s}_1, t_1), \ldots, Z(\boldsymbol{s}_n, t_n))'$ for any set of spatiotemporal locations follows a multivariate normal distribution. When not stated explicitly, the indexes i and j will go from 1 to n.

The spatiotemporal r.f. $Z(\boldsymbol{s}, t)$ is said to have a *spatially stationary covariance* function if, for any two pairs $(\boldsymbol{s}_i, t_i)$ and $(\boldsymbol{s}_j, t_j)$ on $\mathbb{R}^d \times \mathbb{R}$, the covariance

$C((\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j))$ only depends on the distance between the locations ($\boldsymbol{s}_i$ and $\mathbf{s}_j$) and the times t_i and t_j. And the spatiotemporal r.f. $Z(\boldsymbol{s}, t)$ is said to have a *temporarily stationary covariance* function if, for any two pairs $(\mathbf{s}_i, t_i)$ and $(\mathbf{s}_j, t_j)$ on $\mathbb{R}^d \times \mathbb{R}$, the covariance $C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$ only depends on the distance between the times (t_i and t_j) and the spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$. If the spatiotemporal r.f. $Z(\boldsymbol{s}, t)$ has a *stationary covariance function* in both spatial and temporal terms, then it is said to have a stationary covariance function. In this case, the covariance function can be expressed as

$$C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = C(\mathbf{h}, u) \tag{1.11}$$

with $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$ and $u = t_i - t_j$ the distances in space and time, respectively.

A spatiotemporal r.f. $Z(\mathbf{s}, t)$ has a *separable covariance function* if there is a purely spatial covariance function $C_s(\boldsymbol{s}_i, \boldsymbol{s}_j)$ and a purely temporal covariance function $C_t(t_i, t_j)$ such that

$$C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = C_s(\mathbf{s}_i, \mathbf{s}_j) C_t(t_i, t_j) \tag{1.12}$$

for any pair of spatiotemporal locations $(\mathbf{s}_i, t_i)$ and $(\mathbf{s}_j, t_j) \in \mathbb{R}^d \times \mathbb{R}$.

A spatiotemporal r.f. $Z(\mathbf{s}, t)$ has a *fully symmetrical covariance function* if

$$C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = C_s(\mathbf{s}_i, t_j) C_t(\mathbf{s}_j, t_i) \tag{1.13}$$

for any pair of spatiotemporal locations $(\mathbf{s}_i, t_i)$ and $(\mathbf{s}_j, t_j) \in \mathbb{R}^d \times \mathbb{R}$.

Separability is a particular case of complete symmetry and, as such, any test to verify complete symmetry can be used to reject separability. In the case of stationary spatiotemporal covariance functions, the condition of full symmetry reduces to

$$C(\mathbf{h}, u) = C(\mathbf{h}, -u) = C(-\mathbf{h}, u) = C(-\mathbf{h}, -u), \ \forall (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}. \tag{1.14}$$

A spatiotemporal r.f. has a *compactly supported covariance function* if, for any pair of spatiotemporal locations $(\mathbf{s}_i, t_i)$ and $(\mathbf{s}_j, t_j) \in \mathbb{R}^d \times \mathbb{R}$, the covariance function $C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$ tends toward zero when the spatial or temporal distance is sufficiently large.

If $C(\boldsymbol{s}_i - \boldsymbol{s}_j, t_i - t_j)$ depends only on the distance between positions, that is $\left(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|, t_i - t_j\right)$, the r.f., apart from being stationary, is also *isotropic* in space and time. Note that if the covariance function of a stationary r.f. is isotropic in space and time, then it is fully symmetrical.

The spatiotemporal *variogram* is defined as the function

$$2\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = V(Z(\mathbf{s}_i, t_i) - Z(\mathbf{s}_j, t_j)), \tag{1.15}$$

where V is the variance, and half this quantity is called a semivariogram.

In the case of a r.f. with a zero mean,

$$2\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = E[(Z(\mathbf{s}_i, t_i) - Z(\mathbf{s}_j, t_j))^2]. \tag{1.16}$$

Whenever it is possible to define the covariance function and the variogram, they will be related by means of the following expression:

$$2\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = V(Z(\mathbf{s}_i, t_i)) + V(Z(\mathbf{s}_j, t_j)) - 2C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)). \tag{1.17}$$

If the spatiotemporal r.f. $Z(\boldsymbol{s}, t)$ has an intrinsically stationary variogram in both space and time, then it is said to have an intrinsically stationary variogram. In this case, the variogram can be expressed as

$$2\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = 2\gamma(\mathbf{h}, u). \tag{1.18}$$

The marginals $2\gamma(\cdot, u)$ and $2\gamma(\mathbf{h}, \cdot)$ are called purely spatial and purely temporal variograms, respectively.

A r.f. $Z(\boldsymbol{s}, t)$ is *strictly stationary* if its probability distribution is translation invariant. Second-order stationarity is a less-demanding condition than strict stationarity. A spatiotemporal r.f. $Z(\boldsymbol{s}, t)$ is *second-order stationary* in the broad sense or weakly stationary if it has a constant mean and the covariance function depends on $\mathbf{h}$ and u.

A spatiotemporal r.f. $Z(\boldsymbol{s}, t)$ is said to be intrinsically stationary if it has a constant mean and an intrinsically stationary variogram. Intrinsic stationarity is less restrictive than second-order stationarity. Another widely used function when modeling implicit spatiotemporal dependence in a stationary r.f. is the correlation function. Let $Z(\boldsymbol{s}, t)$ be a second-order stationary r.f. with a priori variance $\sigma^2 = C(\mathbf{0}, 0) > 0$. The autocorrelation function of this r.f. is defined as

$$\rho(\mathbf{h}, u) = \frac{C(\mathbf{h}, u)}{C(\mathbf{0}, 0)}. \tag{1.19}$$

If $\rho(\mathbf{h}, u)$ is a correlation function on $\mathbb{R}^d \times \mathbb{R}$, then its marginal functions $\rho(\mathbf{0}, u)$ and $\rho(\mathbf{h}, 0)$ will, respectively, be the spatial correlation function on $\mathbb{R}^d$ and the temporal correlation function on $\mathbb{R}$.

A function $C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$ of real values, defined on $\mathbb{R}^d \times \mathbb{R}$ is a *covariance function* if it is symmetrical, $C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = C((\mathbf{s}_j, t_j), (\mathbf{s}_i, t_i))$ and positive-definite, that is,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) \geq 0 \tag{1.20}$$

for any $n \in \mathbb{N}$, $(\mathbf{s}_i, t_i) \in \mathbb{R}^d \times \mathbb{R}$, and $a_i \in \mathbb{R}$, $i = 1, \ldots, n$. The condition (1.20) is *sufficient* if the covariance function can take complex values. Similarly, one *necessary and sufficient* condition for a nonnegative function of real values $\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$ defined on $\mathbb{R}^d \times \mathbb{R}$ to be a *semivariogram* is that it is a symmetrical function and conditionally negative-definite, that is,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) \leq 0 \tag{1.21}$$

with $\sum_{i=1}^{n} a_i = 0$.

Schoenberg [10] proved the following theorem characterizing the spatiotemporal semivariogram. Let $\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$ be a function defined on $\mathbb{R}^d \times \mathbb{R}$, with $\gamma((\mathbf{s}, t), (\mathbf{s}, t)) = 0$, $\forall (\boldsymbol{s}, t) \in \mathbb{R}^d \times \mathbb{R}$. Then the following statements are equivalent:

- $\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$ is a semivariogram on $\mathbb{R}^d \times \mathbb{R}$.
- $\exp\big(-\theta\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))\big)$ is a covariance function on $\mathbb{R}^d \times \mathbb{R}$, for any $\theta > 0$.
- $C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = \gamma((\mathbf{s}_i, t_i), (\mathbf{0}, 0)) + \gamma((\mathbf{s}_j, t_j), (\mathbf{0}, 0)) - \gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$ is a covariance function on $\mathbb{R}^d \times \mathbb{R}$.

In case of stationarity, the above results reduce to functions depending on spatial and temporal lags. Another seminal result that characterizes covariance functions is that given in [11]. A function $C(\mathbf{h}, u)$ defined on $\mathbb{R}^d \times \mathbb{R}$ is a stationary covariance function if, and only if, it has the following form

$$C(\mathbf{h}, u) = \iint e^{i(\boldsymbol{\omega}'\mathbf{h} + \tau u)} dF(\boldsymbol{\omega}, \tau), \ (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}, \tag{1.22}$$

where the function F is a nonnegative distribution function with a finite mean defined on $\mathbb{R}^d \times \mathbb{R}$, which is known as a *spectral distribution function*. Therefore, the class of stationary spatiotemporal covariance functions on $\mathbb{R}^d \times \mathbb{R}$ is identical to the class of Fourier transforms of nonnegative distribution functions with finite means on that domain. If the function C can also be integrated, then the spectral distribution function F is absolutely continuous and the representation (1.22) simplifies to

$$C(\mathbf{h}, u) = \iint e^{i(\boldsymbol{\omega}'\mathbf{h} + \tau u)} f(\boldsymbol{\omega}, \tau) d\boldsymbol{\omega} \, d\tau, \ (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R} \tag{1.23}$$

where f is a nonnegative, continuous, and integrable function that is known as a *spectral density function*. The covariance function C and the spectral density function f then form a pair of Fourier transforms, and

$$f(\boldsymbol{\omega}, \tau) = (2\pi)^{-d-1} \iint e^{-i(\boldsymbol{\omega}'\mathbf{h} + \tau u)} C(\mathbf{h}, u) d\mathbf{h} \, du. \tag{1.24}$$

The decomposition (1.22) can be specialized for fully symmetrical covariance functions. Let $C(\cdot, \cdot)$ be a continuous function defined on $\mathbb{R}^d \times \mathbb{R}$, then $C(\cdot, \cdot)$ is a fully symmetrical stationary covariance function if, and only if, the following decomposition is possible

$$C(\mathbf{h}, u) = \iint \cos(\boldsymbol{\omega}'\mathbf{h}) \cos(\tau u) dF(\boldsymbol{\omega}, \tau), \ (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}, \tag{1.25}$$

where F is the nonnegative and symmetrical spectral distribution function defined on $\mathbb{R}^d \times \mathbb{R}$.

Cressie and Huang [12] provide a theorem for characterizing the class of stationary spatiotemporal covariance functions under the additional hypothesis of integrability. Let $C(\cdot, \cdot)$ be a continuous, bounded, symmetrical, and integrable

function defined on $\mathbb{R}^d \times \mathbb{R}$, then $C(\cdot, \cdot)$ is a stationary covariance function if, and only if, in view of $u \in \mathbb{R}$,

$$C_{\boldsymbol{\omega}}(u) = \int e^{-i\boldsymbol{\omega}'\mathbf{h}} C(\mathbf{h}, u) d\mathbf{h}, \tag{1.26}$$

is a covariance function for every $\boldsymbol{\omega} \in \mathbb{R}^d$ except, at the most, in a set with a null Lebesgue mean. Gneiting [13] generalizes this result for C defined on $\mathbb{R}^d \times \mathbb{R}^l$, from which the previous statement is a particular case for $l = 1$.

Both the covariance function and the spectral density function are important tools for characterizing random stationary spatiotemporal fields. Mathematically speaking, both functions are closely related as a pair of Fourier transforms. Furthermore, the spectral density function is particularly useful in situations where there is no explicit expression of the covariance function. Stein [14] shows the benefit of using smooth covariance functions far from the origin, which can be tested by verifying whether their spectral densities have derivatives of certain orders.

1.3.2 Spatiotemporal Kriging

Kriging is aimed at predicting an unknown point value $Z(\boldsymbol{s}_0, t_0)$ at a point $(\boldsymbol{s}_0, t_0)$ that does not belong to the sample. To do so, all the information available about the regionalized variable is used, either at the points in the entire domain or in a subset of the domain called *neighborhood*.

Assume that the value of the r.f. has been observed on a set of n spatiotemporal locations $\{Z(\boldsymbol{s}_1, t_1), \ldots, Z(\boldsymbol{s}_n, t_n)\}$. If we want to predict the value of the r.f. on a new spatiotemporal location $(\boldsymbol{s}_0, t_0)$, we use the linear predictor

$$Z^*(\boldsymbol{s}_0, t_0) = \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{s}_i, t_i) \tag{1.27}$$

constructed from the random variables $Z(\boldsymbol{s}_i, t_i)$. As in the spatial case, spatiotemporal kriging equations will depend on the degree of stationarity attributed to the r.f. that supposedly generates the observed realization. The most widely used kriging techniques in the spatiotemporal case are simple spatiotemporal kriging, ordinary spatiotemporal kriging, and universal spatiotemporal kriging. In the case of *simple spatiotemporal kriging*, we assume that $Z(\boldsymbol{s}, t)$ is a second-order stationary spatiotemporal r.f., with a constant and known mean $\mu(\boldsymbol{s}, t)$, constant and known variance $C(\mathbf{0}, 0)$, and a known covariance function $C(\mathbf{h}, u)$. The kriging equations (n equations with n unknown elements) are of the form

$$\sum_{j=1}^{n} \lambda_j C\left(\boldsymbol{s}_i - \boldsymbol{s}_j, \mathrm{t}_i - \mathrm{t}_j\right) = C\left(\boldsymbol{s}_i - \boldsymbol{s}_0, \mathrm{t}_i - \mathrm{t}_0\right), \quad \forall i = 1, \ldots, n \tag{1.28}$$

from which we obtain the values λ_i that minimize the prediction error variance, which is given by

$$V\left[Z^*(\boldsymbol{s}_0, t_0) - Z(\boldsymbol{s}_0, t_0)\right] = C(\mathbf{0}, 0) - \sum_{i=1}^{n} \lambda_i C(\boldsymbol{s}_i - \boldsymbol{s}_0, t_i - t_0). \tag{1.29}$$

In the case of *ordinary spatiotemporal kriging*, the constant mean $\mu(\boldsymbol{s}, t)$ is not known, and the covariance function $C(\mathbf{h}, u)$ is known, under second-order stationarity. In the case of an intrinsic r.f., the variance is unbounded. In these two cases, simple kriging cannot be performed as the mean cannot be subtracted. We must therefore impose a condition of unbiasedness. In these situations, ordinary spatiotemporal kriging equations can be expressed, in the first case, in terms of the covariance function, and in the second case, in terms of the semivariogram, as there is no covariance at the origin.

In the *universal kriging* approach, assume $Z(\boldsymbol{s}, t)$ is a r.f. with drift, and so the mean of the r.f. is not constant, but depends on the pairs $(\boldsymbol{s}, t)$. In this situation, the so-called "condition of unbiasedness" is affected substantially. In this case, the r.f. can be disaggregated into two components: one deterministic $\mu(\boldsymbol{s}, t)$ and the other stochastic $e(\boldsymbol{s}, t)$ which can be treated as an intrinsically stationary r.f. with zero expectation, $E[e(\boldsymbol{s}, t)] = 0$

$$Z(\boldsymbol{s}, t) = \mu(\boldsymbol{s}, t) + e(\boldsymbol{s}, t). \tag{1.30}$$

We can assume that the mean, even unknown, can be expressed locally by

$$\mu(\boldsymbol{s}, t) = \sum_{h=1}^{p} a_h f_h(\boldsymbol{s}, t), \tag{1.31}$$

where $\{f_h(\boldsymbol{s}, t),\ h = 1, \ldots, p\}$ are p known functions, a_h constant coefficients, and p the number of terms used in the approximation. It must be taken into account that this expression is only valid locally. In this case, the equations that yield the prediction of the weights are obtained from the prediction error conditions of zero expectation and minimum variance.

1.3.3 Spatiotemporal Covariance Models

One key stage in the spatiotemporal prediction procedure is choosing the covariance function (covariogram or semivariogram) that models the structure of the spatiotemporal dependence of the data. However, while the semivariogram is normally chosen for this purpose in the spatial case, in a spatiotemporal framework, the covariance function is the most commonly chosen tool. By referring to a valid covariographic spatiotemporal model, we are implicitly stating that the covariance function must be positive-definite. The purely spatial and temporal covariance models have been widely studied, and there is a long list of those which can be

used to model spatial or spatiotemporal dependence that guarantee the (spatial or temporal) covariance function is positive-definite. However, this is not the case in the spatiotemporal scenario, in which constructing valid spatiotemporal covariance models is one of the main research activities. In addition, while it is difficult to demonstrate that a spatial or temporal function is positive-definite, it is even more so when seeking to determine valid spatiotemporal covariance models. For this reason, many authors began to study how to combine valid spatial and temporal models to obtain (valid) spatiotemporal covariance models.

By way of introduction, the first approximations to modeling spatiotemporal dependence using covariance functions were nothing more than generalizations of the stationary models used in the spatial scenario. In this sense, early studies often modeled the spatiotemporal covariance using metric models by defining a metric in space and time that allowed researchers to directly use isotropic models that are valid in the spatial case. Such metric models were characterized by being nonseparable, isotropic, and stationary. The next step in this initial stage consisted of configuring spatiotemporal covariance functions by means of the sum or product of a spatial covariance and a temporal covariance, both of which were stationary, giving rise to separable, isotropic and stationary models. Later, realizing the limitations of the two procedures detailed above in terms of capturing the spatiotemporal dependence that really exists in the large majority of the phenomena studied, interest shifted toward including the interaction of space and time, in covariance models, giving rise to the so-called "nonseparable models" (while remaining isotropic and stationary). Development continued with the search for nonseparable spatiotemporal, spatially anisotropic, and/or temporally asymmetrical models, together with general nonstationary models. There is a long list of papers dealing with these topics in the literature, and here we refer to [6] and all references therein.

The empirical determination of the covariance function or the variogram of a spatiotemporal process can be generalized naturally using the procedures for merely spatial processes. Let $Z(\cdot,\cdot)$ be an intrinsically stationary process observed on a set of n spatiotemporal pairs $\{(\boldsymbol{s}_1, t_1), \ldots, (\boldsymbol{s}_n, t_n)\}$. Two direct and popular alternatives to obtain an estimation of the variogram $2\gamma(\cdot,\cdot)$ (and its covariance function $C(\cdot,\cdot)$, if the process is also second-order stationary) are the classical estimator based on the method-of-moments (MoM), and the robust estimator proposed by Cressie and Hawkins [15].

1.4 Functional Data Analysis in Brief

In an increasing number of applications in many disciplines of science, the data collected corresponds to curves or surfaces. Such data can be generated by repeated

measurements in space–time or by automatic recordings of a quantity of interest. Functional data analysis (FDA) [16] has been used since the early 1990s to model this type of information. FDA is a general way of thinking, where the basic unit of information is an entire function rather than a set of values.

A random variable $X(t), t \in T \subset \mathbb{R}$ is called a functional variable if it takes values in an infinite-dimensional space (or functional space). A functional random sample $X_1(t), \ldots, X_n(t)$ corresponds to the collection of n functional variables independent and identically distributed (iid) as $X(t)$. An observation of $X(t)$, denoted as $x(t)$, is called a functional observation. For example, in a medical study $x_i(t), i = 1, \ldots, n$, could represent the electrocardiogram of the ith patient of the sample. In this case, it is reasonable to assume the independence assumption because it is natural to think that the responses of the patients are not related. FDA tools allow to estimate models based on a set of random variables taking values in a space of functions (functional variables), i.e. it concerns with the statistical analysis of multiple realizations of one (univariate FDA) or several (multivariate FDA) functional variables. If there is no observational noise, a functional observation $x_i(t)$, $t \in T$, is usually represented as a finite set of pairs $(t_j, x_i(t_j))$, $t_j \in T$, $j = 1, \ldots, M$. The set of points $\{t_j\}_{j=1}^{M}$ can be considered the same for all the functions in a functional data set, and usually, they form a fine evenly spaced grid in T. Nowadays, the number M in real applications is usually in the order of several hundred or thousands. Interpolation methods (if there is no noise) or nonparametric smoothing methods (in the opposite case) are commonly used to represent the data $(t_j, x_i(t_j))$, $j = 1, \ldots, M$, as a real function $x_i(t)$. In this sense, we can say that FDA inherits its methodology from nonparametric estimation. Note that, actually, the curves are not observed, instead only points of the curves are observed. However, when the number of points in a curve is dense for simplicity we talk about "observed curves."

Since the pioneering work by Deville [17] on harmonic analysis, there has been a lot of interest in developing statistical models for functional data. Examples of such methods include exploratory and descriptive data analysis [18], linear models [19], generalized linear models [20, 21], quantile regression [22], analysis of variance [23], nonparametric methods [24], longitudinal data [25], additive models [26] or multivariable techniques [27] such as principal components [28], canonical correlation [29], discriminant analysis [30] or cluster analysis [31]. An overview of inference for functional data is shown in [32]. Some new developments in this field are given in [33] and [34]. A relatively recent problem in FDA is the modeling of univariate and multivariate misaligned functional data. This one arises when the functional samples have systematic differences in shape. Some references in this topic are [35, 36].

Modern technology for acquiring and storing information in real-time often allows getting data that can be considered as functions. It is also possible to obtain

a finite and therefore incomplete amount of information regarding a function. For example, when collecting daily temperature data at weather stations. In this case, it makes sense fitting curves (or surfaces) to obtain functional observations. Generally, this stage is accomplished using smoothing and nonparametric methods [37]. This is the first step in FDA. The purpose is to convert discrete data into a smoothly varying function. In applied FDA, basis functions to obtain curves from the discrete records $(t_j, x_i(t_j))$ are generally used [38]. Basis functions procedures approximate a function by using a fixed truncated basis expansion $x(t) = \sum_{l=1}^{K} c_l B_l(t) = \mathbf{c}^T \mathbf{B}(t)$ in terms of K known basis functions. Once the representation by basis functions is adopted, three types of inquiries need to be answered for computational issues. Which basis functions are appropriate, how many basis functions are used to fit the data, and how the coefficients of the vector $\mathbf{c}$ are estimated. Generally, Fourier (for periodic data) and B-splines (for nonperiodic data) basis functions are applied to this purpose [16]. However, other basis or nonparametric smoothing methods can be used [24]. The number of basis is estimated by cross-validation. A roughness penalty can be included in the minimization problem.

Usually, the approaches for modeling functional data are focused on the assumption that the functions are iid, i.e. it is considered that all the functional observations correspond to realizations of the same stochastic process. However, in many fields of science, it is required to model correlated functional data (temporally or spatially correlated). In these cases, the traditional approaches based on the iid assumption may not be appropriate. An example of temporally correlated functional data is that of daily curves of financial transaction data (time series of functional data). The functions (curves of financial transactions) form a time series $\{X_k\}; k \in \mathbb{Z}$ where each X_k is a (random) function $X_k(t), t \in [a, b]$ (a collection of curves temporally indexed). On the other hand, in the spatial context suppose that penetration resistances (MPa) at different depths (meters) are recorded in many sites (points) of an experimental farm. The set of curves $X_{\boldsymbol{s}(d)}, \boldsymbol{s} \in \mathbb{Z}$ and $d \in [a, b]$ define a realization of a spatial stochastic process, where the response is functional (a collection of curves spatially indexed). Methods for analyzing correlated data have been adapted to the context of functional data. Some works in time series analysis and spatial statistics for functional data are proposed by Zhao et al. [39] and Delicado et al. [40]. A large number of methods of spatial statistics have been adapted to the functional realm, and this book is a good and modern example of this. Indeed, when combining spatial (geostatistical) methods with functional data we enter the field of *geostatistical functional data*, which is the core of this book.

FDA has become a rapidly developing discipline in the statistical field given its wide range of possible applications. Agronomy [41], biology [42], biomedicine

[33, 43], criminology [44], economy [45], medicine [46], meteorology [47], oceanography [48], psychology [49], and veterinary [50], among others, are areas where this relatively recent and novel field of statistics is useful. Ullah and Finch [51] presented a nice compilation of case studies analyzed from a FDA perspective. To apply FDA to a real dataset, there is a need for appropriate software with up-to-date methodological implementation and easy addition of new theoretical developments [52]. **R** R Core Team [53] and python libraries [52] can be used for this purpose.

In the geostatistical setting, kriging and cokriging methods have been also extended to deal with functional data, and methods such as ordinary kriging for functional data (OKFD), continuous time-varying kriging for functional data (CTVKFD), and functional kriging total model (FKTM) are now easily available. The simplest predictor abovementioned is OKFD, where each curve is weighted by a scalar parameter. The second option (CTVKFD) is founded in the theory of the functional linear concurrent (pointwise) regression model. In this case, the parameters are also curves. Finally, the predictor (FKTM) considers double indexed parameters. Here, in order to carry out the prediction at a particular time, each observed curve is weighted by a functional parameter. This approach follows the philosophy of the functional linear total model. Recently, in a similar framework, these methods have been adapted for Hilbert Spaces.

All the predictors before referenced are based on the assumption that the mean function is homogeneous into the region of interest. However, in practice, we often found realizations of nonstationary functional processes (because there exists an explicit spatial trend). To give solution to the problem of spatial prediction of functional data in the absence of stationarity some alternatives have been proposed. All of these have arisen as extensions to the functional framework of some classical kriging methods for nonstationary data.

Geostatistical FDA has been a field of constant growing in the last years. New modeling requirements in this area are opening many new research avenues. The definition of predictors considering realizations of multivariate (possibly nonstationary) functional random fields or the spatial prediction of data belonging to Riemannian manifolds are only a few new developments which indicate that there is still a long way ahead in this area of statistics. This book shows a state-of-the-art with recent contributions in this new environment of the spatial and functional modeling.

New theories emerge every day that allow adapting and extending traditional statistical methods to the treatment of functional data. However, there is still a long way ahead, given the wide range of theoretical and applied possibilities not yet explored. There are undoubtedly many challenges for the statistical community in this area.

References

1 Diggle, P. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press.
2 Cressie, N. and Wikle, C. (2019). *Statistics for Spatio-Temporal Data*. Wiley.
3 Diggle, P. and Giorgi, E. (2019). *Model-Based Geostatistics for Global Public Health: Methods and Applications*. CRC Press.
4 Wikle, C., Zammit-Mangion, A., and Cressie, N. (2019). *Spatio-Temporal Statistics with R*. Chapman & Hall/CRC.
5 Cressie, N. (1993). *Statistics for Spatial Data*. Wiley.
6 Fernandez-Aviles, G., Montero, J., and Mateu, J. (2015). *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. Wiley.
7 Emery, X. (2000). *Geoestadística Lineal*. Departamento de Ingeniería de Minas. Facultad de Ciencias Fisicas and Matematicas. Universidad de Chile.
8 Chiles, J. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. Wiley.
9 Müller, W. and Mateu, J. (2012). *Spatio-Temporal Design. Advances in Efficient Data Acquisition*. Wiley.
10 Schoenberg, I. (1938). Metric spaces and completely monotone functions. *Annals of Mathematics* 39: 811–841.
11 Bochner, S. (1933). Monotone funktionen, stiltjes integrale und harmonische analyse. *Mathematische Annalen* 108: 378–410.
12 Cressie, N. and Huang, H.C. (1999). Classes of nonseparable, spatiotemporal stationary covariance functions. *Journal of the American Statistical Association* 94: 1330–1340.
13 Gneiting, T. (2002). Stationary covariance functions for space-time data. *Journal of the American Statistical Association* 97: 590–600.
14 Stein, M. (2005). Space-time covariance functions. *Journal of the American Statistical Association* 100: 310–321.
15 Cressie, N. and Hawkins, D. (1980). Robust estimation of the variogram. *Journal of the International Association for Mathematical Geology* 12 (2): 115–125.
16 Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. New York: Springer-Verlag.
17 Deville, J. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE* 15 3–101.
18 Ramsay, J. (2000). Functional components of variation in handwriting. *Journal of the American Statistical Association* 95: 9–15.
19 Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters* 45: 11–22.
20 Müller, H. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics* 33 (2): 774–805.

21 Febrero-Bande, M. and González-Manteiga, W. (2013). Generalized additive models for functional data. *Test* 22 (2): 278–292.

22 Kato, K. (2012). Estimation in functional linear quantile regression. *The Annals of Statistics* 40 (6): 3108–3136.

23 Zhang, J. (2014). *Analysis of Variance for Functional Data*. Boca Ratón, FL: CRC Press.

24 Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media.

25 Park, S. and Staicu, A. (2015). Longitudinal functional data analysis. *Stat* 4 (1): 212–226.

26 Greven, S. and Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling* 17 (1–2): 1–35.

27 Górecki, T., Krzyśko, M., Waszak, L., and Wołyński, W. (2018). Selected statistical methods of data analysis for multivariate functional data. *Statistical Papers* 59 (1): 153–182.

28 Berrendero, J., Justel, A., and Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics and Data Analysis* 55 (9): 2619–2634.

29 Shin, H. and Lee, S. (2015). Canonical correlation analysis for irregularly and sparsely observed functional data. *Journal of Multivariate Analysis* 134: 1–18.

30 Gardner-Lubbe, S. (2020). Linear discriminant analysis for multiple functional data analysis. *Journal of Applied Statistics* 48, 1917–1933.

31 Clarkson, D., Fraley, C., Gu, C., and Ramsey, J. (2005). Functional cluster analysis. *S+ Functional Data Analysis: Users Manual for Windows*, pp. 155–164.

32 Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, vol. 200. Springer Science & Business Media.

33 Margaritella, N., Inácio, V., and King, R. (2021). Parameter clustering in Bayesian functional principal component analysis of neuroscientific data. *Statistics in Medicine* 40 (1): 167–184.

34 Yu, J., Park, J., Choi, T. et al. (2021). Nonparametric Bayesian functional meta-regression: applications in environmental epidemiology. *Journal of Agricultural, Biological and Environmental Statistics* 26 (1): 45–70.

35 Olsen, N., Markussen, B., and Raket, L. (2016). Simultaneous inference for misaligned multivariate functional data. *arXiv preprint arXiv:1606.03295*.

36 Zeng, P., Shi, J., and Kim, W.S. (2019). Simultaneous registration and clustering for multidimensional functional data. *Journal of Computational and Graphical Statistics* 28 (4): 943–953.

37 Ferraty, F., Mas, A., and Vieu, P. (2007). Nonparametric regression on functional data: inference and practical aspects. *Australian & New Zealand Journal of Statistics* 49 (3): 267–286.

38 Wang, J., Chiou, J., and Müller, H. (2016). Functional data analysis. *Annual Review of Statistics and Its Application* 3: 257–295.
39 Zhao, X., Marron, J., and Wells, M. (2004). The functional data analysis view of longitudinal data. *Statistica Sinica* 14 789–808.
40 Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics: The Official Journal of the International Environmetrics Society* 21(3–4): 224–239.
41 Giraldo, R., Caballero, W., and Camacho-Tamayo, J. (2018). Mantel test for spatial functional data. *Advances in Statistical Analysis* 102 (1): 21–39.
42 Cremona, M., Xu, H., Makova, K. et al. (2019). Functional data analysis for computational biology. *Bioinformatics* 35 (17): 3211–3213.
43 Lopez-Pintado, S. and Qian, K. (2021). A depth-based global envelope test for comparing two groups of functions with applications to biomedical data. *Statistics in Medicine*. https://doi.org/10.1002/sim.8861.
44 Park, K., Suk, H., Hwang, H., and Lee, J.H. (2013). A functional analysis of deception detection of a mock crime using infrared thermal imaging and the concealed information test. *Frontiers in Human Neuroscience* 7: 70.
45 Frois Caldeira, J., Gupta, R., Suleman, M., and Torrent, H. (2020). Forecasting the term structure of interest rates of the BRICS: evidence from a nonparametric functional data analysis. In: *Emerging Markets Finance and Trade*, 1–18. https://doi.org/10.1080/1540496X.2020.1808458.
46 Strzalkowska-Kominiak, E. and Romo, J. (2021). Censored functional data for incomplete follow-up studies. *Statistics in Medicine*. https://doi.org/10.1002/sim.8930/
47 Suhaila, J., Jemain, A., Hamdan, M., and Zin, W. (2011). Comparing rainfall patterns between regions in Peninsular Malaysia via a functional data analysis technique. *Journal of Hydrology* 411 (3–4): 197–206.
48 Assunção, R., Silva, A., Roy, A. et al. (2020). 3D characterisation of the thermohaline structure in the southwestern tropical Atlantic derived from functional data analysis of in situ profiles. *Progress in Oceanography* 187: 102399.
49 Levitin, D., Nuzzo, R., Vines, B., and Ramsay, J. (2007). Introduction to functional data analysis. *Canadian Psychology* 48 (3): 135.
50 Gruen, M., Alfaro-Córdoba, M., Thomson, A. et al. (2017). The use of functional data analysis to evaluate activity in a spontaneous model of degenerative joint disease associated pain in cats. *PLoS ONE* 12 (1): e0169576.
51 Ullah, S. and Finch, C. (2013). Applications of functional data analysis: a systematic review. *BMC Medical Research Methodology* 13 (1): 1–12.

52 Golovkine, S. (2021). FDApy: a Python package for functional data. *arXiv preprint arXiv:2101.11003*.

53 R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org`.

Part I

Mathematical and Statistical Foundations

2

Mathematical Foundations of Functional Kriging in Hilbert Spaces and Riemannian Manifolds

Alessandra Menafoglio[1], *Davide Pigoli*[2], *and Piercesare Secchi*[3,4]

[1] *MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy*
[2] *King's College London, UK*
[3] *MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy*
[4] *CADS, Center for Analysis Decisions and Society, Human Technopole, Via Cristina Belgioioso, 171, 20157, Milano, Italy*

2.1 Introduction

Modern field studies yield diverse types of observations, in the form of highly heterogeneous and high-dimensional data. In this context, environmental observations are routinely available in the form of functional or distributional data. For instance, these kinds of variables are found in climatic investigations, where complex data are regularly collected at different sites in the study region. Examples are temperature profiles along the year, or the precipitation distributions, which are key to characterize and classify the domain of interest from the climatic viewpoint.

In these cases, the object of the analysis is often infinite-dimensional, i.e. it would need an infinity of point evaluations to be fully characterized. In some cases, constraints can be included – e.g. positivity or convexity – particularly when distributional data are concerned. In fact, the full interpretation and statistical treatment of such kinds of complex data poses relevant challenges for geoscience applications.

In this broad context, a relatively large body of recent literature has been devoted to the mathematical foundations of geostatistics for complex data, with particular reference to data embedded in Hilbert spaces and Riemannian manifolds. We focus here on the approach developed within the area of *object-oriented spatial statistics* (O2S2, [1]), which was developed starting from the works [2–4]. The foundational idea of the approach is to interpret data as *objects*: the *atom* of the geostatistical analysis is the entire object, which is seen as an indivisible unit

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

rather than a collection of features. In this view, the observations are interpreted as random points within a space of objects – called *feature space* – whose dimensionality and geometry should properly represent the data features and their possible constraints. The O2S2 approach follows the funding ideas of object oriented data analysis (OODA, [5]), and generalizes the theory of functional geostatistics developed from the seminal works of [6–8], mainly for functional data in L^2.

Among the challenges related to the spatial analysis of complex data, we focus here on the problems of spatial prediction. Similarly as in classical geostatistics (e.g. [9]), in O2S2, the latter problem is addressed by formulating optimal unbiased predictors linear in the data. We review here the mathematical framework for kriging Hilbert and manifold data, in stationary or nonstationary settings and discuss the estimators that can be used for the mean and the covariance structure.

The remaining of this chapter is organized as follows. Section 2.2 introduces the main definitions and assumptions which may be formulated to perform a geostatistical analysis of Hilbert-space valued random fields. Section 2.3 describes a global approach to kriging, interpreted as optimal linear combinations of the data. Here, we show an example of application to climate data and thoroughly discuss on the importance of selecting an appropriate feature space for the analysis. In Section 2.4, we briefly review an alternative approach to kriging, which arises when the predictor is interpreted in a more general sense, grounding on the theory of measurable linear transformations. Interestingly, this general theory allows to draw connections between several different formulations of functional kriging available in the literature. Section 2.5 introduces the methodologies to perform geostatistical analysis of manifold-valued random fields, based on the local-approximation property of such spaces. Here, for illustration, we consider the case of positive definite matrices that are used to analyze and predict the field of covariance matrices between temperature and precipitations in a region of Canada.

2.2 Definitions and Assumptions

We call $(\Omega, \mathfrak{F}, \mathbb{P})$ a probability space and $\mathcal{H}$ a separable Hilbert space, endowed with operations $(+, \cdot)$, and an inner product $\langle \cdot, \cdot \rangle$. The space $\mathcal{H}$ will indicate the feature space for the geostatistical analysis: we will consider the data as realizations of random points in $\mathcal{H}$. In several cases in this chapter, the space $\mathcal{H}$ will represent a space whose elements are real-valued functions defined over a compact interval. Nevertheless, the theory presented in this chapter is entirely general and may involve manifold data, as we shall show in Section 2.5.

In the following, we denote by χ a random element in $\mathcal{H}$, that is a measurable function defined on $(\Omega, \mathfrak{F}, \mathbb{P})$ and valued in $\mathcal{H}$, $\chi : \Omega \to \mathcal{H}$. We indicate a realization of χ – that is a nonrandom element of $\mathcal{H}$ – with the symbol χ, i.e. $\chi = \chi(\omega)$,

for $\omega \in \Omega$. We call $\mathcal{L}(\mathcal{H}, \mathcal{H}_1)$ the Banach space of continuous linear operators on $\mathcal{H}$ in $\mathcal{H}_1$. We say that two random elements χ_1, χ_2 are equivalent (indicated by $\chi_1 =_{\mathcal{H}} \chi_2$, or $\chi_1 = \chi_2$ for short) if $\chi_1 = \chi_2$ almost surely.

Given a set of locations $s_1, \dots, s_n$ in a spatial domain $D \subset \mathbb{R}^d$ (usually $d = 2, 3$), we denote by $\chi_{s_1}, \dots, \chi_{s_n}$ the set of observations collected at these locations, that form our dataset of spatially dependent objects. As in classical geostatistics (e.g. [9]), we assume this dataset to be a partial observation of a random field $\{\chi_s, s \in D\}$ on $(\Omega, \mathfrak{F}, \mathbb{P})$ in $\mathcal{H}$. The latter is defined as a collection of random elements χ_s, indexed by a continuous spatial vector s varying in D.

In this chapter, we will always assume that, for all $s \in D$, the element χ_s, satisfies $\mathbb{E}[\| \chi_s\|^2] < \infty$. Under the latter assumption, one can define the expected value of the field in terms of Bochner integral as

$$m_s = \int_\Omega \chi_s(\omega)\mathbb{P}(d\omega), \quad s \in D. \tag{2.1}$$

In $\mathcal{H}$, the expected value (2.1) can be equivalently defined as the element m_s of $\mathcal{H}$ such that, for any $x \in \mathcal{H}$, $\langle x, m_s\rangle = \mathbb{E}[\langle x, \chi_s\rangle]$. The elements m_s, $s \in D$, describe the first-order structure of the field.

The second-order structure can be fully characterized through the *spatial covariance function* C, which is the map that associates each pairs of locations (s_1, s_2) with the cross-covariance operator between the random elements at those locations, i.e.

$$\begin{aligned} &C\colon D \times D \to \mathcal{L}(\mathcal{H}, \mathcal{H}) \\ &(s_1, s_2) \mapsto \{C(s_1, s_2)\colon \mathcal{H} \to \mathcal{H}, x \mapsto \mathbb{E}[\langle(\chi_{s_1} - m_{s_1}), x\rangle(\chi_{s_2} - m_{s_2})]\}. \end{aligned} \tag{2.2}$$

A global measure of spatial dependence is provided by the *trace-covariogram* C. The latter is defined as the real-valued function $C\colon D \times D \to \mathbb{R}$ that associates a pair of locations (s_1, s_2) in D with the real value:

$$C(s_1, s_2) = \mathbb{E}[\langle \chi_{s_1} - m_{s_1}, \chi_{s_2} - m_{s_2}\rangle]. \tag{2.3}$$

The trace-covariogram can be interpreted as the direct generalization of the classical covariogram, the inner product in $\mathbb{R}$ being replaced by the inner product in $\mathcal{H}$. From the mathematical viewpoint, for any pair of locations (s_1, s_2), the trace-covariogram $C(s_1, s_2)$ coincides with the trace of the corresponding cross-covariance operator $C(s_1, s_2)$, i.e. $\sum_{k=1}^{\infty}\langle C(s_1, s_2)e_k, e_k\rangle$ (see [2] for details). Intuitively, if $\mathcal{H}$ was the Euclidean space $\mathbb{R}^p$, the cross-covariance operator $C(s_1, s_2)$ would be the linear operator associated with the covariance matrix between the random elements χ_{s_1}, χ_{s_2}, and $C(s_1, s_2)$ its trace. In this sense, the trace-covariogram provides a global measure of dependence. The trace-covariogram can be proven to fulfill all the properties of a classical covariogram, e.g. it is a positive-definite function [9]:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C(s_i, s_j) \geq 0, \quad \forall s_i, s_j \in D, \forall \lambda_1, \dots, \lambda_n \in \mathbb{R}.$$

The trace-covariogram was defined in the context of L^2 data in [8], and then generalized to object data in any separable Hilbert space $\mathcal{H}$ in [2].

The trace-covariogram is strictly related with a counterpart of the classical variogram, named *trace-variogram*, that is defined as the function $2\gamma : D \times D \to \mathbb{R}^+$ that maps any pair of locations (s_1, s_2) as

$$2\gamma(s_1, s_2) = \mathbb{E}[\| \chi_{s_1} - \chi_{s_2} \|^2] - \| m_{s_1} - m_{s_2} \|^2. \tag{2.4}$$

The trace-variogram fulfills classical properties, such as being a conditionally negative definite function (e.g. [2]):

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(s_i, s_j) \leq 0, \quad \forall s_i, s_j \in D, \forall \lambda_1, \dots, \lambda_n \in \mathbb{R} \text{s.t.} \sum_{i=1}^{n} \lambda_i = 0.$$

On these premises, definitions of stationarity can be stated for the random field $\{\chi_s, s \in D\}$. In particular, we focus here on definitions of second-order stationarity in a strong sense (Definition 2.1) and in a global sense (Definition 2.2). The interested reader can find weaker definitions of stationarity in [2].

Definition 2.1 A process $\{\chi_s, s \in D\}$ is said to be *strongly second-order stationary* if the following conditions hold:

(i) $\mathbb{E}[\chi_s] = m$, for all $\boldsymbol{s} \in D \subseteq \mathbb{R}^d$ (spatially constant mean);
(ii) $\mathbb{E}[\langle \chi_{s_1} - m, \cdot \rangle (\chi_{s_2} - m)] = C(s_1 - s_2)$ for all $s_1, s_2 \in D \subseteq \mathbb{R}^d$ (spatial covariance function depending only on the increment vector).

Definition 2.2 A process $\{\chi_s, s \in D\}$ is said to be *(globally) second-order stationary* if the following conditions hold:

(i) $\mathbb{E}[\chi_s] = m$, for all $\boldsymbol{s} \in D \subseteq \mathbb{R}^d$ (spatially constant mean);
(ii′) $\mathbb{E}[\langle \chi_{s_1} - m, \chi_{s_2} - m \rangle] = C(s_1 - s_2)$ for all $s_1, s_2 \in D \subseteq \mathbb{R}^d$ (trace-covariogram depending only on the increment vector).

Second-order stationarity thus concerns a spatial homogeneity in the first and second-order structure of the field. It should be noted that stationarity does not imply the existence of a directional homogeneity, which concerns the concept of isotropy instead. Indeed, a strongly second-order stationary field is said to be isotropic if condition (ii) is reinforced by the following condition (iii):

(iii) $\mathbb{E}[\langle \chi_{s_1} - m, \cdot \rangle (\chi_{s_2} - m)] = C(\| s_1 - s_2) \|_d)$, for all $s_1, s_2 \in D \subseteq$, $\| \cdot \|_d$ being the norm on $\mathbb{R}^d$ (spatial covariance function depending only on the distance between locations).

A globally second-order stationary field is said to be isotropic if condition (ii′) is reinforced by the following condition (iii′):

(iii′) $\mathbb{E}[\langle \chi_{s_1} - m, \chi_{s_2} - m \rangle] = C(\| s_1 - s_2 \|_d)$ for all $s_1, s_2 \in D$ (trace-covariogram depending only on the distance between locations).

Both strong and global second-order stationarity are of interest from the application-oriented viewpoint. Indeed, the methods introduced in Sections 2.3.1 and 2.3.2 rely upon the assumption of global stationarity, while the methods devised in Section 2.4 assume the stronger condition of strong second-order stationarity (and Gaussianity). We finally remark that, although assuming isotropy greatly simplifies the notation, it should not be considered as essential for the development of the methods described in Sections 2.3–2.5.

2.3 Kriging Prediction in Hilbert Space: A Trace Approach

A key goal of a typical geostatistical analysis is to perform spatial prediction at a target (unobserved) location. As long as one-dimensional Euclidean fields are concerned, classical geostatistics literature advocates the use of a kriging predictor, that is the best linear unbiased predictor (BLUP) $\chi^*_{s_0} = \sum_{i=1}^n \lambda_i^* \cdot \chi_{s_i}$, whose weights minimize the variance of prediction error under the unbiasedness constraint (e.g. [9]). In fact, in the scalar case, no ambiguity exists in the definition of linear predictor, as this can be equivalently interpreted either as a linear combination of the observations, or as a linear transformation of the vector of observations. Instead, when the feature space $\mathcal{H}$ is an infinite-dimensional Hilbert space, several possible definitions of kriging are available. In this section, we focus on the so-called *trace-approach* that defines the kriging predictor as the best linear combination of the data, as presented in [8] for the stationary L^2 setting, and further generalized in [2] for possibly nonstationary Hilbert data.

2.3.1 Ordinary and Universal Kriging in Hilbert Spaces

Given $s_1, \dots, s_n$ in D and the observations of the field $\chi_{s_1}, \dots, \chi_{s_n}$ at these location, we look for the BLUP $\chi^*_{s_0} = \sum_{i=1}^n \lambda_i^* \cdot \chi_{s_i}$ for χ_{s_0}, where the weights $\lambda_1^*, \dots, \lambda_n^*$ solve the minimization problem

$$\min_{\lambda_1,\dots,\lambda_n \in \mathbb{R}} \mathbb{E}\left[\left\| \chi_{s_0} - \sum_{i=1}^n \lambda_i \cdot \chi_{s_i} \right\|^2 \right] \text{ subject to } \mathbb{E}\left[\sum_{i=1}^n \lambda_i \cdot \chi_{s_i} \right] = \mathbb{E}[\chi_{s_0}]. \quad (2.5)$$

In the presence of second-order stationarity, one may employ an ordinary (trace-) kriging predictor, while for nonstationary data, universal (trace-) kriging may be employed instead. We here consider universal kriging in $\mathcal{H}$, following [2], since ordinary kriging is obtained as a special case.

We represent the elements of the field $\{\chi_s, s \in D\}$ as $\chi_s = m_s + \delta_s$, where m_s is the drift – which describes a possibly nonconstant mean variation – whereas δ_s is assumed to be a globally second-order stationary and isotropic random field with zero-mean and trace-covariogram C. Following the approach of universal kriging for scalar data, we describe the drift term through a linear model with scalar regressors and functional coefficients

$$m_s = \sum_{l=0}^{L} f_l(s) \cdot a_l, \quad s \in D, \tag{2.6}$$

where $f_0(s) = 1$ for all $s \in D, f_l, l = 1, \dots, L$, are known over the entire domain and $a_l, l = 0, \dots, L$ are (possibly unknown) coefficients in $\mathcal{H}$. Note that the stationary case is obtained when $L = 0$ as in that case the mean is spatially constant. Further, the spatial variation is assumed to be entirely captured by the regressors $\{f_l, l = 1, \dots, L\}$, since the coefficients do not depend on the location $s \in D$. Note that other approaches to model the nonstationarity of the mean are possible, e.g. based on (scalar or functional) covariates collected together with the data (i.e. the kriging with external drift proposed in [10], and discussed in Chapter 3).

In our setting, the unbiasedness constraint in (2.5) reads

$$\sum_{i=1}^{n} \lambda_i f_l(s_i) = f_l(s_0), \quad l = 0, \dots, L,$$

which is included in the minimization functional through $L + 1$ Lagrange multipliers. Hence, problem (2.5) becomes that of minimizing, with reference to λ_i, ζ_l, $i = 1, \dots, n, l = 0, \dots, L$,

$$\begin{aligned} \Phi = &\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(\| s_i - s_j \|_d) + C(0) - 2\sum_{i=1}^{n} \lambda_i C(\| s_i - s_0 \|_d) \\ &+ 2\sum_{l=0}^{L} \zeta_l \left(\sum_{i=1}^{n} \lambda_i f_l(s_i) - f_l(s_0) \right). \end{aligned} \tag{2.7}$$

Denote by $\Sigma \in \mathbb{R}^{n \times n}$ the variance–covariance matrix of the observations (in the trace sense), whose (i,j)-th element is $\Sigma_{i,j} = C(\| s_i - s_j \|_d)$ for $i, j = 1, \dots, n$, $C(\| s_i - s_j \|_d)$ appearing in (2.5). Indicate with $\mathbb{F} = (f_l(\boldsymbol{s}_i)) \in \mathbb{R}^{n \times (L+1)}$ the design matrix of the linear model (2.6), by $\vec{\lambda} = (\lambda_1, \dots, \lambda_n)^T$ the vector of weights and $\vec{\zeta} = (\zeta_0, \dots, \zeta_L)^T$ the vector of Lagrange multipliers. Call $\vec{\sigma}_0 = (C(\| s_1 - s_0 \|_d), \dots, C(\| s_n - s_0 \|_d)^T$ the vector of (trace-) covariances between the observations and the target, and $\vec{f}_0 = (f_0(\boldsymbol{s}_0), \dots, f_L(\boldsymbol{s}_0))^T$ the vector of regressors at

the target location (both appearing in (2.5)). With this notation, and under usual assumptions on the sampling design – namely Σ positive definite and $\mathbb{F}$ of full rank – the functional (2.7) admits a unique minimum, found by solving the linear system

$$\begin{pmatrix} \Sigma & \mathbb{F} \\ \mathbb{F}^T & 0 \end{pmatrix} \begin{pmatrix} \vec{\lambda} \\ \vec{\zeta} \end{pmatrix} = \begin{pmatrix} \vec{\sigma}_0 \\ \vec{f}_0 \end{pmatrix}. \tag{2.8}$$

The latter system is easily found by equating to zero the differential of Φ with reference to the λ's and the ζ's.

The variance of the prediction error (i.e. the value of (2.7) at the optimum) is called *universal kriging variance* and is obtained as

$$\begin{aligned} \sigma^2_{UK}(s_0) &= C(0) - \sum_{i=1}^{n} \lambda_i^* C(\| s_i - s_0 \|_d) - \sum_{l=0}^{L} \zeta_l^* f_l(s_0) \\ &= \sum_{i=1}^{n} \lambda_i^* \gamma(\| s_i - s_0 \|_d) + \sum_{l=0}^{L} \zeta_l^* f_l(s_0), \quad s_0 \in D. \end{aligned} \tag{2.9}$$

The latter quantifies the uncertainty of the prediction and can be used to build prediction bands (e.g. by using the Chebychev inequality). Nonetheless, it should be noted that it does not consider the variability of the possible estimate of the covariance structure, as the latter is assumed to be known over the entire construction. Instead, in most cases, the trace-covariogram needs to be estimated as well, and the estimated covariance is eventually plugged-in in (2.8). Classical geostatistics advocates the estimate of the (trace-)variogram in place of the (trace-)covariogram. The two functions are linked by the relation:

$$2\gamma(h) = C(0) - C(h), \quad h \in D.$$

Estimators of the variogram are generally more robust than those of covariogram, hence preferred.

To estimate the variogram, a two-step procedure is generally employed: (i) estimate an empirical variogram, and (ii) fit a parametric model (e.g. spherical, exponential, Matérn) to the estimate at point (i), in order to guarantee that the properties of a valid variogram (e.g. conditional negative definiteness) are fulfilled. If global second-order stationarity and isotropy holds true, one can use the method-of-moment estimator to achieve point (i) [2, 8]

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} \| \chi_{s_i} - \chi_{s_j} \|^2, \tag{2.10}$$

where $N(h) = \{(i,j) \mid h - \Delta h \leq \| s_i - s_j \|_d \leq h + \Delta h\}$, and $|N(h)|$ is its cardinality. Estimator (2.10) provides an unbiased estimate of the trace-variogram 2γ only if the assumptions of stationarity and isotropy are verified. Otherwise, such

estimator should not be employed because it considers only the first term of the variogram definition in (2.4) (i.e. $\mathbb{E}[\| \chi_{s_1} - \chi_{s_2} \|^2]$), while it neglects the additional term $\| m_{s_1} - m_{s_2} \|$ (which is null under stationarity). When considering the nonstationary model here introduced, one can use estimator (2.10) on the residuals $\delta_{s_i}, i = 1, \dots, n$, which are a partial observation of a globally second-order stationary and isotropic process. Nevertheless, these are usually latent and can be estimated by difference once the drift has been assessed. Hence, although in principle one could perform universal kriging without having estimated the drift in advance, whenever the trace-covariogram is unknown, providing a good estimate of the drift is essential. Section 2.3.2 will be dedicated to this point. Instead, in case of stationarity (and isotropy), estimating the drift is not required for the purpose of performing spatial prediction.

2.3.2 Estimating the Drift

The problem of estimating the drift for the spatial model here considered consists in estimating a linear model in the presence of spatially correlated random errors. Indeed, under model (2.6), the model for the data is

$$\chi_{s_i} = \sum_{l=0}^{L} f_l(s_i) \cdot a_l + \delta_{s_i}. \tag{2.11}$$

To set the notation, we call $\mathcal{H}^n$ the Hilbert space $\mathcal{H} \times \cdots \times \mathcal{H}$, with the inner product $\langle \vec{x}, \vec{y} \rangle = \sum_{i=1}^{n} \langle x_i, y_i \rangle$, for $\vec{x} = (x_1, \dots, x_n)^T \in \mathcal{H}^n$, $\vec{y} = (y_1, \dots, y_n)^T \in \mathcal{H}^n$. We denote by $\vec{\chi} = (\chi_{s_1}, \dots, \chi_{s_n})^T \in \mathcal{H}^n$, $\vec{a} = (a_0, \dots, a_L)^T \in \mathcal{H}^{L+1}$, and $\vec{\delta} = (\delta_{s_1}, \dots, \delta_{s_n})^T \in \mathcal{H}^n$. In this setting, model (2.11) can be expressed in matrix form as $\vec{\chi} = \mathbb{F}\vec{a} + \vec{\delta}$.

The theory of functional linear models was developed under the founding hypothesis of independent and identically distributed residuals. As a consequence, in the presence of correlated residuals, the ordinary least squares approach developed in that framework turns out to provide suboptimal results. To properly account for the spatial dependence, a generalized least squares (GLS) criterion can be used instead [2]. The latter seeks to minimize the functional Mahalanobis distance between the observations and the evaluation of the drift at the sampled locations. That is, the GLS estimator for vector $\vec{a}$ is found as the solution of the minimization problem:

$$\min_{\widehat{\vec{a}} \in \mathcal{H}^{L+1}} \sum_{i=1}^{n} \left\| \left[\Sigma^{-1/2} \circ \left(\vec{\chi} - \mathbb{F}\, \widehat{\vec{a}} \right) \right]_i \right\|^2, \tag{2.12}$$

$\circ$ indicating the matrix multiplication in $\mathcal{H}$, that is $[A\vec{x}]_i = \sum_{j=1}^{m} A_{ij} \cdot x_j$, for $i = 1, \dots, q$, with $\vec{x} \in \mathcal{H}^m$, $A \in \mathbb{R}^{q \times m}$. If $\operatorname{rank}(\mathbb{F}) = L + 1 \leq n$ and $\operatorname{rank}(\Sigma) = n$,

problem (2.12) is well posed and its unique solution is found as [2]

$$\hat{\vec{a}}^{GLS} = (\mathbb{F}^T\Sigma^{-1}\mathbb{F})^{-1}\mathbb{F}^T\Sigma^{-1}\circ\vec{\chi}. \tag{2.13}$$

Estimator (2.13) was proved to be the best linear unbiased estimator for $\vec{a}$. However, it depends on the matrix Σ, hence on the trace-covariogram, which can be estimated only once the residuals have been assessed, and the latter residuals can be only computed by difference based on the drift estimate. To jointly assess the residuals and the trace-(co)variogram, one can then resort to an iterative algorithm, initialized, e.g. to the ordinary least squares (OLSs) estimate of the drift. Such algorithm usually converges within five iterations, although theoretical arguments on its convergence are yet to be proved. Having computed the drift and the trace-variogram, the universal kriging system can be solved by plugging-in the estimated variance-covariance matrix in (2.8).

2.3.3 An Example: Trace-Variogram in Sobolev Spaces

We discuss here through an example the key importance of the choice of the ambient space for the (geo)statistical analysis of functional or object data. Although there are cases in which a natural ambient space is available (e.g. suggested by dynamical equations governing the system), the choice of the feature space for the analysis is indeed a crucial modeling step. As a way of example, Figure 2.1 shows a set of spatially dependent curves, simulated from two random fields $\{\chi_s^{(m)}, s \in D\}$, $m = 1, 2$, $D = [0, 2] \times [0, 3]$. The latter fields were built in [2] by combining in different ways the same set of independent, zero mean, second-order stationary, and isotropic scalar random fields $\{\xi_{s,j}, s \in D\}$, $j = 1, \ldots, 7$, as

$$\chi_s^{(1)} = \sum_{k=1}^{7} \xi_{s,k} e_k; \quad \chi_s^{(2)} = \sum_{k=19}^{25} \xi_{s,k-18} e_k, \tag{2.14}$$

where $\{e_k, k \geq 1\}$ denotes the Fourier basis. A detailed description of the simulation setting is provided in [[2], Supplementary material]. Figure 2.1 evidences the very different patterns displayed by the two groups of curves. Indeed, the realizations associated with the field $\{\chi_s^{(2)}, s \in D\}$ show a much higher amplitude variability (i.e. along the vertical direction) than those associated with the field $\{\chi_s^{(1)}, s \in D\}$. This is due to the fact that the second field is built upon a higher-order truncation of the basis and involves only the 10th–12th frequencies, while only the first three frequencies are included in the construction of the first field.

Despite these apparent diversities between the two fields, no difference exists in their spatial dependence structure if the fields are embedded in the space L^2

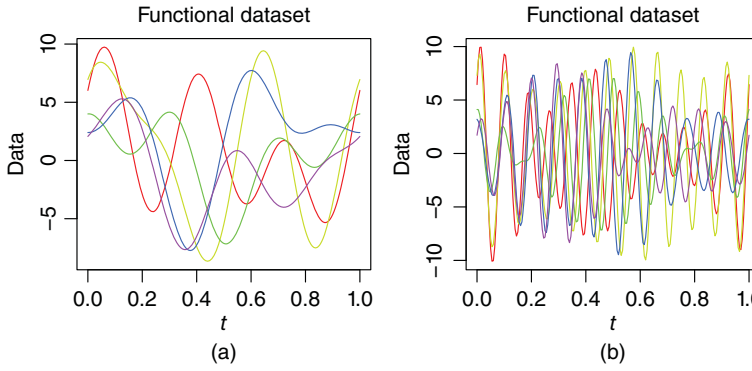


Figure 2.1 Spatially dependent curves simulated from the fields $\{\chi_s^{(1)}, s \in D\}$ (a) and $\{\chi_s^{(3)}, s \in D\}$, (b). (a) 7 basis functions and (b) 25 basis functions. Source: Modified from Menafoglio et al. [2].

of square-integrable functions. Indeed, straightforward computations yield, for $m = 1, 2$,

$$2\gamma_{L^2}^{(m)}(s_i, s_j) = \mathbb{E}\left[\left\|\chi_{s_i}^{(m)} - \chi_{s_j}^{(m)}\right\|_{L^2}^2\right] = \sum_{k=1}^{N_m} \mathbb{E}\left[\left|\xi_{s_i,k}^{(m)} - \xi_{s_j,k}^{(m)}\right|^2\right] = \sum_{k=1}^{7} 2\gamma_{\xi_k},$$

γ_{ξ_k} indicating the variogram of the scalar field of coefficients $\{\xi_{s,k}, s \in D\}$, $k = 1, \dots, 7$. Instead, when modeling the data as objects in the Sobolev space H^1 – i.e. the space of functions in L^2 whose derivatives (in a weak sense) are in L^2 – one can capture the diverse behavior of the fields, through the geometry of the space. The latter choice entails the use of a norm based not only on pointwise evaluations but also on the differential properties of the elements. In such a case, the variogram is indeed different in the two fields, being

$$2\gamma_{H^1}^{(1)} = 2\gamma_{L^2}^{(1)} + \sum_{k=2}^{7} \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2 2\gamma_{\xi_k} = \sum_{k=1}^{7} \left(1 + \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2\right) 2\gamma_{\xi_k};$$

$$2\gamma_{H^1}^{(2)} = 2\gamma_{L^2}^{(2)} + \sum_{k=19}^{25} \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2 2\gamma_{\xi_{k-18}} = \sum_{k=19}^{25} \left(1 + \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2\right) 2\gamma_{\xi_{k-18}}.$$

Notice that, for $k = 1, \dots, 7$, the weights associated with the variogram $2\gamma_{\xi_k}$ depend on the frequency associated with ξ_k, a greater weight being assigned to a higher frequency.

Figure 2.2 shows the empirical trace-variograms estimated in L^2 (Figure 2.2a) and in H^1 (Figure 2.2b) from a sample of 100 observations $\chi_{s_1}^{(m)}, \dots, \chi_{s_{100}}^{(m)}$ from each field $\{\chi_s^{(m)}, s \in D\}$, $m = 1, 2$, the same sites $s_1, \dots, s_{100}$ being uniformly

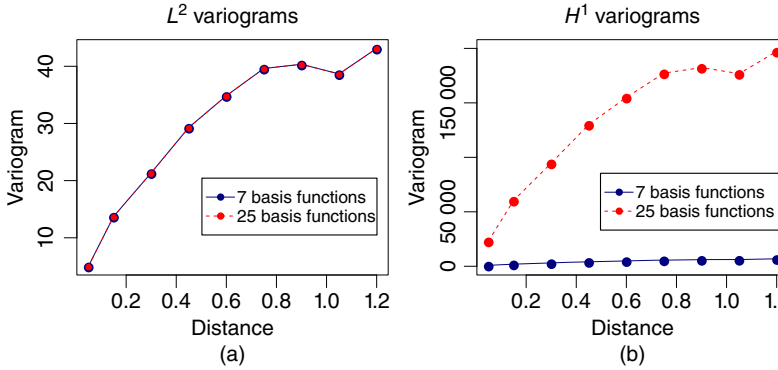


Figure 2.2 Empirical trace-variograms in L^2 (a) and $\mathcal{H}^1$ (b). Source: Modified from Menafoglio et al. [2].

sampled in D. Although the shapes of the variograms are quite similar in the two cases, the orders of magnitude of the horizontal asymptotes – twice the global variance of the process – are significantly different. Indeed, the variogram $2\gamma_{H^1}^{(2)}$ (dashed light line) assumes much higher values than $2\gamma_{H^1}^{(1)}$ (solid dark line), since the random field $\{\chi_s^{(2)}, s \in D\}$ has a much higher *energy*. Indeed, in dynamical system theory, the square of the Sobolev norm of the state (i.e. $\|\chi_s\|^2$) coincides with (twice) the energy of the system. Hence, the ambient space for geostatistical analysis not only provides the feature space for the object data but also implies a precise physical meaning for the measure of stochastic variability: the global variance represents (twice) the mean energy of the system, while the trace-variogram (twice) the mean energy of the increments between two states.

In conclusion, one should pay close attention to the choice of the feature space for the analysis. The latter should be guided by the dataset structure, the possible physical laws governing the system, and by the purposes of the analysis.

2.3.4 An Application to Nonstationary Prediction of Temperatures Profiles

We show here an example of application of the trace-approach to nonstationary environmental data. Following [2], the data we consider are daily mean temperature profiles, collected during 1980 at 27 locations in the Maritime Provinces of Canada (data source: Natural Resources of Canada; http://atlas.nrcan.gc.ca/). Raw data were smoothed by using a Fourier basis of 65 elements, obtaining the set of curves displayed in Figure 2.3.

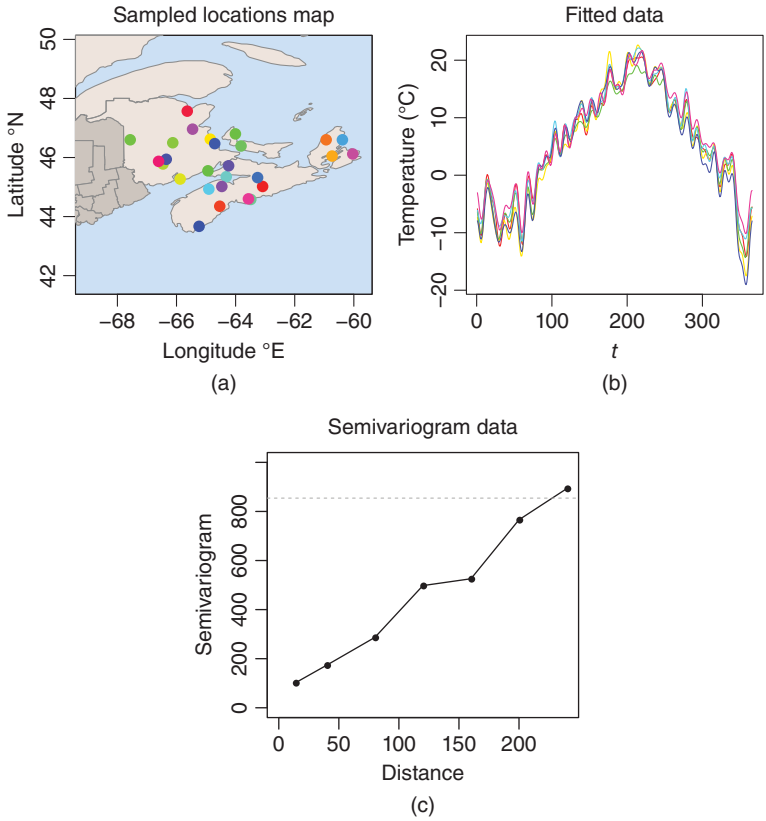


Figure 2.3 Canada's Maritime Provinces Temperatures dataset, year 1980. (a) Map of Maritime Provinces and sampled locations; (b) six fitted temperature curves; (c) empirical estimate of the trace-variogram. Source: Modified from Menafoglio et al. [2].

For illustration purposes, we performed the geostatistical analysis in L^2, using on the spatial domain a geodesic distance, since coordinates are given in latitude and longitude. The graphical inspection of the trace-semivariogram estimated from the data suggests the presence of a nonconstant drift model. Indeed, the empirical estimate does not show any apparent finite sill (i.e. horizontal asymptote, see Figure 2.3). To select the drift model, we considered the polynomial models of degree 2 in the coordinates and sought the one minimizing the kriging prediction error, assessed by leave-one-out cross-validation [2]. On this basis, we found as optimal model

$$m(s,t) = a_0(t) + a_1(t)y + a_2(t)x^2 + a_3(t)y^2 + a_4(t)xy,$$

for $s = (x,y) = (\text{Longitude}, \text{Latitude})$, $t \in \mathcal{T} = [0{,}366]$.

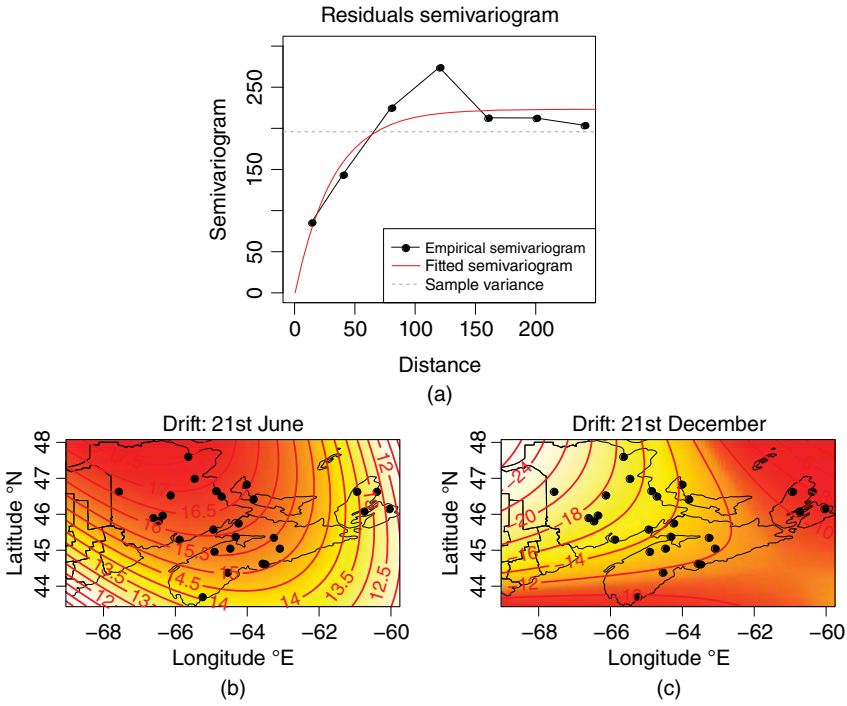


Figure 2.4 Estimated trace-semivariogram from the residuals (a) and estimated drift for the Summer Solstice (21st June; b) and the Winter Solstice (21st December; c). The drift maps are extracted from the drift temperature profiles estimated in L^2. In (a), (geodesic) distances are given in kilometers.

Figure 2.4 displays the estimate of the drift for the days of summer solstice and the winter solstice. Note that we may have chosen any day of the year for such representation: the theory presented here allows obtaining joint and consistent estimates (and predictions) for all the days of the year. The maps in Figure 2.4 have a clear climatical interpretation, as they represent the presence of currents from the Ocean, whose circulation causes a change in the gradient of temperature along the year. In Figure 2.5, we represent the universal kriging maps for the same days considered in Figure 2.4. From the analysis of the maps, one can conclude that the drift term tends to drive the estimates in the colder seasons (Figure 2.4c), whereas during the summer season (Figure 2.4b), the temperature map displays evident local patterns, due to the peculiar geographical configuration of the area – particularly for the Bay of Fundy. Although the spatial patterns evidenced during the year tend to be different, the universal kriging predictor allows to properly capture them, thanks to its flexibility.

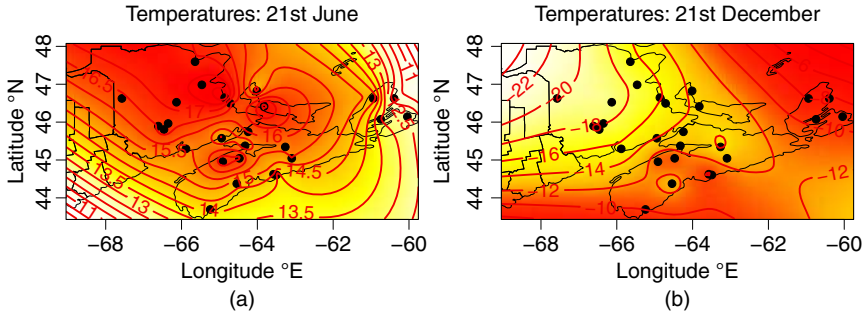


Figure 2.5 Universal kriging maps for the Summer Solstice (21st June; a) and the Winter Solstice (21st December; b), extracted from the temperature profiles predicted via universal kriging in L^2.

2.4 An Operatorial Viewpoint to Kriging

In this section, we briefly review a second approach to kriging, namely the operatorial ordinary kriging. Here, the kriging predictor is built upon a linear transformation of the vector of data $\mathcal{X}^{\Lambda}_{s_0} = \Lambda\vec{\chi}$, for some linear operator $\Lambda\colon \mathcal{H}^n \to \mathcal{H}$.

The development of such classes of predictors is motivated by the observation that, despite its simplicity, predictor $\chi^*_{s_0} = \sum_{i=1}^n \lambda_i \cdot \chi_{s_i}$ does not provide, in general, the best linear unbiased transformation of the vector of observations. The operatorial viewpoint was first considered in [7] in reproducing Kernel Hilbert spaces (RKHSs), that are functional spaces whose elements are more regular than general Hilbert spaces (see also Chapter 4). These authors addressed the problem of finding the best predictor over the class of linear unbiased Hilbert–Schmidt transformations of the observations, of the form $\chi^{\boldsymbol{B}}_{s_0} = \sum_{i=1}^n B_i\chi_{s_i}$, where $B_i\colon \mathcal{H} \to \mathcal{H}$ are linear Hilbert–Schmidt operators and $\mathcal{H}$ a RKHS. Although all finite-dimensional Hilbert spaces are RKHS, other widely used spaces, such as the space L^2, need not be a RKHS. In [4], a more general theory was introduced based upon the idea of working with measurable linear transformations instead of linear Hilbert–Schmidt operators: in this particular setting, these authors showed that the operatorial kriging problem is well posed for any separable Hilbert space $\mathcal{H}$.

We now formally introduce the latter predictor and discuss its relation with the other kriging predictors discussed here.

Given two separable Hilbert spaces, $\mathcal{H}_1, \mathcal{H}_2$, we denote by L a Borel-measurable map from $(\mathcal{H}_2, \mathfrak{B}(\mathcal{H}_2))$ to $(\mathcal{H}_1, \mathfrak{B}(\mathcal{H}_1))$, $\mathfrak{B}(\mathcal{H}_i)$ being the Borel σ-algebra of $\mathcal{H}_i$, for $i = 1, 2$. Given a measure μ on $(\mathcal{H}_2, \mathfrak{B}(\mathcal{H}_2))$, we say that L is a *measurable linear transformation with respect to* μ (μ-mlt), if L is linear on a subspace $\mathcal{D}_L \in \mathfrak{B}(\mathcal{H}_2)$ with $\mu(\mathcal{D}_L) = 1$.

Given a set of locations $s_1, \dots, s_n$ and the observation of the process at these locations, $\chi_{s_1}, \dots, \chi_{s_n}$, we consider the operatorial ordinary kriging predictor $\chi_{s_0}^{\Lambda^*} = \Lambda^* \vec{\chi}$ for χ_{s_0}. Here, $\Lambda^* \colon \mathcal{H}^n \to \mathcal{H}$ is a measurable linear transformation with respect to the law $\mu_{\vec{\chi}}$ of $\vec{\chi}$, and minimizes

$$\mathbb{E}\left[\left\|\chi_{s_0} - \chi_{s_0}^{\Lambda}\right\|^2\right] \quad \text{subject to} \quad \mathbb{E}\left[\chi_{s_0}^{\Lambda}\right] = m, \tag{2.15}$$

over all $\Lambda : \mathcal{H}^n \to \mathcal{H}$ a $\mu_{\vec{\chi}}$-mlt and where in the objective functional $\chi_{s_0}^{\Lambda}$ stands for $\chi_{s_0}^{\Lambda} = \Lambda \vec{\chi}$.

To tackle this problem, throughout the section, we assume $\{\chi_s, s \in D\}$ to be a Gaussian random field on $(\Omega, \mathfrak{F}, \mathbb{P})$ in $(\mathcal{H}, \mathfrak{B}(\mathcal{H}))$, that is, we assume that all its finite-dimensional laws are Gaussian in $\mathcal{H}$. Recall that a random element in $\mathcal{H}$ is Gaussian if $\langle x, \chi \rangle$ has a Gaussian distribution for every $x \in \mathcal{H}$. Note that this assumption is crucial for the validity of the results presented here, because a full characterization of the properties of mlts is only available under Gaussianity. We further assume that the field is strongly second-order stationary; we call m its (spatially constant) mean, and C its (stationary) spatial covariance function, defined as in (2.2).

Under this assumptions, the ordinary kriging problem can be shown to be well posed [4]. To state such result, we need the following further notation. We call $1 : \mathcal{H} \to \mathcal{H}^n$ the linear operator acting on $x \in \mathcal{H}$ as $1x = (x, \dots, x)^T$, and $1'$ its adjoint. We denote by $\mu_{\vec{\chi}_0} = N\left(\boldsymbol{m}_{\vec{\chi}_0}, C_{\vec{\chi}_0}\right)$ the law of the random vector $\vec{\chi}_0 = \left(\chi_{s_0}, \vec{\chi}^T\right)^T$ in $\mathcal{H}^{n+1}$, with expected value $\boldsymbol{m}_{\vec{\chi}_0} = \left(m, (1\ m)^T\right)^T$ and covariance operator $C_{\vec{\chi}_0} : \mathcal{H}^{n+1} \to \mathcal{H}^{n+1}$. The latter can be expressed in block form as

$$C_{\vec{\chi}_0} = \begin{pmatrix} C_{\chi_{s_0}} & C_{\chi_{s_0}\vec{\chi}} \\ C_{\vec{\chi}\chi_{s_0}} & C_{\vec{\chi}} \end{pmatrix}.$$

Here $C_{\vec{\chi}}$ indicates the covariance operator of $\vec{\chi}$, i.e. $(C_{\vec{\chi}}\vec{x})_i = \sum_{j=1}^n C(s_i - s_j)x_j$, for $\vec{x} \in \mathcal{H}^n$, $i = 1, \dots, n$, and $C_{\vec{\chi}\chi_{s_0}}$ is the cross-covariance operator between $\vec{\chi}$ and χ_{s_0}, i.e. $C_{\vec{\chi}\chi_{s_0}}\vec{x} = \sum_{j=1}^n C(s_0, s_j)x_j$, for $\vec{x} \in \mathcal{H}^n$.

The following Theorem 2.1 – proved in [4] – states that the operatorial ordinary kriging problem is well posed.

Theorem 2.1 *([4])* Under the previous assumptions and notation, *(*2.15*)* admits a unique minimizer $\chi_{s_0}^{\Lambda *} = \Lambda^* \vec{\chi}$, where Λ^* is the $\mu_{\vec{\chi}}$-mlt solving

$$\begin{cases} \Lambda C_{\vec{\chi}} - C_{\chi_{s_0}\vec{\chi}} + \zeta_1\ 1' = 0; \\ \Lambda\ 1 - I = 0, \end{cases} \tag{2.16}$$

with $1\ x = (x, x, \dots, x)^T$, *for* $x \in \mathcal{H}$, $I \colon \mathcal{H} \to \mathcal{H}$ the identity operator and ζ_1 a μ_{χ_0}-mlt that represents the Lagrange multiplier associated with the unbiasedness

constraint. Moreover, for $x \in \mathcal{H}^n$, one has

$$\Lambda^* x = M^* x + L(x - 1\, M^* x), \tag{2.17}$$

where M^* is the $\mu_{\vec{\chi}}$-mlt defined, for $x \in \mathcal{H}^n$, as $M^* x = \left(1' C_{\vec{\chi}}^{-1} 1\right)^{-1} 1' C_{\vec{\chi}}^{-1} x$, and L is the $\mu_{\vec{\chi}}$-mlt of conditional expectation that acts on $x \in \mathcal{H}^n$ as $Lx = C_{\chi_{s_0}\vec{\chi}} C_{\vec{\chi}}^{-1} x$.

We refer the interested reader to [4] for the proof of Theorem 2.1. We note however that (2.16) can be expressed in matrix form as

$$\begin{pmatrix} \Lambda & \zeta_1 \end{pmatrix} \begin{pmatrix} C_{\mathcal{X}} & 1 \\ 1' & 0 \end{pmatrix} = \begin{pmatrix} C_{\mathcal{X}_{s_0}\mathcal{X}} & I \end{pmatrix}, \tag{2.18}$$

which as the very same form as (2.8), but in an operatorial setting. Moreover, a second key element of Theorem (2.1) is the explicit expression for the optimal $\mu_{\vec{\chi}}$-mlt Λ^* in (2.17). The latter involves two parts: the first related with an operatorial version of the GLS estimator for the mean function, analogue to that described in Section 2.3.2; the second part exploits the operator of conditional expectation L of χ_{s_0} given $\vec{\chi}$, applied to the estimated residuals. As shown in [11, 12], the latter operator L is the $\mu_{\vec{\chi}}$-mlt that allows to obtain the conditional expectation $\mathbb{E}[\chi_{s_0}|\vec{\chi}]$ when applied to the centered observations $(\vec{\chi} - 1m)$, under the assumption that the mean m is known, i.e.

$$\mathbb{E}[\chi_{s_0}|\vec{\chi}] = m + L(\chi - 1\, m). \tag{2.19}$$

Note that (2.19) has the very same form of the familiar expression of conditional expectation in the multivariate setting. Hence, (2.17) shows an interesting relation of the operatorial kriging predictor with the conditional expectation, which has a very analogous counterpart in the finite dimensional case [9]. Indeed, kriging coincides with the conditional expectation only when the mean is known (i.e. simple kriging [4]). In all other cases, it is a plug-in estimator that is built upon the conditional expectation, by employing the GLS estimate of the mean.

We finally mention that both the operatorial kriging predictor in RKHSs and the trace-kriging predictor can be seen as particular cases of the operatorial kriging predictor defined by Theorem 2.1. Indeed, the kriging predictor proposed in [7] can be found by embedding Theorem 2.1 in a RKHS. A particular case is then obtained for finite-dimensional Hilbert spaces, already explored by Nerini et al. [7], that are interpreted as finite-dimensional approximations of the BLUP $\chi_{s_0}^{\Lambda^*}$. Similarly, in the stationary Gaussian case, the trace-kriging predictor can be interpreted as the finite-dimensional approximation of the operatorial kriging predictor within the n-dimensional Hilbert space generated by the observations. We refer the reader to [4] for further details.

2.5 Kriging for Manifold-Valued Random Fields

While spatial statistics of functional data has recently received much attention, as proved by the many contributions in this book, the extension to the case of non-Euclidean data is even a greater challenge because they do not belong to a vector space. A kriging procedure has been recently proposed in [3] for data belonging to Riemannian manifolds, where local tangent space approximations can be used. Indeed, any Riemannian manifold admits an approximation based on a Hilbert tangent space, where linear geostatistical methods can be developed. Thus, it is possible to use the local geometry of the manifold to find a data-driven linearization, i.e. looking for the tangent space where the parametric model provides the best possible fit for the available data. Then, the spatial dependence can be modeled in the tangent space using the methods for Hilbert spaces described above. In this section, we describe the method introduced in [3] and we discuss some possible generalization.

2.5.1 Residual Kriging

We first need to introduce some definitions and notations to model data that take values in a Riemannian manifold. Let $\mathcal{M}$ be a Riemannian manifold and, given a point P in $\mathcal{M}$, let $\mathcal{H}$ be the tangent space at the point P, $\mathcal{H} = T_P\mathcal{M}$. This is a Hilbert space when equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ in $\mathcal{H}$. Since our aim is to model the spatial variation in the local tangent space in P, we need a way to map elements of the tangent space to the manifold and vice versa. Thus, two important objects are the exponential map and its inverse, the logarithmic map. The exponential map is a smooth map from $T_P\mathcal{M}$ to $\mathcal{M}$, which is defined via the geodesics (the shortest paths between points on the manifold) passing through P: it maps a tangent vector $T \in T_P\mathcal{M}$ to an element in $\mathcal{M}$ by traveling on the manifold, for a unit of time, along the geodesic starting in P in direction T. Indeed, under some technical assumptions on $\mathcal{M}$, for every pair $(P; T) \in \mathcal{M} \times T_P\mathcal{M}$, there is a unique geodesic curve $g(t)$ such that $g(0) = P$ and $g'(0) = T$. The exponential map of $\mathcal{M}$ in P is defined as the point at $t = 1$ of this geodesic, i.e. $g(1)$. We indicate here with $\exp_P$ the exponential map in P, and with $\log_P$ its inverse. Figure 2.6 shows a visualization of these concepts for the case of a sphere. More details on these definitions and on the properties of Riemannian manifolds can be found, e.g. in [13] and a detailed example for the case of the manifold of positive definite symmetric matrices is discussed in Section 2.5.2.

For a location s in the spatial domain D, we can now model the random element $\boldsymbol{S}_s$, taking value in $\mathcal{M}$, as

$$\boldsymbol{S}_s(\vec{a}, P) = \exp_P(A\vec{f}(s); \vec{a}) + \boldsymbol{\delta}_s). \tag{2.20}$$

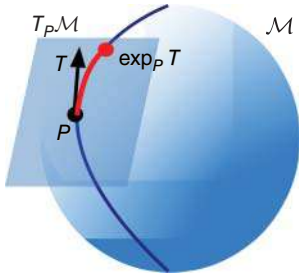


Figure 2.6 Visual representation of the tangent space in P on a sphere and of the exponential map of a vector T in the tangent space.

Here, $A(\vec{f}(s);\vec{a})$ is a drift term defined in the tangent space $\mathcal{H}$, described by a linear model with $\vec{a} = (a_0, \dots, a_L)$ a vector of coefficients belonging to $\mathcal{H}$ and $\vec{f}(s)$ a vector of scalar regressors:

$$A(\vec{f}(s);\vec{a}) = \sum_{l=0}^{L} f_l(s) \cdot a_l,$$

where $f_0(s) = 1$. Instead, $\{\delta_s, s \in D\}$ denotes a zero-mean globally second-order stationary and isotropic random field in the Hilbert space $\mathcal{H}$, with covariogram C and semivariogram γ.

Let now $s_1, \dots, s_n$ be n locations in D, and let $\boldsymbol{S}_1, \dots, \boldsymbol{S}_n$ be the observations of (2.20) at these locations. The goals are to estimate the parameters P and $\vec{a}$ in model (2.20) and to perform spatial prediction at an unobserved location s_0. We denote by $\Sigma \in \mathbb{R}^{n\times n}$ the covariance matrix of the array $\vec{\delta} = (\delta_{s_1}, \dots, \delta_{s_n})^T$ in $\mathcal{H}^n$, that is $\Sigma_{ij} = C(\|s_i - s_j\|_d^2)$, and call $\vec{\boldsymbol{R}} \in \mathcal{H}^n$ the array of residuals $\boldsymbol{R}_i = A(\vec{f}(s);\vec{a}) - \log_P(\boldsymbol{S}_i)$. To estimate $(P, \vec{a})$ accounting for the spatial dependence, the GLS functional

$$\left(\widehat{P}, \widehat{\vec{a}}\right) = \underset{P\in\mathcal{M},\vec{a}\in\mathcal{H}^{L+1}}{\operatorname{argmin}} \left\|\Sigma^{-1/2}\vec{\boldsymbol{R}}\right\|_{\mathcal{H}^n}^2 \tag{2.21}$$

needs to be minimized. When Σ is known, problem (2.21) can be solved iteratively, by alternatively minimizing the GLS functional in (2.21) with respect to P given $\vec{a}$ and to $\vec{a}$ given P. The minimizer in $\vec{a}$ given P can be explicitly determined as detailed in Section 2.3.2, i.e.

$$\widehat{\vec{a}}^{GLS}(P) = (\mathbb{F}^T\Sigma^{-1}\mathbb{F})^{-1}\mathbb{F}^T\Sigma^{-1} \circ \vec{\boldsymbol{Y}}(P), \tag{2.22}$$

where $\mathbb{F} \in \mathbb{R}^{n\times(L+1)}$ is the design matrix, $\mathbb{F}_{il} = f_l(s_i)$, and $\vec{\boldsymbol{Y}}(P)$ is the array $\vec{\boldsymbol{Y}}(P) = (\log_P(\boldsymbol{S}_1), \dots, \log_P(\boldsymbol{S}_n))^T \in \mathcal{H}^n$. On the other hand, an expression for the minimizer in P given $\vec{a}$ is not available, in general. The complexity of such minimization is problem dependent and may require the development of specific optimization techniques.

When $(P, \vec{a})$ is known, it is possible to estimate Σ by estimating the semivariogram $\gamma(h)$, for example following the strategy of Section 2.3.2. That is (i) estimate

the empirical semivariogram from the residuals as

$$\widehat{\boldsymbol{\gamma}}(h) = \frac{1}{2|N(h)|} \sum_{(s_i,s_j)\in N(h)} ||\boldsymbol{\delta}_{s_i} - \boldsymbol{\delta}_{s_j}||_H^2,$$

and (ii) fit a parametric model to the empirical estimate to obtain a valid model. A good estimate of the spatial dependence (including the choice of the model semivariogram) is crucial in the applications. Note that, the tangent space being Hilbert, most of the existing methods in linear geostatistics can be used in this step of the analysis (see, e.g. [14, 15], reference therein). Since in practice both the parameters and the spatial dependence are unknown, there is the need to resort to a nested iterative algorithm, where the semivariogram is estimated from the residuals at the current iteration.

Estimated the parameters of model (2.20) as $(\widehat{\boldsymbol{P}}, \widehat{\vec{\boldsymbol{a}}}, \widehat{\boldsymbol{\Sigma}})$, the kriging prediction can be performed as follows: in the Hilbert space $\mathcal{H}$, the simple kriging predictor for $\boldsymbol{\delta}_{s_0}$ is well defined, and it is obtained as $\sum_{i=1}^n \lambda_i^0 \widehat{\boldsymbol{\delta}}_{s_i}$, where $\widehat{\boldsymbol{\delta}}_{s_i}$ stands for the estimated residual at s_i, $\widehat{\boldsymbol{\delta}}_{s_i} = A(\vec{f}(s_i); \widehat{\vec{\boldsymbol{a}}}) - \log_{\widehat{\boldsymbol{P}}}(\boldsymbol{S}_i)$, and the vector of kriging weights $\vec{\lambda}_0 = \left(\lambda_1^0, \ldots, \lambda_n^0\right)$ is found as $\vec{\lambda}_0 = \widehat{\boldsymbol{\Sigma}}^{-1}\vec{\boldsymbol{c}}$, with $\vec{\boldsymbol{c}} = (\widehat{\boldsymbol{C}}(||s_1 - s_0||_d), \ldots, \widehat{\boldsymbol{C}}(||s_n - s_0||_d))^T$. The spatial prediction of $\boldsymbol{S}$ at the target location s_0 is then

$$\widehat{\boldsymbol{S}}_0 = \exp_{\widehat{\boldsymbol{P}}}\left(\widehat{\boldsymbol{a}}_0^{GLS}(\widehat{\boldsymbol{P}}) + \sum_{l=1}^{L} \widehat{\boldsymbol{a}}_l^{GLS}(\widehat{\boldsymbol{P}}) f_l\left(s_0\right) + \sum_{i=1}^{n} \lambda_i^0 \widehat{\boldsymbol{\delta}}_{s_i}\right),$$

where $\vec{f}(s_0)$ is the vector of covariates given at the location s_0.

2.5.2 An Application to Positive Definite Matrices

Positive definite matrices are an example of manifold-valued data, the modeling of positive definite matrices random field being relevant in applications such as Diffusion Tensor Imaging [16] and covariances between meteorological variables [3, 17]. The space $\mathrm{PD}(p)$ of positive definite matrices of dimension p is a convex subset of $\mathbb{R}^{p(p+1)/2}$, but it is not a linear space: in general, a linear combination of elements of $\mathrm{PD}(p)$ does not belong to $\mathrm{PD}(p)$. The tangent space $T_P\,\mathrm{PD}(p)$ to the manifold of positive definite symmetric matrices of dimension p in the point $P \in \mathrm{PD}(p)$ can be identified with the space $\mathrm{Sym}(p)$, the space of symmetric matrices of dimension p, which is linear and can be equipped with an inner product. A Riemannian metric in $\mathrm{PD}(p)$ is then induced by the inner product in $\mathrm{Sym}(p)$ and the choice of the inner product in the tangent space determines the form of the geodesic (i.e. the shortest path between two elements on the manifold) and thus the expression of the geodesic distance between two positive definite symmetric matrices. A possible choice for the Riemannian metric is generated by the scaled Frobenius inner product in $\mathrm{Sym}(p)$, which is

defined as $\langle A, B\rangle_P = \text{trace}(P^{-\frac{1}{2}}A^TP^{-1}BP^{-\frac{1}{2}})$, where $A, B \in \text{Sym}(p)$. This choice is very popular for covariance matrices because it generates a distance which is invariant under affine transformation of the random variables. For every pair $(P, A) \in \text{PD}(p) \times \text{Sym}(p)$, there is then a unique geodesic curve $g(t)$ such that

$$g(0) = P,$$
$$g'(0) = A.$$

When the Riemannian metric is generated by the scaled Frobenius inner product, the expression of the geodesic becomes

$$g(t) = P^{\frac{1}{2}} \exp\left(tP^{-\frac{1}{2}}AP^{-\frac{1}{2}}\right) P^{\frac{1}{2}},$$

where $\exp(C)$ indicates the exponential matrix of $C \in \text{Sym}(p)$. The exponential map of $\text{PD}(p)$ in P is defined as the point at $t = 1$ of this geodesic:

$$\exp_P(A) = P^{\frac{1}{2}} \exp\left(P^{-\frac{1}{2}}\text{AP}^{-\frac{1}{2}}\right) P^{\frac{1}{2}}.$$

Thus, the exponential map takes the geodesic passing through P with "direction" A and follows it until $t = 1$. The exponential map has an inverse which is called logarithmic map and is defined as

$$\log_P(D) = P^{\frac{1}{2}} \log\left(P^{-\frac{1}{2}}DP^{-\frac{1}{2}}\right) P^{\frac{1}{2}},$$

where $\log(C)$ is the logarithmic matrix of $C \in \text{PD}(p)$. The logarithmic map returns the tangent element A that allows the corresponding geodesic to reach D at $t = 1$.

With this structure, it is possible to apply the residual kriging method described above to positive definite matrix-valued data. Kriging interpolation for the covariance matrices between temperature and precipitation in Quebec has been explored in [3], using data from Canadian meteorological stations made available by Environment Canada on the website http://climate.weatheroffice.gc.ca. The seven meteorological stations where all monthly temperature and precipitation data are available from 1983 to 1992 are considered. For each station and for each month from January to December, these 10-year measures are used to estimate the 2×2 covariance matrix between temperature and precipitation, obtaining and separately analyzing 12 datasets, each composed by $n = 7$ spatially dependent sample covariance matrices (with the previous notation, $n = 7$ and $p = 2$). Pigoli et al. [3] found out that the covariation between temperature and precipitation changes across the calendar year. We report here the results obtained for the month of January. Including only a constant term in the tangent space model (i.e. assuming that the matrix random field has a constant mean) leads to an estimate of the empirical semivariogram that suggests to move toward a nonstationary model, as it can be seen in Figure 2.7. The simplest drift model

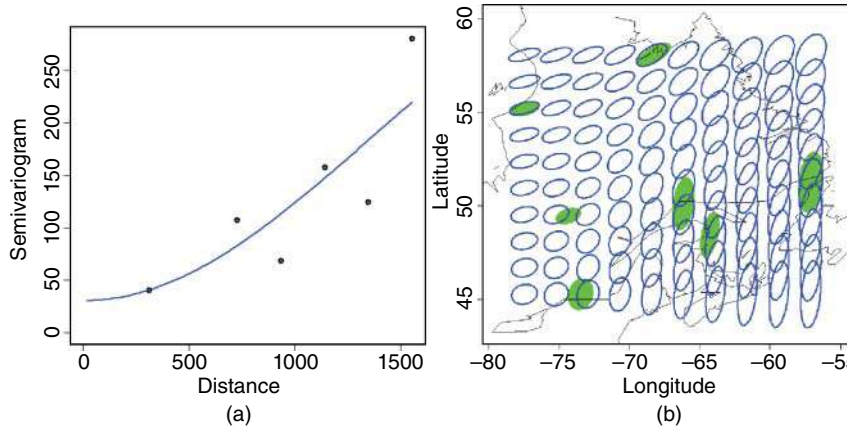


Figure 2.7 (a) Empirical semivariogram (symbols) and fitted exponential model (solid line). The geodesic distances in the spatial domain are measured in kilometers. (b) Ordinary kriging for the (temperature, precipitation) covariance matrix field for the month of January; filled ellipses indicate original data. A covariance matrix $\boldsymbol{S}$ at location s is represented as an ellipse centered in s and with axis $\sqrt{\sigma_j}\vec{e}_j$, where $\boldsymbol{S}\vec{e}_j = \sigma_j\vec{e}_j$ for $j = 1, 2$. Horizontal and vertical axes of the ellipses represent temperature and precipitation, respectively. Source: Modified from Pigoli et al. [3].

which guarantees the stationarity of the residuals is found to be the following: linear model depending on longitude:

$$A(\phi_i, \lambda_i) = a_0 + a_1\phi_i, \tag{2.23}$$

(ϕ, λ) denoting longitude and latitude. A possible meteorological interpretation is associated with the exposition of the region toward the sea, since model (2.23) accounts for the distance between the location of interest and the Atlantic Ocean. This is likely to influence temperatures, precipitations, and their covariability. The estimates of the semivariogram and of the drift and the kriging prediction can be seen in Figure 2.8.

2.5.3 Validity of the Local Tangent Space Approximation

The method introduced in [3] relies on a local Euclidean approximation and, albeit the choice of the best local approximation is data-driven, one may question about the suitability of the model for the data at hand. The authors presented a simulation study to explore this issue in the case of positive-definite matrices, by evaluating the performance of the kriging predictor when data are generated from a model different from (2.20) (i.e. the local linear approximation is not valid). In particular, they generate a nonstationary matrix field according to a probabilistic model with mean $G_s = \exp_P(A(\vec{f}(s); \vec{a}))$, where P and $A(\vec{f}, \vec{a})$ are set parameters. This random

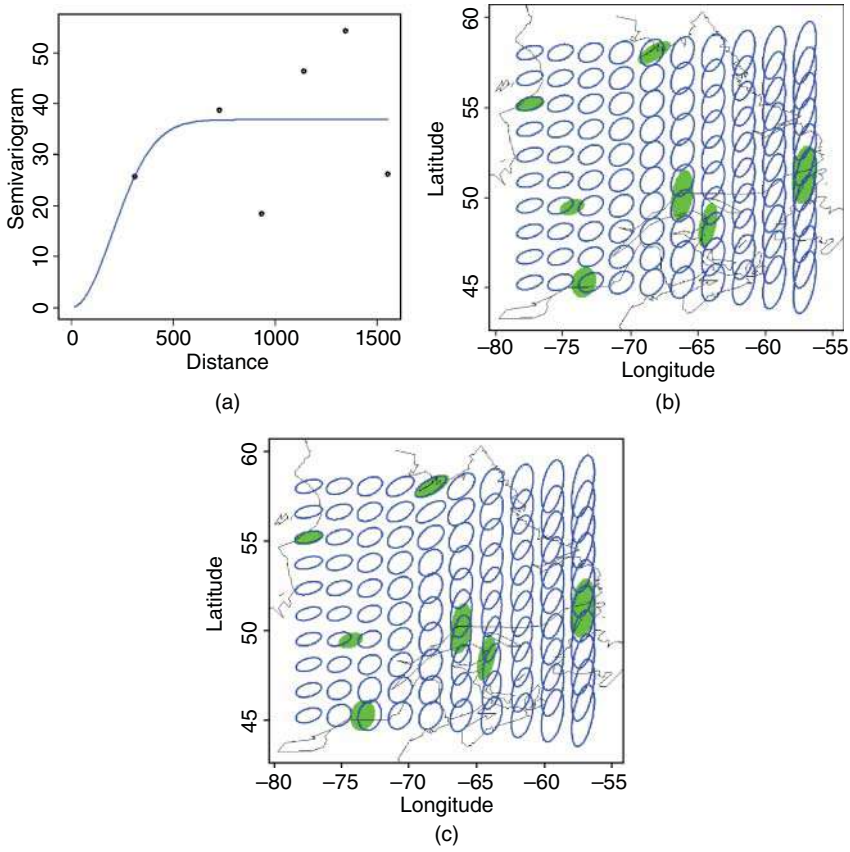


Figure 2.8 Kriging of the (temperature, precipitation) covariance matrix field during January, with a drift term depending on longitude. A covariance matrix $\boldsymbol{S}$ at location s is represented as an ellipse centered in s and with axis $\sqrt{\sigma_j}\vec{\boldsymbol{e}}_j$, where $\boldsymbol{S}\vec{e}_j = \sigma_j\vec{\boldsymbol{e}}_j$ for $j = 1, 2$. Horizontal and vertical axes of the ellipses represent temperature and precipitation, respectively. In subfigures (b) and (c), filled ellipses indicate the data, empty ellipses the estimated drift and the kriging interpolation, respectively. In subfigure (a), the residual empirical semivariogram (symbols) and the fitted exponential model (solid line) are reported. The geodesic distances in the spatial domain are measured in kilometers. Source: Modified from Pigoli et al. [3].

matrix field is obtained through the sample covariance matrices generated by the realizations of a Gaussian random vector field $\vec{\boldsymbol{v}}$ in the following way. Let $D \subset \mathbb{R}^2$ indicate the common spatial domain of two independent Gaussian random fields $\boldsymbol{w}_s, \boldsymbol{y}_s,\ s \in D$. Both random fields $\boldsymbol{w}_s$ and $\boldsymbol{y}_s$ have zero mean and Gaussian spatial covariance with decay $\phi = 0.1$, sill equal to 1 and zero nugget. For $s \in D$, the covariance matrix (between components) of the random vector field $\vec{\boldsymbol{v}}_s = (G_s)^{\frac{1}{2}}(\boldsymbol{w}_s, \boldsymbol{y}_s)'$

is equal to G_s. Then, N independent realizations of the random vector field $\vec{\boldsymbol{v}}$ are generated and, for $s \in D$, the realization of the manifold-valued random field is given by the sample covariance matrices:

$$\boldsymbol{S}_s = \frac{1}{N-1}\sum_{k=1}^{N}\left(\vec{\boldsymbol{v}}_{s,k} - \bar{\vec{\boldsymbol{v}}}_s\right)\left(\boldsymbol{v}_{s,k} - \bar{\vec{\boldsymbol{v}}}_s\right)^T \sim \text{Wishart}_2\left(\frac{1}{N-1}G_s, N-1\right), \quad (2.24)$$

$\bar{\vec{\boldsymbol{v}}}_s$ being the sample mean in $s \in D$. This simulation process is therefore defined on the manifold of positive definite symmetric matrices and the parameter N controls the variability of the matrix random field $\boldsymbol{S}$ in s. When N is large, the data will be concentrated around the drift (which satisfies the tangent space approximation). Therefore, N also controls the violation of the tangent space approximation. We want to evaluate the performance of the kriging procedure when applied to these simulated fields by comparison with the case when data are generated by model (2.20). Data are then generated on an equally spaced 10×10 grid, 15 locations are taken as known and the prediction error $\overline{\mathbf{p}} = \frac{1}{85}\sum_{i=1}^{85} d\left(\mathbf{S}_{s_i}, \widehat{\mathbf{S}}_{s_i}\right)^2$ in the remaining 85 locations is measured. Here $d(\cdot,\cdot)$ denotes the Riemannian distance between two positive definite matrices (see [3]). This experiment is repeated with different values of the model parameters. Since the two models are controlled by different parameters, to compare them on the same footing, we can measure the sample marginal variability, defined as $\varsigma = \frac{1}{100}\sum_{i=1}^{100} d(\mathbf{S}_{s_i}, G_{s_i})^2$, i.e. the variation of the realization of the field with respect to the true mean field G_s.

Figure 2.9 compares the performances of the kriging prediction when data are generated with model (2.24) and when data are generated with the tangent space model (2.20), for a range of values of the sample marginal variability.

This suggests that the higher the value of the sample marginal variability, the worse is the relative performance of the kriging predictor between the two cases. This is to be expected because a high dispersion on the manifold means that no tangent space can accurately describe the data. However, for low values of the sample marginal variability, the performance of the kriging predictor in the two settings is comparable, supporting its robustness to the violation of the model provided that the tangent space approximation is able to describe the observations in a fairly accurate way. More details on this simulation study can be found in [3].

By way of example, Figure 2.9b,c represents two realizations of the matrix field generated from (2.24) for high and low values of N, respectively, i.e. low or high values of the sample marginal variability. It can be seen that values of sample marginal variability where the performance of the kriging predictor gets worse correspond to random fields too noisy to be of any use in applied scenarios. However, other examples of manifold-valued data may present cases where the tangent space approximation is not suitable and a kriging procedure defined directly on the

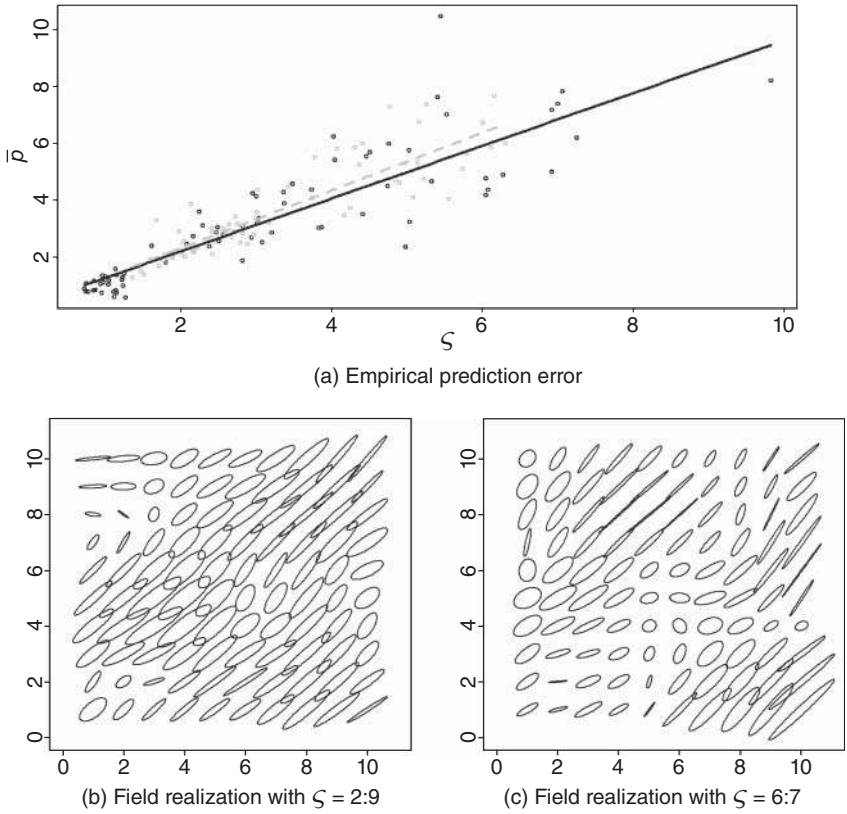


Figure 2.9 (a) Empirical prediction error as a function of the sample marginal variability ς, with a local polynomial smoothing added to help visual comparison, for data generated from the tangent space model (2.20) (black points and solid black line) and from procedure (2.24) (light gray points and dashed line), both with Gaussian covariance function. (b, c) Examples of simulated fields from procedure (2.24) for $N = 6$ (b) and $N = 4$ (c) and Gaussian covariance function, with the respective values of sample marginal variability ς. Source: Modified from Pigoli et al. [3].

manifold would be needed. Ordinary kriging for a stationary manifold-valued random field can be achieved with a weighted Fréchet mean. For example, Pigoli and Secchi [17] used this approach to estimate the mean from a spatially dependent sample. However, the optimal choice of the weights for the ordinary kriging predictor is still an open problem. When the field is nonstationary, the problem is even more complex since the nonlinear nature of the data does not allow the removal of a non-stationary mean function. A possibility currently investigated by the authors is to model the response field by segmenting the spatial domain into regions where the field can be assumed to be stationary and ordinary kriging prediction can be

used. The challenge is of course the identification of the correct scale for these subregions and how to deal with the discontinuities that may be introduced in the predicted field, e.g. via randomized approaches. Alternatively, one can think to extend to this setting locally stationary models in the same vein as [18].

2.6 Conclusion and Further Research

In this chapter, we introduced the ideas and mathematical foundations upon which object spatial statistics is grounded. We recalled the model and methods for data which can be embedded in Hilbert spaces, such as L^2 data, or Sobolev data. In this regard, we mention that constrained data such as distributional data can be dealt with in this context, by properly choosing the Hilbert space embedding for the data (see Chapter 5).

Whenever data cannot be embedded into a Hilbert space, extensions of the framework proposed in Sections 2.2 and 2.3 need to be developed. Here, we presented a possible extension to manifold-valued data, based on the idea of locally approximating the manifold with its tangent space.

The methods presented in this chapter are all based on global hypotheses of stationarity, or on global drift models to describe the nonstationarity of the field. A recent extension of the methodology presented here addresses the problem of prediction when the field cannot be assumed to be stationary on a *global* scale, but it is stationary on a *local* scale. In this setting, Menafoglio et al. [19] propose a *divide-et-impera* strategy based upon repeated Random Domain Decompositions, each defining a set of homogeneous subregions where to perform local object-oriented spatial analyses, under stationarity assumptions, to be then aggregated into a final global analysis. Besides being entirely general, and prone to be used with numerous types of object data (e.g. functional data, density data, or manifold data), the method allows to deal with complex domains, such as large and highly textured regions, with holes or barriers. This class of methods naturally finds application in several environmental applications, such as those focused on the characterization of complex estuarine systems.

References

1 Menafoglio, A. and Secchi, P. (2017). Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. *European Journal of Operational Research* 258 (2): 401–410.

2 Menafoglio, A., Secchi, P., and Dalla Rosa, M. (2013). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics* 7: 2209–2240.

3 Pigoli, D., Menafoglio, A., and Secchi, P. (2016). Kriging prediction for manifold-valued random fields. *Journal of Multivariate Analysis* 145: 117–131.

4 Menafoglio, A. and Petris, G. (2016). Kriging for Hilbert-space valued random fields: the operatorial point of view. *Journal of Multivariate Analysis* 146: 84–94. https://doi.org/10.1016/j.jmva.2015.06.012.

5 Marron, J.S. and Alonso, A.M. (2014). Overview of object oriented data analysis. *Biometrical Journal* 56 (5): 732–753.

6 Goulard, M. and Voltz, M. (1993). Geostatistical interpolation of curves: a case study in soil science. In: *Geostatistics Tróia '92*, vol. 2 (ed. A. Soares), 805–816. Dordrecht: Kluwer Academic.

7 Nerini, D., Monestiez, P., and Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis* 101 (2): 409–418.

8 Giraldo, R., Delicado, P., and Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics* 18 (3): 411–426.

9 Cressie, N. (1993). *Statistics for Spatial data*. New York: Wiley.

10 Ignaccolo, R., Mateu, J., and Giraldo, R. (2014). Kriging with external drift for functional data for air quality monitoring. *Stochastic Environmental Research and Risk Assessment* 28 (5): 1171–1186. https://doi.org/10.1007/s00477-013-0806-y.

11 Mandelbaum, A. (1984). Linear estimators and measurable linear transformations on a Hilbert space. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 65 (3): 385–397.

12 Luschgy, H. (1996). Linear estimators and radonifying operators. *Theory of Probability & Its Applications* 40 (1): 167–175.

13 Lee, J. (2012). *Introduction to Smooth Manifolds*, vol. 218. Springer Science & Business Media.

14 Diggle, P. and Ribeiro, P. (2007). *Model-Based Geostatisticss*. Springer.

15 Chiles, J.P. and Delfiner, P. (2009). *Geostatistics: Modeling Spatial Uncertainty*, vol. 497. Wiley.

16 Yuan, Y., Zhu, H., Lin, W., and Marron, J.S. (2012). Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74 (4): 697–719.

17 Pigoli, D. and Secchi, P. (2012). Estimation of the mean for spatially dependent data belonging to a Riemannian manifold. *Electronic Journal of Statistics* 6: 1926–1942.

18 Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics* 25 (1): 1–37.

19 Menafoglio, A., Gaetani, G., and Secchi, P. (2018). Random domain decompositions for object-oriented kriging over complex domains. *MOX-report 10/2018*, Politecnico di Milano.

3

Universal, Residual, and External Drift Functional Kriging

Maria Franco-Villoria[1] *and Rosaria Ignaccolo*[2]

[1] *Department of Economics "Marco Biagi", Università di Modena e Reggio Emilia, Modena, Italy*
[2] *Department of Economics and Statistics "Cognetti de Martiis", Università di Torino, Torino, Italy*

The previous chapters cover ordinary functional kriging, which allows to predict a curve at an unmonitored site under the assumption of a constant mean over the spatial domain. However, in many applications, this is an unrealistic assumption. Then, in order to realize spatial prediction for nonstationary processes, in classical geostatistics a spatial trend (also called "drift") is modeled as a function of the coordinates only or defined "externally" through some auxiliary/exogenous variables. It is common to reserve the term *Universal kriging* (UK) for the case when only the coordinates are used as regressors, whereas many authors refer to *Kriging with External Drift* (KED) model when other covariates are also considered. When the drift is estimated by Generalized Least Squares (GLS), UK, and KED prediction at a new site can be obtained by adding up the predicted drift and the result of simple kriging applied on GLS residuals because of a mathematical equivalence [1]. For this reason, some authors use the term *Residual Kriging* or the more general *Regression Kriging*.

This chapter introduces methods for nonstationary functional data; in particular, in Sections 3.2–3.4, three different approaches are presented in increasing order of trend complexity: we will refer to them as UK [2, 3], residual kriging [4], and KED [5], where the mean function depends on the spatial coordinates, scalar covariates, and scalar and functional covariates, respectively. Section 3.5 illustrates an iterative algorithm to take into account spatial dependence in drift estimation, while a bootstrap method to evaluate prediction uncertainty is presented in Section 3.6.

Application of the different kriging methods will be illustrated on a case study of PM_{10} concentrations in the region of Piemonte, Italy (see [5, 6]) in Section 3.7. The dataset and R code are freely available upon request to the authors.

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

3.1 Introduction

Let $\{\chi_s(t); t \in T\}$ be a functional random variable observed at location $s \in D \subseteq \mathbb{R}^d$ ($d = 2$ or 3 generally), where T is a compact subset of $\mathbb{R}$. Assume that we observe a sample of curves $\chi_{s_i}(t)$, for $t \in T$ and $s_i \in D$, $i = 1, \dots, n$, taking values in a separable Hilbert space H of square integrable functions, that is in L^2. The set $\{\chi_s, s \in D\}$ constitutes a functional random field or a *spatial functional process* [7] that is not necessarily stationary. With the aim of predicting the curve $\chi_{s_0}(t)$ at an unmeasured location $s_0 \in D$, the following model is assumed:

$$\chi_s(t) = m_s(t) + \epsilon_s(t), \quad s \in D, \tag{3.1}$$

where $m_s(t)$ is the drift describing a spatial trend and $\epsilon_s(t)$ is a zero-mean, second-order stationary, and isotropic residual random field, with covariance function $\mathbb{C}(\epsilon_{s_i}(t), \epsilon_{s_j}(u)) = \text{Cov}(\epsilon_{s_i}(t), \epsilon_{s_j}(u)) = C(h; t, u), \forall s_i, s_j \in D$, where $h = ||s_i - s_j||$ represents the Euclidean distance between locations s_i and s_j, $t, u \in T$. Denote $\mathbf{\Sigma} = \mathbb{V}ar\left(\boldsymbol{\epsilon}_s(t)\right)$, $\sigma_0^2(t) = \mathbb{V}ar(\epsilon_{s_0}(t))$ and $\mathbf{c} = C(\boldsymbol{\epsilon}_s(t), \epsilon_{s_0}(t)), t \in T$.

Sections 3.2–3.4 describe different kriging approaches depending on the complexity of the trend term $m_s(t)$.

3.2 Universal Kriging for Functional Data (UKFD)

In the case where the trend $m_s(t)$ only depends on the spatial coordinates s, Model (3.1) can be rewritten as a spatial functional regression model:

$$\chi_{s_i}(t) = \sum_{p=1}^{P} \beta_p(t) f_p(s_i) + \epsilon_{s_i}(t), \quad i = 1, \dots, n, \quad t \in T, \tag{3.2}$$

where $f_p(s)$ are functions of the spatial coordinates. Given the functional vector of observations $\chi_s(t) = \{\chi_{s_1}, \dots, \chi_{s_n}\}$, the model can be written in matrix form as follows:

$$\chi_s(t) = \mathbf{X}(s)\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}_s(t),$$

where $\mathbf{X}(s)$ is a matrix of size $n \times P$ with generic element $f_p(s_i)$, $\boldsymbol{\beta}(t) = \left[\beta_1(t), \dots, \beta_P(t)\right]^T$ is the vector of functional regression coefficients, and $\boldsymbol{\epsilon}(t) = \left[\epsilon_{s_1}(t), \dots, \epsilon_{s_n}(t)\right]^T$ is the vector of functional residuals.

Caballero et al. [2] propose the functional UK predictor of $\chi_{s_0}(t)$ as follows:

$$\hat{\chi}_{s_0}(t) = \lambda^T \chi_s(t), \tag{3.3}$$

where $\lambda = \{\lambda_1, \dots, \lambda_n\} \in \mathbb{R}^n$ is the vector of kriging weights, chosen so that predictor (3.3) is unbiased with minimum variance. The unbiasedness condition

means that $\mathbb{E}\left[\hat{\chi}_{s_0}(t)\right] = \mathbb{E}\left[\chi_{s_0}(t)\right]$ for all $t \in T$, i.e.

$$\lambda^T \mathbf{X}(s)\boldsymbol{\beta}(t) = \mathbf{X}^T(s_0)\boldsymbol{\beta}(t)$$

or, equivalently,

$$\left[\lambda^T \mathbf{X}(s) - \mathbf{X}^T(s_0)\right]\boldsymbol{\beta}(t) = \mathbf{0};$$

so that the predictor is unbiased if

$$\lambda^T \mathbf{X}(s) = \mathbf{X}^T(s_0). \tag{3.4}$$

The kriging weights $\{\lambda_1, \ldots, \lambda_n\}$ are obtained by minimizing the variance of the prediction error subject to the unbiasedness constraint (3.4), that is solving the minimization problem:

$$\min_{\lambda_1,\ldots,\lambda_n} \mathrm{MSE}(s_0) \quad \text{subject to} \quad \lambda^T \mathbf{X}(s) = \mathbf{X}^T(s_0),$$

where $\mathrm{MSE}(s_0) = \int_T \mathrm{Var}\{\hat{\chi}_{s_0}(t) - \chi_{s_0}(t)\}dt$.

The variance $\mathrm{Var}\{\hat{\chi}_{s_0}(t) - \chi_{s_0}(t)\}$ can be rewritten as follows:

$$\begin{aligned}\mathrm{Var}\{\hat{\chi}_{s_0}(t) - \chi_{s_0}(t)\} &= \mathrm{Var}[\hat{\chi}_{s_0}(t)] - 2\mathbb{C}(\hat{\chi}_{s_0}(t), \chi_{s_0}(t)) + \mathrm{Var}[\chi_{s_0}(t)] \\ &= \lambda^T \mathrm{Var}(\chi_s(t))\lambda - 2\lambda^T \mathbb{C}(\chi_s(t), \chi_{s_0}(t)) + \sigma_0^2(t) \\ &= \lambda^T \boldsymbol{\Sigma}\lambda - 2\lambda^T \mathbf{c}(t) + \sigma_0^2(t). \end{aligned} \tag{3.5}$$

The function to be minimized becomes

$$\begin{aligned}\phi(\lambda, \mathbf{v}) = &\int_T \left(\lambda^T \boldsymbol{\Sigma}\lambda\right)(t)dt - 2\int_T \lambda^T \mathbf{c}(t)dt \\ &+ \int_T \sigma_0^2(t)dt - 2\left[\lambda^T \mathbf{X}(s) - \mathbf{X}^T(s_0)\right]\mathbf{v},\end{aligned} \tag{3.6}$$

where $\mathbf{v}$ is the vector of Lagrange multipliers for the unbiasedness condition.

Deriving equation (3.6) partially with respect to λ and $\mathbf{v}$ and equating to zero, we obtain the system of equations:

$$\begin{cases} \left[\int_T \boldsymbol{\Sigma}dt\right]\lambda - \mathbf{X}(s)\mathbf{v} &= \int_T \mathbf{c}(t)dt, \\ \mathbf{X}^T(s)\lambda + \mathbf{0}\mathbf{v} &= \mathbf{X}(s_0), \end{cases}$$

that needs to be solved in order to have an explicit analytical expression for the predictor (3.3) and for the variance of the prediction error $\mathrm{MSE}(s_0)$. In particular, from the first line, we have

$$\lambda = \left[\int_T \boldsymbol{\Sigma}dt\right]^{-1}\left[\int_T \mathbf{c}(t)dt + \mathbf{X}(s)\mathbf{v}\right] \tag{3.7}$$

and then substituting in the second line, we get

$$\left\{\left[\int_T \boldsymbol{\Sigma}dt\right]^{-1}\left[\int_T \mathbf{c}(t)dt + \mathbf{X}(s)\mathbf{v}\right]\right\}^T \mathbf{X}(s) = \mathbf{X}^T(s_0)$$

or alternatively,

$$\left[\int_T \mathbf{c}^T(t)dt + \mathbf{v}^T\mathbf{X}^T(s)\right]\left[\int_T \mathbf{\Sigma}dt\right]^{-1}\mathbf{X}(s) = \mathbf{X}^T(s_0).$$

Solving for $\mathbf{v}$, we obtain

$$\mathbf{v}^T = \left\{\mathbf{X}^T(s_0) - \int_T \mathbf{c}^T(t)dt\left[\int_T \mathbf{\Sigma}dt\right]^{-1}\mathbf{X}(s)\right\}\left\{\mathbf{X}^T(s)\left[\int_T \mathbf{\Sigma}dt\right]^{-1}\mathbf{X}(s)\right\}^{-1}$$

and finally,

$$\mathbf{v} = \mathbf{W}^{-1}\left\{\mathbf{X}(s_0) - \mathbf{Y}\right\}, \tag{3.8}$$

where $\mathbf{W} = \mathbf{X}^T(s)\left[\int_T \mathbf{\Sigma}dt\right]^{-1}\mathbf{X}(s)$ and $\mathbf{Y} = \mathbf{X}^T(s)\left[\int_T \mathbf{\Sigma}dt\right]^{-1}\int_T \mathbf{c}(t)dt$.

Substituting (3.8) into (3.7) we obtain

$$\lambda = \left[\int_T \mathbf{\Sigma}dt\right]^{-1}\left[\int_T \mathbf{c}(t)dt\right] + \left[\int_T \mathbf{\Sigma}dt\right]^{-1}\mathbf{X}(s)\mathbf{W}^{-1}\left\{\mathbf{X}(s_0) - \mathbf{Y}\right\} \tag{3.9}$$

so that the predictor (3.3) can be expressed as

$$\hat{\chi}_{s_0}(t) = \left\{\left[\int_T \mathbf{c}^T(t)dt\right]\left[\int_T \mathbf{\Sigma}dt\right]^{-1} + \left(\mathbf{X}(s_0) - \mathbf{Y}\right)^T\mathbf{W}^{-1}\mathbf{X}^T(s)\left[\int_T \mathbf{\Sigma}dt\right]^{-1}\right\}\chi_s(t).$$

Given (3.5), we have $\mathrm{MSE}(s_0) = \int_T \left(\lambda^T\mathbf{\Sigma}\lambda\right)(t)dt - 2\lambda^T\int_T \mathbf{c}(t)dt + \int_T \sigma_0^2(t)dt$, where it is possible to substitute the value of λ in order to obtain the variance of the predictor $\sigma^2_{s_0} = \mathrm{Var}\left(\hat{\chi}_{s_0}(t)\right)$, that is

$$\sigma^2_{s_0} = \int_T \sigma_0^2(t)dt - \left[\int_T \mathbf{c}(t)dt\right]^T\left[\int_T \mathbf{\Sigma}dt\right]^{-1}\left[\int_T \mathbf{c}(t)dt\right] + \left(\mathbf{X}(s_0) - \mathbf{Y}\right)^T\mathbf{W}^{-1}\left(\mathbf{X}(s_0) - \mathbf{Y}\right).$$

It is clear that in order to evaluate the kriging coefficients and the variance of the predictor, we need to know $\mathbf{\Sigma}$ and $\mathbf{c}$. To this end, the so-called "trace-variogram" has been introduced and a method-of-moments estimator proposed in Giraldo et al. [8] (see Chapter 1).

Remark. The UK described above is valid for spatial functional processes valued in L^2. A more general theory valid for Hilbert spaces in general has also been developed by Menafoglio et al. [3]. This general theory is presented in Chapter 2 and coincides exactly with what has been described in Section 3.2 [2] when the Hilbert space considered is L^2.

3.3 Residual Kriging for Functional Data (ResKFD)

A more general situation of nonstationarity can be modeled by allowing the trend component $m_s(t)$ in Eq. (3.1) to depend on scalar covariates different from the

spatial coordinates. In this case, residual functional kriging – as proposed by Reyes et al. [4] – can be used to predict a curve at an unknown location s_0. As already stated, estimating the drift coefficients by GLS and applying simple kriging to GLS residuals is equivalent to UK prediction in classical geostatistics [1]; indeed, this is also true for the functional case universal kriging for functional data (UKFD) with the kriging coefficients λ derived in Section 3.2. In practice, residual kriging is easier to implement than UK, because it consists in applying kriging to detrended curves. Note that in UKFD, in order to evaluate λ in (3.9) it is necessary to estimate $\mathbf{\Sigma}$ and $\mathbf{c}$. For doing so, first the trend component has to be estimated and subsequently a theoretical model for the trace-variogram of the detrended data (regression residuals) has to be chosen; the parameters of the latter are then estimated from the regression residuals and not from the original data. Thus, from the computational point of view, this initial detrending step is common to UKFD and residual kriging for functional data (ResKFD).

To proceed with ResKFD in a general way, Model (3.1) can be written as follows:

$$\chi_{s_i}(t) = \sum_{p=1}^{P} \beta_p(t) C_{p,i} + \epsilon_{s_i}(t), \quad i = 1, \dots, n, \tag{3.10}$$

where C_p, $p = 1, \dots, P$ are scalar covariates observed at sites s_i and not only functions of the spatial coordinates as in (3.2). The trend is estimated using a functional linear regression model and one of the three kriging alternatives for stationary spatial functional processes (see Chapter 1) is applied on the detrended data to obtain $\hat{\epsilon}_{s_0}(t)$: (1) ordinary kriging for functional data (OKFD) [8], where $\hat{e}_{s_0}(t) = \sum_{i=1}^{n} \lambda_i e_{s_i}(t)$ and $\lambda_i \in \mathbb{R}$; (2) continuous time-varying kriging for functional data (CTKFD) [9], where $\hat{e}_{s_0}(t) = \sum_{i=1}^{n} \lambda_i(t) e_{s_i}(t)$ so that the kriging coefficients $\lambda_i(t)$ are functional too, and (3) functional kriging total model (FKTM) [10, 11], where $\hat{e}_{s_0}(t) = \sum_{i=1}^{n} \int_T \lambda_i(\tau, t) e_{s_i}(\tau) d\tau$ and the kriging coefficients are defined on $T \times T$. In doing so, the trace-variogram is used for estimating the kriging weights in ordinary kriging, while for the continuous time-varying kriging and the FKTM a Linear Model of Coregionalization (LMC) [12] is needed.

The final prediction $\hat{\chi}_{s_0}(t)$ is obtained as the sum of the estimated trend at s_0 plus the krigged residual at s_0:

$$\hat{\chi}_{s_0}(t) = \sum_{p=1}^{P} \hat{\beta}_p(t) C_{p,0} + \hat{\epsilon}_{s_0}(t), \tag{3.11}$$

where $C_{p,0}$ is the pth scalar covariate at site s_0.

When a number of potential covariates are available, interest may be not only in prediction but in assessing the effect of those covariates. In this sense, ResKFD is advantageous as a trend estimation, and spatial prediction of residuals is performed separately, allowing interpretation of each of the two components. Further, the trend does not necessarily have to be linear but more complex forms or regression is possible.

3.4 Functional Kriging with External Drift (FKED)

The third modeling strategy considers a trend component $m_s(t)$ that is allowed to depend on both scalar and functional covariates, as proposed by Ignaccolo et al. [5]. At a fixed site s_i, $i = 1, \dots, n$, and domain point t Model (3.1) can be seen as a functional concurrent linear model:

$$\chi_{s_i}(t) = m_{s_i}(t) + \epsilon_{s_i}(t), \tag{3.12}$$

where the trend or drift term can be expressed in terms of a set of scalar and functional covariates:

$$m_{s_i}(t) = \alpha(t) + \sum_p \beta_p(t) C_{p,i} + \sum_q \gamma_q(t) X_{q,i}(t), \tag{3.13}$$

where $\alpha(t)$ is a functional intercept, $C_{p,i}$ is the pth scalar covariate at site s_i, $X_{q,i}$ is the qth functional covariate at site s_i, and $\beta_p(t)$ and $\gamma_q(t)$ are the covariate functional coefficients. As in the previous models, $\epsilon_{s_i}(t)$ represents the zero-mean, stationary, and isotropic residual spatial functional process $\{\epsilon_s(t), t \in T, s \in D\}$ at site s_i.

Prediction is carried out following a three-step procedure [5]. At the *first step*, we need to estimate the drift coefficients of the functional regression model (3.12) with functional response and scalar and functional covariates. The functional coefficients in equation (3.13) can be estimated by means of a generalized additive model (GAM) representation using the R package `mgcv` by re-expressing the functional linear model as a standard additive model [13–15]. To rewrite Model (3.12) as a GAM, we assume that the functional coefficients in (3.13) are expandable as

$$\alpha(t) = \sum_{l=1}^{k_0} A_{0,l}(t) c_{0,l}, \quad \beta_p(t) = \sum_{l=1}^{k_p} a_{p,l}(t) c_{p,l}, \quad \text{and} \quad \gamma_q(t) = \sum_{l=1}^{k_q} a_{q,l}(t) c_{q,l},$$

where $A_{0,l}(t)$, $a_{p,l}(t)$, and $a_{q,l}(t)$ are known basis functions, while $c_{0,l}$, $c_{p,l}$, and $c_{q,l}$ are the related coefficients (to be estimated). Then by plugging-in these quantities, we have

$$\beta_p(t) C_{p,i} = \sum_{l=1}^{k_p} a_{p,l}(t) C_{p,i} c_{p,l} = \sum_{l=1}^{k_p} A_{p,l,i}(t) c_{p,l},$$

and

$$\gamma_q(t) X_{q,i}(t) = \sum_{l=1}^{k_q} a_{q,l}(t) X_{q,i}(t) c_{q,l} = \sum_{l=1}^{k_q} A_{q,l,i}(t) c_{q,l}$$

and finally, the functional linear model (3.12) can be rewritten as a standard additive model:

$$\chi_{s_i}(t) = \sum_{l=1}^{k_0} A_{0,l}(t) c_{0,l} + \sum_p \sum_{l=1}^{k_p} A_{p,l,i}(t) c_{p,l} + \sum_q \sum_{l=1}^{k_q} A_{q,l,i}(t) c_{q,l} + \epsilon_{s_i}(t),$$

where $A_{p,l,i}(t) = a_{p,l}(t)C_{p,i}$ and $A_{q,l,i}(t) = a_{q,l}(t)X_{q,i}(t)$ are known because $C_{p,i}$ and $X_{q,i}(t)$ are "observed" without noise. To fit this model, a penalized regression spline approach can be used, where the smoothing parameters can be automatically chosen using the Generalized Cross Validation (GCV) criterion or estimated using Restricted Maximum Likelihood (REML) [16, 17], as implemented in the `mgcv` package. To take the spatial dependence into account when estimating the drift term, an iterative algorithm can be implemented as described in Section 3.5.

Once the drift coefficients have been estimated, the functional residuals can be obtained as

$$e_{s_i}(t) = \chi_{s_i}(t) - \hat{m}_{s_i}(t) = \chi_{s_i}(t) - \left[\hat{\alpha}(t) + \sum_p \hat{\beta}_p(t)C_{p,i} + \sum_q \hat{\gamma}_q(t)X_{q,i}(t)\right].$$

At the *second step*, the resulting functional residuals (at the last iteration of the algorithm described in Section 3.5) $e_{s_i}(t)$ can be used to predict the residual curve at an unmonitored site s_0 via one of three kriging options described in detail in Chapter 1 and already introduced for ResKFD: ordinary kriging, continuous time-varying kriging, or the FKTM.

Finally, at the *third step* the two terms are added – as in the classical regression kriging – to obtain the prediction at the unmonitored site s_0, that is

$$\hat{\chi}_{s_0}(t) = \hat{m}_{s_0}(t) + \hat{e}_{s_0}(t),$$

where $\hat{m}_{s_0}(t) = \hat{\alpha}(t) + \sum_p \hat{\beta}_p(t)C_{p,0} + \sum_q \hat{\gamma}_q(t)X_{q,0}(t)$ depends on the covariate values $C_{p,0}$ and $X_{q,0}(\cdot)$ at the site s_0.

3.5 Accounting for Spatial Dependence in Drift Estimation

To take into account the spatial correlation between functional observations when estimating the drift term, Menafoglio et al. [3] and Franco-Villoria and Ignaccolo [18] propose an iterative algorithm to adjust the estimated functional coefficients for spatial dependence. We refer here the details of the latter proposal [18] for being a more general strategy that considers a drift term that may depend on both scalar and functional covariates, while the former [3] only includes the spatial coordinates as covariates.

Model (3.1) can be written as a GAM and subsequently as a mixed effects model [19, 20], where the drift term $m_s(t)$ may be more or less complex depending on the kriging model as specified in Sections 3.2–3.4. In this framework, the parameters can be estimated using REML [17] assuming normally distributed functional errors $\epsilon_{s_i}(t)$ (from a longitudinal point of view). The spatial dependence can be incorporated considered $\epsilon_{s_i}(t)$ as a functional random intercept whose covariance

structure can be estimated in terms of the trace-variogram [21]. This can be implemented using the following algorithm:

(1) Estimate the drift term $m_{s_i}(t)$ under the assumption of independent observations and obtain the functional residuals $e_{s_i}(t) = \chi_{s_i}(t) - \hat{m}_{s_i}(t)$.
(2) Estimate the correlation matrix $K = \left\{ \text{Corr}(\epsilon_{s_i}(t), \epsilon_{s_j}(t)) \right\}_{i,j=1,\ldots,n}$ of the residual spatial functional process using the trace-variogram [8], introduced in Chapter 1.
(3) Consider the term $\epsilon_s(t)$ as a functional random effect with precision matrix $\hat{K}^{-1}$ [21] and re-estimate Model (3.1).
(4) Iterate Steps 1–3 until convergence, defined in terms of the Akaike Information Criterion (AIC); convergence is reached when the *AIC* rate, defined for the *j*th iteration as

$$\text{AICrate} = \left| \frac{\text{AIC}^j - \text{AIC}^{j-1}}{\text{AIC}^{j-1}} \right|,$$

is smaller than a preset tolerance value (e.g. 0.1%).

3.5.1 Drift Selection

An iterative algorithm to select the most appropriate drift term could be used, which chooses the "best" drift from a set of candidate drifts based on cross-validation mean square error (MSE) as proposed by Menafoglio et al. [3] (see Chapter 2). Clearly, the selection of the drift could be performed in classical ways, by using a validation set approach or other variable selection indexes. When a GAM model is fitted the AIC criterion could be also adopted for model comparison.

3.6 Uncertainty Evaluation

Quantifying the uncertainty associated with a predicted curve $\hat{\chi}_{s_0}(t)$ is fundamental; in the case of functional data, ideally, one would want an uncertainty measure that may vary over the domain of the predicted curve, while the classic functional kriging variance provides a unique value. Given the lack of an analytic expression of a domain-varying kriging variance for a curve, resampling methods (bootstrapping) need to be used for prediction interval calculation as proposed by Franco-Villoria and Ignaccolo [18]. Given that the functional observations are not independent, resampling directly would not be appropriate; however, if one decorrelates the data so that they become spatially independent, sampling with replacement can be done to obtain a bootstrap sample that has to be transformed

again to incorporate the spatial dependence. The idea is to approximate F_n, the distribution of $\hat{\chi}_{s_0}(t) - \chi_{s_0}(t)$, using bootstrapping, to then build a $1-\alpha$ prediction interval for $\chi_{s_0}(t)$ using the quantiles of the approximated distribution $\hat{F}_n^*$. This can be achieved following the algorithm described below.

(1) Estimate the drift term $m_{s_i}(t)$ using the algorithm described in Section 3.5 and remove it to obtain the functional residuals.
(2) Decompose the covariance matrix of the functional residuals (that can be estimated using the trace-variogram in the OKFD case) using Cholesky decomposition as $\hat{\Sigma}_{n\times n} = \hat{L}_{n\times n}\hat{L}_{n\times n}^T$ and transform the functional residuals multiplying by $\hat{L}^{-1}$ so that they become (spatially) uncorrelated.
(3) Sample with replacement from the vector of uncorrelated functional residuals obtained in Step 2 to generate B bootstrap samples of size $n+1$.
(4) Transform the bootstrap samples obtained in Step 3 to reincorporate the spatial dependence. This can be achieved using the covariance matrix $\hat{\Lambda} = \begin{bmatrix} \hat{\Sigma} & \hat{c}_n^T \\ \hat{c}_n & \hat{\sigma}^2 \end{bmatrix}$, where $\hat{c}_n = \{\hat{C}(s_i - s_0)\}_{i=1}^n$, $\hat{C}$ is the estimated covariance function and $\hat{\sigma}^2 = \hat{C}(0)$ is the estimated scale. Using the Cholesky decomposition once again, the matrix can be decomposed as $\hat{\Lambda} = \hat{R}\hat{R}^T$. The (independent) bootstrap samples are then multiplied by the matrix $\hat{R}$.
(5) Add back the drift term estimated in Step 1 to the functional residuals obtained in Step 4 to obtain the final bootstrap sample, that includes data not only at the $s_1, \dots, s_n$ locations but also at the unmonitored location s_0.

B prediction curves $\hat{\chi}_{s_0}^{*j}$ are then obtained by fitting a kriging model (universal, residual, functional kriging with external drift [FKED]) to each of the bootstrap samples $\{\chi_{s_1}^{*j}, \dots, \chi_{s_n}^{*j}\}_{j=1}^B$ and the differences between the predicted and "observed" curves $\{\hat{\chi}_{s_0}^{*j} - \chi_{s_0}^{*j}\}_{j=1}^B$ are considered. The $(1-\alpha)\%$ prediction interval for $\chi_{s_0}(t)$ can be obtained as

$$\left(\hat{\chi}_{s_0}(t) - q_{1-\alpha/2}^*, \hat{\chi}_{s_0}(t) - q_{\alpha/2}^*\right),$$

where q_α^* is the αth percentile of $\hat{F}_n^*$, that can be obtained ordering the set of curves $\{\hat{\chi}_{s_0}^{*j} - \chi_{s_0}^{*j}\}_{j=1}^B$. These can be ordered using different approaches, such as modified band depth (MBD) [22] or on L^2 distance between curves [23]. The former provides a measure of centrality by calculating the proportion of times that each curve is (fully or partially) contained in all possible bands; a band is defined as the area contained within two curves. With this scheme, the bigger the band depth value, the more central the curve is. In the latter, the bootstrap-based predicted curves are ordered based on how distant they are from the zero curve, according to the L^2 distance; with this scheme, the smaller the distance, the more central the curve is.

Further details, a simulation study and discussion can be found in Franco-Villoria and Ignaccolo [18].

3.7 Implementation Details in R

Conversion from discrete data to functional data is carried out using the `fda` package [24]. The drift term can be estimated using the `fRegress` function from the `fda` package; however, this does not allow to take into account the spatial dependence. Implementation of the algorithm described in Section 3.5 requires estimating the model as a generalized additive mixed model, for which the function `gamm` from the `mgcv` package can be used. As seen in Chapter 1, the `geofd` R package can be used to implement ordinary kriging; the function `okfd` includes the automatic choice of a model (among spherical, exponential, Gaussian or Matérn) for the trace-variogram, by minimizing the sum of square errors (SSE) between the theoretical variogram and the empirical one. Instead CTKFD and FKTM can be carried out taking advantage of the package `gstat` [25] for fitting a linear coregionalization model.

We illustrate how to do it in practice on a case study of PM_{10} concentration in the region of Piemonte, Italy. In practice for any of the three cases, we can always do residual kriging – i.e. fitted trend at s_0 plus krigged residuals (using ordinary kriging) as explained in Sections 3.3 and 3.4. Here, we concentrate on an example for the more complex case, i.e. FKED. The remaining two (UKFD, ResKFD) can be easily implemented simplifying the trend term $m_s(t)$ so that it only includes scalar or spatial coordinates as covariates. The code for implementing FKED is available upon request to the authors, while code for the uncertainty evaluation is available as supplementary material of [18].

3.7.1 *Example*: Air Pollution Data

The example concerns air pollution in the Piemonte region (Italy) and has been previously considered in the air quality literature [5, 6, 26]. Daily PM_{10} concentrations ($\mu g/m^3$) measured by the monitoring network of Piemonte are available from October 2005 to March 2006 on 24 sites (light gray triangles in Figure 3.1) plus 10 additional sites (dark gray dots in Figure 3.1) used as validation stations. Daily maximum mixing height (*MMH*, in m), daily total precipitation (*PREC*, in mm), daily mean wind speed (*MWS*, m/s), daily mean temperature (*TEMP*, in $^\circ$K), and daily emission rates of primary aerosols (*EMISS*, in g/s) are available as functional covariates; these were obtained from the output of a nested system of deterministic computer-based models implemented by the environmental agency ARPA Piemonte [27]. Longitude, latitude (*UTMX* and *UTMY*, in km), and altitude (*ALT*, in m) of the measuring stations were used as scalar covariates.

The response variable (PM_{10}) included a number of missing values; the dataset used here is an imputed version with no missing values. The original dataset can

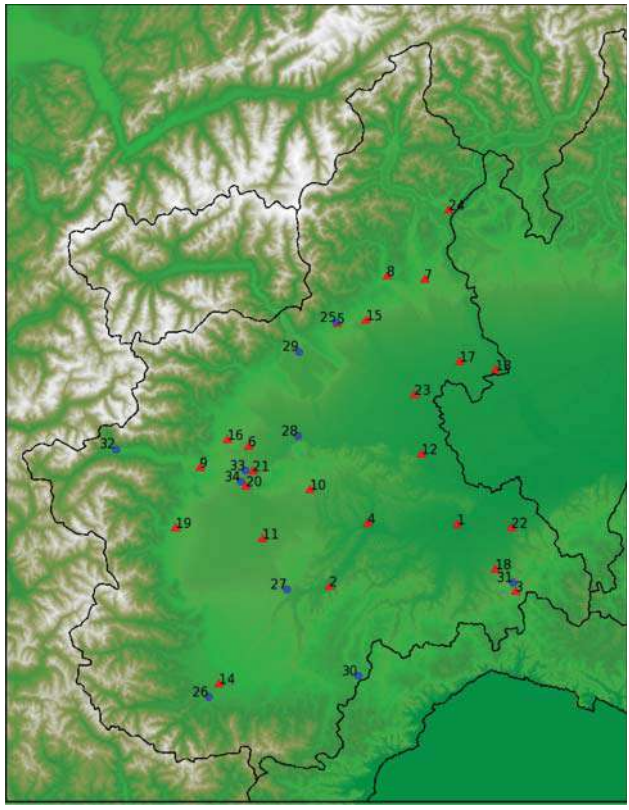


Figure 3.1 Locations of the 24 PM_{10} monitoring sites (light gray triangles) and 10 validation stations (dark gray dots) in the Piemonte region.

be downloaded from the R-INLA website (http://www.r-inla.org/examples/case-studies/cameletti-et-al) as this dataset was analyzed in the paper by Cameletti et al. [26].

Prior to modeling, PM_{10} data were log transformed (Figure 3.2) and both $\log(PM_{10})$ and the functional covariates were smoothed using cubic B-splines (146 basis functions, no penalty); these were chosen using functional cross-validation [5]). All covariates were standardized.

Prediction at the validation stations was done using the FKED model introduced in Section 3.4. All three alternatives (ordinary, time varying, total kriging model) were considered but ordinary kriging (with an exponential variogram model and no nugget) turned out to be the best option in terms of model performance. The trace-variogram cloud and the corresponding fitted variogram model (exponential with estimated sill $\hat{\sigma}^2 = 46.0156$ and range $\hat{\phi} = 25.1335$) are shown in Figure 3.3.

The estimated functional coefficients (Model 3.13) are shown in Figure 3.4; the black curves correspond to estimates obtained under the assumption of

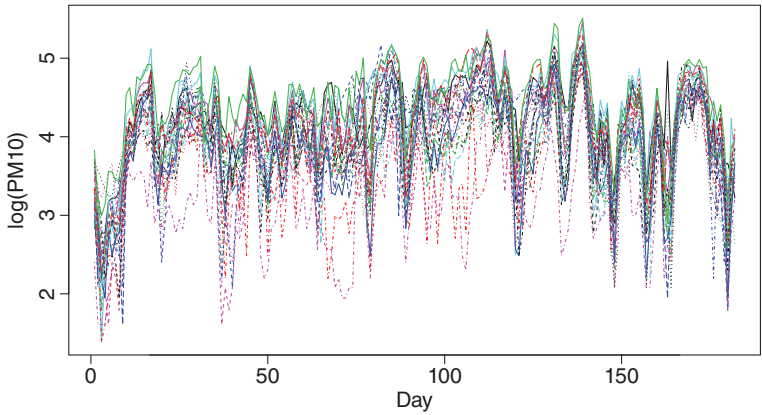


Figure 3.2 PM_{10} raw data (in log scale) observed at the 24 monitoring sites. Source: Modified from Ignaccolo et al. [5].

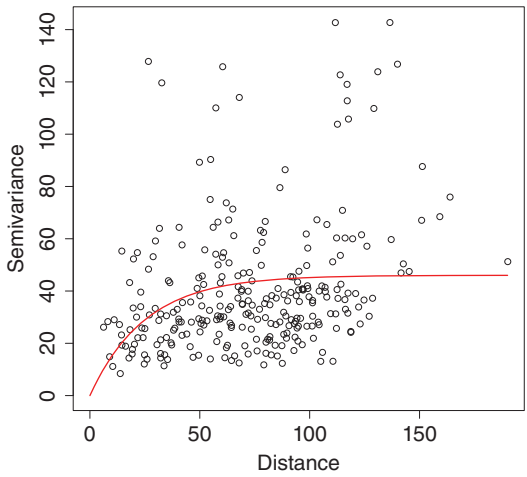


Figure 3.3 Trace-variogram cloud and estimated trace-variogram.

independence. The iterative algorithm described in Section 3.5 was implemented to take into account the spatial dependence and the resulting estimates are shown in light gray. Even though there are no great changes in the shape of the estimated coefficients, the pointwise confidence bands become a bit wider once the spatial dependence is taken into account.

Model performance was evaluated by comparing the (raw) observed and predicted data at the 10 validation sites using 4 different indexes: the normalized mean bias factor (NMBF), the root mean square error (RMSE), the weighted normalized mean squared error of the normalized ratios (WNNR), and the correlation coefficient (ρ); these are summarized in Table 3.1.

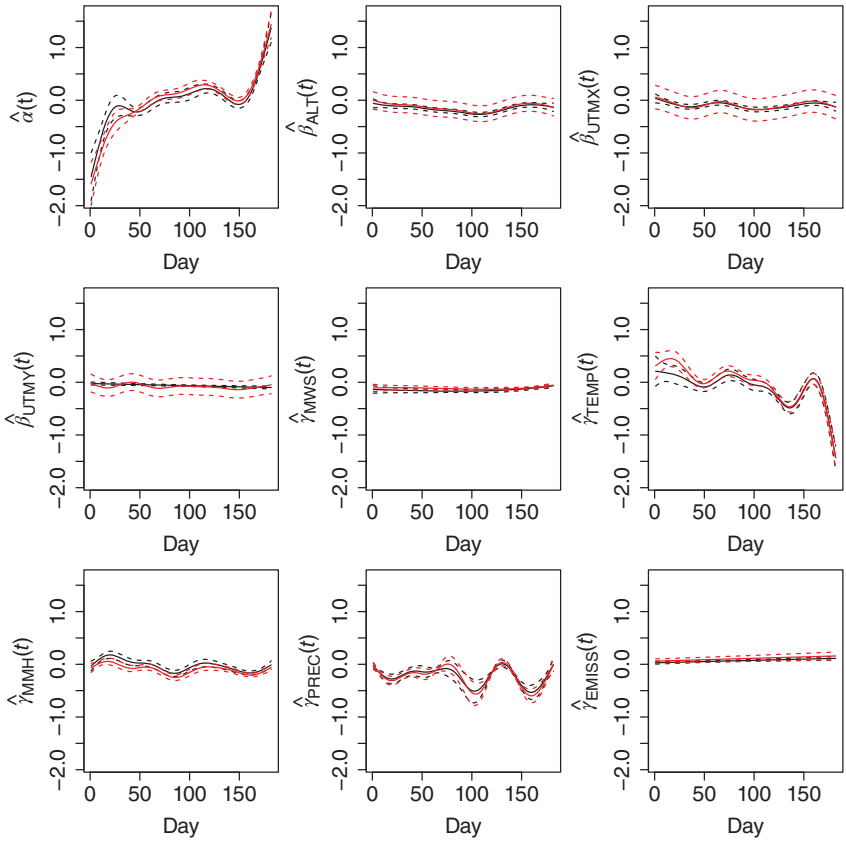


Figure 3.4 Estimated functional coefficients assuming independent observations (black) and adjusted for spatial dependence (gray).

Table 3.1 Performance indexes over the 10 validation sites.

	Site	NMBF	RMSE	WNNR	ρ
25	BI – Largo Lamarmora	−0.020	0.196	0.003	0.943
26	Borgo San Dalmazzo	−0.048	0.460	0.018	0.762
27	Bra	−0.059	0.313	0.006	0.922
28	Chivasso – Edipower	0.004	0.250	0.004	0.885
29	Ivrea	−0.058	0.331	0.008	0.914
30	Saliceto	−0.130	0.719	0.055	0.612
31	Serravalle Scrivia	0.038	0.463	0.012	0.793
32	Susa	0.005	0.481	0.016	0.785
33	TO – Piazza Rivoli	−0.005	0.199	0.002	0.940
34	TO – Via Gaidano	−0.015	0.267	0.004	0.903

For a fixed location s_i, let z_j and $\hat{z}_j$ be the observed and predicted time series (in our case y_{ij} and $\hat{Y}_{s_i}(t_j)$), respectively, with $j = 1, \ldots, M$ and let $\bar{z}$ and $\bar{\hat{z}}$ be the corresponding mean values. The NMBF is defined on $\mathbb{R}$ by

$$\text{NMBF} = \begin{cases} \frac{\sum_j \hat{z}_j}{\sum_j z_j} - 1 & \text{if } \bar{\hat{z}} \geq \bar{z} \\ 1 - \frac{\sum_j z_j}{\sum_j \hat{z}_j} & \text{if } \bar{\hat{z}} < \bar{z} \end{cases}$$

and has the advantage of both avoiding inflation due to low values of observations and overcoming the asymmetry problem between overestimation and underestimation, as discussed in [28]. The WNNR is defined by

$$\text{WNNR} = \frac{\sum_j s_j^2 (1 - k_j)^2}{\sum_j s_j k_j},$$

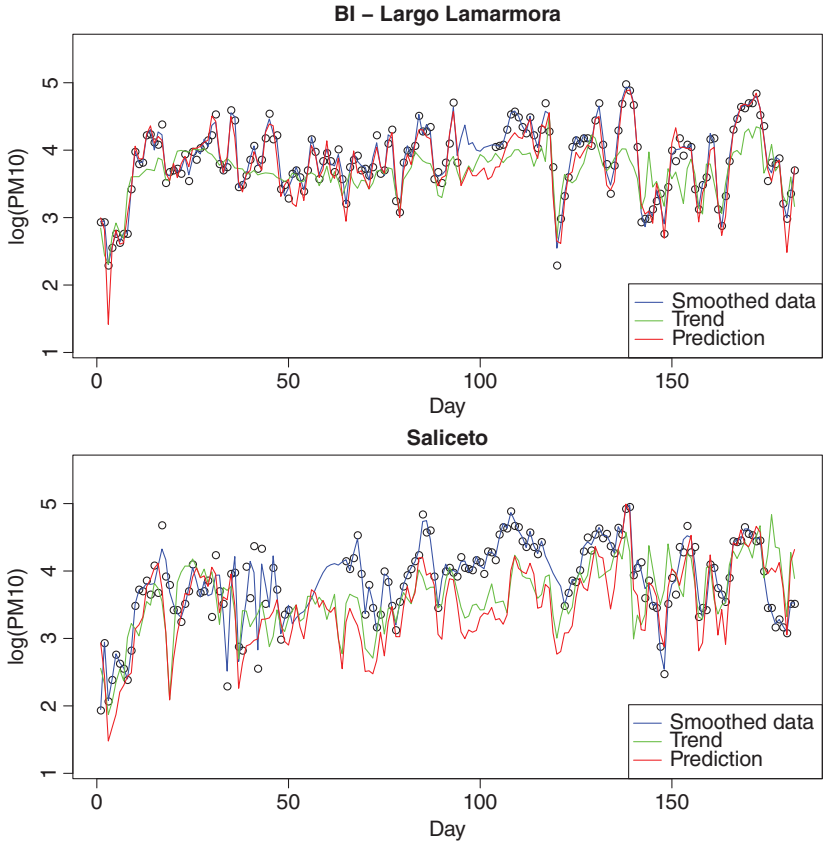


Figure 3.5 Raw data (dots), smoothed data (dashed line), predicted drift (light gray) and predicted curve (dark gray) for two validations stations.

where $s_j = z_j/\overline{z}$ is the weight and $k_j = \exp\left(-|\ln(\hat{z}_j/z_j)|\right)$ is the Normalized ratio. WNNR is positive and has the advantage of taking properly into account the peaks of observed data (see the discussion in [29]).

Results for two validation stations are shown in Figure 3.5, where the original data are represented as dots, the smoothed data as a dashed line, the predicted drift $\hat{m}_{s_i}(t)$ as a light gray, and the predicted curve (i.e. the predicted drift plus the krigged residual) as a dark gray. These two stations correspond to the best and the worst in terms of predictive performance (especially looking at ρ).

A total of 95% prediction bands can be obtained following the algorithm described in Section 3.6. We generated 500 bootstrap samples and a FKED model (including the same covariates as the model for the original data set) was fitted to each of them to obtain 500 prediction curves at each validation station. Prediction bands based on both the MBD and on L^2 distance are shown in Figure 3.6. It can be seen that overall, the two prediction bands agree well, although in some cases, the depth based band appears to be slightly wider than the distance based one. The domain coverage defined as the proportion over the domain T of the observed curve within the prediction band, varies from 97.3% to 100%.

This example illustrates the potential of using kriging for functional data under nonstationarity conditions. On one hand, it is possible to have a flexible model for the drift, identifying the effect of available covariates as shown in Figure 3.4. On the other hand, we can obtain a predicted curve at an unmonitored location with a related uncertainty band.

3.8 Conclusions

In this chapter, we have used the argument $t \in T$ to define the domain of the functional observations, notation that might lead the reader to think about time, in particular, in the example considered this argument is precisely temporal. However, let us highlight that the domain of the function does not necessarily have to be temporal. The three kriging alternatives introduced here include a trend or drift component that can be more or less complex depending on the covariate information available. As in classical geostatistics, including a complex trend in the model might fully account for the spatial structure in the data leading to uncorrelated residuals. Depending on the aim of the study, one can choose to specify a rich model for the drift part or a simpler one that leaves spatial structure in the residuals. Finally, the methods discussed in this chapter are not the only way of approaching this kind of data. Alternative approaches for dealing with spatially correlated functional data are illustrated in the following chapters.

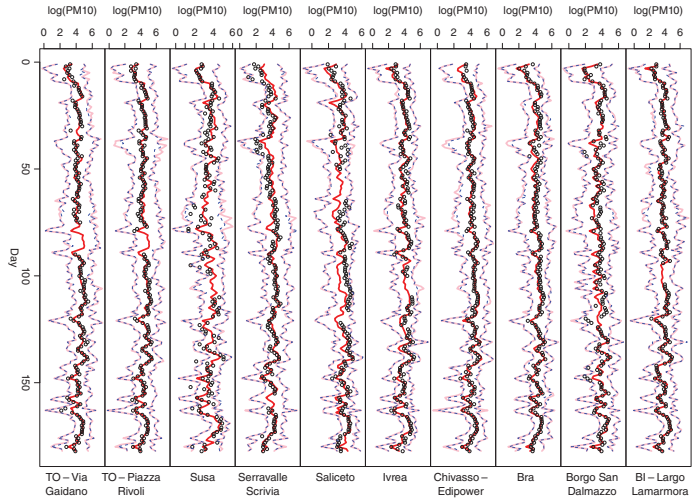


Figure 3.6 Original PM_{10} data (black dots), FKED predicted curve (dark gray line), 95%prediction band based on L^2 distance (dashed light gray) and on *MBD* (light gray) for validation stations.

References

1 Hengl, T., Heuvelink, G., and Rossiter, D. (2007). About regression-kriging: from equations to case studies. *Computers & Geosciences* 33 (10): 1301–1315.
2 Caballero, W., Giraldo, R., and Mateu, J. (2013). A universal kriging approach for spatial functional data. *Stochastic Environmental Research and Risk Assessment* 27 (7): 1553–1563.
3 Menafoglio, A., Secchi, P., and Rosa, M.D. (2013). A universal kriging predictor for spatially dependent functional data of a Hilbert space. *Electronic Journal of Statistics* 7: 2209–2240.
4 Reyes, A., Giraldo, R., and Mateu, J. (2015). Residual kriging for functional spatial prediction of salinity curves. *Communications in Statistics - Theory and Methods* 44 (4): 798–809.
5 Ignaccolo, R., Mateu, J., and Giraldo, R. (2014). Kriging with external drift for functional data for air quality monitoring. *Stochastic Environmental Research and Risk Assessment* 28: 1171–1186.
6 Cameletti, M., Ignaccolo, R., and Bande, S. (2011). Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics* 22 (8): 985–996.
7 Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics* 21: 224–239.
8 Giraldo, R., Delicado, P., and Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics* 18 (3): 411–426.
9 Giraldo, R., Delicado, P., and Mateu, J. (2010). Continuous time-varying kriging for spatial prediction of functional data: an environmental application. *Journal of Agricultural, Biological, and Environmental Statistics* 15 (1): 66–82.
10 Giraldo, R., Delicado, P., and Mateu, J. (2009). Geostatistics with Infinite Dimensional Data: A Generalization of Cokriging and Multivariable Spatial Prediction. *Tech. Rep*., Reporte Interno de Investigacion No. 14. Universidad Nacional de Colombia.
11 Nerini, D., Monestiez, P., and Mantè, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis* 101: 409–418.
12 Wackernagel, H. (1995). *Multivariable Geostatistics: An Introduction with Applications*. Springer.
13 Wood, S. (2015). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*, R package version 1.8.6. http://CRAN.R-project.org/package=mgcv (accessed 16 April 2021).
14 Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.
15 Ivanescu, A., Staicu, A., Greven, S. et al. (2015). Penalized function-on-function regression. *Computational Statistics* 30 (2): 539–568.

16 Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* 99 (467): 673–686.

17 Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society B* 73 (1): 3–36.

18 Franco-Villoria, M. and Ignaccolo, R. (2017). Bootstrap based uncertainty bands for prediction in functional kriging. *Spatial Statistics* 21: 130–148. https://doi.org/10.1016/j.spasta.2017.06.005.

19 Robinson, G. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* 6: 15–32.

20 Speed, T. (1991). [That BLUP is a good thing: the estimation of random effects]: Comment. *Statistical Science* 6 (1): 42–44.

21 Scheipl, F., Staicu, A., and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics* 24 (2): 477–501.

22 Lopez-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association* 104 (486): 718–734.

23 Cuevas, A., Febrero, M., and Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis* 51: 1063–1074.

24 Ramsay, J.O., Wickham, H., Graves, S., and Hooker, G. (2014). *FDA: Functional Data Analysis*, R package version 2.4.4. http://CRAN.R-project .org/package=fda (accessed 16 April 2021).

25 Pebesma, E. (2004). Multivariable geostatistics in S: the gstat package. *Computational Geosciences* 30: 683–691.

26 Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis* 97 (2): 109–131.

27 Finardi, S., DeMaria, R., D'Allura, A. et al. (2008). A deterministic air quality forecasting system for Torino urban area, Italy. *Environmental Modelling and Software* 23 (3): 344–355.

28 Yu, S., Eder, B., Dennis, R. et al. (2006). New unbiased symmetric metrics for evaluation of air quality models. *Atmospheric Science Letters* 7 (1): 26–34.

29 Poli, A. and Cirillo, M. (1993). On the use of the normalized mean square error in evaluating dispersion model performance. *Atmospheric Environment* 27 (15): 2427–2434.

4

Extending Functional Kriging When Data Are Multivariate Curves: Some Technical Considerations and Operational Solutions

David Nerini[1], Claude Manté[1], and Pascal Monestiez[2]

[1] *Aix-Marseille Université, CNRS/INSU, Université de Toulon, IRD, Mediterranean Institute of Oceanography UM 110, Marseille, France*
[2] *INRAE, BioSP, Avignon, France*

4.1 Introduction

Suppose that we dispose of a spatial domain $\mathcal{D}$ where two variables are sampled as curves (Figure 4.1).

On some sampling stations, both functional variables are observed. On other locations, only one or the other is available. The problem that must be addressed can be summarized by the following questions:

- Can we predict one functional variable, or the other, or both functions at unknown location on domain $\mathcal{D}$ using the sample on hand?
- How can we use the cross-information given by few stations on both curves for the prediction on a station where one variable is missing?

To illustrate our purpose, the following example will be developed as a guideline. A collection of $N = 90$ sampling locations in France (essentially airport places) is available where weather stations have recorded temperature T (°C) and precipitation levels P (mm) for 20 years since 1991. We are interested in (i) reconstituting the annual profiles of both temperature and precipitation from 12 monthly mean of these climate variables, (ii) predicting temperature and precipitation curves at an unknown location in France (Figure 4.2).

The proposed approach to answer the previous questions relies on the decomposition of the observed curves in a particular basis. Section 4.2 focuses on some nice properties of variance operators and provides a basis decomposition of the observed curves using functional principal component analysis (FPCA). After recalling some basics in functional kriging, Sections 4.3 and 4.4 give some operational solutions to estimate the spatial covariance models required when

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

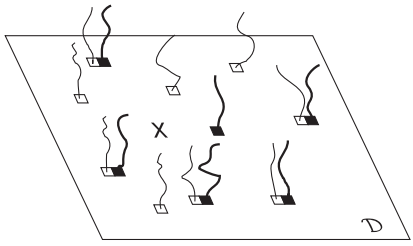


Figure 4.1 On a sampled domain $\mathcal{D}$, two functional variables are observed sometimes pairwise and sometimes alone. How both of them can be predicted on the cross-marked position?

kriging curves. An example is proposed where predictions of precipitation curves are made both using a classical approach and a principal component analysis (PCA)-based method. Following the ideas previously presented, Sections 4.5 and 4.6 are devoted to an extension of kriging in the multivariate case and give operational solutions to perform kriging of several curves simultaneously. A discussion is engaged in Section 4.7 about the capabilities and limits of kriging when dealing with multivariate functional data.

4.2 Principal Component Analysis for Curves

This section briefly reminds basics of FPCA presented as a projection method for dimension reduction. More details of this famous method can be found in [1].

4.2.1 Karhunen–Loève Decomposition

Let $Z = (Z_t, t \in \tau)$ be a continuous-time real stochastic process where τ is an interval in $\mathbb{R}$. Under some regularity properties of its sample paths, Z can be viewed as a random function belonging to a separable Hilbert space $\mathcal{H}$ endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the associated norm $\| \cdot \|_{\mathcal{H}}$. Denote by $\mu = \mathbb{E}(Z)$ the expectation of Z which is a function and define $Y = Z - \mu$, the centered version of Z. The covariance operator Γ applied to function $f \in \mathcal{H}$ is defined as

$$\Gamma(f) = \mathbb{E}\left[Y \otimes Y\right](f) = \mathbb{E}\left[\langle Y, f \rangle_{\mathcal{H}} Y\right] = \int_{\tau} \gamma(s,t) f(s) ds,$$

where γ is the associated symmetrical bivariate covariance function of Z. This linear operator is nuclear and hence Hilbert–Schmidt (HS). This means that Γ admits a decomposition of the form:

$$\Gamma(f) = \sum_{j=1}^{\infty} \lambda_j \langle f, \xi_j \rangle_{\mathcal{H}} \xi_j, \tag{4.1}$$

where (ξ_j) forms an orthonormal basis of eigenfunctions in $\mathcal{H}$ and (λ_j) is the associated ranked sequence of positive eigenvalues such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and such

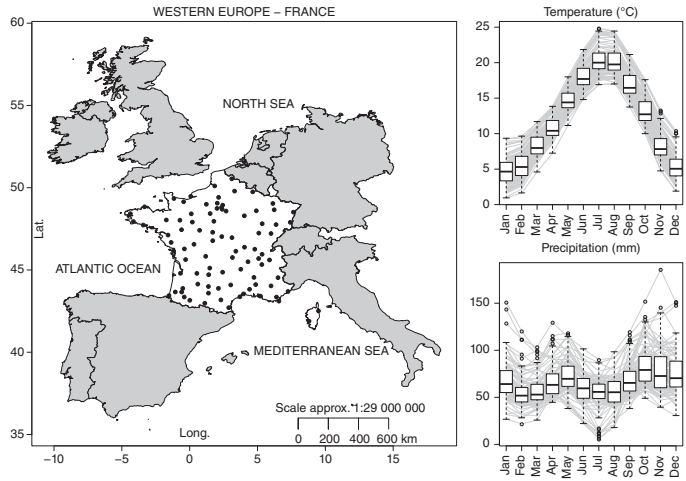


Figure 4.2 Map of France and climate dataset. On each point, annual curves (gray lines) of temperature and precipitation are available. Boxplots are observed data (1991–2010 monthly means). Source: Data from Meteo France.

that $\sum_{j\geq 1}\lambda_j = \mathbb{E}\|Y\|^2_{\mathcal{H}} < \infty$. The operator Γ also belongs to a Hilbert space, and its HS-norm is given with $\|\Gamma\|^2_{HS} = \sum_{j\geq 1}\lambda_j^2$. Equation (4.1) is the diagonal form of Γ, where the eigenelements (λ_j, ξ_j) verify

$$\Gamma(\xi_j) = \lambda_j \xi_j.$$

The associated spectral decomposition of the covariance function is given with

$$\gamma(s,t) = \sum_{j\geq 1}\lambda_j \xi_j(s)\xi_j(t), (s,t) \in \tau \times \tau.$$

Thanks to this decomposition, an approximation $\tilde{Z}$ of Z is constructed in a finite Q-dimensional space generated by $(\xi_q)_{1\leq q\leq Q}$ associated with the first Q eigenvalues ranked in descending order and $\tilde{Z}$ is expressed as a linear combination of the eigenfunctions:

$$\tilde{Z} = \mu + \sum_{q=1}^{Q} c_q \xi_q,$$

with the centered random coordinates $c_q = \langle Z - \mu, \xi_q \rangle_{\mathcal{H}}$ called principal components and which check $\mathrm{Var}(c_q) = \lambda_q$. This expansion is known as the Karhunen–Loève decomposition of Z truncated at order Q or FPCA. For sufficient regularity conditions on Z, it is expected that the sequence (λ_j) rapidly decreases to 0 which means that the value of Q can be selected small. These decomposition properties of covariance operators will be widely used through this work.

4.2.2 Dealing with a Sample

Suppose that we dispose of a i.i.d. sample $\{Z_1, \ldots, Z_N\}$ of Z, and that these functions are expressed as a linear combination of known basis functions $\{\phi_1, \ldots, \phi_L\}$ such that

$$Z_n(t) = \sum_{l=1}^{L} \alpha_{nl}\phi_L(t) = \boldsymbol{\alpha}_n' \boldsymbol{\Phi}(t),$$

where $\boldsymbol{\alpha}_n = (\alpha_{n1}, \ldots, \alpha_{nL})'$ is the vector of coefficients of the decomposition into the ϕ-basis, and $\boldsymbol{\Phi}(t) = (\phi_1(t), \ldots, \phi_L(t))'$, the vector of basis functions evaluated in $t \in \tau$. Here, the space of function $\mathcal{H}$ is of finite dimension L, and the following equality holds:

$$\|Z_n\|^2_{\mathcal{H}} = \|\boldsymbol{\alpha}_n\|^2_{\mathbf{W}} = \boldsymbol{\alpha}_n' \mathbf{W} \boldsymbol{\alpha}_n,$$

where

$$\mathbf{W} = \int_{\tau} \boldsymbol{\Phi}(t)\boldsymbol{\Phi}'(t)dt$$

is the symmetrical Gram matrix of the basis functions with entries being the inner product of the basis functions $\langle \phi_k, \phi_l \rangle_{\mathcal{H}}$. Once the basis has been chosen, it is easy to resort to numerical integration to compute this matrix. An estimator $\widehat{\mu}$ of the expectation μ is chosen as

$$\widehat{\mu}(t) = \frac{1}{N}\sum_{n=1}^{N} Z_n(t) = \overline{\boldsymbol{\alpha}}'\boldsymbol{\Phi}(t),$$

where $\overline{\boldsymbol{\alpha}}$ is the empirical mean vector of the coefficients. A classical empirical estimator of the covariance function γ is then provided with

$$\widehat{\gamma}(s,t) = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\Phi}'(s)(\boldsymbol{\alpha}_n - \overline{\boldsymbol{\alpha}})(\boldsymbol{\alpha}_n - \overline{\boldsymbol{\alpha}})'\boldsymbol{\Phi}(t) = \boldsymbol{\Phi}'(s)\widehat{\boldsymbol{\Gamma}}\boldsymbol{\Phi}(t).$$

The $L \times L$ matrix

$$\widehat{\boldsymbol{\Gamma}} = \frac{1}{N}\mathbf{C}'\mathbf{C} \tag{4.2}$$

is the empirical matrix of variance between coefficients of the decomposition, where $\mathbf{C}$ if the $N \times L$ matrix of centered coefficients. Estimation of the eigenfunctions and associated eigenvalues is computed by solving

$$\int_{\tau} \widehat{\gamma}(s,t)\widehat{\xi}_l(s)ds = \widehat{\lambda}_l\widehat{\xi}_l(t),$$

where $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_L \geq 0$ are positive eigenvalues and $\widehat{\xi}_l$ are orthonormal eigenfunctions. This equation can be written in the ϕ-basis, and solutions are obtained by performing the eigen-decomposition of matrix $\widehat{\boldsymbol{\Gamma}}\mathbf{W}$ such that

$$\widehat{\boldsymbol{\Gamma}}\mathbf{W}\mathbf{b}_l = \widehat{\lambda}_l\mathbf{b}_l, \quad l = 1, \ldots, L,$$

where the $\mathbf{b}_l$ are eigenvectors of the decomposition. The eigenfunctions can now be reconstituted with

$$\widehat{\xi}_l(t) = \boldsymbol{\Phi}'(t)\mathbf{a}_l,$$

where $\mathbf{a}_l = \mathbf{W}^{-\frac{1}{2}}\mathbf{b}_l$ are $\mathbf{W}$-orthonormal eigenvectors. Notice that $\mathbf{W}$ is a symmetric positive definite matrix so it has a unique Cholesky decomposition:

$$\mathbf{W} = \mathbf{W}^{\frac{1}{2}\prime}\mathbf{W}^{\frac{1}{2}},$$

where $\mathbf{W}^{\frac{1}{2}}$ is an invertible upper triangular matrix.

Finally, the coefficients of the curve Z_n may also be approximated in a Q-dimensional space ($Q < L$) as a linear combination of the eigenvectors with

$$\tilde{\boldsymbol{\alpha}}_n = \overline{\boldsymbol{\alpha}} + \mathbf{A}_Q\mathbf{c}_n,$$

where $\mathbf{A}_Q = [\mathbf{a}_1, \ldots, \mathbf{a}_Q]$ is a $L \times Q$ matrix of normalized eigenvectors and

$$\mathbf{c}_n = \mathbf{A}_Q'\mathbf{W}(\boldsymbol{\alpha}_n - \overline{\boldsymbol{\alpha}})$$

is the vector of the Q first principal coordinates of Z_n. If $Q = L$, it is easy to check that the norm of the centered variable $Y_n = Z_n - \widehat{\mu}$ is equal to

$$\|Y_n\|_{\mathcal{H}} = \|\mathbf{A}_L \mathbf{c}_n\|_{\mathbf{W}} = \|\mathbf{c}_n\|_L,$$

where $\|\cdot\|_L$ is the usual norm in $\mathbb{R}^L$. If the sequence of eigenfunctions (λ_j) rapidly decreases to zero, it is expected that a few number $Q < L$ of principal components is sufficient to get a good approximation of Y_n. FPCA is then a nice way to reduce the dimension of the system.

4.3 Functional Kriging in a Nutshell

Suppose now that there is spatial dependency between functional observations $\{Z_1, \ldots, Z_n\}$. From a more formal point of view, $Z = (Z_x, x \in \mathcal{D})$ can be seen as a $\mathcal{H}$-valued random spatial process where weak stationarity conditions are added over the domain $\mathcal{D}$. First, we will consider that the expectation of Z is the same on any position $x \in \mathcal{D}$:

$$\mathbb{E}(Z_x) = \mu, \ \mu \in \mathcal{H}.$$

The spatial covariance operator of Z defined as:

$$\Gamma_{x,y} = \mathbb{E}[(Z_x - \mu) \otimes (Z_y - \mu)],$$

is also supposed to satisfy

$$\Gamma_{x+h,y+h} = \Gamma_{x,y}, \quad x, y, h \in \mathcal{D}.$$

The spatial covariance of Z between position x and position y is invariant for any translation of the pair (Z_x, Z_y) into the spatial domain. Notice that a spatial covariance operator is also HS but is nonsymmetrical i.e. $\Gamma_{x,y} \neq \Gamma_{y,x}$. However, one may associate its adjoint $\Gamma^{\star}_{x,y}$ defined by

$$\langle \Gamma^{\star}_{x,y}(f), g \rangle_{\mathcal{H}} = \langle f, \Gamma_{x,y}(g) \rangle_{\mathcal{H}}, \quad f, g \in \mathcal{H},$$

that always verifies that $\Gamma^{\star}_{x,y} = \Gamma_{y,x}$.

Consider now that the sample $\{Z_1, \ldots, Z_N\}$ comes from observations of Z on N spatial positions $\{x_1, \ldots, x_N\}$. The problem of kriging functional data consists in estimating the curve Z_0 in an unknown position $x_0 \in \mathcal{D}$ using a linear model of the form

$$\widehat{Z}_0 = \sum_{n=1}^{N} B_n(Z_n), \tag{4.3}$$

where $B_n : \mathcal{H} \to \mathcal{H}$ are HS linear operators such that

$$B_n(f)(t) = \int_{\tau} \beta_n(s,t) f(s) ds, \quad f \in \mathcal{H},$$

with β being a bivariate function that forms the kriging weights. A sufficient condition for $\widehat{Z}_0$ to be unbiased is given with

$$\sum_n B_n = K.$$

The linear operator K is such that

$$K(f)(t) = \int_\tau \kappa(s,t)f(s)ds = f(t), \quad f \in \mathcal{H},\ t \in \tau,$$

where the function κ is the bivariate kernel function of the identity operator in $\mathcal{H}$ called reproducing kernel. The operator K plays the role of an identity operator but with the additional property that $\|K\|^2_{HS} < \infty$ which is not the case for the classical identity in infinite dimension. This technical point matters because the existence of an unbiased functional kriging estimator relies on that property. In practice and under some mild regularity conditions, it is easy to find a useful function space $\mathcal{H}$ where K exists, this will be the topics of Section 4.3.1 (see [2, 3] for more details on reproducing kernels).

Our objective is now to find the operators B_n that minimize

$$\mathbb{E}\|\widehat{Z}_0 - Z_0\|^2_{\mathcal{H}}$$

under the constraint $\sum_n B_n = K$.

Using Lagrange multiplier method and properties of HS spaces, it has been shown in [4] that the best linear unbiased predictor of Z_0 is solution of the following linear system:

$$\begin{cases} \sum_m B_m \Gamma_{nm} + \Lambda &= \Gamma_{n0}, \quad n = 1, \dots, N \\ \sum_n B_n &= K \end{cases} \tag{4.4}$$

with Γ_{nm} being the covariance operator between Z_n and Z_m, Λ a Lagrange multiplier. Section 4.3.1 gives technical solutions to find kriging weights B_n and to compute the functional kriging variance.

4.3.1 Solution Based on Basis Functions

As previously presented, each Z_n is expanded in a ϕ-basis of L known functions such that

$$Z_n(t) = \boldsymbol{\alpha}_n' \boldsymbol{\Phi}(t).$$

The functional space $\mathcal{H}$ being of finite dimension L, it, therefore, admits a reproducing kernel of the form:

$$\kappa(s,t) = \boldsymbol{\Phi}'(s)\mathbf{W}^{-1}\boldsymbol{\Phi}(t), \quad (s,t) \in \tau \times \tau.$$

The matrix $\mathbf{W}^{-1}$ is the inverse of the Gram matrix and constitutes the matrix form of the operator K. This suggests that the basis is nondegenerated and that the Gram

matrix $\mathbf{W}$ may be inverted which is not systematic from a numerical point-of-view. If we turn back to the kriging Eq. (4.3), it may be easily expressed in the space of the coefficients with

$$\widehat{\boldsymbol{\alpha}}_0 = \sum_{n=1}^{N} \mathbf{B}_n' \mathbf{W} \boldsymbol{\alpha}_n, \tag{4.5}$$

where the $L \times L$ matrix $\mathbf{B}_n$ is the discrete version of the operator B_n containing kriging coefficients. These coefficients are estimated when minimizing

$$\mathbb{E} \left\| \widehat{\boldsymbol{\alpha}_0} - \boldsymbol{\alpha}_0 \right\|_{\mathbf{W}}^2$$

under the constraint $\sum_n \mathbf{B}_n = \mathbf{W}^{-1}$.

Following the matrix formalism proposed by Myers [5] in the multivariate case, Nerini et al. [4] and Giraldo [6] have shown that kriging curves using a basis expansion boils down to solve the following system of $L(N+1)$ equations:

$$\begin{bmatrix} \mathbf{W}\boldsymbol{\Gamma}_{11}\mathbf{W} & \cdots & \mathbf{W}\boldsymbol{\Gamma}_{1N}\mathbf{W} & \mathbf{I}_L \\ & \ddots & & \vdots \\ \mathbf{W}\boldsymbol{\Gamma}_{N1}\mathbf{W} & \cdots & \mathbf{W}\boldsymbol{\Gamma}_{NN}\mathbf{W} & \mathbf{I}_L \\ \mathbf{I}_L & \cdots & \mathbf{I}_L & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_N \\ \boldsymbol{\Lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{W}\boldsymbol{\Gamma}_{10} \\ \vdots \\ \mathbf{W}\boldsymbol{\Gamma}_{N0} \\ \mathbf{W}^{-1} \end{bmatrix}, \tag{4.6}$$

where the matrices

$$\boldsymbol{\Gamma}_{nm} = \mathbb{E}\left[\left(\boldsymbol{\alpha}_n - \boldsymbol{\mu} \right) \left(\boldsymbol{\alpha}_m - \boldsymbol{\mu} \right)' \right]$$

are the $L \times L$ matrices of spatial covariances between coefficients with

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{\alpha}_n), \quad \forall x_n \in \mathcal{D},$$

the mean vector, $\mathbf{I}_L$ the identity matrix in $\mathbb{R}^L$ and $\boldsymbol{\Lambda}$ a $L \times L$ matrix of Lagrange multipliers. The associated functional kriging variance (see Appendix 4.A.1 for computational details) is given with

$$\sigma_{FKG}^2 = \mathrm{Tr}\left(\mathbf{W}\boldsymbol{\Gamma}_{00} \right) - \sum_n \mathrm{Tr}\left(\mathbf{B}_n \mathbf{W}\boldsymbol{\Gamma}_{n0}\mathbf{W} \right) - \mathrm{Tr}\left(\boldsymbol{\Lambda} \right).$$

Once the kriging coefficients $\mathbf{B}_1, \ldots, \mathbf{B}_N$ have been found, the function Z_0 is estimated with

$$\widehat{Z}_0(t) = \widehat{\boldsymbol{\alpha}}_0' \boldsymbol{\Phi}(t) = \sum_{n=1}^{N} \boldsymbol{\alpha}_n' \mathbf{W} \mathbf{B}_n \boldsymbol{\Phi}(t).$$

Using a basis decomposition, the problem of kriging curves is solved in finite dimension, in the isotopic case (all coefficients available in all stations). If the chosen basis in orthonormal is ($\mathbf{W} = \mathbf{I}_L$), it reduces to a straightforward cokriging using the coefficients of the decomposition as raw data.

4.3.2 Estimation of Spatial Covariances

The challenge is now to find admissible estimators of $\mathbf{\Gamma}_{nm}$ in (4.6). And the problem is that we dispose of a unique realization of $\{Z_1, \ldots, Z_n\}$: the classical empirical estimator (see Eq. (4.2)) of covariance cannot be used in that case. Some additional properties over the spatial functional process must be considered. We will assume that the sample of functions, reduced to their coefficients $\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N\}$, comes from a L-multivariate random field $\boldsymbol{\alpha}(x),\ x \in \mathcal{D}$. In addition to second-order stationarity conditions:

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{\alpha}(x)\right] = \boldsymbol{\mu}, \quad \forall x \in \mathcal{D},\\
&\mathbf{\Gamma}_{x,y} = \mathbf{\Gamma}_{x+h,y+h}, \quad x, y, h \in \mathcal{D},
\end{aligned}$$

where the spatial covariance

$$\mathbf{\Gamma}(h) = \mathbb{E}\left[(\boldsymbol{\alpha}(x+h) - \boldsymbol{\mu})(\boldsymbol{\alpha}(x+h) - \boldsymbol{\mu})'\right]$$

only depends on lag h, another hypothesis is required. The spatial variation of Z (of its coefficients) can be handled by defining the spatial increment $(Z_{x+h} - Z_x)$ as the difference between values of Z at locations $x \in \mathcal{D}$ and $x + h \in \mathcal{D}$ separated by a lag vector h on the increments. We will suppose that the average value of the increments is the same over the whole domain:

$$\mathbb{E}\left[\boldsymbol{\alpha}(x) - \boldsymbol{\alpha}(x+h)\right] = \mathbf{0}, \quad \forall x, x+h \in \mathcal{D},$$

and that the variance of the increments possesses a finite value $2\mathbf{G}(h)$ that depends on the length $|h|$ and the orientation of lag vector h, but not on the position of h in $\mathcal{D}$. Under these hypotheses, the random field is said to be intrinsically stationary because it can be expressed through a multivariate variogram $L \times L$ matrix $\mathbf{G}(h)$ related to the variance of the increments with

$$\mathbf{G}(h) = \frac{1}{2}\mathbb{E}\left[(\boldsymbol{\alpha}(x+h) - \boldsymbol{\alpha}(x))\,(\boldsymbol{\alpha}(x+h) - \boldsymbol{\alpha}(x))'\right],$$

that only depends on the separation vector h. Moreover, if the function $\mathbf{G}(h)$ depends upon the separation vector only through its length $|h|$, then the process is isotropic. Covariance and variogram matrices are thus connected by the relation:

$$\mathbf{G}(h) = \mathbf{\Gamma}(0) - \frac{1}{2}\left[\mathbf{\Gamma}(h) + \mathbf{\Gamma}(-h)\right].$$

As usual in geostatistics, covariance matrices $\mathbf{\Gamma}_{nm}$ between coefficients of the decomposition will be estimated through the fitting of variogram models to experimental ones. Fitting is realized using a linear model of coregionalization (LMC) that states that the matrix $\mathbf{\Gamma}(h)$ of cross-covariances can be modeled using a combination of a small number S of correlation functions $\rho_s(h)$ such that

$$\widehat{\mathbf{\Gamma}}(h) = \sum_{s=1}^{S} \widehat{\mathbf{P}}_s \rho_s(h),$$

where $\widehat{\mathbf{P}}_s$ are estimated positive semidefinite coregionalization matrices. In practice, the estimation of the matrices $\mathbf{P}_s$ is carried out through weighted least squares fitting of variogram models to experimental data (see [7] chapter 26, Goulard and Voltz [8]). Along that study, three nested covariance structures have been used. At short scale (50 km) and large scale (700 km), correlation structures are estimated through the fit of a Gaussian model:

$$g_1(|h|) = 1 - \exp\left(\frac{-|h|^2}{r}\right), \quad r > 0,$$

over the experimental variogram and cross-variograms between coefficients. At medium scale (320 km), a spherical model:

$$g_2(|h|) = \begin{cases} \frac{3|h|}{2r} - \frac{1}{2}\left(\frac{|h|}{r}\right)^3 & \text{for } |h| \le r \\ 1 & \text{for } |h| \ge r \end{cases}$$

has been fitted. Variogram models and number of covariance structure have been chosen as discussed in [8]. Spatial ranges have been fixed by the practitioner so as to provide the most graphically satisfactory fit.

Once the correlation function has been estimated, it can be used to solve the kriging system (4.6) by replacing the covariance matrices with

$$\widehat{\mathbf{\Gamma}}_{nm} = \sum_{s=1}^{S} \widehat{\mathbf{P}}_s \rho_s(x_n - x_m).$$

4.4 An Example with the Precipitation Observations

Functional data do not arrive as entire curves. In the weather data example, considering the annual precipitation curves only, coefficients of the decomposition must be estimated using 12 monthly average values of precipitation. We consider here a B-spline expansion with $L = 9$ coefficients that have been estimated by regression (see [1] for technical details). The number of basis function has been arbitrarily fixed for the example, but other choices can be relevant as well, still dependent on the number of available raw data. For the sake of simplicity, we will note

$$z_n^P(t) = \sum_{l=1}^{L} \alpha_{nl}^P \phi_l(t) = \mathbf{\Phi}'(t)\boldsymbol{\alpha}_n^P, \quad n = 1, \dots, N,$$

the sample of estimated precipitation curves, where the $\boldsymbol{\alpha}_n^P$ are now estimated coefficients instead of random vectors.

4.4.1 Fitting Variogram Model

The $L \times L$ symmetrical variogram matrix:

$$\widehat{\mathbf{G}}(h) = \begin{bmatrix} \widehat{g}_{11}(h) & \dots & \widehat{g}_{1L}(h) \\ \vdots & \ddots & \vdots \\ \widehat{g}_{L1}(h) & \dots & \widehat{g}_{LL}(h) \end{bmatrix}$$

is composed of simple variograms (diagonal elements) and cross-variograms (off-diagonal elements) that have been estimated on experimental variograms using the nested model with three structures (Figure 4.3). Experimental variograms are computed using B-spline coefficients as spatial data (see [7] p. 47 for computational details).

4.4.2 Making Prediction

Predictions are made by leave-one-out cross-validation in each of the N towns by considering the following kriging model:

$$\widehat{z}^P_{(-n)} = \sum_{m \neq n} B_m(z^P_m), \quad n = 1, \dots, N.$$

For curve n, experimental variograms are constructed using data of the $N-1$ remaining locations. For any curve, same nested models of coregionalization are fitted on experimental variograms. The prediction error is estimated with the integrated squared error (ISE):

$$\text{ISE}_n = \|\widehat{z}^P_{(-n)} - z^P_n\|^2_{\mathcal{H}} = \|\widehat{\boldsymbol{\alpha}}_{(-n)} - \boldsymbol{\alpha}_n\|^2_{\mathbf{W}},$$

and the whole error is estimated with the mean ISE:

$$\text{MISE} = \frac{1}{N} \sum_n \text{ISE}_n.$$

Figure 4.4 displays some predicted curves and spatial location of the ISEs. It is interesting to notice that worst predictions are mainly located in southwestern area which is known to get some particular precipitation patterns. Bad predictions also appear in mountainous areas (center and east). In that case, the stationarity assumption does not hold anymore. More generally, bad predictions could appear in case not only where (i) data are far from stationarity conditions but also (ii) when a great number of coefficients is required for a good fitting of observations. We will turn back to this issues in Section 4.7.

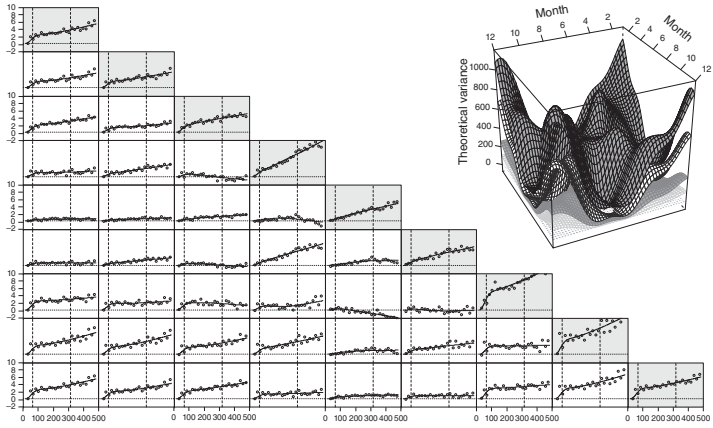


Figure 4.3 An example of computation of the spatial covariance. Experimental variograms (points in gray boxes) and cross-variograms (points in white boxes) are computed using the nine coefficients of the B-spline decomposition (precipitation curves). A theoretical nested variogram model with three structures (50, 320, and 700 km) is fitted by weighted least squares (curves in each panel, horizontal scales in km, vertical dashed lines show the three nested scales of the LMC). Surfaces on the top right display spatial covariance operators $\Gamma(h)$ deduced from variogram model for various distances (from bottom zero surface to top covariance map, $|h| = \infty, 1000, 750, 300, 0$ km).

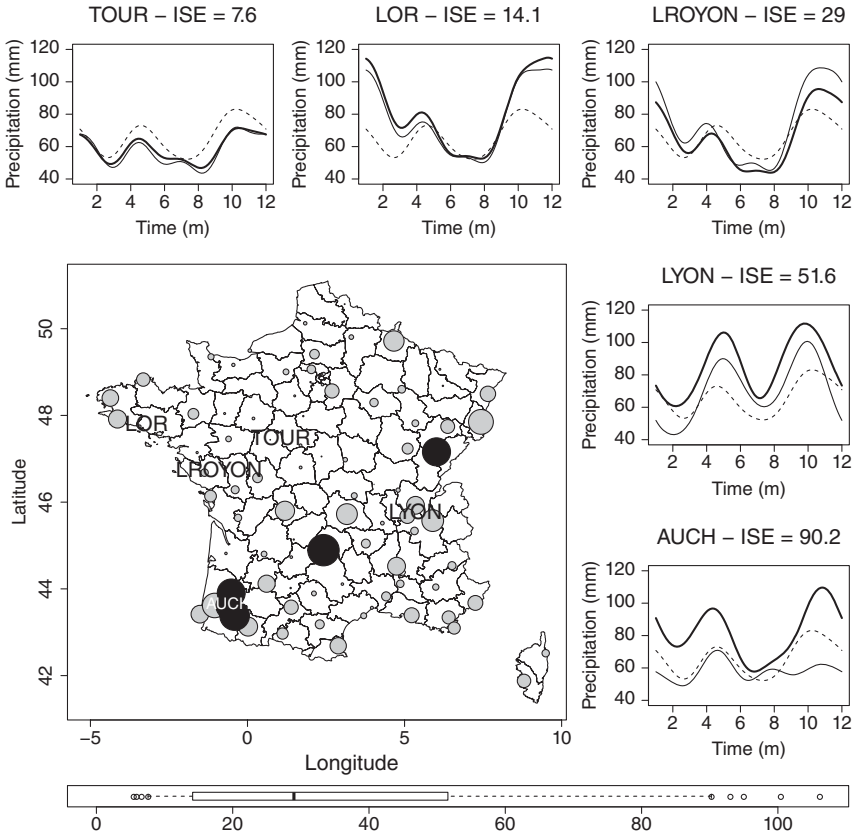


Figure 4.4 Predictions of precipitation curves (dashed line: mean precipitation, thin line: observed curve, thick line: predicted curve), spatial map, and boxplot of the integrated squared errors (ISEs). From top-left to bottom-right panel, curves are, respectively, sorted by ISE quantiles of order 0.05, 0.25, 0.5, 0.75, and 0.95. The ISE median town is the upper-right panel. Bad predictions (black points corresponding to ISE quantiles of order ≥ 0.95) are located in areas where the stationarity hypothesis does not hold. Notice that curve shape is generally well predicted, whereas mean precipitation level is not.

4.5 Functional Principal Component Kriging

One of the main problems with kriging curves arises from the fitting of an unreasonable number of variogram models to the experimental data. An efficient solution to decrease this number of fitted variograms is to take advantage of the properties of the spectral decomposition of covariance operators. Consider the sample of curves $\{Z_1, \ldots, Z_N\}$ expressed in a finite ϕ-basis of dimension L.

Suppose that each function is well reconstituted as a linear combination of $Q < L$ eigenfunctions such that

$$\tilde{Z}_n(t) = \mathbf{\Phi}'(t)\left[\overline{\boldsymbol{\alpha}} + \mathbf{A}_Q \mathbf{c}_n\right],$$

where $\mathbf{A}_Q$ is the matrix of eigenfunction coefficients of size $L \times Q$, $\overline{\boldsymbol{\alpha}}$ is the empirical mean of the curve sample, and $\mathbf{c}_n$ the Q-vector of the principal coordinates associated with an observation Z_n. Reminding that $\mathbf{A}_Q'\mathbf{W}\mathbf{A}_Q = \mathbf{I}_Q$ by properties of orthogonality of eigenvectors and that $\sum_n \mathbf{B}_n = \mathbf{W}^{-1}$, it is easy to show that the kriging estimator becomes

$$\widehat{\mathbf{c}}_0 = \sum_{n=1}^{N} \mathbf{A}_Q'\mathbf{W}\mathbf{B}_n'\mathbf{W}\mathbf{A}_Q\mathbf{c}_n, \tag{4.7}$$

where $\widehat{\mathbf{c}}_0$ is the principal coordinates of the observation Z_0 predicted in position $x_0 \in \mathcal{D}$. If we set

$$\mathbf{U}_n = \mathbf{A}_Q'\mathbf{W}\mathbf{B}_n'\mathbf{W}\mathbf{A}_Q,$$

the kriging solutions are given when minimizing

$$\mathbb{E}\left\|\mathbf{c}_0 - \sum_{n=1}^{N}\mathbf{U}_n\mathbf{c}_n\right\|_L^2$$

under the unbiased condition constraint $\sum_n \mathbf{U}_n = \mathbf{I}_Q$.

Principal component kriging boils down to solve the following system of $Q(N+1)$ equations:

$$\begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1N} & \mathbf{I}_Q \\ & \ddots & & \vdots \\ \Sigma_{N1} & \cdots & \Sigma_{NN} & \mathbf{I}_Q \\ \mathbf{I}_Q & \cdots & \mathbf{I}_Q & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}_1 \\ \vdots \\ \mathbf{U}_N \\ \mathbf{\Lambda} \end{bmatrix} = \begin{bmatrix} \Sigma_{10} \\ \vdots \\ \Sigma_{N0} \\ \mathbf{I}_Q \end{bmatrix}, \tag{4.8}$$

where the matrices

$$\Sigma_{nm} = \mathbb{E}\left[\mathbf{c}_n\mathbf{c}_m'\right]$$

are the $Q \times Q$ matrices of spatial covariances between principal coordinates, $\mathbf{I}_Q$ the identity matrix in $\mathbb{R}^Q$, and $\mathbf{\Lambda}$ a $Q \times Q$ matrix of Lagrange multipliers. Once the kriging coefficients $\mathbf{U}_n$ have been estimated by LMC, the predicted function $\widehat{Z}_0$ can be computed with

$$\widehat{Z}_0(t) = \mathbf{\Phi}'(t)\left[\overline{\boldsymbol{\alpha}} + \sum_n \mathbf{U}_n\mathbf{c}_n\right].$$

Note that the variance of this kriging estimator is given with

$$\sigma_Q^2 = \mathrm{Tr}(\Sigma_{00}) - \sum_n \mathrm{Tr}(\Sigma_{0n}\mathbf{U}_n) - \mathrm{Tr}(\mathbf{\Lambda}).$$

Consider the case where $Q = L$. If one remembers that $\mathbf{A}_L\mathbf{A}_L'\mathbf{W} = \mathbf{I}_L$, then it is easy to verify that

$$\sigma_L^2 = \sigma_{FKG}^2,$$

since $\mathrm{Tr}(\Sigma_{mn}) = \mathrm{Tr}(\mathbf{A}_L'\mathbf{W}\boldsymbol{\Gamma}_{mn}\mathbf{W}\mathbf{A}_L)$. We are in the simplest case where the solution is given by a standard cokriging using all the principal components as variables. Results are the same working with B-spline coefficients except that we gain orthogonality when working with the PCs. Anyway, it is often interesting to consider the case where $Q < L$ i.e. where the sample of curves is approximated in a subspace of Q eigenfunctions.

Figure 4.5 displays the LMC fitting over the first four PCs of the precipitation curves accounting for more than 95% of the entire variability. It is remarkable to note the decrease in magnitude of each variogram (gray boxes) as the number of PCs increases. By properties of the PCA decomposition, the Q principal components are orthogonal at lag $|h| = 0$, i.e. the matrix of empirical variance–covariance of the PCs:

$$\hat{\Sigma}_{00} = \frac{1}{N}\sum_n \mathbf{c}_n\mathbf{c}_n',$$

is diagonal, i.e. the covariance between PCs is zero. The best required configuration for principal component kriging is that this orthogonality property extends

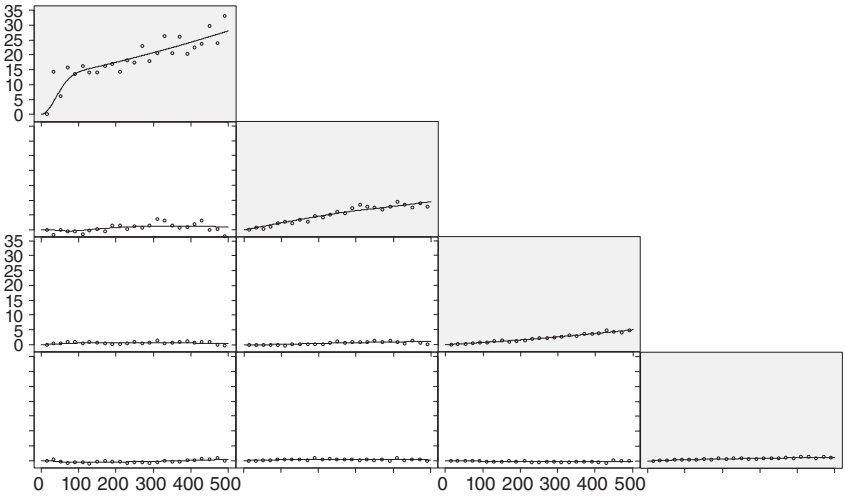


Figure 4.5 LMC fitting of the variogram model on the four PCs of precipitation FPCA. Same models of spatial covariances were used as before. Notice that the cross-variograms (curves in white panels) are very close to zero as if the data were spatially independent at any scale.

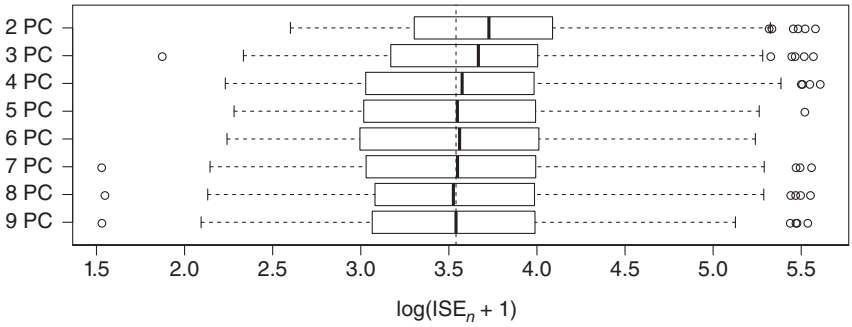


Figure 4.6 Boxplot of the errors when decreasing the number of principal components required for kriging. Taking $Q \geq 4$ principal components gives similar results. Vertical dashed line is the median log-error for nine PCs.

for all other values of lag vector h, i.e. that the spatial covariance matrices $\widehat{\boldsymbol{\Sigma}}(h)$ are also diagonal. This property is not guaranteed and is satisfied if the variables are intrinsically correlated [7, 9]. In practice, a look at the cross-variograms makes it possible to evaluate this property as shown on off-diagonal panels in Figure 4.5. In our case, kriging the PCs independently would give roughly the same results as cross-variograms are very close to zero. And only Q variogram fitting would be needed instead of $\frac{Q(Q+1)}{2}$. This raises an interesting question about the gain in prediction accuracy when increasing the number Q of PCs for principal component kriging.

Figure 4.6 shows the influence of the number of principal components on the prediction error. For the precipitation data, errors start increasing when $Q \leq 4$ which corresponds to the number of eigenvalues with a significant value. But there is no general rule for choosing this number Q which is case-dependent.

4.6 Multivariate Kriging with Functional Data

Now, the problem of kriging can be stated in a multivariate setting. Let the variable $\boldsymbol{Z}_x = ([Z_x^T, Z_x^P]',\ x \in \mathcal{D})$ be a bivariate $\mathcal{H}$-valued spatial random process over a domain $\mathcal{D}$. Denote by $\mathbb{E}(Z_x^T) = \mu_T$ the expectation of Z_x^T and $\mathbb{E}(Z_x^P) = \mu_P$ the expectation of Z_x^P. Under weak stationarity conditions, these expectations are invariant on the whole spatial domain:

$$\mathbb{E}(\mathbf{Z}_x) = \boldsymbol{\mu} = \left[\mu_T, \mu_P\right]', \quad \forall x \in \mathcal{D}.$$

With $\mathbf{Y}_x = \mathbf{Z}_x - \boldsymbol{\mu}$, the 2×2 matrix of spatial covariance operators between $\mathbf{Z}_m$ and $\mathbf{Z}_n$ is defined as

$$\boldsymbol{\Gamma}_{mn} = \mathbb{E}\left[\mathbf{Y}_m \otimes \mathbf{Y}_n\right] = \mathbb{E}\begin{bmatrix} Y_m^T \otimes Y_n^T & Y_m^T \otimes Y_n^P \\ Y_m^P \otimes Y_n^T & Y_m^P \otimes Y_n^P \end{bmatrix} = \begin{bmatrix} \Gamma_{TT}^{mn} & \Gamma_{TP}^{mn} \\ \Gamma_{PT}^{mn} & \Gamma_{PP}^{mn} \end{bmatrix}$$

and only depends on the distance between sampling station. An example of such operator is presented in Figure 4.7.

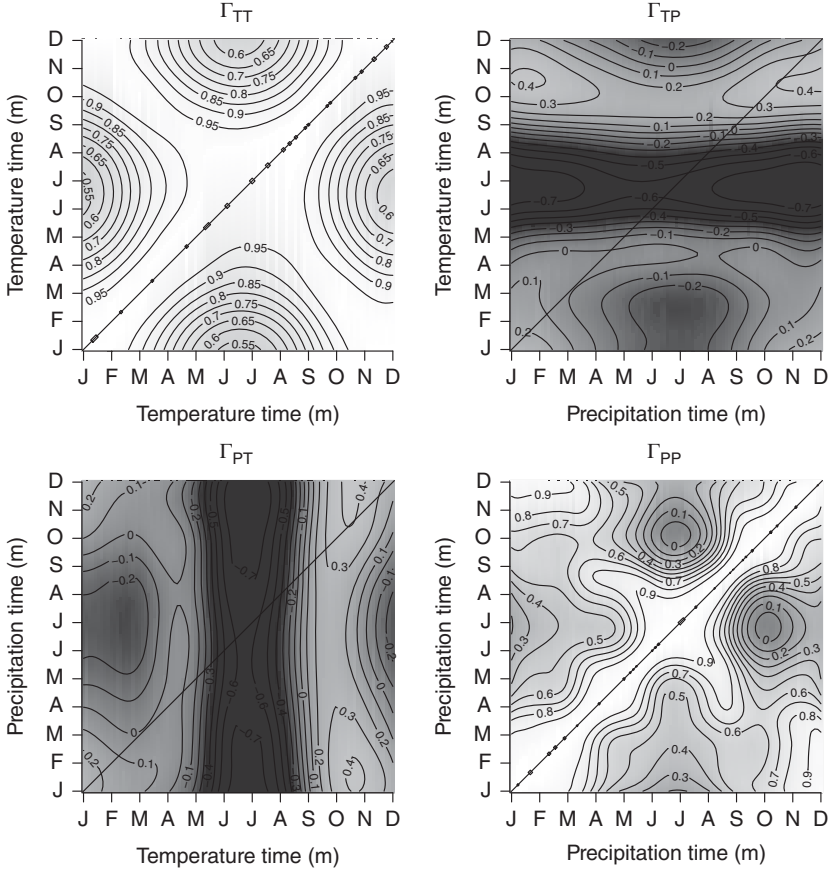


Figure 4.7 Empirical estimate of a 2×2 matrix of correlation operators (normalized version of $\boldsymbol{\Gamma}_{00}$) computed on the entire set of the coefficients of a B-spline expansion of temperature and precipitation curves. In that case, the operator is symmetrical. At each point on these mappings, it is possible to read the correlation or the cross-correlation between both variables for different couples $(t, s) \in \tau \times \tau$. Gray scale from black (minimum negative correlation) to white (maximum positive correlation).

Choose a given spatial position x_0, where no data is available and set

$$\mathbf{Z}_0 = \left[Z_0^T, Z_0^P\right]'.$$

Our aim is to estimate $\mathbf{Z}_0$ given the knowledge of a sample $\left\{\mathbf{Z}_1, \ldots, \mathbf{Z}_N\right\}$ of $\mathbf{Z}_x$. The multivariate functional kriging linear estimator can be written in the following form:

$$\widehat{\mathbf{Z}}_0 = \sum_{n=1}^{N} \mathbf{B}_n(\mathbf{Z}_n), \tag{4.9}$$

where $\mathbf{B}_n$ is a 2×2 matrix of HS linear operators such that

$$\mathbf{B}_n = \begin{bmatrix} B_{TT}^n & B_{TP}^n \\ B_{PT}^n & B_{PP}^n \end{bmatrix}.$$

In nonmatrix form, Eq. (4.9) would be

$$\begin{cases} \widehat{Z}_0^T = \sum_{n=1}^{N} B_{TT}^n(Z_n^T) + \sum_{n=1}^{N} B_{TP}^n(Z_n^P), \\ \widehat{Z}_0^P = \sum_{n=1}^{N} B_{PT}^n(Z_n^T) + \sum_{n=1}^{N} B_{PP}^n(Z_n^P). \end{cases}$$

Kriging each variable separately would correspond to set the operator B_{PT}^n and B_{TP}^n to zero. Looking for an unbiased estimator $\widehat{\mathbf{Z}}_0$ of $\mathbf{Z}_0$ leads to the condition:

$$\sum_n \mathbf{B}_n = \begin{bmatrix} K & 0 \\ 0 & K \end{bmatrix} = \mathbf{K},$$

where $\mathbf{K}$ is a diagonal block matrix of reproducing kernels. The diagonal elements are the same as in the univariate case because Z^T and Z^P are supposed to belong to the same Hilbert space. Now, looking for a BLUP estimator of $\widehat{\mathbf{Z}}_0$ leads to minimize

$$\mathbb{E}\left\|\widehat{\mathbf{Z}}_0 - \mathbf{Z}_0\right\|^2 = \mathbb{E}\left\|\widehat{Z}_0^T - Z_0^T\right\|_{\mathcal{H}}^2 + \mathbb{E}\left\|\widehat{Z}_0^P - Z_0^P\right\|_{\mathcal{H}}^2$$

under the constraint

$$\sum_n \mathbf{B}_n = \mathbf{K}.$$

Following Myers [5] and Nerini et al. [4], it is not difficult to show that the solution is the same as in the univariate case:

$$\begin{cases} \sum_{m=1}^{N} \mathbf{\Gamma}_{nm}\mathbf{B}_m + \mathbf{\Lambda} = \mathbf{\Gamma}_{n0}, \quad n = 1, \ldots, N \\ \sum_{m=1}^{N} \mathbf{B}_m = \mathbf{K} \end{cases}$$

except that the terms above are matrix of operators. A solution of that linear system can be found using FPCA in a multivariate setting.

4.6.1 Multivariate FPCA

If we go back to what we previously did in the univariate case, the observed sample $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_N\}$ is composed with vectors $\mathbf{Z}_n = (Z_n^T, Z_n^P)'$ whose entries are functions, linear combination of L known basis functions such that

$$Z_n^T(t) = \boldsymbol{\Phi}'(t)\boldsymbol{\alpha}_n^T,\ Z_n^P(t) = \boldsymbol{\Phi}'(t)\boldsymbol{\alpha}_n^P.$$

For the sake of simplicity, we suppose that T and P curves belong to the same Hilbert space which is a fairly natural choice and that they have the same number of coefficients. If variability in curve shape may change drastically from a variable to another, it is always possible to increase the number L of basis functions to catch shape variations of interest.

Consider now the $2L$-vector

$$\boldsymbol{\alpha}_n = \left(\alpha_{n1}^T, \ldots, \alpha_{nL}^T; \alpha_{n1}^P, \ldots, \alpha_{nL}^P\right)' = \left(\boldsymbol{\alpha}_n^{T\prime}, \boldsymbol{\alpha}_n^{P\prime}\right)'$$

that merges both coefficients of temperature and precipitation annual curves and the associated mean $2L$-vector of coefficients

$$\overline{\boldsymbol{\alpha}} = (\overline{\boldsymbol{\alpha}}_T', \overline{\boldsymbol{\alpha}}_P')'.$$

Let $\mathbf{C}$ be the $N \times 2L$ matrix of centered coefficients and construct

$$\widehat{\boldsymbol{\Gamma}} = \frac{1}{N}\mathbf{C}'\mathbf{C}$$

the $2L \times 2L$ matrix of empirical covariances between coefficients. The multivariate functional principal component analysis (MFPCA) consists in finding the decomposition of the matrix $\widehat{\boldsymbol{\Gamma}}\mathbf{W}\mathbf{M}$ by solving the following eigenvalue problem:

$$\widehat{\boldsymbol{\Gamma}}\mathbf{W}\mathbf{M}\mathbf{b}_l = \widehat{\lambda}_l\mathbf{b}_l, \quad l = 1, \ldots, 2L$$

where $\mathbf{b}_l$ is the lth eigenvector associated with the positive eigenvalue $\widehat{\lambda}_l$. These $2L$-eigenvectors can be ordered according to their associated eigenvalues. The covariance matrix is structured by blocks:

$$\widehat{\boldsymbol{\Gamma}} = \begin{bmatrix} \widehat{\boldsymbol{\Gamma}}_{TT} & \widehat{\boldsymbol{\Gamma}}_{TP} \\ \widehat{\boldsymbol{\Gamma}}_{PT} & \widehat{\boldsymbol{\Gamma}}_{PP} \end{bmatrix},$$

where matrix $\widehat{\boldsymbol{\Gamma}}_{TP}$ denotes the $L \times L$ covariance matrix between coefficients of variable T and variable P. The $2L \times 2L$ matrix $\mathbf{W}$ ensures the metric equivalence between the functional problem and its discrete version (working on coefficients). It is composed by blocks as follows:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_T & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_P \end{bmatrix},$$

where matrices $\mathbf{W}_T$ and $\mathbf{W}_P$ are Gram matrices. Here, $\mathbf{W}_T = \mathbf{W}_P$ because same B-splines basis has been used for constructing T and P curves, but different basis choice might also be relevant.

Compared to the univariate case, a significant change is carried out with the matrix $\mathbf{M}$. This block diagonal weighting matrix:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_T & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_P \end{bmatrix}$$

of size $2L \times 2L$ is used to balance the coefficient values when decomposing both the variance of temperature and precipitation data. It acts as a normalization step usual in standard PCA when variables do not have the same units or the same orders of magnitude. Diagonal terms of matrices $\mathbf{M}_T$ and $\mathbf{M}_P$ are naturally chosen with, respectively,

$$m_{ll}^T = 1/\sigma_T^2, \quad m_{ll}^T = 1/\sigma_P^2,$$

where

$$\sigma_T^2 = \mathrm{Tr}(\widehat{\mathbf{\Gamma}}_{TT}\mathbf{W}_T) \quad \text{and} \quad \sigma_P^2 = \mathrm{Tr}(\widehat{\mathbf{\Gamma}}_{PP}\mathbf{W}_P).$$

This weighting system will give the same importance to temperature and precipitation curves in the MFPCA decomposition.

If one remarks that the structure of a $2L$-eigenvector is such that

$$\mathbf{b} = \left(b_1^T, \ldots, b_L^T; b_1^P, \ldots, b_L^P\right)' = \left(\mathbf{b}_T', \mathbf{b}_P'\right)',$$

normalized eigenvectors are obtained with

$$\mathbf{a} = \left(\mathbf{a}_T', \mathbf{a}_P'\right)' = \mathbf{M}^{-\frac{1}{2}}\mathbf{W}^{-\frac{1}{2}}\mathbf{b}.$$

The eigenfunctions associated with a principal axis can then be computed for each variable with

$$\widehat{\xi}^T(t) = \mathbf{\Phi}'(t)\mathbf{a}_T, \quad \widehat{\xi}_l^P(t) = \mathbf{\Phi}'(t)\mathbf{a}_P.$$

Observations are projected in a space of small dimension $Q < 2L$ when computing the $N \times Q$ matrix $\mathbf{P}$ of principal coordinates with

$$\mathbf{P}_Q = \mathbf{C}\mathbf{A}_Q$$

with $\mathbf{A}_Q = (\mathbf{a}_1, \ldots, \mathbf{a}_Q)$ the $2L \times Q$ matrix of normalized eigenvectors vertically structured by blocks

$$\mathbf{A}_Q = \begin{bmatrix} \mathbf{A}_Q^T \\ \mathbf{A}_Q^P \end{bmatrix}.$$

Functions are finally reconstructed through their coefficients:

$$\tilde{\boldsymbol{\alpha}}_n^T = \overline{\boldsymbol{\alpha}}^T + \mathbf{A}_Q^T\mathbf{c}_n, \quad \tilde{\boldsymbol{\alpha}}_n^P = \overline{\boldsymbol{\alpha}}^P + \mathbf{A}_Q^P\mathbf{c}_n,$$

where $\mathbf{c}_n$ is the Q-vector of principal coordinates of $\mathbf{Z}_n$ (row n in $\mathbf{P}_Q$) and

$$\tilde{Z}_n^T = \mathbf{\Phi}'(t)\tilde{\boldsymbol{\alpha}}_n^T, \quad \tilde{Z}_n^P = \mathbf{\Phi}'(t)\tilde{\boldsymbol{\alpha}}_n^P.$$

4.6.2 MFPCA Displays

Consider now the sample of pairwise curves $(z_1^T, z_1^P), \ldots, (z_N^T, z_N^P)$ estimated from row data of temperature and precipitation by regression in a same B-spline basis of arbitrary dimension $L = 9$. Once the eigen-decomposition has been achieved over the $2L \times 2L$ covariance matrix of B-spline coefficients, it is possible to represent the main factors of variability as a perturbation of the mean functions such that

$$\widehat{\mu}_T(t) \pm \sqrt{\lambda_l} \times \widehat{\xi}_l^T(t), \quad \widehat{\mu}_P(t) \pm \sqrt{\lambda_l} \times \widehat{\xi}_l^P(t).$$

Figure 4.8 displays of three factors of the PCA associated with the highest eigenvalues $(\lambda_1, \lambda_2, \lambda_3)$ and the associated map of individuals. Size of circles indicates

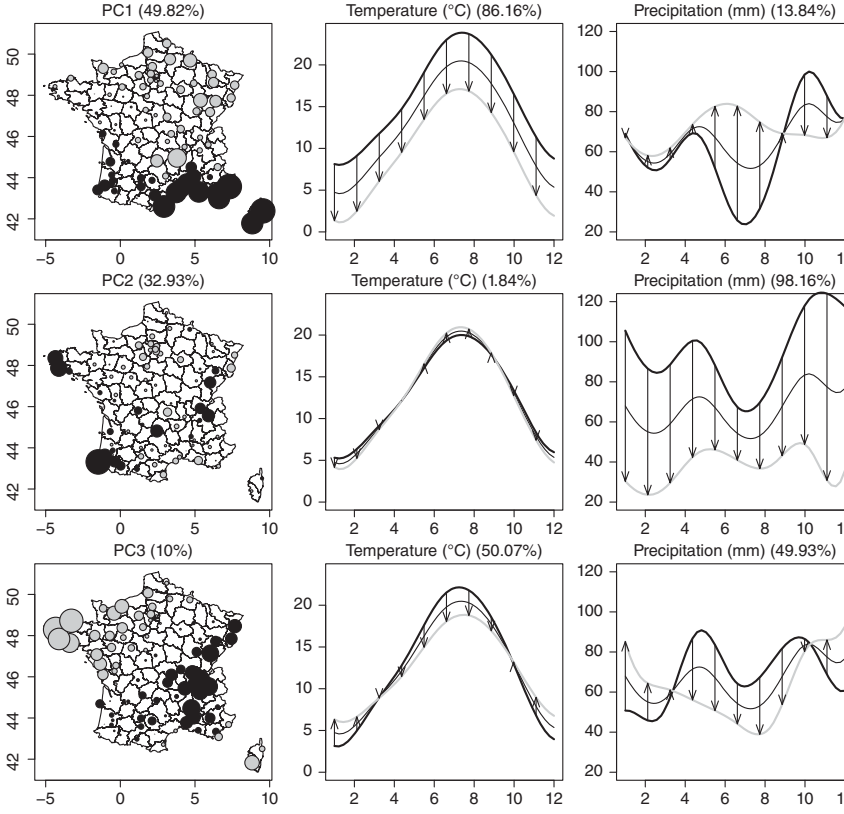


Figure 4.8 First factors of the MFPCA of temperature and precipitation profiles and maps displaying the contribution of each sampled station to an associated factor (gray = negative value, black = positive value). Arrows indicate changes in curve shape when moving from black points to gray ones. Thin curves are the mean functions. Percentages above graphs indicate from left to right: amount of variability, contribution of variable T, contribution of variable P to that variability.

the contribution of each city to the variability of the axis. Color is associated with the sign of the observation when projected of the related principal axis (gray = negative, black = positive).

Small diameter circles indicate locations that are close to the mean curve. It is interesting to note that these three factors account for about 93% of the entire variability. The first factor (50% of the variability) is attached to variation in latitudes. Its effects are mainly related to average temperature variations (86% of the variability, shape of the T profiles does not change) and variations in shape for precipitation curves (14%). These latter present two seasonal modes (maximum values in spring and autumn) for southern towns that vanish when moving to northern towns. The second factor is independent of the temperature (tiny contribution), only attached to variations in precipitation mean (the shape of the curves does not change). Nevertheless, it accounts for 33% of the variability. It only concerns some erratically distributed towns which belong to specific geographical zones (mountains, local climatic conditions). The third factor (10%) is an East-West effect that opposites oceanic to continental climate conditions. It is characterized by changes in shape for both temperature and precipitation curves (same contribution). A direct effect of the ocean is to modify the temperature shape. Amplitude between winter and summer of western temperature curves, close to the Atlantic Ocean, is less important than in continental areas. Precipitation curves are V-shaped (gray curve) under the influence of the ocean, while seasonal variations in spring and autumn are much more pronounced for continental areas (black curve). It is worth reminding that the approximation of any curves of temperature and precipitation as a linear combination of these three factors only

$$\begin{cases} \tilde{z}_n^T(t) = \widehat{\mu}_T(t) + c_{n1}\widehat{\xi}_1^T(t) + c_{n2}\widehat{\xi}_2^T(t) + c_{n3}\widehat{\xi}_3^T(t) \\ \tilde{z}_n^P(t) = \widehat{\mu}_P(t) + c_{n1}\widehat{\xi}_1^P(t) + c_{n2}\widehat{\xi}_2^P(t) + c_{n3}\widehat{\xi}_3^P(t) \end{cases}$$

provides good estimations as shown in Figure 4.9.

4.6.3 Multivariate Functional Principal Component Kriging

Using the unbiased conditions, the multivariate functional kriging system (4.9) can be expressed with the principal components of the MFPCA, such as

$$\begin{bmatrix} \mathbf{A}_Q^T \\ \mathbf{A}_Q^P \end{bmatrix} \widehat{\mathbf{c}}_0 = \sum_{n=1}^{N} \begin{bmatrix} \mathbf{B}_{TT}^n & \mathbf{B}_{PT}^n \\ \mathbf{B}_{TP}^n & \mathbf{B}_{PP}^n \end{bmatrix}' \begin{bmatrix} \mathbf{W}_T & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_P \end{bmatrix} \begin{bmatrix} \mathbf{A}_Q^T \\ \mathbf{A}_Q^P \end{bmatrix} \mathbf{c}_n.$$

Change in metric makes $\mathbf{A}_Q'\mathbf{W}\mathbf{M}\mathbf{A}_Q = \mathbf{I}_Q$ and the system can be written as in (4.7) with

$$\widehat{\mathbf{c}}_0 = \sum_{n=1}^{N} \mathbf{A}_Q'\mathbf{W}\mathbf{M}\mathbf{B}_n'\mathbf{W}\mathbf{A}_Q\mathbf{c}_n,$$

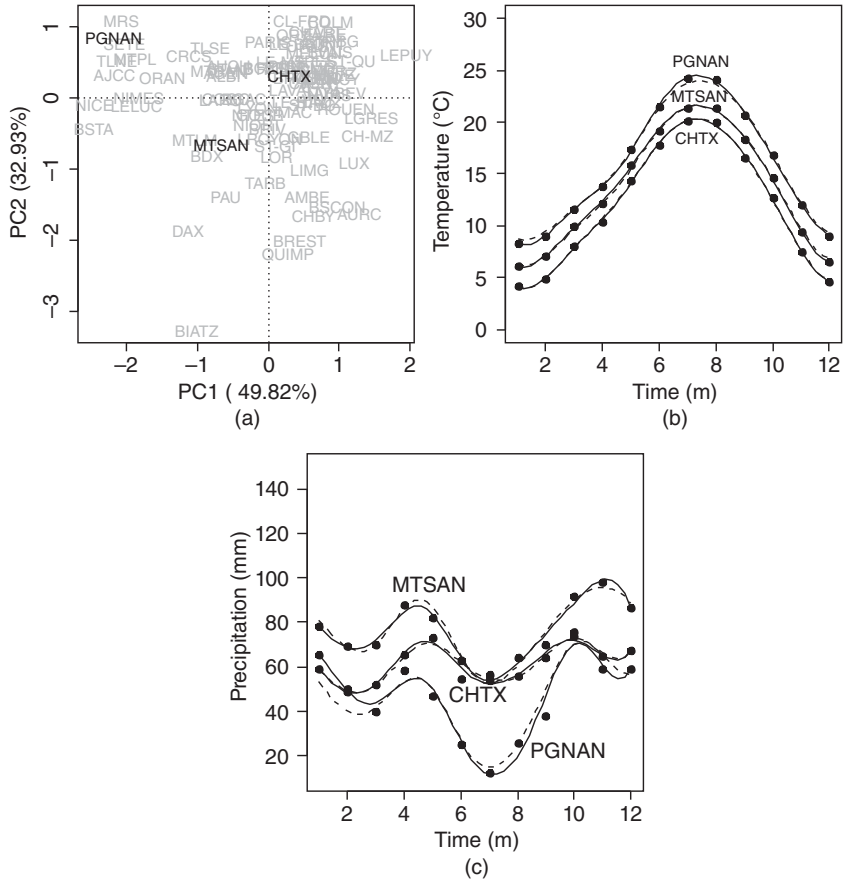


Figure 4.9 Example of PCA 2D-mapping of observations (z_n^T, z_n^P) (a) accounting for 83% of the variability. Panels (b) and (c) show an example of the reconstitution of three observations of temperature and precipitation. Continuous lines are observations, and dashed lines are reconstructions using three PCs of the MFPCA.

where the $\mathbf{c}_n$ is the Q-principal coordinates for observation $\mathbf{Z}_n = \left[Z_n^T, Z_n^P\right]'$ and $\mathbf{B}_n$ the block-matrix of kriging weights. If $Q = 2L$, the search for a BLUP estimator $\widehat{\mathbf{Z}}_0$ is achieved when minimizing

$$\mathbb{E}\|\widehat{\mathbf{c}}_0 - \mathbf{c}_0\|_{2L}^2 = \frac{1}{\sigma_T^2}\mathbb{E}\|\widehat{\boldsymbol{\alpha}}_0^T - \boldsymbol{\alpha}_0^T\|_{\mathbf{W}_T}^2 + \frac{1}{\sigma_P^2}\mathbb{E}\|\widehat{\boldsymbol{\alpha}}_0^P - \boldsymbol{\alpha}_0^P\|_{\mathbf{W}_P}^2$$

still under the same constraints of unbiasedness. The estimator is constructed giving same weights to T and P curves, a direct effect of additional metric $\mathbf{M}$. Once the LMC is fitted on experimental variograms, the weights $\mathbf{U}_n = \mathbf{A}_Q'\mathbf{W}\mathbf{M}\mathbf{B}_n'\mathbf{W}\mathbf{A}_Q$

are determined and an observation $\mathbf{Z}_0$ can be estimated with

$$\widehat{\mathbf{Z}}_0(t) = \begin{bmatrix} \boldsymbol{\Phi}'(t)\widehat{\boldsymbol{\alpha}}_0^T \\ \boldsymbol{\Phi}'(t)\widehat{\boldsymbol{\alpha}}_0^P \end{bmatrix},$$

where

$$\begin{bmatrix} \widehat{\boldsymbol{\alpha}}_0^T \\ \widehat{\boldsymbol{\alpha}}_0^P \end{bmatrix} = \begin{bmatrix} \overline{\boldsymbol{\alpha}}^T \\ \overline{\boldsymbol{\alpha}}^P \end{bmatrix} + \begin{bmatrix} \mathbf{A}_Q^T \\ \mathbf{A}_Q^P \end{bmatrix} \widehat{\mathbf{c}}_0.$$

As in the univariate case, we are generally interested in cases where $Q < 2L$.

4.6.4 Mixing Temperature and Precipitation Curves

Figure 4.10 presents an example of LMC fitting on the first four PCs of the MFPCA. As in the univariate case, the cross-variograms are close to zero, making reasonable the hypothesis of intrinsic correlation. For observation in position n, the integrated prediction error is computed by leave-one-out cross-validation with

$$\text{ISE}_n = \frac{1}{\sigma_T^2}\|\widehat{\boldsymbol{\alpha}}_{(-n)}^T - \boldsymbol{\alpha}^T\|_{\mathbf{W}_T}^2 + \frac{1}{\sigma_P^2}\|\widehat{\boldsymbol{\alpha}}_{(-n)}^P - \boldsymbol{\alpha}^P\|_{\mathbf{W}_P}^2.$$

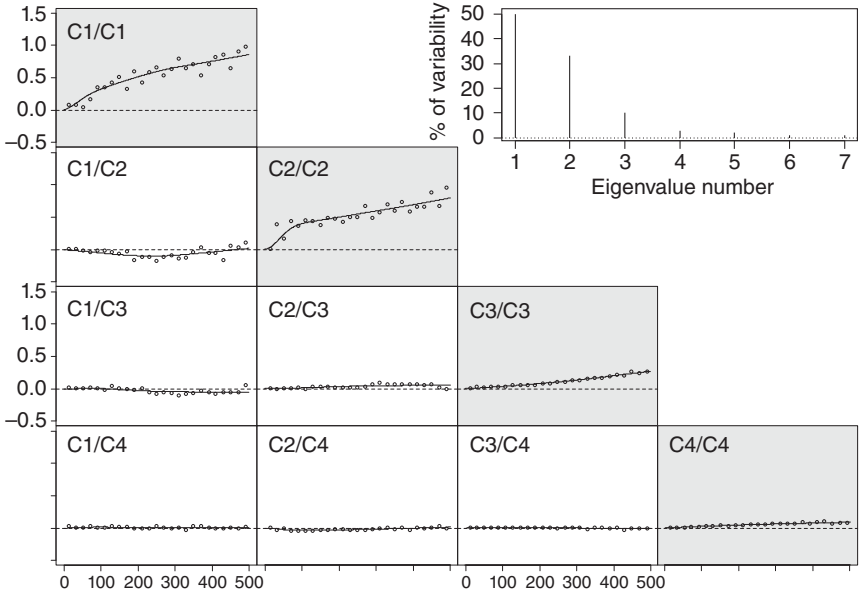


Figure 4.10 Empirical variograms and fitting of a coregionalization model on the first four principal components of the MFPCA. The magnitude of the components is decreasing very fast. It is remarkable to note the quasispatial independence between components as cross-variograms are surprisingly close to zero functions.

In that case, the variances σ_T^2 and σ_P^2 have been estimated using the $N-1$ remaining observed curves. The plot of some predictions (quantiles of the ISEs) is displayed in Figure 4.11. Again, bad predictions appear in areas with specific climatic conditions (mountains, local climatic peculiarities) where stationarity hypothesis

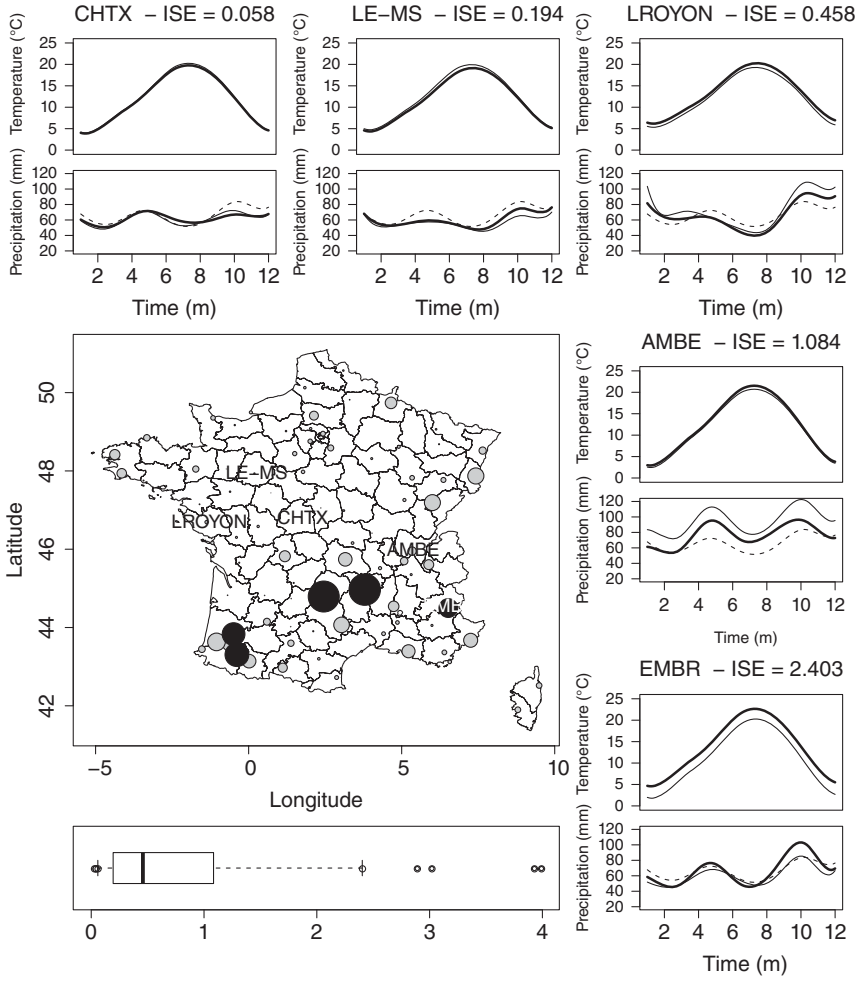


Figure 4.11 Predictions of temperature and precipitation curves (dashed line: mean precipitation, thin line: observed curve, thick line: predicted curve), spatial map, and boxplot of the integrated squared errors (ISEs). From top-left to bottom-right panel, curves are, respectively, sorted by ISE quantiles of order 0.05, 0.25, 0.5, 0.75, and 0.95. The ISE median location is displayed in the top-right panel (LROYON). Bad predictions (black points corresponding to ISE quantiles of order ≥ 0.95) are located in areas where the stationarity hypothesis does not hold. Because of metric **M**, ISEs are balanced between temperature and precipitation.

is not relevant. However, it is worth noting that the shape of the curves is globally well predicted. The main source of errors is attached to bad predictions of temperature or precipitation mean levels. This might suggest that the source of error is essentially due to the assumption of constant mean over the whole domain. It should be interesting to consider in that case kriging with external drift (refer to Chapter 3 of this book).

4.7 Discussion

We propose a kriging method for multivariate functional data which provides an answer to the first question addressed in the introduction: the prediction of both temperature and precipitation curves is possible in the isotopic case, i.e. when both T and P data are available at any position $\{x_1, \ldots, x_n\}$ over the domain. The employed method relies on nice properties of the spectral decomposition of variance operators. The so-called "Karuhen–Loève" decomposition is not new and has been applied in many fields of functional data analysis. For example, He et al. [10] used a canonical correlation analysis, a generalization of the PCA, in the framework of functional linear model. The decomposition of the variance operator is also known as spectral cut regularization method used to find admissible solution for the functional linear model [11].

The pioneering work of [2] on FPCA has proposed a formalism based upon the theory of reproducing kernels. It is a well-suited strategy to get rid of ill-conditioned problems under some hypotheses of regularity of the considered Hilbert space. We place the functional kriging in the same context. This formalism fits with operational situations since work is anyway achieved in finite dimension and a reproducing kernel always exists in that case. Once the choice of the basis has been made, the reproducing kernel attached to the identity operator in that basis has a matrix form straightforwardly given by the inverse of the Gram matrix. Other methods of kriging curves exist and are also presented in this book.

The original point developed in this current work is the use of a multivariate version of the FPCA in order to merge both temperature and precipitation profiles and then achieve kriging on principal components. This approach allows to circumvent the unreasonable estimation of a huge number of variograms when working on the basis coefficients. The functional problem of kriging is then transformed to a multivariate cokriging problem of small dimension within the construction of some metric that makes the functional framework equivalent. The basic idea is that kriging a linear combination of the coefficients of the basis decomposition is the same as working directly on the coefficients themselves. And it should be more efficient in case where it is possible to concentrate a great amount of variability on a few number of dimensions. We have shown that the variance of the proposed kriging estimators is the same when considering all the principal components.

Early works of [12] already proposed conditions for making equivalent the problem of kriging several real variables and kriging a linear combination of these variables. More recently, other approaches for kriging multivariate curves have been proposed and also rely on the use of FPCA [13, 14]. The works of [9] that developed the use of PCs for kriging multivariate real spatial observations, pointed out a number of problems that are still encountered in the functional framework. Even if principal component kriging is less computationally expensive than cokriging the entire set of coefficients, it suffers from several drawbacks:

- Only locations where all variables are jointly sampled can be considered: our approach is limited to the isotopic case.
- The cross-covariance between PCs is not necessarily zero at $|h| \neq 0$. Although for the temperature and precipitation example, it is an acceptable assumption, the intrinsic stationarity is a strong hypothesis.
- One can wonder if the estimation of a spatial covariance model by LMC on a few PCs can capitalize all the spatial correlation structures available in the initial variables.

This last point deserves to be developed. Goovaerts [15] pointed out that the correlation between principal components may be far from negligible, especially when the correlation structure greatly changes from one spatial scale to another. This result must be kept in mind when one intends to replace cokriging by kriging uncorrelated combinations (at least when $|h| = 0$) of the original variables. This implies that the choice of the basis for the representation of the observed curves may matter as shown in Figure 4.12.

It remains for us to fix the problem of kriging several curves in the heterotopic case. How to proceed when both T and P curves are not available on same locations? The proposed approach by MFPCA works well for the isotopic case, i.e. when data of both functional variables are available in all the sampled stations. However, keeping in mind the problems of estimation raised above, FPCA (or any related method) can still be used to summarize functional data in a small number of PCs. Simple co-kriging on several functional variables can then be achieved on a set of merged PCs computed from separate FPCAs and heterotopic cases can be tackled in that case (see [7] for practical details or [14] for an extension of kriging multivariate curves that deals with anisotropy).

To conclude this discussion, a word on competing approaches to kriging. The strength of a kriging method is to propose a model for characterizing spatial covariance. This implies that if the model is the good one, the prediction is efficient especially in areas where no data are available or where ancillary information is available. There exists very little theoretical works which compare kriging approaches to other spatial methods of interpolation. In the paper of [16], the authors concluded that when the data do satisfy the intrinsic random function

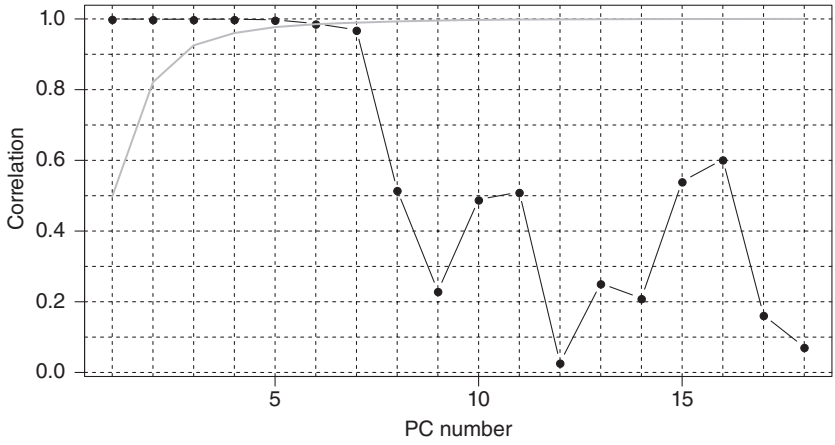


Figure 4.12 Correlation (absolute values) between PCs of a MFPCA realized from a Fourier basis decomposition and a B-spline basis decomposition for temperature and precipitation curves. Gray curve is the percentage of cumulated variability (same values for Fourier or B-splines). Even if eigenvalues are the same between Fourier and B-splines, the structure of the points cloud changes. This can have some bad consequences when estimating the spatial covariance model if correlation structures are strong on PCs associated with small eigenvalues.

hypothesis and when the variogram family is known, then the kriging estimator does perform better than nonparametric approaches, but only marginally better. On the other hand, they pointed out that the kriging approach is not robust when the intrinsic hypothesis does not hold. They also mention that when the data do not come from an intrinsic random function with the right variogram, a nonparametric approach seems consistently more advisable especially when regarding the error estimation. Functional kriging approaches are indeed constituted with a complicated assemblage of nested methods that require caution use and experience, which make *kriging an art rather than an algorithm* [16]. A future challenging task is clearly required for the comparison of functional kriging approaches to those proposed in [17] or in [18] for functional data.

4.A Appendices

4.A.1 Computation of the Kriging Variance

Using a basis expansion, the solution of kriging is given when minimizing the variance of the errors:

$$\mathbb{E}\|\widehat{Z}_0 - Z_0\|^2_{\mathcal{H}} = \mathbb{E}\|\widehat{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_0\|^2_{\mathbf{W}}$$

under the constraint $\sum_n \mathbf{B}_n = \mathbf{W}^{-1}$.

Let $\mathbf{A} \in \mathcal{M}_L$ and $\mathbf{B} \in \mathcal{M}_L$ be $L \times L$ matrices. Denote by

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}'$$

the $L \times L$ matrix with $\mathbf{u}$ and $\mathbf{v}$ vectors in $\mathbb{R}^L$. Reminding that

$$\langle \mathbf{A}\mathbf{u}, \mathbf{B}\mathbf{v} \rangle_{\mathbf{W}} = \mathbf{u}'\mathbf{A}'\mathbf{W}\mathbf{B}\mathbf{v} = \langle \mathbf{A}, \mathbf{B}\,[\mathbf{v} \otimes \mathbf{u}] \rangle_{HS} = \mathrm{Tr}\left(\mathbf{A}'\mathbf{W}\mathbf{B}\,[\mathbf{v} \otimes \mathbf{u}]\right),$$

and thanks to the linearity of the expectation, to the properties of the trace operator, and unbiased conditions of the kriging estimator, the previous expression of the variance can be developed with

$$\begin{aligned}
\mathbb{E}\|\widehat{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_0\|^2_{\mathbf{W}} &= \mathbb{E}\|\widehat{\boldsymbol{\alpha}}_0 - \boldsymbol{\mu} - \boldsymbol{\alpha}_0 + \boldsymbol{\mu}\|^2_{\mathbf{W}} \\
&= \mathbb{E}\|\boldsymbol{\alpha}_0 - \boldsymbol{\mu}\|^2_{\mathbf{W}} + \sum_n \sum_m \mathbb{E}\left\langle \mathbf{B}'_n\mathbf{W}(\boldsymbol{\alpha}_n - \boldsymbol{\mu}), \mathbf{B}'_m\mathbf{W}(\boldsymbol{\alpha}_m - \boldsymbol{\mu}) \right\rangle_{\mathbf{W}} \\
&\quad - 2\sum_n \mathbb{E}\left\langle \mathbf{B}'_n\mathbf{W}(\boldsymbol{\alpha}_0 - \boldsymbol{\mu}), \boldsymbol{\alpha}_n - \boldsymbol{\mu} \right\rangle_{\mathbf{W}} \\
&= \left\langle \mathbf{I}_L, \mathbb{E}[(\boldsymbol{\alpha}_0 - \boldsymbol{\mu}) \otimes (\boldsymbol{\alpha}_0 - \boldsymbol{\mu})] \right\rangle_{HS} \\
&\quad + \sum_n \sum_m \left\langle \mathbf{B}'_n\mathbf{W}, \mathbf{B}'_m\mathbf{W}\mathbb{E}[(\boldsymbol{\alpha}_m - \boldsymbol{\mu}) \otimes (\boldsymbol{\alpha}_n - \boldsymbol{\mu})] \right\rangle_{HS} \\
&\quad - 2\sum_n \left\langle \mathbf{B}'_n\mathbf{W}, \mathbb{E}[(\boldsymbol{\alpha}_n - \boldsymbol{\mu}) \otimes (\boldsymbol{\alpha}_0 - \boldsymbol{\mu})] \right\rangle_{HS} \\
&= \mathrm{Tr}(\mathbf{W}\boldsymbol{\Gamma}_{00}) + \sum_n \sum_m \mathrm{Tr}(\mathbf{W}\mathbf{B}_n\mathbf{W}\mathbf{B}_m{}'\mathbf{W}\boldsymbol{\Gamma}_{nm}) \\
&\quad - 2\sum_n \mathrm{Tr}(\mathbf{W}\mathbf{B}_n\mathbf{W}\boldsymbol{\Gamma}_{n0}). \qquad (4.A.1)
\end{aligned}$$

Following the method of Lagrange multipliers, define the function:

$$\begin{aligned}
&F\left(\mathbf{B}_1, \ldots, \mathbf{B}_n, \boldsymbol{\Lambda}\right) \\
&\quad = \mathrm{Tr}(\mathbf{W}\boldsymbol{\Gamma}_{00}) + \sum_n \sum_m \mathrm{Tr}(\mathbf{W}\mathbf{B}_n\mathbf{W}\mathbf{B}_m{}'\mathbf{W}\boldsymbol{\Gamma}_{nm}) - 2\sum_n \mathrm{Tr}(\mathbf{W}\mathbf{B}_n\mathbf{W}\boldsymbol{\Gamma}_{n0}) \\
&\qquad + 2 \times \mathrm{Tr}\left(\left[\sum_n \mathbf{B}_n - \mathbf{W}^{-1}\right]\mathbf{W}\boldsymbol{\Lambda}\right),
\end{aligned}$$

where the last term of the equation is the Lagrange multiplier term.

The Gâteaux derivative $\delta_{\boldsymbol{\Delta}} F(\mathbf{B}_n)$ of F at $\mathbf{B}_n \in \mathcal{M}_L$ in direction $\boldsymbol{\Delta} \in \mathcal{M}_L$ is given with

$$\delta_{\boldsymbol{\Delta}} F(\mathbf{B}_n) = \lim_{\varepsilon \to 0} \frac{F(\mathbf{B}_1, \ldots, (\mathbf{B}_n + \varepsilon\boldsymbol{\Delta}), \ldots, \mathbf{B}_N, \boldsymbol{\Lambda}) - F(\mathbf{B}_1, \ldots, \mathbf{B}_n, \ldots, \mathbf{B}_N, \boldsymbol{\Lambda})}{\varepsilon}.$$

Kriging weights are found when

$$\delta_{\boldsymbol{\Delta}} F(\mathbf{B}_n) = 0, \quad n = 1, \ldots, N, \;\; \delta_{\boldsymbol{\Delta}} F(\boldsymbol{\Lambda}) = 0.$$

Developing the above expression of the $N+1$ derivatives, solutions are found when

$$\begin{cases} \sum_m \mathrm{Tr}(\mathbf{W}\boldsymbol{\Gamma}_{nm}\mathbf{W}\mathbf{B}_n\mathbf{W}\boldsymbol{\Delta}') - \mathrm{Tr}(\mathbf{W}\boldsymbol{\Gamma}_{n0}\mathbf{W}\boldsymbol{\Delta}') + \mathrm{Tr}(\boldsymbol{\Lambda}\mathbf{W}\boldsymbol{\Delta}') = 0, \quad n = 1, \ldots, N \\ \mathrm{Tr}\left(\left[\sum_n \mathbf{B}_n - \mathbf{W}^{-1}\right]\mathbf{W}\boldsymbol{\Delta}'\right) = 0 \end{cases}$$

in any direction $\boldsymbol{\Delta}$. A sufficient condition to find the minimum is that the $\mathbf{B}_n$s verify the kriging system (4.6) of $L(N+1)$ equations. Now, when introducing

$$\sum_m \mathrm{Tr}(\mathbf{W}\boldsymbol{\Gamma}_{nm}\mathbf{W}\mathbf{B}_n) - \mathrm{Tr}(\mathbf{W}\boldsymbol{\Gamma}_{n0}) + \mathrm{Tr}(\boldsymbol{\Lambda}) = 0$$

in Eq. (4.A.1), the variance of the functional kriging estimator is given with

$$\sigma^2_{FKG} = \mathrm{Tr}(\mathbf{W}\boldsymbol{\Gamma}_{00}) - \sum_n \mathrm{Tr}(\mathbf{B}_n\mathbf{W}\boldsymbol{\Gamma}_{0n}\mathbf{W}) - \mathrm{Tr}(\boldsymbol{\Lambda}).$$

Notice that this variance is roughly the same as in [5] within the metric $\mathbf{W}$.

References

1 Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*, 3e. New York: Springer.
2 Besse, P. and Ramsay, J. (1986). Principal components analysis of sampled functions. *Psychometrika* 51 (2): 285–311.
3 Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers.
4 Nerini, D., Monestiez, P., and Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis* 101: 409–418.
5 Myers, D.E. (1982). A matrix formulation of co-kriging. *Journal of the International Association for Mathematical Geology* 14: 249–257.
6 Giraldo, R. (2014). Cokriging based on curves, prediction and estimation of the prediction variance. *InterStat* 2: 1–30.
7 Wackernagel, H. (2003). *Multivariate geostatistics: an introduction with applications*, 3e. Berlin Heidelberg: Springer-Verlag.
8 Goulard, M. and Voltz, M. (1992). Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Mathematical Geology* 24 (3): 269–286.
9 Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press.
10 He, G., Müller, H.-G., Wang, J.-L., and Yang, W. (2010). Functional linear regression via canonical analysis. *Bernoulli* 16 (3): 705–729.
11 Crambes, C. and Mas, A. (2013). Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli* 19 (5B): 2627–2651.
12 Myers, D.E. (1983). Estimation of linear combinations and co-kriging. *Journal of the International Association for Mathematical Geology* 15: 633–637.
13 Bohorquez, M., Giraldo, R., and Mateu, J. (2017). Multivariate functional random fields: prediction and optimal sampling. *Stochastic Environmental Research and Risk Assessment* 31: 53–70.

14 Grujic, O., Menafoglio, A., and Yang, G. (2017). Cokriging for multivariate hilbert space valued random fields: application to multi-fidelity computer code emulation. *Stochastic Environmental Research and Risk Assessment* 32: 1955–1971.

15 Goovaerts, P. (1993). Spatial orthogonality of the principal components computed from coregionalized variables. *Mathematical Geology* 25 (3): 281–302.

16 Yakowitz, S. and Szidarovszky, F. (1985). A comparison of kriging with non-parametric regression methods. *Journal of Multivariate Analysis* 16 (1): 21–53.

17 Cardot, H. (2007). Conditional functional principal components analysis. *Scandinavian Journal of Statistics* 34 (2): 317–335.

18 Dabo-Niang, S. and Yao, A.F. (2007). Kernel regression estimation for continuous spatial processes. *Mathematical Methods of Statistics* 16 (4): 298–317.

5

Geostatistical Analysis in Bayes Spaces: Probability Densities and Compositional Data

Alessandra Menafoglio[1], *Piercesare Secchi*[1,2], *and Alberto Guadagnini*[3,4]

[1] *MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy*
[2] *CADS - Center for Analysis Decisions and Society, Human Technopole, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy*
[3] *Department of Civil and Environmental Engineering, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy*
[4] *Department of Hydrology and Atmospheric Sciences, The University of Arizona 1133 James E. Rogers Way #122, Tucson, AZ 85721, USA*

5.1 Introduction and Motivations

The availability of complex, high-dimensional, and often constrained data has recently fostered new areas of statistical research. These are typically placed at the intersection between functional data analysis (FDA, [1]), geostatistics, and other fields classically devoted to the analysis of constrained data, such as compositional data analysis (CoDa, [2]). In this context, there is a general consensus that modern geostatistical approaches should always consider the nature of the data. In some cases, this would require resorting to a geometry which is not necessarily the one of the space of square-integrable functions (i.e. L^2).

The focus of this chapter is on functional compositions (FCs), that constitute the generalization to the functional setting of multivariate compositional data [2, 3]. The latter are defined as vector data that only provide relative information, i.e. for which the only relevant information is conveyed by the ratios between their components (termed *parts*). Examples of data that can be interpreted as compositional are discrete distributional data (i.e. probability mass functions), or, more generally, data whose components represent *parts* (e.g. proportions, percentages) of a whole (e.g. they sum up to unity) with respect to a given partition of the domain. For instance, concentration of chemicals adsorbed onto soil samples or distribution of population in age classes is often considered as compositional information.

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

In this broad context, FCs are functional data which only convey relative information. One can envision FCs as positive data, constrained to integrate to a constant – even as this might not be the case for some applications. Informally, in FCs, the ratios between their point evaluations are considered to be informative rather than their absolute values. For instance, probability density functions (PDFs) can be interpreted as FCs, and their point evaluations as infinitesimal parts of a whole, that is the probability of the sample space.

One can readily see that PDFs – as well as FCs in general – cannot simply be considered as square-integrable functions because the geometry of L^2 is not appropriate to treat them (e.g. the L^2-sum of two FCs is meaningless). Instead, the Bayes space geometry, introduced in [4–6] and recalled in Section 5.2, is well suited for FC data, since it was precisely designed to correctly represent the peculiar features of those data.

Throughout the chapter, we will illustrate the geostatistical methods for FCs developed in [7–9], and their application to the field setting which first motivated those works, which deals with particle-size distributions sampled in a heterogeneous aquifer system. These data describe the local distribution of soil particles sizes and are relevant to problems related to groundwater hydrology, soil science, geophysics, petroleum engineering, and geochemistry, with emphasis on applications oriented toward modeling physical and chemical processes occurring in heterogeneous Earth systems. Here, we illustrate methods for the preprocessing, kriging, and assessing uncertainty of such data. These methods need to be framed within a space different from L^2.

The remaining of the chapter is organized as follows. Section 5.2 introduces the Bayes space geometry for FCs, whereas Section 5.3 illustrates the data. The stationary kriging for FCs is addressed in Section 5.4, and the nonstationary approach is addressed in Section 5.5. Section 5.6 concludes the chapter.

5.2 Bayes Hilbert Spaces: Natural Spaces for Functional Compositions

The theory of Bayes spaces [4–6, 10] was introduced as a generalization to density functions of the Aitchison geometry. The latter is commonly employed to deal with compositional data that are multivariate observations carrying only relative information (e.g. [2, 3] and references therein). Compositional data are usually collected in the form of constrained objects summing up to a constant, usually set to 1 or 100, in case of proportions or percentages, respectively. PDFs can be then considered as compositional vectors with infinitely many parts [4], and with the key properties of compositions (e.g. [11]).

We denote by f the density function of an absolutely continuous measure μ with respect to the Lebesgue measure on the Borel space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, with compact support $I \subset \mathbb{R}$. In the following, we will express the properties of μ through those

of f. It should be noted that the theory of Bayes spaces was developed in a completely general framework in [5, 6, 10]. Two density functions f, g are considered as equivalent if they are proportional, and we denote such equivalence relation by $f =_{\mathcal{B}} g$. In this setting, the integral constraint $\int_I f(x)dx = 1$ of PDFs singles out a representative within an equivalence class of FCs that are equivalent from the viewpoint of the *relative* information they provide. Indeed, for any other representative $\tilde{f}$ (i.e. such that $\tilde{f} = c \cdot f$ for $c > 0$), the relative contribution of Borel subsets of $\mathbb{R}$ w.r.t. the measure of the support is the same. This property is known as *scale invariance* and is related to the observation that the probability of an event has no meaning *per se* – as noted in [10]. Otherwise, it is clearly framed in a relative context, as it is related to the probability of the entire sample set, which is set to unity for convenience.

Another relevant feature of FCs is the *relative scale*. The latter indicates that the increase of probability should be understood and measured in a relative sense, rather than on an absolute scale. For instance, the increase of probability over a Borel set from 0.05 to 0.1 (2 multiple) differs from the increase 0.5 to 0.55 (1.1 multiple), although the absolute differences are the same in both cases. This property further motivates the use of the log-ratio approach to deal with density functions.

The abovementioned properties are well-known and recognized in the multivariate setting (e.g. [2]) but are completely neglected when considering PDFs as unconstrained objects. For instance, the notions of sum and product by a constant that would be used for data analysis in L^2 (the space of square-integrable functions) appear to be inappropriate for compositions, their application may yield functions that are no longer compositions. These elements motivated the introduction of a geometry capable of capturing and properly incorporating the properties of FCs. Such a geometry is that of Bayes Hilbert spaces, that generalize the Aitchison geometry [12] to the functional setting.

For ease of notation, and following [7–9, 13, 14], we focus here on density functions with compact support. Note that the theory here presented could be extended to general supports, through the use of reference measures different from the Lebesgue one. However, it should be noted that, in several real datasets, finite values for the inferior and superior extremes of the support can be determined without a substantial loss of generality, or working with conditional distributions.

We term $\mathcal{B}^2(I)$ the Bayes space of (equivalence classes of) positive FCs f on I with square-integrable logarithm. In the following, the representative of an equivalence class will be its element integrating to 1; moreover, we only consider continuous FCs on a closed interval $I = [a, b]$, any compact subset of $\mathbb{R}$ being compatible with our framework. Given two FCs $f, g \in \mathcal{B}^2(I)$ and $\alpha \in \mathbb{R}$, we denote by $f \oplus g$ and $\alpha \odot f$ the perturbation and powering operations, defined as, respectively, [4, 6]:

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s)ds}, \quad (\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_I f(s)^\alpha ds}, \quad t \in I.$$

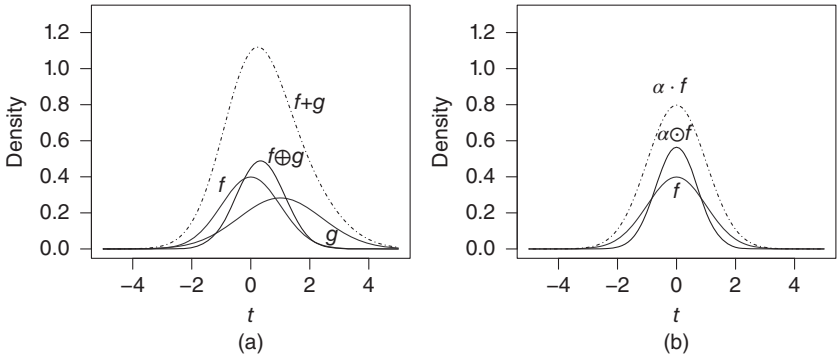


Figure 5.1 Example of perturbation and powering in $\mathcal{B}^2(I)$, compared to the typical operations in $L^2(I)$. (a) Perturbation $f \oplus g$ (solid black curve) of two Gaussian densities f, g restricted to $I = [-5, 5]$ (gray curves), and the sum $f + g$ in the space $L^2(I)$ (dot-dashed curve). (b) Powering of a Gaussian density f restricted to $I = [-5, 5]$ (gray curve) by $\alpha = 2$, $\alpha \odot f$ (solid black curve), and its counterpart $\alpha \cdot f$ in L^2 (dot-dashed curve). Source: Modified from Egozcue et al. [14].

It is then clear that the results of such operations are still PDFs. Note that $\mathcal{B}^2(I)$ endowed with the operations $(\oplus, \odot)$ is a vector space [4] and that the origin of the space $\mathcal{B}^2(I)$ is $e(t) = 1/\eta$, with $\eta = b - a$. Moreover, the difference between two FCs $f, g \in \mathcal{B}^2(I)$ is obtained as perturbation of f with the reciprocal of g, i.e. $f \ominus g = f \oplus [(-1) \odot g]$.

Figure 5.1 depicts an example considered in [14] of the effect of perturbation and powering operations in $\mathcal{B}^2(I)$, as opposed to standard operations of sum and product by a constant in $L^2(I)$. In [14], the authors considered the restriction to $I = [-5, 5]$ of the Gaussian densities $f =_{\mathcal{B}} \exp\{-t^2/2\}$ and $g =_{\mathcal{B}} \exp\{-(t-m)^2/(2s^2)\}$, with $m = 1$ and $s^2 = 2$. Figure 5.1a juxtaposes the perturbation of f by g ($f \oplus g$) to the sum in L^2 of f and g ($f + g$). Note that the latter sum does not result in a PDF, while the former does. Further, the perturbation of f by g yields a density function that is more concentrated than f and shifted toward g: this is the consequence of adding to f the information content in g and viceversa. Notice that the operation of perturbation can be interpreted as a Bayesian update of information, and $\ominus$ as a cancellation of information [10]. As such, all conjugate priors define affine subspaces of $\mathcal{B}(I)$. Within the latter class, we mention the Gaussian family and, more generally, the exponential family. Thus, it is not surprising that the result $f \oplus g$ displayed in Figure 5.1 is still a Gaussian density, as shown in [14].

Figure 5.1b depicts the result of the powering operation $\alpha \odot f$ in $\mathcal{B}^2(I)$, as well as the multiplication $\alpha \cdot f$ in $L^2(I)$, for the same f of Figure 5.1a and the scalar $\alpha = 2$. It is noted that $\alpha \cdot f$ is not a density function, and, as an element of $\mathcal{B}^2(I)$, it belongs to the same equivalence class as f itself. Otherwise, the powering of f by $\alpha = 2$ has

the effect of increasing the concentration of f around its mean (i.e. it decreases the variance of f by a factor 2). In the Bayesian framework, this is interpreted as the increase of information which is obtained by incrementing the "evidence" in f by the "evidence" in f itself.

The space $(\mathcal{B}^2(I), \oplus, \odot)$ is a separable Hilbert space structure if equipped with the inner product [4]

$$\langle f, g\rangle_{\mathcal{B}} = \frac{1}{2\eta}\int_I\int_I \ln\frac{f(t)}{f(s)}\ln\frac{g(t)}{g(s)}dt\ ds, \quad f, g \in \mathcal{B}^2(I), \tag{5.1}$$

which induces the following norm

$$||f||_{\mathcal{B}} = \left[\frac{1}{2\eta}\int_I\int_I \ln^2\frac{f(t)}{f(s)}dt\ ds\right]^{1/2}.$$

Each element of $\mathcal{B}^2(I)$ can be mapped onto an element of $L^2(I)$, preserving its distance and angle with any other element, that is, isometric isomorphisms exist between $\mathcal{B}^2(I)$ and $L^2(I)$. An example of such isometric isomorphism is defined by the *centered log-ratio* (clr) transformation [6, 7], which is defined, for $f \in \mathcal{B}^2(I)$, as

$$\mathrm{clr}(f)(t) = f_c(t) = \ln f(t) - \frac{1}{\eta}\int_I \ln f(s)ds. \tag{5.2}$$

One can see that the operations and inner products among the elements in $\mathcal{B}^2(I)$ can be equivalently computed in $L^2(I)$ as

$$\begin{aligned}&\mathrm{clr}(f \oplus g)(t) = f_c(t) + g_c(t), \quad \mathrm{clr}(\alpha \odot f)(t) = \alpha \cdot f_c(t),\\ &\langle f, g\rangle_{\mathcal{B}} = \langle f_c, g_c\rangle_2 = \int_I f_c(t)g_c(t)dt.\end{aligned} \tag{5.3}$$

Note that clr-transform induces, by construction, a zero-integral constraint, which may yield model-singularities. However, this is not the case of the geostatistical methods presented here.

5.3 A Motivating Case Study: Particle-Size Data in Heterogeneous Aquifers – Data Description

This section illustrates the key features of the field setting within which our theoretical framework is applied. As a showcase scenario, we consider the Lauswiesen site, which is an experimental test site located near the city of Tuebingen, Germany. The aquifer system under consideration has been the subject of an extensive series of experimental campaigns and modeling studies. Among these, the reader is referred to the works of Riva et al. [15–18], Hoffmann and Dietrich [19], Rein et al. [20], Neuman et al. [21, 22], Lessof et al. [23], Barahona-Palomo et al. [24], Handel and Dietrich [25], and Menafoglio et al. [7–9]. Characterization of the site has been based on data acquired through detailed geological,

hydrogeological, hydraulic, sedimentological, and geophysical investigations. The latter have been conducted at the field and laboratory scale.

The lithostratigraphic characterization has been performed through the stratigraphy information stemming from 150 mm-diameter monitoring wells [26, 27]. The aquifer at the site has a saturated thickness of about 5 m and is composed of fluvial geomaterial, overlain by stiff silty clay and underlain by hard silty clay. Available datasets include particle-size curves (PSCs), pumping and tracer tests, direct-push injection logging, and down-hole impeller flowmeter records. A detailed description of the analyses performed at the site is presented by Riva et al. [15, 16] and Lessof et al. [23], to which the reader is referred for details.

Of particular interest to our application are a collection of more than 400 PSCs collected along 12 vertical boreholes at the site. These indicate the presence of very heterogeneous, highly conductive alluvial deposits and were previously employed in [15–17] to provide a stochastic Monte Carlo-based numerical study of flow and transport process at the site. These studies considered diverse conceptual geological models of the structural heterogeneity of the system and analyzed their elative skill to interpret available tracer tests data. The available PSCs were assessed on core samples of characteristic length ranging from 5 to 26.5 cm. They are reconstructed through grain sieve analysis performed with a set of 12 discrete sieve diameters (i.e. 0.063, 0.125, 0.25, 0.50, 1.0, 2.0, 4.0, 8.0, 16.0, 31.5, 63.0, and 100.0 mm). Figure 5.2 depicts the three-dimensional structure of the sampling network at the site.

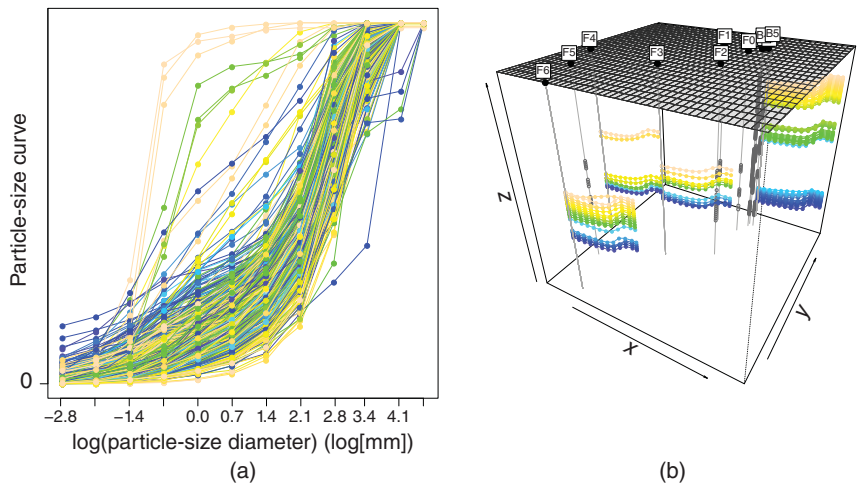


Figure 5.2 Raw particle-size data at the Lauswiesen site. (a) Collection of all available particle-size curves (PSCs) (b) Raw PSCs along boreholes B5, F3, F4, and F6. Gray scale colors correspond to the depth of the sampling locations. Source: Modified from Menafoglio et al. [8].

These PSCs have been employed in [15] to classify the types of geomaterials at the site and to construct geostatistically based models of the internal architecture of the aquifer. In this context, the latter could then be conceptualized as formed by a collection of regions (or blocks), randomly located in space, each formed by a given material type. Hydraulic properties of each of these blocks can then be estimated through available empirical formulations relating, e.g. permeability and porosity to characteristic diameters of a PSC. For example, Riva et al. [15–17] relate d_{10} and d_{60} (respectively, representing the particle size associated with the 10th and 60th percentile of a given PSC) to permeability through the Beyer's formula [28]. A geostatistical analysis of d_{10} and d_{60} or of the associated permeability can then be employed to characterize the heterogeneous distribution of hydraulic properties within the region occupied by each of the materials identified. The details of these analysis can be found in [15, 17]. Barahona-Palomo et al. [24] analyze the relationship between hydraulic conductivity estimates obtained through PSCs and impeller flowmeter measurements, while Riva et al. [18] rely on the available data to demonstrate their analytical study rendering relationships between the spatial covariance of hydraulic conductivity and of representative soil particle sizes and porosity.

5.4 Kriging Stationary Functional Compositions

5.4.1 Model Description

We term D the compact subset of $\mathbb{R}^d$ (usually $d = 2, 3$) corresponding to the spatial domain of the study and denote by $s_1, \dots, s_n$ the sampling locations in the test area. We denote by $\chi_{s_1}, \dots, \chi_{s_n}$ the dataset collected at those locations, formed by a set of positive PDFs on a compact domain I, i.e. $\chi_{s_i} : I \to (0, +\infty)$, such that $\int_I \chi_{s_i}(t)dt = 1$. In Section 5.2, we consider $\chi_{s_1}, \dots, \chi_{s_n}$ as objects of the Bayes Hilbert space $\mathcal{B}^2(I)$ and assume these to be a partial observation from a random field $\{\chi_s, s \in D\}$ valued in $\mathcal{B}^2(I)$. For instance, $\chi_{s_1}, \dots, \chi_{s_n}$ may be the densities of the particle-size distributions described in Section 5.3. Note that any other PDF can be considered for the application of our theoretical framework, including, e.g. rainfall (precipitation) distributions, or population pyramids [13], or dissolved chemical concentrations in groundwater.

In this section, we assume the process to be globally second-order stationary and isotropic, i.e. the following conditions hold:

(i) **Spatially constant mean**: $\mathbb{E}[\chi_s] = m$ for all $s \in D$;
(ii) **Stationary and isotropic trace-covariogram**: $\mathbb{E}[\langle \chi_{s_1} \ominus m, \chi_{s_2} \ominus m \rangle] = C(\|s_1 - s_2\|_d)$ for all $s_1, s_2 \in D$, $\|\cdot\|_d$ denoting a metric in $\mathbb{R}^d$.

Here, the mean and the covariogram are expressed in $\mathcal{B}^2(I)$, according to its geometric structure illustrated in Section 5.2. In such a space, under stationarity and isotropy, one may also define the spatial dependence structure through the trace-variogram of the process as

$$2\gamma(\|s_1 - s_2\|_d) = \mathbb{E}\left[\|\chi_{s_1} \ominus \chi_{s_2}\|^2\right].$$

The ordinary kriging predictor at a target location $s_0 \in D$ assumes in this context the form of the best linear combination of the data, linearity being interpreted in $\mathcal{B}^2(I)$ as $\chi^*_{s_0} = \oplus_{i=1}^n \lambda_i^* \odot \chi_{s_i}$. Informally, and in light of the example in Figure 5.1, such a linear combination is interpreted as a weighted sum of the information collected at each location, higher precisions (i.e. higher weight) being associated with nearby locations. Note that a zero weight $\lambda_i = 0$ powering a data-object χ_{s_i}, yields a contribution to the predictor in terms of a uniform PDF. This is precisely a null contribution in Bayes spaces, as the uniform PDF is the neutral element of the perturbation.

The ordinary kriging predictor is then found as the Best Linear Unbiased Predictor, whose weights minimize the variance of prediction error, under the unbiasedness constraint, i.e.

$$\mathbb{E}\left[\left\|\chi_{s_0} \ominus \oplus_{i=1}^n \lambda_i \odot \chi_{s_i}\right\|^2\right] \quad \text{subject to } \mathbb{E}\left[\oplus_{i=1}^n \lambda_i \odot \chi_{s_i}\right] = \mathbb{E}[\chi_{s_0}]. \tag{5.4}$$

Similar to the general case discussed in Chapter 2 (under mild assumptions on the sampling design), the optimal kriging weights are found by solving a linear system

$$\begin{pmatrix} \Sigma & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \vec{\lambda} \\ \zeta \end{pmatrix} = \begin{pmatrix} \vec{\sigma}_0 \\ 1 \end{pmatrix}. \tag{5.5}$$

Here, $\Sigma \in \mathbb{R}^{n \times n}$ denotes the variance–covariance matrix of the observations, $\Sigma_{i,j} = C(\|s_i - s_j\|_d)$ for $i,j = 1, \dots, n$, $\vec{\lambda} = (\lambda_1, \dots, \lambda_n)^T$ the vector of weights and ζ a Lagrange multiplier, and $\vec{\sigma}_0 = (C(\|s_1 - s_0\|_d), \dots, C(\|s_n - s_0\|_d))^T$ the vector of (trace-) covariances between observations and the random element at the target location.

Whenever the spatial dependence structure is unknown, the trace-covariogram, or the trace-variogram, can be estimated from the data by embedding the general procedure detailed in Chapter 2 in the Bayes Hilbert setting. In particular, the empirical estimator of the trace-semivariogram takes the form of

$$2\widehat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} \|\chi_{s_i} \ominus \chi_{s_j}\|^2, \tag{5.6}$$

where $N(h) = \{(i,j) \mid h - \Delta h \leq \|s_i - s_j\|_d \leq h + \Delta h\}$, and $|N(h)|$ is its cardinality.

Although expression (5.1) could be directly used to estimate (5.6), it involves double integrals, which might pose challenges for their accurate numerical evaluation. For the sake of efficiency, one may perform the computations on a transformed dataset, built upon mapping each data-object from $\mathcal{B}^2(I)$ to $L^2(I)$ through

the clr transformation (5.2). The latter allows expressing operations and inner products in $\mathcal{B}^2$ as operations and inner products in L^2, which markedly simplifies the calculations. We refer the reader to [7] for additional details.

5.4.2 Data Preprocessing

Data preprocessing, or data smoothing, is a first key step of almost any (geo)statistical analysis of functional or object data. Even as a wide body of literature has been devoted to smoothing data in L^2, still limited attention has been given to the problem of smoothing FCs. Since PDFs can be interpreted as instances of FCs, all methods apt to smooth PDFs or cumulative distribution functions (CDFs) can be adopted to deal with a range of FCs as well. This approach was considered in [7], where an extension of a smoothing method based on Bernstein polynomials [29] was proposed to deal with the particle-size densities (PSDs) described in Section 5.3. We briefly review the method, which serves as a basis to smooth the data described in Section 5.3.

Consider the problem of obtaining from raw data a smooth estimate of the jth curve, χ_{s_j}, that represents the PDF at location s_j ($j = 1, \dots, n$). We first note that the underlying distribution can be equivalently represented by the PDF χ_{s_j} (our target), or by the corresponding CDF $\mathcal{Y}_{s_j}(t) = \int_a^t \chi_{s_j}(\tau)d\tau$. As such, one can perform the smoothing either on χ_{s_j} or through the CDF. Bernstein polynomials are used here to provide a smooth estimate of the CDF, a key advantage with respect to other approaches being that these allow to explicitly obtain a smooth estimate also of the PDF.

For convenience of notation, we assume here that χ_{s_j} is supported on the compact domain $[0, 1]$, for $j = 1, \dots, n$; the case of a general compact support $[a, b]$ can be obtained through the variable transformation $x = \frac{(t-a)}{(b-a)}$, with $t \in [a, b]$. Recall that, given a sample $\vec{x}_j = (x_{1j}, \dots, x_{N_j j})$ of i.i.d. observations from (the distribution whose PDF is) χ_{s_j}, a (discontinuous) nonparametric estimator for the CDF $\mathcal{Y}_{s_j}$ is given by the Empirical Cumulative Distribution Function (ECDF), denoted by $\overline{\mathcal{Y}}_{s_j}(t; N_j)$ and defined as

$$\overline{\mathcal{Y}}_{s_j}(t; N_j) = \frac{1}{N_j} \sum_{i=1}^{N_j} I_{x_{ij} < t}. \tag{5.7}$$

Equation (5.7) renders estimates with jump discontinuities in correspondence of the data. Bernstein polynomials can then be introduced to obtain a smooth estimate of $\mathcal{Y}_{s_j}$ from $\overline{\mathcal{Y}}_{s_j}(t; N_j)$. In [29], the following estimator was proposed

$$\hat{\mathcal{Y}}_{s_j}(t; N_j, B_j) = \sum_{k=0}^{B_j} \overline{\mathcal{Y}}_{s_j}(k/B_j; N_j) b_{k,B_j}(t), \tag{5.8}$$

where $b_{k,B_j}(t) = B_j k t^k (1-t)^{B_j-k}$, $k = 0, \ldots, B_j$, and B_j denotes the number of Bernstein polynomials used to smooth the jth ECDF. Estimators (5.7) and (5.8) are strongly consistent for $\mathcal{Y}_{s_j}$, but the latter is also continuous and allows obtaining a smooth estimate of the PDF χ_{s_j} as

$$\tilde{\chi}_{s_j}(t; N_j, B_j) = B_j \sum_{k=0}^{B_j-1} \left(\overline{\mathcal{Y}}_{s_j}((k+1)/B_j; N_j) - \overline{\mathcal{Y}}_{s_j}(k/B_j; N_j) \right) b_{k,B_j-1}(t). \tag{5.9}$$

Unlike the well-known kernel smoothing estimators, estimator (5.9) is suitable to be adopted for compactly supported PDFs. It was adapted to smooth PSCs collected through grain sieve analysis, by considering a modified yet consistent estimator, based on a preprocessing of partially observed ECDF. We note however that other smoothing methods based on Bernstein polynomials have been developed for the same purpose, e.g. [30–32].

A different approach to smooth FCs was proposed in [33], by combining the approaches of FDA and CoDa. These authors developed a B-spline representation for the clr-transformation of an FC, estimated from a discrete clr-transformation applied to the histogram of raw data. This idea is closer to the typical viewpoint employed in the main literature on FDA [1]. Extending FDA methods to the Bayes space setting is often nontrivial. For instance, in the case addressed in [33], the B-spline representation had to imbue through appropriate conditions the zero-integral constraints characterizing clr-transformations. Basis expansions are, however, very useful from the computational viewpoint: the B-spline representation of [33] was used to markedly simplify computations in [14, 34].

In general, most geostatistical methods for FCs developed in the literature are based on the assumption that the data have been already smoothed. As such, the smoothing procedure is seen as a separate step of the analysis, for which the technique of choice – possibly data-driven – can be applied.

5.4.3 An Example of Application

As an illustration of the approach, we consider here the analysis of the dataset of PSCs illustrated in Section 5.3. Here, we focus on the data observed at borehole B5, as in [7].

Menafoglio et al. [7] preprocessed the raw data described in Section 5.3 by smoothing the PSCs through the use of 140 Bernstein polynomials. The density functions of the PSCs were then explicitly computed from the smoothed PSCs (see Section 5.4.2). The latter densities, hereafter called *PSDs*, were interpreted as FCs, and embedded in the Bayes space $\mathcal{B}^2(I)$. In [7], the interval I was set to $I = [\log(0.001), \log(200)]$, considered as the largest range of observation consistent with the type of lithology at the site. Other choices are possible: for instance,

one may consider the distribution of grain-sizes conditional to the range of observation, as proposed in [8] and discussed in Section 5.5.

In [7], a stationary spatial model was considered for the data. Note that in this case, the spatial domain is one-dimensional, as the data at borehole B5 were observed along the vertical coordinate in the range $D = [301.0, 308.3]$ meters above the sea level (m a.s.l.).

Figure 5.3 depicts the smoothed data at borehole B5, together with the empirical estimate of the variogram, estimated through (5.6). The curve in Figure 5.3c denotes the exponential structure with nugget which was fitted to the empirical estimate. Note that, although the empirical variogram might show some degree of nonstationarity, prior knowledge on the field site supports adopting a stationary hypothesis at B5, and was thus considered as a basis assumption of the study. From the application viewpoint, the estimated variogram displays a rapid growth up to a lag of about 0.6 m, where it reaches a sill around a value of 2.4. As such, the range of spatial dependence appears quite small if compared with the width of D (7.3 m). This has a direct impact on predictions, as the ordinary kriging sets

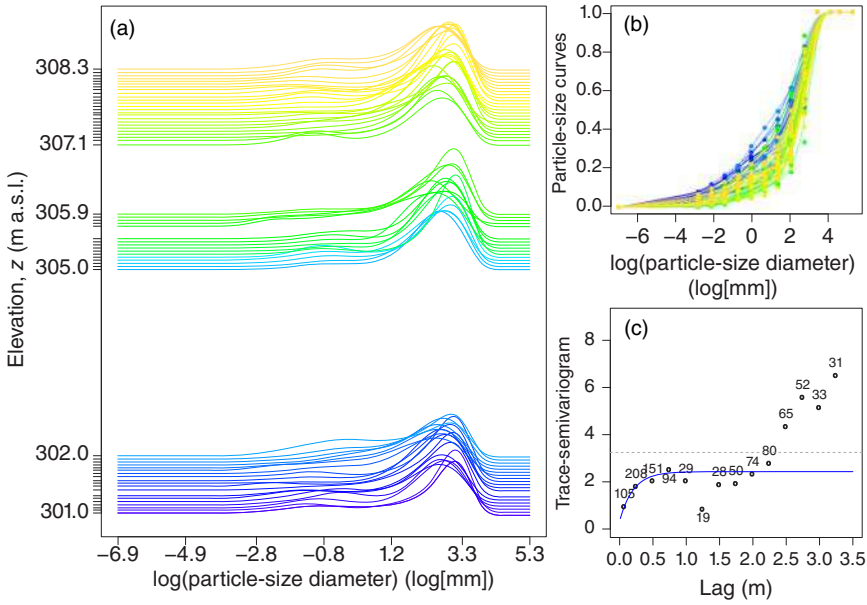


Figure 5.3 (a) Vertical distribution of smoothed densities; (b) raw particle-size curves (symbols) and particle-size curves smoothed by Bernstein polynomials with $m = 140$ (solid curves); (c) estimated trace-semivariogram of the particle-size densities: empirical trace-semivariogram (symbols), fitted model (solid curve), and sample variance (dotted curve); the number of pairs associated with each lag is reported. Source: Modified from Menafoglio et al. [7].

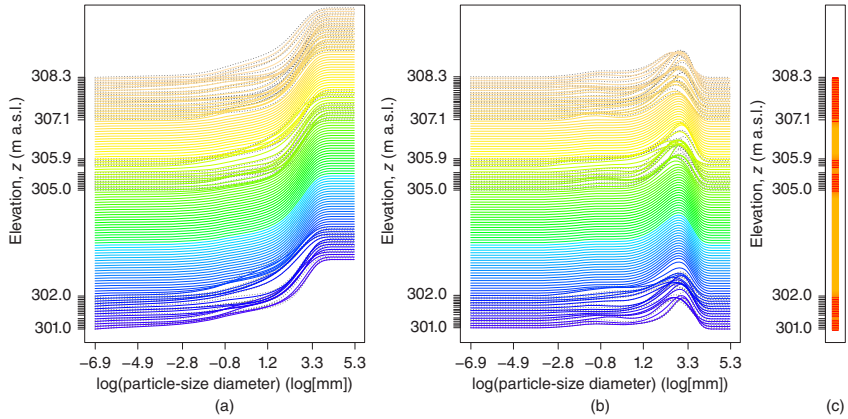


Figure 5.4 Vertical distribution of ordinary kriging predictions results: (a) PSCs: kriged curves (solid curves) and smoothed (dotted curves); (b) PSDs: kriged curves (solid curves) and smoothed (dotted curves); (c) kriging variance, ranging between 0 (darkest shade, corresponding to vertical locations where observations are available) and 2.53 (lightest shade). Source: Modified from Menafoglio et al. [7].

the predictions to the generalized least squares (GLS) estimated mean when the target location is at a distance higher than the variogram range (in this case 0.6 m) from the closest observed site.

Figure 5.4 shows the results of the ordinary kriging in $\mathcal{B}^2(I)$, for a fine grid of target locations along the vertical direction. One can clearly notice that, consistent with our previous remark, the two widest gaps between the sample locations (i.e. the ranges [302.0, 305.0] and [305.9, 307.1]) are mostly predicted with the mean PSD. It is noted that, in cases of such short ranges, the experimental design, i.e. the distribution of the sampling points within the domain (here the vertical dimension), is key to the performance of our predictions. As such, a rigorous assessment of the extent to which the collection of additional information about the system can (i) reduce predictive uncertainty and (ii) yield potential benefits in terms of, e.g. reduced sampling cost and/or risk reduction, is key to improve our understanding of complex natural systems such as groundwater reservoirs. The value of additional information can be quantified through a variety of approaches (see, e.g. [35] and references therein). An example of these – which is relevant to our application – is the multimodel data-worth assessment framework proposed by Neuman et al. [35] and [36] and references therein. The approach is based on a Maximum Likelihood version of the Bayesian Model Averaging (MLBMA) and is consistent with modern statistical methods of parameter estimation. Implementations of MLBMA data-worth assessments considered the geostatistical characterization of aquifer hydraulic conductivity fields in the presence of multiple variogram models (and eventually measured values) [35, 37].

Dealing with functional data through an approach of the kind we illustrate here is of interest, for example, in the context of the hydrogeological characterization of heterogeneity of aquifers and reservoirs. PSCs are routinely assessed from soil samples in modern laboratories through simple and inexpensive procedures. These typically involve the successive use of a series of sieves of decreasing grid size, which are regulated by appropriate international standards. A variety of other methods are also available to extract PSCs from soil samples, including sedigraph, laser diffraction, and dry and wet sieving. The PSCs enable one to characterize a number of effective grain diameters, d_e, defined as the representative particle size diameter in terms of percent in mass, corresponding to the eth percentile of a measured PSC. Having the ability to treat the whole PSC in a consistent geostatistical framework enables us to transfer information not only on hydraulic but also on sedimentological and eventually geochemical parameters which can control solute fluxes in the subsurface.

5.4.4 Uncertainty Assessment

A kriging prediction is optimal in terms of mean squared error within the class of linear unbiased predictors. However, it does not always represent the natural variability of the process: the field realization is usually much "rougher" than a typical kriging map. Quantifying the uncertainty associated with predictions is then key to provide a full characterization of the phenomenon. For this purpose, one may employ the kriging variance, that is the variance of prediction error explicitly expressed at a target location s_0 as

$$\sigma^2_*(s_0) = C(0) - \sum_{i=1}^{n} \lambda_i^* C(\|s_i - s_0\|_d) - \zeta^*, \tag{5.10}$$

where (λ^*, ζ^*) are the solutions of the kriging system (5.5). Indeed, on these bases one can provide Chebyshev bands on the norm of the prediction errors by using the following inequality:

$$P\big(\|\chi_{s_0} \ominus \chi^*_{s_0}\| > \kappa \cdot \sigma_*(s_0)\big) < \frac{1}{\kappa^2}. \tag{5.11}$$

Even though expression (5.11) provides a useful bound on the prediction error, it often proves to be very conservative, as shown in [7]. Indeed, in the study presented in [7] and recalled in Section 5.4.3, the authors estimated via cross-validation that the 75% prediction bands constructed through the Chebyshev inequality (5.11) were associated with an empirical level of 98.3%.

We also remark that the kriging variance does not take into account the uncertainty associated with the estimate of the cross-variogram, as the latter is assumed to be known when formally developing the kriging predictor (see Chapter 2). Hence, prediction bands built on these bases inevitably suffer from being approximate.

Another perspective in assessing the uncertainty of the estimate is that of generating multiple realizations of the field, compatible with the data. This approach was recently pursued in [9], that proposed a methodology for geostatistical simulation in Bayes spaces. The idea upon which the method is grounded is to reproduce the variability of the phenomenon – which is only partially represented by kriging maps – by drawing samples from the conditional distribution of χ_{s_0} given $\chi_{s_1}, \dots, \chi_{s_n}$. Accordingly, if the procedure is performed for multiple target locations in D, one can obtain a set of maps that, although suboptimal, provide an improved representation of the natural variability and are still "compatible" with the data, in the sense that they coincide with the data at the measurement locations (as well as kriging maps do).

Before briefly describing the method, we illustrate the results on the field data of Sections 5.3 and 5.4.3. Figure 5.5b displays an example of a realization from the conditional field $\{\chi_s | \vec{\chi}, s \in D\}$, with $\vec{\chi} = (\chi_{s_1}, \dots, \chi_{s_n})^T$. It is apparent that the spatial variability associated with the realization is much higher than that of the kriged field, displayed in Figure 5.5a.

Performing repeated conditional simulations leads to generate a wide range of scenarios that could have been observed with the same data. As a way of example, Figure 5.5c,d depicts a sample of 1000 conditional simulations at elevations 303.0 and 306.0 m a.s.l., the corresponding prediction being depicted as black curves in Figure 5.5a. The amplitude of the gray shade can be used to qualitatively represent the variability of the predictions at the target location. Note that, although the predicted curves at elevations 303.0 and 306.0 m a.s.l. show some similarities, the associated uncertainty is indeed different: at an elevation of 303.0 m a.s.l., the variability is much higher due to the absence of data nearby that location.

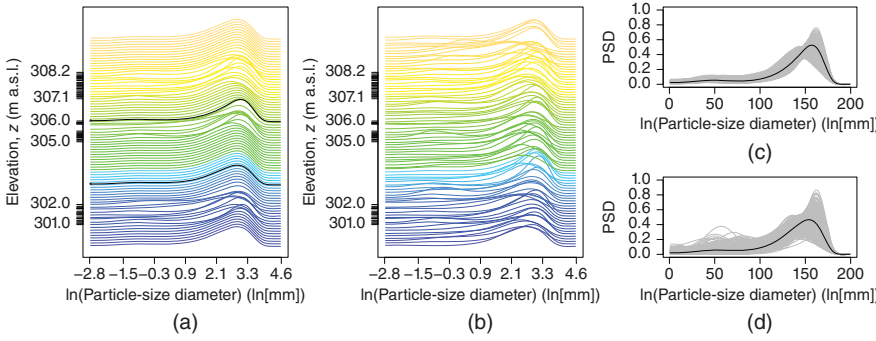


Figure 5.5 Kriged field and conditional realizations. (a) Kriging estimation over a grid along the vertical direction; black curves indicate predictions at elevations of 303.0 and 306.0 m a.s.l. (b) A conditional realization on the same grid considered in panel (a). (c, d) Kriging estimation at elevations of 306.0 and 303.0 m a.s.l. (black curve) and a sample of 1000 conditional simulations at the same sites (gray curves).

From a theoretical viewpoint, generating realizations from the distribution of $\chi_{s_0}|\vec{\chi}$ is a problem of random generation in infinite dimension. It is clear that a strategy based on joint simulations of pointwise values of the curves would not be affordable either from the theoretical or the computational viewpoint. It is also noticed that a global approach as that used for ordinary kriging did not prove to be successful, as the trace-covariogram seems to be insufficient for the characterization of the spatial dependence structure for simulation purposes.

In [9], the authors proposed a simulation strategy based on an optimal dimensionality reduction of the problem in the Bayes Hilbert space. Specifically, to provide a conditional realization at a target location s_0, they proposed to

(i) perform a functional principal component in $\mathcal{B}^2(I)$ [14] and compute the scores along the first K principal components (where K is sufficiently high to represent the data variability);
(ii) model the spatial dependence of the multivariate random field of the scores and perform geostatistical simulation of the latter, through any of the widely employed geostatistical techniques for the simulation of multivariate random fields.

The reason that led the authors to choose a dimensionality reduction step based on principal component analysis is that it provides nested optimal approximations of the observations for any finite order K. The optimal choice for K is critical because it controls the quality of the approximation of the data through the principal components, and the complexity of step (ii) (thus the computational effort involved in the actual computations). To set K, well-known methods in principal component analysis can be employed, e.g. looking for an elbow in the scree-plot, compatible with the computational power available.

Our methodology and type of results can be readily transferred to the general context of numerically based Monte Carlo simulations of flow and transport processes in environmentally and industrially relevant scenarios, including, e.g. groundwater systems, oil reservoirs, and shale gas formations. A critical element in these applications is to have at our disposal multiple realizations of (i) the heterogeneous structure of the porous/fractured system (in terms of the spatial arrangements of geomaterials/hydrofacies), and (ii) the distribution of properties such as porosity and hydraulic conductivity within each of the identified facies. This enables us to propagate uncertainty associated with the reconstruction of the subsurface onto uncertainties characterizing target variables of interest such as local composition of soil, pressure heads, dissolved chemical concentrations, reaction rates, and fluid saturations. All of these elements will constitute avenues of future development and exploitation of the approach we present in this work.

5.5 Analyzing Nonstationary Fields of FCs

In several real cases, the field data cannot be consider either stationary or isotropic. For instance, one may have prior information about possible secondary variables which have an influence on the response. The latter need to be taken into account in the geostatistical model and exploited for prediction purposes, e.g. in a Universal Kriging setting. A particular case in this broad context is the situation in which data are featured by a grouping structure. This case was addressed in [8] and was motivated by the analysis of the entire dataset of PSCs described in Section 5.3 for which the existence of different soil types was observed. In this case, the field was also found to be anisotropic. In this section, we recall the kriging method of [8] – termed *class-kriging* – for the prediction of anisotropic random fields of grouped FCs.

Throughout the section, we consider the setting in which the field of FCs $\{\chi_s, s \in D\}$, is observed together with a secondary field $\{T_s, s \in D\}$, whose elements represent random labels associated with the grouping structure of the data. For instance, they may represent soil types, in case PSCs are observed in a heterogeneous system, but may also represent climatic regions, if weather data over a large region are concerned instead.

The random elements T_s, $s \in D$, are discrete variables. We call $\tau^{(1)}, \dots, \tau^{(K)}$ the K values which may be taken by the T_s (i.e. the labels of the K possible groups), and denote by $(\chi_{s_1}, \tau^{(k_1)}), \dots, (\chi_{s_n}, \tau^{(k_n)})$ the pairs of FCs and labels observed at the measurement locations $s_1, \dots, s_n$. In [8], the authors proposed to model the field $\{\chi_s, s \in D\}$, conditional to the field of labels $\{T_s, s \in D\}$ as the sum (in $\mathcal{B}^2$) of a drift term dependent on the label at s, and a stationary residual, independent of the grouping structure. Formally,

$$\chi_s | \{T_s = \tau^{(k)}\} = m^{(k)} \oplus \delta_s,$$

where $m^{(k)} = \mathbb{E}[\chi_s | T_s = \tau^{(k)}]$ denotes the drift, and $\{\delta_s, s \in D\}$ is a random field of FCs, with "zero-mean" in $\mathcal{B}^2$, i.e. with mean coinciding with the neutral element of perturbation $\mathbb{E}[\boldsymbol{\delta}_s] = 0_{\oplus} = 1/\eta$. The random field $\{\delta_s, s \in D\}$ is also assumed to be (i) independent of the field of labels $\{T_s, s \in D\}$, and (ii) globally second-order stationary (possibly anisotropic), with trace-covariogram C and trace-variogram γ:

$$\begin{aligned} C(s_1 - s_2) &= \mathbb{E}[\langle \boldsymbol{\delta}_{s_1}, \boldsymbol{\delta}_{s_2} \rangle], \\ 2\gamma(s_1 - s_2) &= \mathbb{E}\left[\| \boldsymbol{\delta}_{s_1} \ominus \boldsymbol{\delta}_{s_2} \|^2\right], \quad s_1, s_2 \in D. \end{aligned}$$

This model can be framed in the Universal Kriging setting introduced in Chapter 2. Indeed, denote by $\{\psi_k(\boldsymbol{s}), k = 1, \dots, K-1\}$ a set of binary variable,

which represent indicators associated with the labels: for $k = 1, \dots, K-1$, $\psi_k(\boldsymbol{s}) = 1$ if $T_{\boldsymbol{s}} = \tau^{(k)}$, and $\psi_k(\boldsymbol{s}) = 0$ otherwise; if $T_{\boldsymbol{s}} = \tau^{(K)}$, then $\psi_k(\boldsymbol{s}) = 0$ for every $k = 1, \dots, K-1$. The drift in $\boldsymbol{s} \in D$ can then be described through a linear model in $\mathcal{B}^2$, with these indicators as regressors

$$\mathbb{E}[\mathcal{Y}_{\boldsymbol{s}} \big| \Pi_{\boldsymbol{s}} = \boldsymbol{\pi}_{\boldsymbol{s}}, T_{\boldsymbol{s}} = \tau^{(k)}] = a_0 \oplus \oplus_{l=1}^{K-1} \psi_l(\boldsymbol{s}) \odot a_l, \tag{5.12}$$

where $a_0, \dots, a_{K-1}$ are (possibly unknown) deterministic coefficients in $\mathcal{B}^2$. In the light of model (5.12), one has

$$\begin{cases} m^{(k)} = a_0 \oplus a_k, & k = 1, \dots, K-1, \\ m^{(k)} = a_0, & k = K. \end{cases} \tag{5.13}$$

Coefficients $a_0, \dots, a_{K-1}$ thus represent how different the drift in the kth group is from that of a reference group, which is here set to the Kth group, without loss of generality.

In [8], the authors relied on the Universal Kriging results introduced in Chapter 2, to propose a class-kriging predictor for χ_{s_0} at a target location s_0, given the realization of $\{T_s, s \in D\}$ in D (i.e. the grouping structure over the entire spatial domain). The class-kriging predictor is defined as $\chi_{s_0} = \oplus_{i=1}^{n} \lambda_i^* \cdot \chi_{s_i}$, whose weights minimize the (conditional) variance of prediction error under the unbiasedness constraint, that is, solve

$$\begin{aligned} &\min_{\substack{\lambda_1, \dots, \lambda_n \in \mathbb{R}: \\ \chi_{s_0}^{\vec{\lambda}} = \oplus_{i=1}^{n} \lambda_i \odot \chi_{s_i}}} \mathbb{E}\left[\| \chi_{s_0}^{\vec{\lambda}} \ominus \chi_{s_0} \|^2 \Big| T_{s_0} = \tau^{(k_0)}, T_{s_i} \in \tau^{(k_i)}, i = 1, \dots, n \right] \\ &\text{subject to } \mathbb{E}\left[\chi_{s_0}^{\vec{\lambda}} \Big| T_{s_0} = \tau^{(k_0)}, T_{s_i} \in \tau^{(k_i)}, i = 1, \dots, n \right] = m^{(k_0)}. \end{aligned} \tag{5.14}$$

The drift being linear, the optimal weights are found by solving the system of $n + K$ linear equations, obtained by embedding in the Universal Kriging setting model (5.12):

$$\begin{pmatrix} \Sigma & \Psi \\ \Psi^T & 0 \end{pmatrix} \begin{pmatrix} \vec{\lambda} \\ \vec{\zeta} \end{pmatrix} = \begin{pmatrix} \vec{\sigma}_0 \\ \vec{\psi}_0 \end{pmatrix}, \tag{5.15}$$

where $\vec{\zeta} = (\zeta_0, \dots, \zeta_{K-1})^T$ are K Lagrange multipliers associated with the unbiasedness constraint, whereas $\vec{\sigma}_0 = (C(\boldsymbol{h}_{i,0})) \in \mathbb{R}^n$, and $\vec{\psi}_0 = (\psi_k(\boldsymbol{s}_0)) \in \mathbb{R}^K$.

From the application viewpoint, several critical points may be encountered in class-kriging. First, the estimate of the spatial dependence structure is crucial to solve (5.15). Here, all the methods described in Chapter 2 can be employed. In particular, one may resort to an iterative algorithm to estimate the drift via generalized least squares and jointly estimate the residual variogram 2γ. More delicate is the case in which the field $\{T_s, s \in D\}$ is only observed at the measurement locations,

or if it is completely latent. In [8], methods to deal with both the situations were developed. For the first case (i.e. the labels are only observed at $s_1, \ldots, s_n$), one needs to formulate a model for the stochastic distribution of $\{T_s, s \in D\}$, and then to employ such a model to predict the T_s at unsampled locations. For instance, the T_s may be modeled as independent realizations from a multinomial variable, and the interpolation can be consistently performed via indicator kriging, as in [8]. When the field is completely latent, one needs additionally to cluster the data. Although several methods are available for spatial clustering of scalar data, little attention has been paid so far to the problem of spatial clustering of FCs. In [8], the authors proposed a spatial K-mean clustering, which is an extension of the K-mean method, tailored on model (5.12). Other methods could be applied to this purpose, for instance the Bagging Voronoi method illustrated in Chapter 9.

As an illustration of the class-kriging method, we illustrate the results of its application to the dataset described in Section 5.3, following [8]. Unlike the case discussed in Section 5.4.3, the authors focused on these restriction of the PSDs to the actual domain of observation, because for most data no information was available on the left tail of the PSC, due to the sieve measurement procedure. Figure 5.6 displays the full dataset of smoothed PSDs at the site. The colors of the symbols in Figure 5.6b denote the three soil types identified in the study region, associated with as many groups in the data.

A geometric anisotropy was found by the authors when looking at directional variograms. This was corrected by scaling the vertical dimension by a factor $r = 25$, thus working in the modified spatial domain where an isotropic model

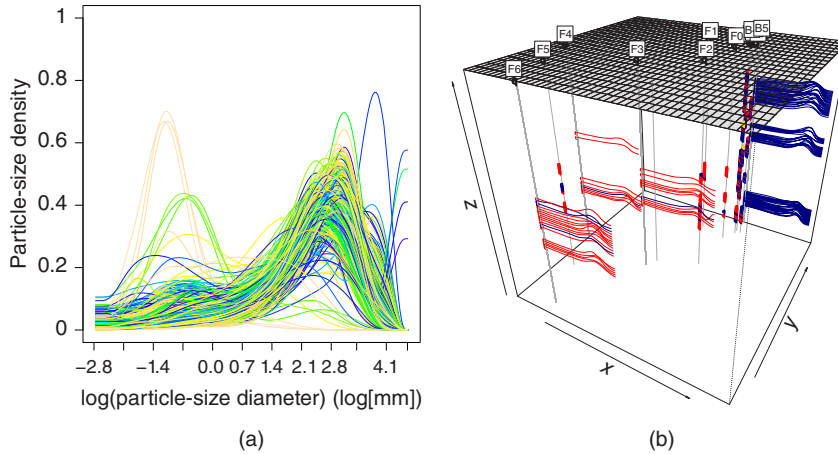


Figure 5.6 Field data: (a) smoothed PSDs; (b) soil types at the field site (denoted with gray colors) and smoothed PSDs along the boreholes B5, F3, F4, and F6. Source: Modified from Menafoglio et al. [8].

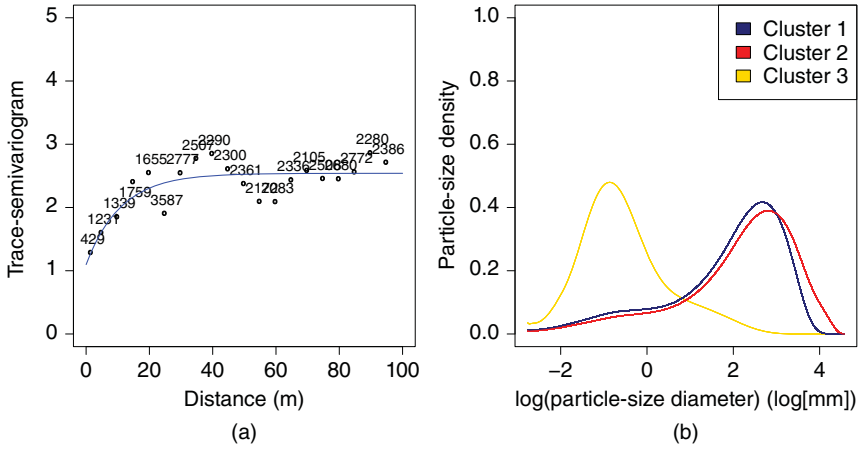


Figure 5.7 (a) Estimated trace-semivariogram of the residuals; (b) estimated drift for the three soil types. Source: Modified from Menafoglio et al. [8].

can be used. The estimated omnidirectional trace-semivariogram of the residuals is displayed in Figure 5.7a, together with the fit of an exponential model with nugget. Similarly, as in Section 5.4.3, the range of the variogram appears to be quite short when compared to the extension of the domain. The drift estimated within the groups is reported in Figure 5.7b. Here, the first two groups are interpreted as a characterization of two diverse behaviors within the right tail of the PSD, the first cluster featuring a lighter tail than the other one. The third group, formed by 1% of the sample, is associated with a drift displaying its main peak at a grain size of about 0.4 mm. As shown in Figure 5.6b, the first group is mainly associated with the boreholes B1–B5 and the second group with the boreholes F0–F6. The former group of boreholes is located in an area where the Neckar River displays a bend, and thus favors the accumulation of the finer sediments in this area. The PSDs at borehole B5 – considered in Section 5.4.3 – belongs to the first group, consistent with the stationarity assumption considered before.

Figure 5.8 finally displays the prediction of the field in some unsampled locations. The kriged field is a smooth interpolation of the available data. The outlying observations, such as the curve at $z = 305.5$ at borehole F6, influence prediction results only locally. For distances higher than the estimated range, the kriged field is representative of mean particle-size distribution associated with the soil type at the target location.

Uncertainty assessment of such predictions is nontrivial, as it should take into account the prediction variance as well as the uncertainty in parameters estimate (variogram and drift). For this purpose, an extension of the simulation methods discussed in Section 5.4.4 can be considered.

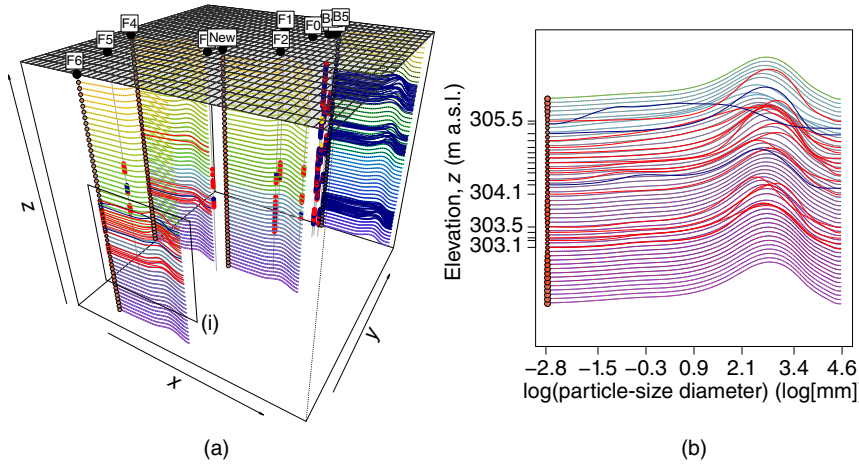


Figure 5.8 Class-kriging of PSDs: (a) results at boreholes B5, F4, and F6 and at an unsampled vertical (denoted as "New"). (b) Vertical distribution of predicted PSDs, for the group of samples at elevations $301 \leq z \leq 306$ m above sea level (a.s.l.), at borehole F6. In both panels: gray colors of the solid curves indicate depth; gray-scale colors of the symbols denote soil type. Smoothed data are represented with solid curves colored according to the cluster assignment. Source: Modified from Menafoglio et al. [8].

5.6 Conclusions and Perspectives

We here illustrated methods for the geostatistical analysis of FCs, which combine the perspective of Object Oriented Spatial Statistics (O2S2, [38]), with that of Compositional Data Analysis in Bayes spaces. The main points addressed in this chapter can be summarized as follows:

(1) FCs, such as PDFs, should not be considered just as data in the space L^2, but one should pay close attention to treat them through an appropriate geometry. A possible sensible geometry is that of Bayes Hilbert spaces. Although we focused on FCs with compact support, the theory of Bayes spaces is available for FCs with support in noncompact set, provided that a reference measures other than the Lebesgue one is considered.
(2) Stationary and nonstationary methods are available to predict FCs in Bayes spaces – and particularly PDFs – by relying on the theory of Universal Kriging in Hilbert spaces of [39]. Unlike traditional methods based on selected features of distributional data (e.g. moments or quantiles), predicting the entire PDF allows taking into account the entire information content of the data, and project it to unsampled location in the system.

(3) Uncertainty assessment is key for a full characterization of the field under study. Here, we illustrated a method for stochastic simulation grounded upon dimensionality reduction in Bayes spaces. Although only the stationary case was considered, the extension of the strategy to the nonstationary setting can be readily envisaged.
(4) Throughout the chapter, we illustrated the methodologies with a real case study, dealing with PSDs. Advancements of the work illustrated here include embedding our theoretical and operational framework in the context of (forward and inverse) stochastic analyses of subsurface flow and transport in heterogeneous media by way of numerical Monte Carlo simulations or groundwater flow and transport Moment Equations (e.g. [36, 40–42] and references therein).
(5) Future perspectives for application of the approach include the analysis of environmental and Earth system variables whose main characteristics can be encapsulated in terms of a functional behavior. In addition to PSCs of the kind we examine here, these might comprise, for example, relative permeability curves (which are relevant in multiphase flow settings), mineralogic composition of rocks (for the geochemical characterization of a host reservoir), as well as breakthrough curves of dissolved chemical migrating in water bodies and/or chemical composition of fluids sampled at multiple locations in an aquifer system (with implication on human exposure and health hazards).

References

1 Ramsay, J. and Silverman, B.W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer-Verlag.
2 Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications*. Wiley.
3 Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)* 44 (2): 139–177.
4 Egozcue, J.J., Díaz-Barrero, J.L., and Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica, English Series* 22 (4): 1175–1182.
5 van den Boogaart, K., Egozcue, J.J., and Pawlowsky-Glahn, V. (2010). Bayes linear spaces. *SORT* 34 (2): 201–222.
6 van den Boogaart, K.G., Egozcue, J., and Pawlowsky-Glahn, V. (2014). Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics* 56: 171–194.
7 Menafoglio, A., Guadagnini, A., and Secchi, P. (2014). A Kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment* 28 (7): 1835–1851.

8 Menafoglio, A., Secchi, P., and Guadagnini, A. (2016). A class-kriging predictor for functional compositions with application to particle-size curves in heterogeneous aquifers. *Mathematical Geosciences* 48: 463–485.

9 Menafoglio, A., Guadagnini, A., and Secchi, P. (2016). Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a Bayes space approach. *Water Resources Research* 52 (8): 5708–5726.

10 Egozcue, J., Pawlowsky-Glahn, V., Tolosana-Delgado, R. et al. (2013). Bayes spaces: use of improper distributions and exponential families. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas* 107 (2): 475–486.

11 Egozcue, J. (2009). Reply to "On the Harker variation diagrams; …" by J.A. Cortes. *Mathematical Geosciences* 41 (7): 829–834.

12 Pawlowsky-Glahn, V. and Egozcue, J.J. (2001). Geometric approach to statistical analysis in the symplex. *Stochastic Environmental Research and Risk Assessment* 15: 384–398.

13 Delicado, P. (2011). Dimensionality reduction when data are density functions. *Computational Statistics and Data Analysis* 55 (1): 401–420.

14 Hron, K., Menafoglio, A., Templ, M. et al. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics and Data Analysis* 94: 330–350.

15 Riva, M., Sánchez-Vila, X., Guadagnini, A. et al. (2006). Travel time and trajectory moments of conservative solutes in two-dimensional convergent flows. *Journal of Contaminant Hydrology* 82: 23–43.

16 Riva, M., Guadagnini, A., Fernández-Garcáa, D. et al. (2008). Relative importance of geostatistical and transport models in describing heavily tailed breakthrough curves at the Lauswiesen site. *Journal of Contaminant Hydrology* 101: 1–13.

17 Riva, M., Guadagnini, L., and Guadagnini, A. (2010). Effects of uncertainty of lithofacies, conductivity and porosity distributions on stochastic interpretations of a field scale tracer test. *Stochastic Environmental Research and Risk Assessment* 24: 955–970.

18 Menafoglio, A., Guadagnini, A., and Secchi, P. (2016). Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a Bayes space approach. *Water Resources Research* 52: 5708–5726.

19 Hoffmann, R. and Dietrich, P. (2004). An approach to determine equivalent solutions to the geoelectrical 2D inversion problem. *Journal of Applied Geophysics* 56 (2): 79–91.

20 Rein, A., Hoffmann, R., and Dietrich, P. (2004). Influence of natural time - dependent variations of electrical conductivity on dc resistivity measurements. *Journal of Hydrology* 285 (1–4): 215–232.

21 Neuman, S.P., Blattstein, A., Riva, M. et al. (2007). Type curve interpretation of late-time pumping test data in randomly heterogeneous aquifers. *Water Resources Research* 43 (10): W10421.

22 Neuman, S.P., Riva, M., and Guadagnini, A. (2008). On the geostatistical characterization of hierarchical media. *Water Resources Research* 44 (2): W02403.

23 Lessoff, S.C., Schneidewind, U., Leven, C. et al. (2010). Spatial characterization of the hydraulic conductivity using direct-push injection logging. *Water Resources Research* 46: W12502. https://doi.org/10.1029/2009WR008949.

24 Barahona-Palomo, M., Riva, M., Sánchez-Vila, X. et al. (2011). Quantitative comparison of impeller flowmeter and particle-size distribution techniques for the characterization of hydraulic conductivity variability. *Hydrogeology Journal* 19 (3): 603–612.

25 Händel, F. and Dietrich, P. (2012). Relevance of deterministic structures for modeling of transport: the Lauswiesen case study. *Groundwater* 50: 935–942. https://doi.org/10.1111/j.1745-6584.2012.00948.x.

26 Sack-Kühner, B. (1996). Einrichtung des naturmessfeldes "lauswiesen tübingen", erkundung der hydraulischen eigenschaften, charakterisierung der untergrundheterogenität und vergleich der ergebnisse unterschiedlicher erkundungsverfahren. MSc thesis. University of Tübingen, Geological Institute.

27 Martac, E. and Ptak, T. (2003). Data Sets for Transport Model Calibration/Validation, Parameter Upscaling Studies and Testing of Stochastic Transport Models/Theory. Report D16 of Project "Stochastic Analysis of Well-Head Protection and Risk Assessment - W-SAHaRA". *EU contract EVK1-CT-1999-00041*. Milan, Italy.

28 Beyer, W. (1964). Zur bestimmung der wasserdurchlässigkeit von kiesen und sanden aus der kornverteilungskurve. *Wasserwirtschaft-Wassertechnik (WWT)* 14 (6): 165–168.

29 Babu, G.J., Canty, A.J., and Chaubey, Y.P. (2002). Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference* 105: 377–392.

30 Manté, C. (1999). The use of regularization methods in computing Radon–Nikodým derivatives. Application to grain-size distributions. *SIAM Journal on Scientific Computing* 21 (2): 455–472.

31 Manté, C. (2012). Application of iterated Bernstein operators to distribution function and density approximation. *Applied Mathematics and Computation* 218: 9156–9168.

32 Manté, C. (2015). Iterated Bernstein operators for distribution function and density estimation: balancing between the number of iterations and the polynomial degree. *Computational Statistics and Data Analysis* 84: 68–84.

33 Machalová, J., Hron, K., and Monti, G.S. (2016). Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics* 43 (8): 1419–1435.

34 Talska, R., Menafoglio, A., Machalova, J. et al. (2017). Compositional Regression with Functional Response, *Mox report 27/2017*, Politecnico di Milano.

35 Neuman, S. P., Xue, L., Ye, M., and Lu, D. (2012). Bayesian analysis of data-worth considering model and parameter uncertainties. *Advances in Water Resources* 36: 75–85.

36 Xue, L., Zhang, D., Guadagnini, A., and Neuman, S. (2014). Multimodel bayesian analysis of groundwater data worth. *Water Resources Research* 50: 8481–8496.

37 Lu, D., Ye, M., Neuman, S. P., and Xue, L. (2012). Bayesian analysis of data-worth applied to unsaturated fractured tuffs. *Advances in Water Resources* 35: 69–82.

38 Menafoglio, A. and Secchi, P. (2017). Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. *European Journal of Operational Research* 258 (2): 401–410.

39 Menafoglio, A., Secchi, P., and Dalla Rosa, M. (2013). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics* 7: 2209–2240.

40 Guadagnini, A. and Neuman, S. (2001). Recursive conditional moment equations for advective transport in randomly heterogeneous velocity fields. *Transport in Porous Media* 42 (1/2): 37–67.

41 Morales Casique, E., Neuman, S., and Guadagnini, A. (2006). Nonlocal and localized analyses of nonreactive solute transport in bounded randomly heterogeneous porous media: Computational analysis. *Advances in Water Resources* 29: 1399–1418.

42 Panzeri, M., Riva, M., Guadagnini, A., and Neuman, S. (2015). EnKF coupled with groundwater flow moment equations applied to Lauswiesen aquifer, Germany. *Journal of Hydrology* 521: 205–216.

6

Spatial Functional Data Analysis for Probability Density Functions: Compositional Functional Data vs. Distributional Data Approach

Elvira Romano[1], *Antonio Irpino*[1], *and Jorge Mateu*[2]

[1] *Department of Mathematics and Physics, University of Campania "Luigi Vanvitelli", Caserta, Italy*
[2] *Department of Mathematics, University Jaume I of Castellon, Spain*

Spatial functional statistics (SFS) is a recent research area combining functional and spatial statistics for dealing with large, complex, and high-dimensional spatiotemporal data [1]. Starting from the functional data analysis (FDA) paradigm [2, 3], SFS also takes into account the spatial structure of the data. Differently from the spatiotemporal framework, the key characteristics of spatiofunctional techniques consist in defining methods making no parametric assumptions about time effects. The emerging characteristics about the developed methods are related to the three possible spatial data structures (geostatistical, spatial point pattern, and lattice data) that can be combined with functional data.

Geostatistical data refer to spatial data varying into a continuous space [4]. Examples can be found in many fields of science, for example dealing with soil ph in water.

Spatial point pattern data are related to data in the form of a set of points, or "events" of interest. Examples include locations of trees in a forest, of nests in a breeding colony of birds, or of nuclei in a microscopic section of tissue [5]. Lattice data involves regular or irregular spaced points linked to a nearest neighbor structure [6]. Remote sensing data from satellite offers a wealth example. Differences among geostatistical, spatial point pattern, and lattice data consist in the different nature of the spatial domain of definition. In analogy, in SFS, data can be distinguished into geostatistical functional methods, functional marked point data, and functional areal data [7]. Formally, following Delicado et al. [8], a spatial functional process can defined as

$$\left\{\chi_s : s \in D \subseteq \Re^d\right\}, \tag{6.1}$$

where s is a generic data location in the d-dimensional Euclidean space and χ_s are functional random variables, defined as random elements taking values in a

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

functional space. When $D \subseteq \mathfrak{R}^d$ has a positive volume and the n functions are observed in n points $s_1, \ldots, s_n$, it is usual to refer to them as geostatistical functional data. When a function is observed at each point s generated by a point process, it corresponds to a functional marked point process. In the case of D being a fixed and countable set, and there is a bijection between D and a partition of a geographical area, we are then in the context of functional areal (or lattice) data.

In many applied contexts, the functional variables are distributions of quantity of interest. Let us think, for instance in income distributions for several cities of a country, in distributions of the fraction of bids submitted in an online purchasing related to different proxy locations, and in distributions of functional magnetic resonance imaging (fmRI) signals on medical images. In this case, the analysis needs to give rise to the functional structure of the georeferenced distribution that can be seen a special case of a functional variable.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and consider the process:

$$\left\{f_s, s \in D \subseteq \mathfrak{R}^d\right\}$$

and such that for $s \in D \subseteq \mathfrak{R}^d,\ \omega \in \Omega$, the data structure can be expressed by

$$f_s : I \to (0, +\infty), \text{ s.t. } \int_I f_s(t)dt = 1,$$

where $I = [a, b]$ is a compact interval of $\mathfrak{R}$.

In these situations, where the functions are probability density functions (PDFs) located in the space, the analysis becomes difficult since constrains on density functions need to be considered [9, 10]. Density functions are nonnegative and integrate to 1, thus they do not live in a vector space, such as the Hilbert space usually assumed in FDA methods. A different space of support is needed.

Even though methods for dealing with density functions without considering any information on the spatial dependence are a growing area, methodologies developed for carrying out their analysis in the geostatistical field are very few, and they are mainly related to prediction problems [10, 11].

Two main approaches encompass statistical methodologies for data described by density functions: Symbolic Data Analysis (SDA) [12], using a distributional approach [13], and FDA, through a compositional interpretation [9, 14–16].

A distinctive feature of these types of analysis is that they make use of the information inherent to PDF in the case of distributional data (SDA approach), or quantile functions, and by the infinite dimensional compositional data (namely, functional compositional data) (FDA approach).

Suppose we have a realization of $f_s(\cdot), s \in D$ at a finite set of locations $s_1, \ldots, s_n$, and we focus on a geostatistical method to describe their spatial correlation. In particular, we aim to compare and to illustrate the advantages and peculiarities of SDA and FDA approaches by implementing a local indicator of spatial association.

Thus, at first, we present a theoretical comparison of the FDA and SDA analysis when data are density functions without the spatial constrains. Then, we focus on the potentials and the drawbacks of the two approaches by evaluating their performances in spatial FDA. Especially, we extend a local spatial association index, the local Moran's I [17], to evaluate spatial dependence among density functions

The rest of the chapter is organized as follows: in Section 6.1, we introduce the current state-of-the-art and the formal framework of the two different statistical approaches: SDA by a distributional approach and FDA by means of a compositional approach.

Section 6.2 shows an extension of the local Moran's I for density functions that are spatially located. In Section 6.3, we apply the proposed indexes on an aerial data set coming from official statistics in United States. The chapter ends with some final conclusions.

6.1 FDA and SDA When Data Are Densities

As for the statistical techniques on standard data, the analysis of density functions aims at representing or visualizing and at explaining or predicting the variability of data. However, two main challenges arise when dealing with this kind of data. First of all, since a generic density function $f : I \to (0, +\infty)$, with $I = [a, b] \subset \mathfrak{R}$ is constrained by $\int_I f(t)dt = 1$, the functional space where densities live is convex but not linear, and the classic FDA methods as well as the classic data analysis methods result to be inappropriate. The main difficulty consists in the fact that the observations belong to a constrained infinite dimensional space and constrains that are not properly taken into account in the L_1 space where density functions live.

As shown by Petersen and Muller [15], the set of probability functions is a convex subset of L_1 that do not have a linear space structure when using the ordinary sum and multiplication by real constants.

Moreover, all the usual distances, such as the L_1 distance and the L_2 distance between square root density, are not invariant under relevant transformations of densities.

In the last two decades, two main approaches are facing these challenges: the distributional and FDA.

These two different approaches have proposed methods and techniques for the analysis of PDFs starting by different characterizations of distributions. A distribution can be described by a density (or probability) function f_i or by a cumulative distribution function (*cdf*) F_i or by a quantile function (*qf*) F_i^{-1}.

In a "functional" perspective, the specific features of density functions are accounted through the generalization to the infinite dimensional setting of the Aitchison geometry for compositional data, the so-called "Bayes spaces" [18].

A Bayes space is a linear space of equivalence classes of proportions, including probability measures. It can be seen as the sample space of random compositions, and thus, a generalization of the Euclidean structure of the simplex. This space ensures an interesting geometric representation of density functions that permits further analysis.

In a "distributional" perspective, density functions are analyzed by means of quantile functions. Properties and characteristics of these functions are captured by the use of a Wasserstein metric enlarging methods based on a measure of inertia equivalent to inertia in the Euclidean space.

In the following, we introduce features of density functions as compositional functional data and as distributional data.

6.1.1 Features of Density Functions as Compositional Functional Data

Density functions can be considered as a special case of functional data [3], and their proper statistical treatment represents a challenging task in FDA.

The Aitchison geometry in a simplex, defined as a particular case of a more general space with $\mathcal{P}$-densities and σ-measures called Bayes linear space, and has been a convenient and a simple way for dealing with their compositional nature.

In this section, focusing on the definition of the Aitchison geometry in a simplex, we recall and discuss the main characteristics of the Bayes space for dealing with density functions in a functional space.

Denote the space of continuous and strictly positive density functions by

$$F(I) = \left\{ f : I \to \mathbb{R}, \text{such that } f \geq 0 \ \forall t \in I, \int_I f(t)dt = 1 \right\}$$

with $I = [a, b]$ a compact subset of $\mathbb{R}$.

Due to the inherent constraints related to the unit–integral constraint and to the relative contributions of the Borel sets of real line to the overall probability, the usual operations of the classic $L_2(I)$ defined as

$$L_2(I) = \left\{ f : I \to \mathbb{R}, \text{such that } \|f(t)\| = \int_I (f(t))^2 dt < \infty \right\}, \tag{6.2}$$

are not appropriate.

Considering the compositional nature of density functions, Egozcue et al. [14] proposed to extend the compositional Aitchison geometry to the subset of bounded PDF over I defined by

$$\mathcal{A}_B{}^2(I) = \left\{ f : I \to \mathbb{R}, \text{ such that } f \geq 0 \text{ and } \log(f) \in L_2(I) \right\}$$

that is the set of nonnegative real functions on a compact domain I whose logarithm is square integrable.

Assuming a uniform reference measure, with the aim of finding a generating system and to express compositions in coefficients of coordinate system, a transformation of a density function is obtained by a log-ratio transformation of the original data by an isomorphism $\psi_A : F(I) \to {\mathcal{A}_B}^2(I)$.

Let f, g be two elements of ${\mathcal{A}_B}^2(I)$ and $\eta = b - a$, to obtain a Euclidean space structure, the following inner product, with associated norm and distance are defined.

The inner product $\langle \cdot, \cdot \rangle_{\mathcal{A}} : {\mathcal{A}_B}^2(I) \times {\mathcal{A}_B}^2(I) \to \mathbb{R}$ is defined through the functional

$$\langle f, g \rangle_A = \frac{1}{2\eta} \int_I \int_I \log \frac{f(t)}{f(s)} \log \frac{g(t)}{g(s)} dtds \tag{6.3}$$

that can be rewritten as

$$\int_I [\log f(t) \log g(t)] dt - \frac{1}{\eta} \int_I \log f(t) dt \int_I \log f(s) ds.$$

For $f \in {\mathcal{A}_B}^2(I)$, the norm $\|\cdot\|_{\mathcal{A}} : {\mathcal{A}_B}^2(I) \to \mathbb{R}$ associated with the inner product defined in (6.3) is given by

$$\|f\|_{\mathcal{A}} = \left[\int_I \log^2 f(t) dt - \frac{1}{\eta} \left(\int_I \log f(t) dt \right)^2 \right]^{\frac{1}{2}}.$$

In addition, for $f, g \in {\mathcal{A}_B}^2(I)$, the associate distance $d_{\mathcal{A}}(f, g)$ can be defined as

$$d_{\mathcal{A}}(f, g)(x) = \left[\frac{1}{2\eta} \int_I \int_I \left(\log \frac{f(t)}{f(s)} - \log \frac{g(t)}{g(s)} \right)^2 dtds \right]^{1/2}. \tag{6.4}$$

The obtained metric and normed space $({\mathcal{A}_B}^2(I), \|\cdot\|_{\mathcal{A}}, d_{\mathcal{A}}(\cdot, \cdot))$ is shown to be an Hilbert space of infinite dimension isometric to the L_2 space. This basically says that operations in $L^2(I)$ can be replicated in ${\mathcal{A}_B}^2(I)$ via an inverse operation given by the isomorphism. The distance (6.4) can then be rewritten as

$$d_{\mathcal{A}}(f, g)(t) = \frac{1}{\sqrt{2\eta}} d_{L_2(IxI)}(f^*, g^*),$$

where f^* and g^* are defined analogously as $f^* : I \times I \to \mathbb{R}$ with $f^*(t, s) = \log\left(\frac{f(t)}{f(s)}\right)$.

This distance, based on the principles of the standard Aitchison distance for compositions, is invariant under relevant transformations of densities. That means that when two probability densities are updated by means of the Bayes' theorem, the distance between them remains invariant [14].

In this space, the perturbation (the analogous to the addition) and the powering (the scalar multiplication by a constant) are defined by

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s)ds} \quad \alpha \odot f = \frac{f(t)^\alpha}{\int_I f(s)^\alpha ds}.$$

The defined Aitchison geometry is obtained by the isometric log-ratio (ilr) transformation, although this is not the only possible transformation of data.

One of the advantages of the ilr transformation is that it moves the whole Aitchison geometry to the Euclidean one.

Sharing the principles of Compositional Data Analysis (CoDa), other possible transformations of the density function such as the additive log-ratio (alr), or the centered log-ratio (clr) transformations can be considered. Note that different alr transformations are related by linear transformations [19]. Like the ilr, also clr has been proposed in FDA [16], is defined as

$$\operatorname{clr}(f)(t) = f_c(t) = \ln f(t) - \frac{1}{\eta}\int_a^b \ln f(\tau)d\tau, \tag{6.5}$$

where η stands for length of the interval I, in particular $\eta = b - a$ and

$$\frac{1}{\eta}\int_a^b \ln f(\tau)d\tau$$

is the geometric mean of the functional part.

Such an isometry allows to define the following operations among the clr-transformations in terms of their expression in L^2

$$\operatorname{clr}(f \oplus g)(t) = f_c(t) + g_c(t) \quad \operatorname{clr}(\alpha \odot f)(t) = \alpha f_c(t)$$

and the following inner product

$$\langle f, g\rangle_{\mathcal{A}} = \langle f_c, g_c\rangle_2 = \int_I f_c(t)g_c(t)dt.$$

Introducing the clr-transformation for density functions has, as in CoDa, the advantage to give good interpretability of the transformed data and a better visualization by means of a biplot for compositional data [20]. However, as in CoDa, it is verified the constraint

$$\int_I \operatorname{clr}(f)(t)dt = \int_I \ln f(t)dt - \int_I \frac{1}{\eta}\int_a^b \ln f(s)dsdt = 0.$$

That led to the loss of one dimension and needs to be taken into account for the analysis of transformed data.

As observed in [21], this constraint may lead to computational problems for some statistical methods, and it is still well acceptable for distance-based methods or functional principal component analysis.

The crucial point of the introduced Aitchison geometry with both the ilr and clr transformations is that it presents the problem that the density functions must be strictly positive, because the log ratios, and thus the distance computed between compositions located at the boundaries of the simplex, could degenerate.

Indeed, density functions cannot be always extrapolated as compositional data because the set where the density functions are zero does not necessarily have zero measure. This implies that the log ratios in the integral of (6.4) can be annulled in a

set of measure greater than zero, which implies that the logarithm tends to infinity, which causes the integral to diverge.

In these cases, if some values of the function are null, they are usually substituted by some positive values using different methods (i.e. the support of the functions should be homogenized) [22, 23].

The statistical theory concerning the analysis of density functions according to the Aitchison geometry and, more in general, to the Bayes space, has been developed in some few works.

We refer to [9] where a well-suited dimensionality reduction tool has been proposed when functional data are densities. In particular, a multidimensional scaling approach is evaluated with respect to three different distances: the L_2 distance between log-densities, the symmetric Kullback–Leibler divergence, and finally the Aitchison distance between densities using a clr transformation of the data.

The ideas proposed in the approach rely on multivariate techniques that can be applied to a special class of functional data. However, as the same author claims, one must care about the choice of the right density function transformation and an appropriate distance needs to be defined.

For many commonly used metrics on the space of densities, it does not exist an isomorphism between the space of densities and the Hilbert space L^2, and this is the reason why the obtained results cannot be simply generalized.

With the aim of proposing an approach that can be generalizable, the work of [15] proposes a representation of the densities by applying the inverse map from the linear functional space to the density space. Log-quantile density and log-hazard transformations are introduced and rates of convergence are derived for their representations.

In the field of multidimensional techniques giving a key rule to the spline approximation of the clr-transformation, a functional principal component on density functions is proposed in [16]. The spline approximation constrained to the clr-transformation is introduced by granting a regular covariance matrix of the observations. However, this provides little flexibility to the approach since it depends on the adopted transformation.

Once the data have been transformed, the smoothing splines (used for the subsequent analysis) cannot always be well adapted to raw (discretized) distributional observations.

As pointed out by Machalova et al. [21], although the methodology of Bayes spaces was successfully applied to theoretical problems related with the Bayesian approach to statistical analysis, its application to statistical processing of density functions is still limited.

Smoothing the original discretized densities [9] is not coherent with the Bayes space methodology. This is due to the absence of an approximation tool that enables to proceed from compositional functional data to smooth functions.

Optimal smoothing splines for clr-transformed density functions has been proposed by Machalova et al. [21]. It takes all their specific features into account and provides a concise methodology for reasonable preprocessing of raw (discretized) distributional observations.

Further arguments in [11] and [10] consider the possibility of combining functional geostatistics and compositional data in infinite dimensional spaces by working in the framework of the Aitchison geometry. In particular, [11] proposes a kriging predictor in a Hilbert space, Salazar et al. [10] introduces two alternative kriging predictors that depend on the characteristics of the density function to be predicted.

In conclusion, the compositional approach to the data analysis described by densities, via the log-ratio approach and the theory of Bayes space, provides a good mathematical framework when density functions are positive along all the shared support. Indeed, once data have been transformed using the clr, for other suitable transformations, the analysis is performed with the classical techniques of FDA. On the other hand, no solutions are provided when densities assume structural zero values. For example, if one wants to compute the distance between two daily temperature distributions observed, let us say, the first at the North Pole and the second at the Death Valley, it is expected that the first ranges between -32 and $0\ ^\circ$C, while the second between 11 and $39\ ^\circ$C. For avoiding the zero density problem, it would be hard to justify any assignment. So this problem may affect the final results of an analysis on densities that are considered only in a portion of their support and/or are modified in order to avoid the zero density problem.

All these methods have been developed for phenomena that are usually related on a single variable observed over a continuum [9]. The problem could be also extended to multivariate case.

6.1.2 Features of Density Functions as Distributional Data

Density functions are rarely found in a continuous form in practice. Often, the aggregation of individual observations leads to a discretized forms in terms of histogram data, frequency distributions, and so on. In a probabilistic perspective, these can be seen as a realization of distributional data.

In Section 6.1.1, we noted that functional compositional data can provide a description of density functions exploiting its compositional interpretation. However, considering a density function in a compositional view implies doing a transformation of the data that must be done in an accurate way. In this section, we discuss a SDA approach for dealing with density functions, and we refer to it as SDA approach.

Following the terminology adopted in SDA, the variables, which allow distributions as description of individuals, are termed *modal-numeric* (probabilistic)

variables or distributional symbolic variables. As defined by Bock and Diday [24], a modal variable Y, with basic domain $\mathcal{Y}$ on the set E of objects is a mapping $E \to \mathcal{M}$ of all possible measures π on $\mathcal{Y}$ (completed by a σ-field):

$$i \to Y(i) \in \mathcal{M}, \quad \text{for } i \in E.$$

Typically, $E = \{1, \dots, n\}$, π_i is a frequency, probability or weight distribution, thus, for ease of notation, we may write

$$Y(i) = \{S(i), \pi_i\},$$

where $S(i) \subseteq \mathcal{Y}$ is the support of π_i in $\mathcal{Y}$.

Let f be a density function of a variable Y, F be the corresponding distribution function, and F^{-1} its inverse function (the quantile function).

In its distributional version proposed by Irpino and Verde [13], a modal variable is called a distributional symbolic variable if for all i, the measure π_i has a density f_i, and is written as

$$Y(i) = f_i.$$

For the generic ith object, we recall that the quantile function, being the inverse of the cdf, is a function $[0, 1] \to S(i)$ such that

$$F_i^{-1}(t) = \inf \left\{ y \in S(i) : t \leq F_i(y) \right\}. \tag{6.6}$$

It is also defined when the distribution is not continuous on $S(i)$. If the distribution is continuous on the domain $S(i)$, the quantile function corresponds to its classical inverse function.

This general definition includes several and different types of distributional symbolic variables such as multivalued discrete data, interval data, and histogram valued data that permit to consider a unique approach for their analysis.

The classical summary statistics for univariate data can be applied equally to distributional data. As pointed out in [13], it is possible to define the so-called *Fréchet* and *Chisini* means starting, respectively, from proximity relations among data from the definition of a function of the observed data.

The definition of a *Fréchet* and *Chisini* means of distributional variables requires two conditions: the definition of a distance between distributions (or random variables) and the definition of, at least, the sum of distributions and the product of a distribution by a scalar.

Many metrics or semimetrics in the space of the density functions can be defined to compare distributional data [25]. Among them the most promising, from the practical point of view, is the metric based on quantile functions that are in a one-to-one correspondence with the respective density functions.

Let $f_i(y)$ and $f_{i'}(y)$ be two density functions and let $F_i^{-1}(t), F_{i'}^{-1}(t)$ be the two corresponding quantile functions associated with *cdf*s $F_i(y)$ and $F_{i'}(y)$. The distance

between the probability distributions is defined by means of quantile functions as

$$d_W(f_i(y), f_{i'}(y)) = \sqrt{\int_0^1 \left[F_i^{-1}(t) - F_{i'}^{-1}(t)\right]^2 dt}. \tag{6.7}$$

This definition introduced by Ruschendorf [26] corresponds to the Wasserstein distance with interesting interpretative results related to the first two finite moments of the density functions.

Let $f_i(y)$ and $f_{i'}(y)$ be two density functions having finite the first two moments. The density function $f_i(y)$ is in a one-to-one correspondence with the cdf $F_i(y)$ and the quantile function $F_i^{-1}(t)$ (the inverse of the distribution function). Denoted by μ_i the expected value of f_i, and by σ_i the standard deviation, Irpino and Verde [27] has been shown the following equivalence:

$$\begin{aligned} d_W^2(f_i(y), f_{i'}(y)) &= \int_0^1 \left[F_i^{-1}(t) - F_{i'}^{-1}(t)\right]^2 dt \\ &= \underbrace{\left(\mu_i - \mu_{i'}\right)^2}_{Location} + \underbrace{\underbrace{\left(\sigma_i - \sigma_{i'}\right)^2}_{Size} + \underbrace{2\sigma_i\sigma_{i'}(1-\rho_{i,i'})}_{Shape}}_{Variability}, \end{aligned} \tag{6.8}$$

where $\rho_{i,i'}$ is defined as

$$\begin{aligned} \rho_{i,i'} &= \frac{\int_0^1 \left(F_i^{-1}(t) - \mu_i\right)\left(F_{i'}^{-1}(t) - \mu_{i'}\right) dt}{\sqrt{\left[\int_0^1 \left(F_i^{-1}(t) - \mu_i\right)^2 dt\right]\left[\int_0^1 \left(F_{i'}^{-1}(t) - \mu_{i'}\right)^2 dt\right]}} \\ &= \frac{\int_0^1 \left(F_i^{-1}(t) - \mu_i\right)\left(F_{i'}^{-1}(t) - \mu_{i'}\right) dt}{\sigma_i\sigma_{i'}} = \int_0^1 \frac{\left(F_i^{-1}(t) - \mu_i\right)}{\sigma_i}\frac{\left(F_{i'}^{-1}(t) - \mu_{i'}\right)}{\sigma_{i'}} dt \\ &= \frac{\int_0^1 F_i^{-1}(t)F_{i'}^{-1}(t)dt - \mu_i\mu_{i'}}{\sigma_i\sigma_{i'}} \end{aligned} \tag{6.9}$$

and provides the correlation of two series of data. It is represented, respectively, by the t-th quantile of the first distribution and the t-th quantile of the second. In this sense, we may consider it as the correlation between quantile functions represented by the curve of the infinite quantile points in a Q–Q plot. It is worth noting that, as σ_i and $\sigma_{i'}$ are positive, $0 < \rho_{i,i'} \leq 1$ and is equal to 1 when the two standardized series of quantiles are the same, or, in other words, when the two distributions are identical except for the means and the standard deviations (i.e. they are two uniforms, two normal distributions, etc.).

The L_2 Wasserstein distance has been object of many works dealing with the analysis of distributional data. The reasons are mainly related to the key rule that its property of decomposition plays in the interpretation of results as introduced by Irpino and Verde [28].

First, as shown in [27], it can be used for defining an inertia measure among distributional variables, providing a way of looking at variability in the same manner of the Euclidean distance. In addition, according to its decomposition (6.8), it permits to investigate the kind of variability present in the data.

Starting from the properties of this distance, a number of basic statistical tools have been extended to distributional variables [13, 29].

As for the compositional functional methods, but in a "symbolic" perspectives in the context of Distributional Data Analysis (DDA), the specific features of density functions are accounted for their complex characteristic as "informative object" like distributional data, or multivalued data. In this case, each entity corresponds to a distribution that may be represented by a histogram or a quantile function [30].

Within this context, proposals regarding statistics for histogram (or distributional) valued data [13], clustering strategies [27, 31–33], regression methods [29, 34], and time series forecasting techniques [35] have been developed.

SDA methods, contrary to FDA techniques, have been developed for more than one single variable.

Among the mentioned methods, a good review and discussion about the many points of view can be found in [13], where histogram-valued, or distribution-valued data analysis is introduced by means of quantile functions.

Under this paradigm, quantile functions are seen as a way to describe information about density functions.

Among the SDA and FDA methods, we focus here on two special cases, the simple log-ratio and the quantile approach, and we illustrate the potential of both the methods. In Section 6.2, a measure of spatial association for georeferenced density functions is presented.

Even if this approach lacks some mathematical properties related to the space of quantile functions, on the other hand, has some interpretative advantages in explaining the variability of a distributional variable. Further, it is insensitive to the zero density problem and does not require that the density function must be positive on a common support.

6.2 Measures of Spatial Association for Georeferenced Density Functions

All the elements to introduce a spatial correlation measure for evaluating performances of SDA and FDA methods on density functions can now be introduced.

We now illustrate and compare the main characteristics of the two approaches FDA and SDA for dealing with density functions by using the two different recalled measures of distances the Aitchison and the Wasserstein metrics.

According to Tobler's First Law of Geography, the main characteristic of spatial data is that the units are correlated; in this framework, spatial association and spatial autocorrelation measures are used to indicate the coincidence of values similarity with location proximity. Although the two concepts are similar, spatial autocorrelation is a weaker form based on second moments of a joint distribution.

The concept of spatial autocorrelation may be viewed as a special case of correlation, with a meaning related to the space of reference. Correlation statistics were classically designed to show relationships between variables; spatial autocorrelation shows the correlation within variables across a georeferenced space. Whereas the concept of spatial correlation is related to "neighboring observations" that can be defined using contiguity or distance matrices among the spatial units expressed through the definition of a spatial weights matrix. A large number of tools to investigate the nature and extent of spatial correlation between spatial variables exist. One of the most used is the Moran's index I [36].

Originally introduced in 1950, it is the first measure of spatial autocorrelation introduced to study stochastic phenomena distributed in space in two or more dimensions. Moran's index has been subsequently used in almost all studies employing spatial autocorrelation.

Moran's index I is used to estimate the strength of the correlation between observations as a function of the spatial distance separating them (correlograms). It can be defined as an indicator of both global and local spatial association. It shares many similarities with Pearson's correlation coefficient since its numerator is a covariance, while its denominator is the sample variance. And as a correlation coefficient, its values range from +1 (meaning strong positive spatial autocorrelation), to 0 (meaning a random pattern) and to −1 (indicating strong negative spatial autocorrelation).

Formally, a spatial correlation index measures the spatial association in the data considering simultaneously both locational and attribute information. Two types of measure of spatial correlation can be defined: global measures summarizing the spatial association with respect to the whole region, and local measures related to the association of a single location with respect to all its neighborhood.

In the following, we will focus on two main types of spatial association indices: the Moran's index I as measure of global spatial autocorrelation, and the local indicator of spatial association (LISA) function.

6.2.1 Identification of Spatial Clusters by Spatial Association Measures for Density Functions

Let $f_{s_1}, \ldots, f_{s_n}$ be a set of n georeferenced density functions observed in $s_1, \ldots, s_n$ arbitrary locations in $D \subseteq \mathfrak{R}^d$.

We can define the well-known Moran's index I as a global measure of spatial autocorrelation by

$$I = \frac{\sum_{s_i,s_j} w_{s_i,s_j} f^*_{s_i} f^*_{s_j}}{n\sigma^2(f_s)}, \tag{6.10}$$

where

- $w_{s_i s_j}$ is an element of a $n \times n$ contiguity matrix W defined as follows:

$$w_{s_i s_j} = \begin{cases} 1, & \text{if a region } s_i \text{ is next to region } s_j, \\ 0, & \text{other cases;} \end{cases}$$

- $f^*_{s_i} = f_{s_i} - \overline{f}^*_{s_i}$ and $f^*_{s_j} = f_{s_i} - \overline{f}^*_{s_j}$ are the centered density functions with respect to the mean functions $\overline{f}_{s_i} = \frac{1}{n}\sum_1^n f_i$ and $\overline{f}^*_{s_j} = \frac{1}{n}\sum_1^n f_j$;
- $\sigma^2(f_s) = \frac{1}{n}\sum_{i=1}^n \int_T (f_{s_i}(t) - \overline{f}_{s_i}(t))^2 dt$ is the sample variance.

The proximity matrix W is everywhere 0 except for contiguous locations s_i and s_j, where it takes the value 1. This is the simplest formalization, however, an extended definition of this contiguity matrix can be considered by allowing for the computation of Moran's I at various "levels" of distance. This formalization provides a global measure of spatial autocorrelation by the impact of distance on the strength of spatial autocorrelation for each variable [17]. The strength of Moran's I lies in its simplicity; however, it presents the limitations associated with averaging local variations. In order to overcome this problem, to examine local autocorrelation a LISA function [17], which can be seen as the local equivalent of Moran's I, has been introduced. It is defined such that the sum of all local indices is proportional to the (global) value of Moran's statistic. Formally, it can be expressed by

$$I_{s_i} = \frac{\sum_{i,j} w_{s_i s_j} f^*_{s_i} f^*_{s_j}}{f^2_{s_i *}}. \tag{6.11}$$

According to LISA values, it is possible to compute, for each location, its similarity with its neighbors and also to test its significance. In addition, it is possible to define a clustering structure by observing the observed values against the averaged value of their neighbors. The clustering configuration can be identified from a scatterplot. This is the so-called "Moran scatterplot" where four possible situations can be configured:

- Locations with similar neighbors and high values, also named "hot spots" with high–high relations;
- Locations with similar neighbors, but low values: low–low. Also known as "cold spots."

- Locations with low-value neighbors and with high values: low–high. Potential "spatial outliers."
- Locations with local autocorrelation not significant.

We propose to use both indices for dealing with density functions by using the two approaches, the SDA and FDA.

6.3 Real Data Analysis

We compare the proposed approaches using data coming directly in form of distributions (histograms) from the US Census. We limited our data for the 254 counties of Texas (the state with the largest number of counties in United States) considering two variables: the age distribution of the population and the income distribution of households. Data have been extracted from the American Community Survey (ACS) with the five years estimates of the distributions for 2015.[1] In particular, data comes from the S0101 and S1901 table of the survey program. For each county, from the S0101 table, we have the estimate of the proportions of population broken down in classes of 5 years from 0 to 85 years and a last open class, 85 years and more, that we decided to close at 100 years, for a total of 18 age classes (the bins of the histograms). About the income of households, from the S1901 table, we observed the distribution of households broken down in 10 classes of unequal length, as follows less than $ 10K, from $ 10K to $ 15K, from $ 15K to $ 25K, from $ 25K to $ 35K, from $ 35K to $ 50K, from $ 50K to $ 75K, from $ 75K to $ 100K, from $ 100K to $ 150K, from $ 150K to $ 200K, $ 200K or more. For the application, we decided to close the first class at 0, as lower bound, and the last one at $ 500K, as an upper bound. It is worth noting that the income distribution has bins of different length and that, as it is commonly known, is very skewed. The last aspect may affect the use of the functional–distributional approach that is based on a preprocessing step where distributions are smoothed by means of B-spline functions. In order to take into account this aspect, we considered also a Box–Cox transformation of data with λ parameter equal to 0.2.[2]

For the analysis, we used the `HistDAWass` package, developed in R for the analysis using the SDA-distributional approach, while for the compositional–functional one, we followed the strategy proposed in [16] consisting of a preprocessing step of the density functions by means of particular B-splines computed

1 Source: U.S. Census Bureau, 2011–2015 ACS 5-Year Estimates.

2 After performing a cross-validation step on a grid of values ranging from -2 to 2, we observed that the choice of $\lambda = 0.2$ allows to obtain income distribution with a low skewness for all the densities and quasi-Gaussian distributions.

on the clr transformations of the densities as proposed in [21], followed by a simplicial functional principal component analysis (SFPCA) [16].

For both the approaches, we computed the LISA functions using the local Moran index, and we plotted the resulting clusters on a map considering only those counties having a significant value (p-value less than 0.05%) of the index. As contiguity matrix, we used an adjacency matrix that is a binary matrix, where the value 1 indicates that two counties have at least one point in common on the borders.

In Figure 6.1, we show a part of the input data matrix. Consistently with the SDA approach, each cell contains a distribution (a histogram in this case).

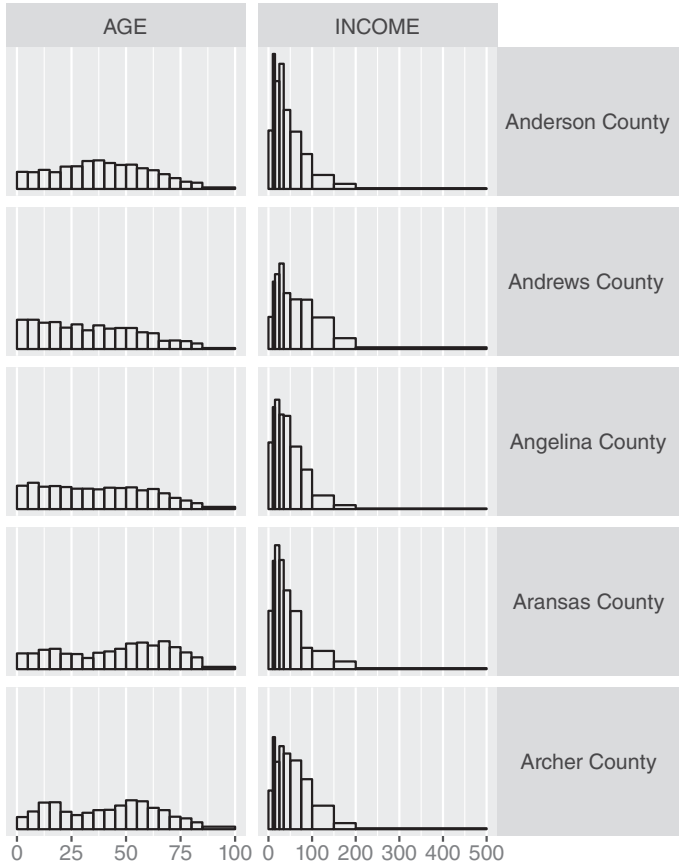


Figure 6.1 ACS-5y 2015, Texas data: first five counties (of 254) of the input data table: age is in years, income is in thousands of dollars. Source: Data from American Community Survey 2015.

6.3.1 The SDA Distributional Approach

For the spatial analysis of the two variables, we use the statistics developed in [13]. For data represented by histograms, it is possible to compute such basic statistics in a finite time and without approximations taking advantage from the fact that quantile functions of histograms are piecewise linear functions [27].[3]

For the *AGE* variable, we obtained a global Moran's index equal to 0.325, while for *INCOME* variable, the index is equal to 0.344. We computed the LISA functions using the local Moran's indices for the two variables, and we colored those counties having a significant index different from zero (see Figure 6.2).

Further, we considered a modified version of the Moran plot, where we represented the observed residual quantile functions from the mean quantile one vs. the weighted ones (with respect to the mean weighted quantile function). In Figure 6.3, the residual functions are shown, and we labeled the countries using the classical Moran way of labeling into four categories (HH, LH, HL, and LL). Finally, in Figure 6.4, we represent the counties with a significant local index, colored accordingly with their label and nothing that some clusters are obtained. Looking at clusters, we note that counties closer at the Mexico borders are characterized by a younger population and a low income, while the richest zones are around Austin, Houston, and Dallas that are the main cities of the state.

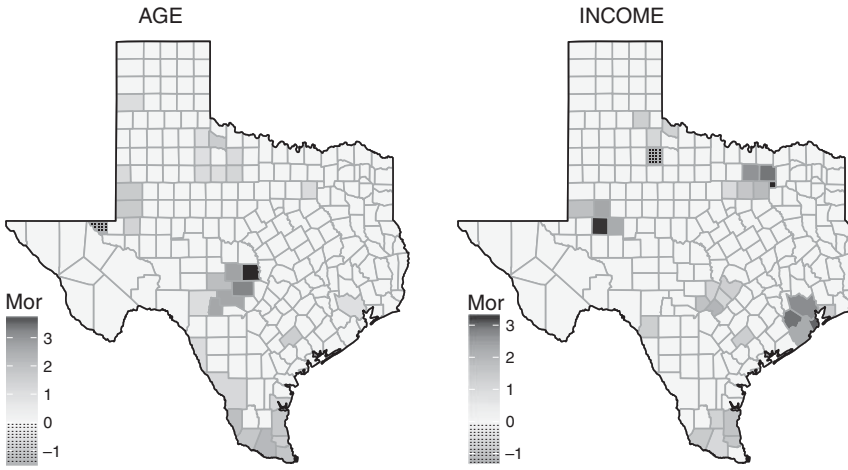


Figure 6.2 ACS-5y 2015, Texas data: counties with a significant local Moran's *I*.
Source: Data from American Community Survey 2015.

3 The procedures have been implemented in the `HistDAWass` package developed in R.

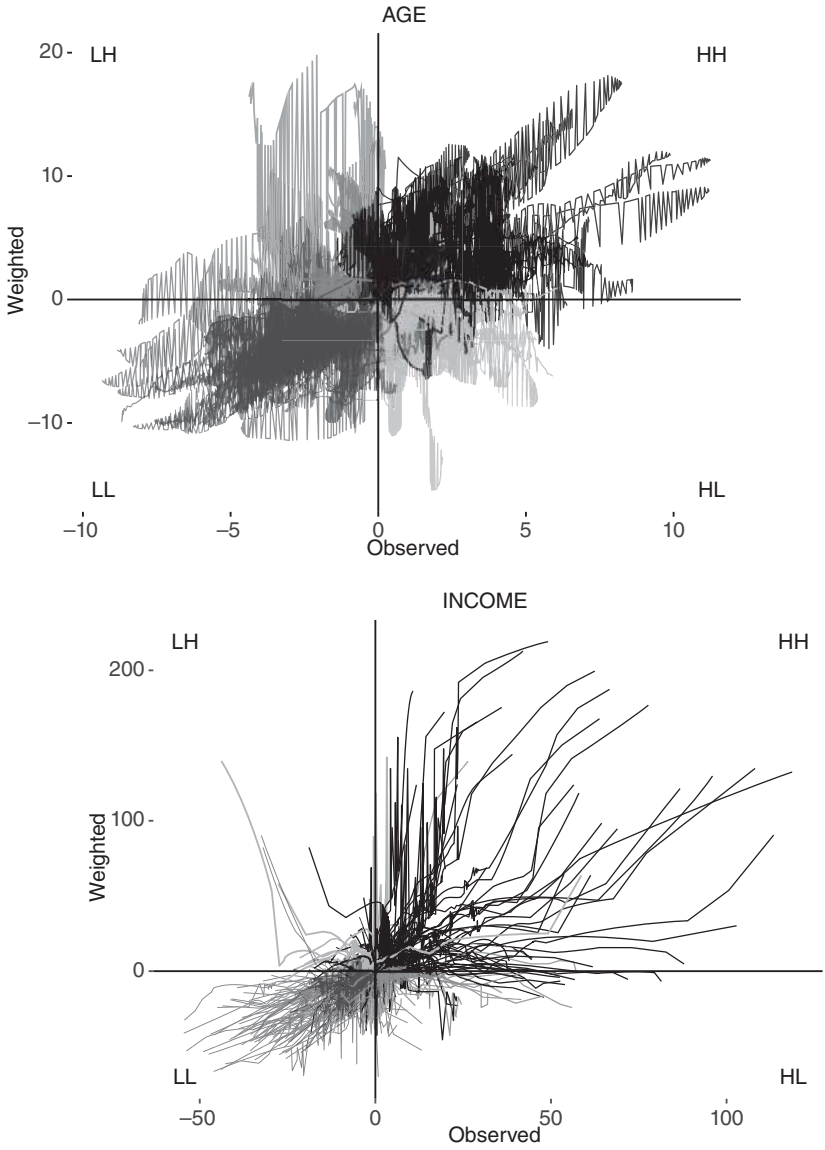


Figure 6.3 ACS-5y 2015, Texas data: Moran's plot for residual functions.
Source: Data from American Community Survey 2015.

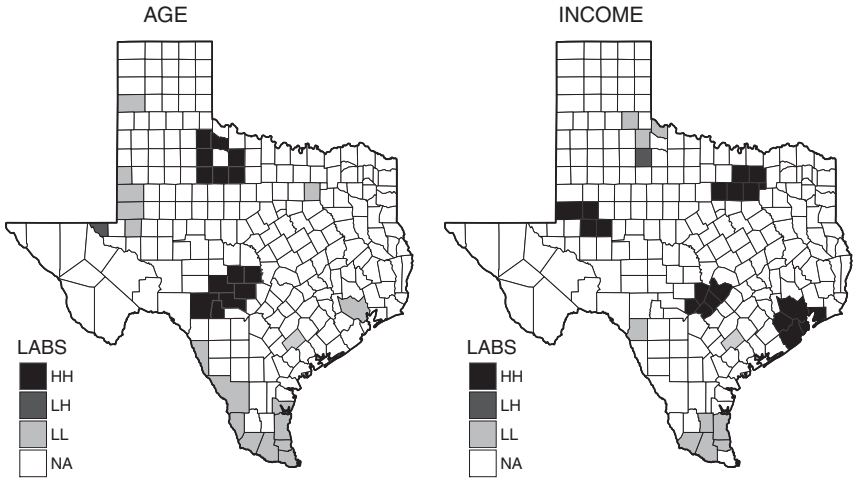


Figure 6.4 ACS-5y 2015, Texas data: cluster of counties with a significant local Moran's *I*. Source: Data from American Community Survey 2015.

Observing the *AGE*, we note also that Harris county (the one of Houston) has younger population, while older population form some clusters in the north and at the east of Dallas.

6.3.2 The Compositional–Functional Approach

Using the compositional–functional approach suggested in [16], we first performed a preprocessing step for smoothing the clr transformed functions of the observed densities. We used a particular procedure for B-spline smoothing of density functions as proposed in [21]. For the *AGE* variable, we used a cubic B-spline of the clr transformation of the densities using five equispaced knots [0 25 50 75 100]. We obtained a 254×7 matrix of coefficients and then performed a functional principal components using the *fda* package in Matlab. We stored the first two dimensions describing, respectively, the 64% and the 18% of the total variation (82% of variation explained by the first two eigenfunctions). We considered the scores for the first and the second factor for computing the local Moran's *I*.

In Figure 6.5 are represented the two harmonics representing the first two eigenfunctions. For each dimension, we computed the global Moran's indexes that are, respectively, 0.378 and 0.008. We also derived the local Moran's indices, and we plotted in Figure 6.6, the counties with the significant indices, and additionally, we also plot the clusters obtained from the first and the second factor.

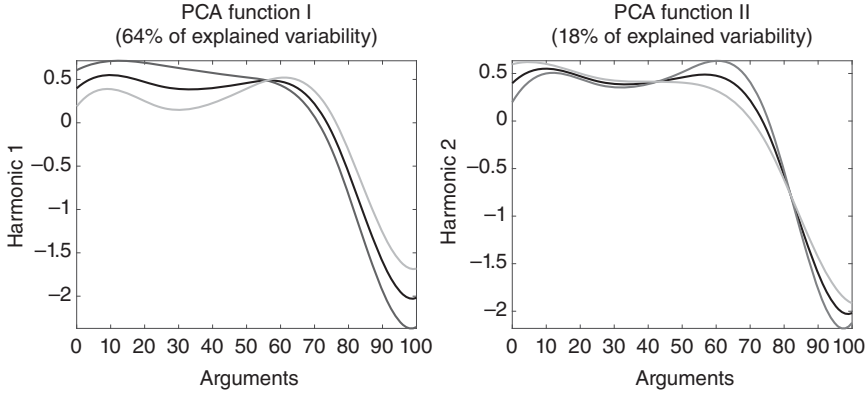


Figure 6.5 ACS-5y 2015, Texas data: *AGE* variable, first two harmonics after the functional principal component analysis (FPCA) of the smoothed clr functions. In black, the main harmonics, while in dark and light gray are drawn, respectively, the ± 2 standard deviations of each are drawn. Source: Data from American Community Survey 2015.

For the *INCOME* variable, we used a cubic B-spline of the clr transformation of the densities using five knots [0 35 75 150 500]. We obtained a 254×7 matrix of coefficients, and we then performed a functional principal components using the *fda* package in Matlab. We stored the first two dimensions describing respectively 61% and the 33% of the total variation (94% of variation explained by the first two eigenfunctions). We considered the scores for the first and the second factor for computing the local Moran's I.

In Figure 6.7 the two harmonics representing the first two eigenfunctions are shown. For each dimension we computed the global Moran's indices that are respectively: 0.303 and 0.109. We also derived the local Moran's indices, and we plotted in Figure 6.8 the counties with the significant indexes together with the plot with the clusters obtained from the first and the second factor.

Considering that the income distributions are skewed, we performed a Box–Cox [37] transformation of the data using a value of λ equal to 0.2. This value has been chosen because it allows a transformation of the distributions into Gaussian ones. For the *INCOME* variable transformed with the above-mentioned Box–Cox transformation, we used a cubic B-spline of the clr transformation of the densities using five equal-spaced knots [0 3.08 6.16 9.25 12.33]. We obtained a 254×7 matrix of coefficients and then we performed a functional principal components using the *fda* package in Matlab. We stored the first two dimensions describing, respectively, 83% and 12% of the total variation (95% of variation explained by the first two eigenfunctions, a bit more than the variation explained in the previous case). We considered the scores for the first and second factors for computing the local Moran's I.

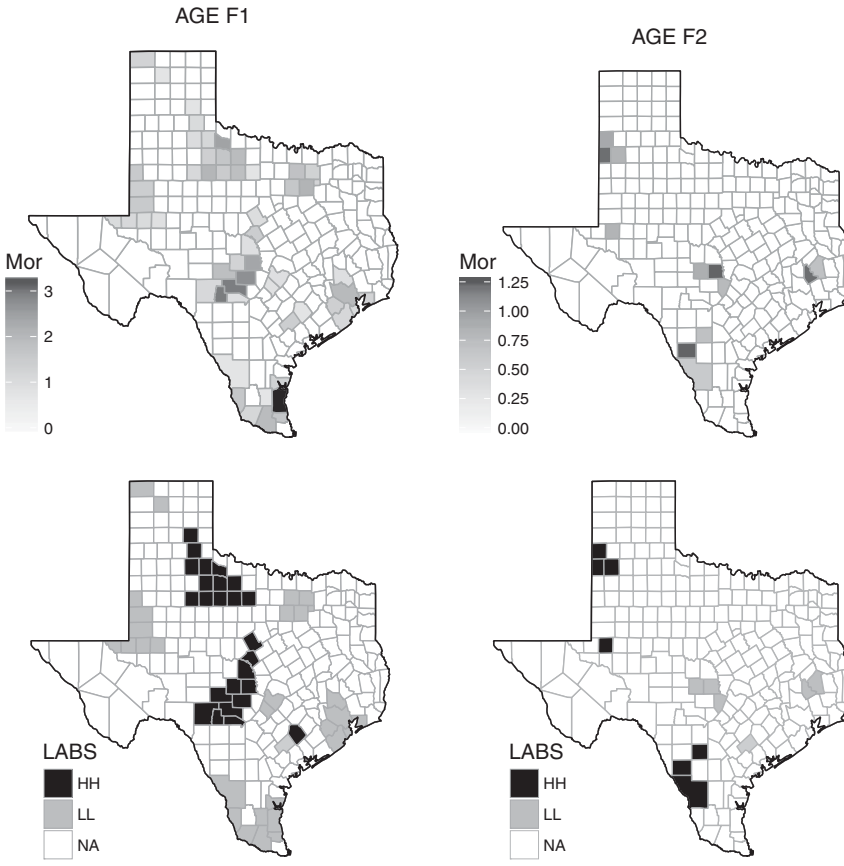


Figure 6.6 ACS-5y 2015, Texas data: *AGE* variable. On top, the two maps with counties colored in gray scale accordingly to significant local Moran's indexes. The clusters are represented on the bottom. Source: Data from American Community Survey 2015.

In Figure 6.9, the two harmonics representing the first two eigenfunctions are shown. For each dimension, we computed the global Moran's indexes that are, respectively, 0.458 and 0.124. We also derived the local Moran's indexes, and we plotted in Figure 6.10, the counties with the significant indexes together with the clusters obtained from the first and second factors.

6.3.3 Discussion

Looking at the results, we note that clusters identified by using the compositional–functional approach are greater in size with respect to the ones identified with the SDA-distributional one. It is important to note that using the

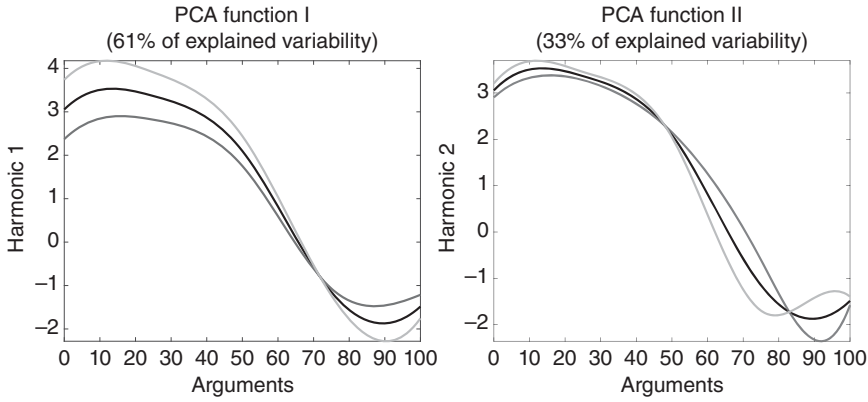


Figure 6.7 ACS-5y 2015, Texas data: *INCOME* variable, first two harmonics after the FPCA of the smoothed clr functions. In black the main harmonics, while in dark and light gray are drawn, respectively, the ± 2 standard deviations of each harmonic. Source: Data from American Community Survey 2015.

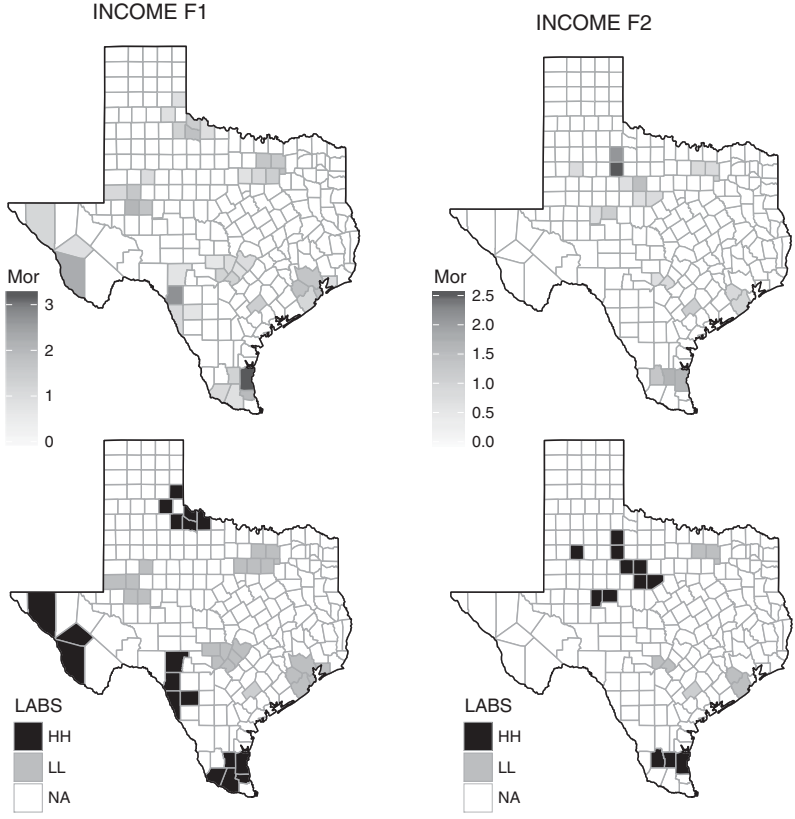


Figure 6.8 ACS-5y 2015, Texas data: *INCOME* variable. On the top, the two maps with the counties colored in gray scale accordingly to significant local Moran's indices. The clusters are represented on the bottom. Source: Data from American Community Survey 2015.

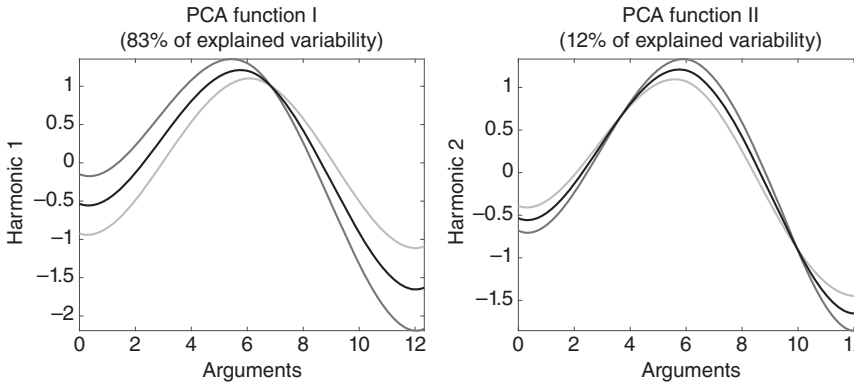


Figure 6.9 ACS-5y 2015, Texas data: *INCOME* variable transformed using a Box–Cox procedure, first two harmonics after the FPCA of the smoothed clr functions. In black, the main harmonics, while in dark and light gray, the ± 2 standard deviations of each harmonic are drawn. Source: Data from American Community Survey 2015.

compositional–functional approach, the preprocessing step plays an important role in modifying the original data.

It appears that the choice of the number of knots and their position leads to smoothed functions that tend to be closer with respect to the original ones. This induces a general reduction of Aitchison distances, and, as a consequence, transformed data may appear more similar, and thus, the spatial correlation is general higher than the ones observed when densities that are not preprocessed. Further, we note that input data are approximations of the population density. For example, as we observed in the application, input data are histograms that can be considered as density estimators of the true densities. Thus, using a preprocessing step for smoothing histograms introduces an additional approximation of the true density that is made not directly on the raw data, so the combination of the two smoothers (the histogram and the B-spline of the clr transformed histograms) may add artificial noise in the data. Finally, for the sake of brevity, we did not experiment with different choices for the parameters of the B-spline smoothing step, but we believe that an optimal choice may be hard to achieve.

6.4 Conclusion

In this chapter, we compared two different approaches for dealing with georeferenced data described by density functions. While the compositional–functional approach related to the Aitchison geometry has nice mathematical properties,

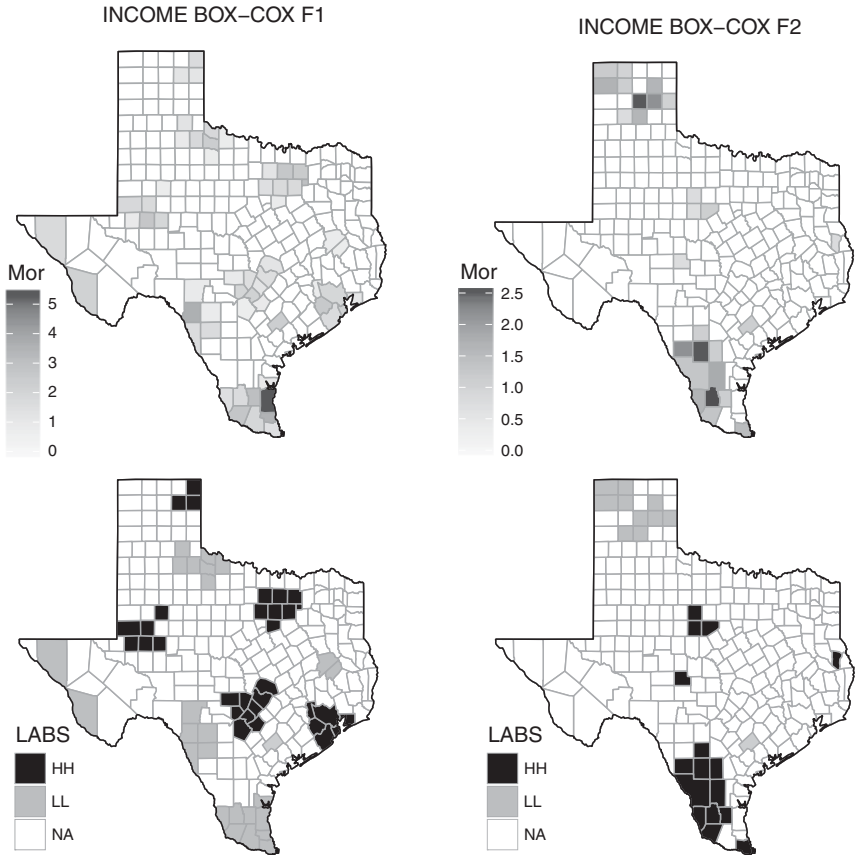


Figure 6.10 ACS-5y 2015, Texas data: *INCOME* variable transformed using Box–Cox procedure. On top, the two maps with the counties colored in gray scale accordingly to significant local Moran's indices. The clusters are represented on the bottom. Source: Data from American Community Survey 2015.

we note that it can be used after a preprocessing smoothing step, where the user choice of the parameter may affect the whole analysis. On the other hand, even if the SDA-distributional approach based on the Wasserstein distance between densities has not the mathematical properties of the Bayes spaces, it is able to work directly on the data represented as a density. In both cases, a source of error is due to the definition of the input data, but in the SDA-distributional approach, no transformation or smoothing of densities is performed. Finally, even if the compositional–functional approach appears to be more elegant, it seems to be less clear from a practical point of view.

Acknowledgments

The authors thank Professors. Hron and Mrs. Machalova for having provided the Matlab routines for the B-spline code of clr-transformed densities.

References

1 Mateu, J. and Romano, E. (2016). Advances in spatial functional statistics. *Stochastic Environmental Research and Risk Assessment* 31: 1–6.
2 Ferraty, F. and Vieu, P. (2011). *Non Parametric Functional Data Analysis: Theory and Practice*. Springer.
3 Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer.
4 Materon, G. (1963). Principles of geostatistics. *Economic Geology* 58: 1246–1266.
5 Diggle, P. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press.
6 Saveliev, A.A., Mukharamova, S.S., and Zuur, A.F. (2007). Analysis and Modelling of lattice data. In: *Analysing Ecological Data*. Statistics for Biology and Health. Springer, New York, NY.
7 Delicado, P. (2007). Functional k-sample problem when data are density functions. *Computational Statistics and Data Analysis* 22: 391–410.
8 Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics* 21: 224–239.
9 Delicado, P. (2011). Dimensionality reduction when data are density functions. *Computational Statistics and Data Analysis* 55: 401–420.
10 Salazar, E., Giraldo, R., and Porcu, E. (2015). Spatial prediction for infinite-dimensional compositional data. *Environmental Research and Risk Assessment* 29: 1737–1749.
11 Menafoglio, A., Guadaglini, A., and Secchi, P. (2014). A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment* 28: 1835–1851.
12 Billard, L. and Diday, E. (2002). *Symbolic Data Analysis*. Chichester: Wiley.
13 Irpino, A. and Verde, R. (2015). Basic statistics for distributional symbolic variables: a new metric-based approach. *Advances in Data Analysis and Classification* 9: 143–175.
14 Egozcue, J., Díaz-Barrero, J., and Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica (English Series)* 22: 1175–1182.
15 Petersen, A. and Muller, H. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics* 44: 183–218.

16 Hron, K., Menafoglio, A., Templ, M. et al. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics and Data Analysis* 94: 330–350.

17 Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis* 27: 93–115.

18 Van den Boogaart, K., Egozcue, J., and Pawlowsky-Glahn, V. (2014). Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics* 56: 171–194.

19 Filzmoser, P. and Hron, K. (2008). Outlier detection for compositional data using Robust methods. *Mathematical Geosciences* 40: 233–248.

20 Aitchison, J. and Greenacre, M. (2002). Biplots of compositional data. *Applied Statistics* 51: 375–391.

21 Machalova, J., Hron, K., and Monti, G.S. (2016). Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics* 43(8): 1419–1435.

22 Aitchison, J. (1992) On criteria for measures of compositional difference. *Mathematical Geology* 24: 365–379.

23 Martín-Fernandez, J., Barcelo-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology* 35: 253–278.

24 Bock, H. and Diday, E. (2000). Analysis of symbolic Data. Exploratory methods for extracting statistical information from complex data. *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer.

25 Gibbs, A. and Su, F. (2002). On choosing and bounding probability metrics. *International Statistical Review* 7: 419–435.

26 Ruschendorf, L. (2001). Wasserstein metric. In: *Encyclopedia of Mathematic* (ed. M. Hazewinkel). Springer.

27 Irpino, A. and Verde, R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: (ed. V. Batagelj, H.-H. Bock, A. Ferligoj, A. Žiberna). *Data Science and Classification*, 185–192. Springer.

28 Irpino, A. and Verde, R. (2007). Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. *EGC*, volume RNTI-E-9 of Revue des Nouvelles Technologies de l'Information 2: 99–110.

29 Irpino, A. and Verde, R. (2015). Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein distance. *Advances in Data Analysis and Classification* 9: 81–106.

30 Noirhomme-Fraiture, M. and Brito, P. (2011). Far beyond the classical data models: symbolic data analysis. *Statistical Analysis and Data Mining* 4: 157–170.

31 Brito, P. and Chavent, M. (2012). Divisive monothetic clustering for interval and histogram-valued data. *Proceedings of the ICPRAM'2012*, Volume 1, SciTePress, pp. 229–234.

32 Montanari, D. and Viroli, C. (1996). A hierarchical modeling approach for clustering probability density functions. *Computational Statistics and Data Analysis* 71: 79–91.

33 Irpino, A., Verde, R., and De Carvalho, F. (2014). Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. *Expert Systems with Applications* 41: 3351–3366.

34 Dias, S. and Brito, P. (2015). Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining* 8: 75–113.

35 Arroyo, J. and Maté, C. (2009). Forecasting histogram time series with k-nearest neighbors methods. *International Journal of Forecasting* 25: 192–207.

36 Moran, P.A.P. (1950). Notes on continuous stochastic phenomena. *Biometrika* 37: 17–23.

37 Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26: 211–252.

Part II

Statistical Techniques for Spatially Correlated Functional Data

7

Clustering Spatial Functional Data

Vincent Vandewalle[1], Cristian Preda[2], and Sophie Dabo-Niang[3]

[1] *University of Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, Inria, Lille F-59000, France*
[2] *Institute of Statistics and Applied Mathematics of the Romanian Academy, 050711 Bucharest, Romania*
[3] *University of Lille, CNRS, UMR 8524, Inria - Laboratoire Paul Painlevé, Lille F-59000, France*

7.1 Introduction

The purpose of this chapter is to present two techniques for clustering spatial functional data. Generally, in any clustering framework, data inside each cluster should be as similar as possible, but different from those in other clusters. Recent researches on the clustering of independent functional data are available in the literature devoted to functional data analysis (FDA). In particular, k-means techniques are adjusted to functional data, hierarchical algorithm, and some of its variants are proposed as well, mainly for independent data (e.g. [1–10]). A review of clustering methods for functional data under the independent model is provided in [11]. Other model-based approaches for clustering functional data are given in [12, 13]. In several domains, data are of spatiofunctional nature, observations may be dependent curves at some spatial locations, and clustering these data taking into account the spatial dependency can be more accurate. The independence hypothesis does not hold in this case. Few works exist on such dependent data: Dabo-Niang et al. [14] and Giraldo et al. [15] extended some approaches on hierarchical clustering to the context of spatially correlated functional. Giraldo et al. [15] measured the similarity between two curves by the trace-variogram [16], while the spatial variation is taken into account by using kernel mode and density estimation in [14]. Other approaches for clustering spatial functional data are given recently in [8, 17].

An appropriate clustering approach should lead to homogeneous clusters and heterogeneity between them. Consequently, the number of clusters is an important issue (e.g. [18]). Some clustering methods do not automatically determine the number of clusters. Techniques are developed in the literature to overcome this

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

difficulty. Most of them propose to estimate or to select the number of clusters by solving an optimization problem involving some cluster homogeneity index (e.g. [18–20]).

We deal with a measurable spatial process $X = \left(X_s,\ s \in \mathbb{R}^N\right),\ N \geq 1$, defined on some probability space $(\Omega,\ \mathcal{A}, \mathbf{P})$. Assume that the process X is observed on some spatial region $\mathcal{I} \subseteq \mathbb{R}^N$ of cardinal n, $\mathcal{I} = \{s_1, \ldots, s_n\}$, $s_i \in \mathbb{R}^N$, $i = 1 \ldots n$. We assume also that for each location $s \in \mathcal{I}$, the random variables X_s are valued in a metric space $(\mathcal{E},\ d)$ of eventually infinite dimension and are locally identically distributed (see for instance [21]). Here $d(.,.)$ is some measure of proximity, for instance a metric or a semimetric. This means that when a site u is close enough to site v, the variables X_u and X_v have same or similar distributions. This assumption is less restrictive that strict stationarity. It is motivated by the fact that one can imagine that variables located at neighbors sites may be similar and have the same local distribution that may be different to the local distribution of another set of variables at other locations. In the classical framework of FDA, the space $\mathcal{E}$ is a space of functions, typically the space of squared integrable functions defined on some finite interval $\mathcal{T} = [0, T]$, $T > 0$.

Let S denote the set of the n curves, $S = \{X_s, s \in \mathcal{I}\}$ (renamed sometime in an arbitrary way, $S = \{X_1, \ldots, X_n\}$).

First, we present the problem of clustering spatial functional data generated by a mixture of Gaussian processes with logistic prior weights depending on the location. Second, we present an extension to spatial data of the method studied by Dabo-Niang et al. [2] which is a descendant hierarchical classification (HC) procedure based on distances between the modal and mean curves of a set of curves. The two approaches are illustrated with pollution data.

7.2 Model-Based Clustering for Spatial Functional Data

In the framework of clustering, the model-based techniques assume that there exists a latent categorical random variable Z defining G clusters of data such that probability distribution of data is a mixture of cluster distributions. Let f denote the probability distribution of X and f_g denote the probability distribution of X, given $Z = g$. Then, the mixture model is written as follows:

$$f(x) = \sum_{g=1}^{G} \pi_g f_g(x), \tag{7.1}$$

where $\pi_g = P(Z = g)$ is the prior probability of cluster g.

In the particular case of spatial dependency, we extend the model given in Eq. (7.1) by involving the location s ($s \in \mathcal{I}$) into the priors probabilities of clusters.

The mixture model becomes

$$f(x|s) = \sum_{g=1}^{G} \pi_g(s;\beta) f_g(x), \tag{7.2}$$

where β is some parametrization of the spatial prior. Thus, conditional to the cluster $Z = g$, the distribution of observations within the cluster is independent of the location, all spatial dependency being captured by the priors $\pi_g(s;\beta)$. This idea is used in [22] for clustering spatiotemporal data. The authors propose the multinomial logistic regression as a model for the $\pi_g(s;\beta)$,

$$\ln \frac{\pi_g(s;\beta)}{\pi_G(s;\beta)} = \beta_{0g} + \langle \beta_g, s \rangle_{\mathbb{R}^N}. \tag{7.3}$$

In a parametric framework, the conditional distribution f_g is depending on parameters θ_g. For example, in the Gaussian model, θ_g is the mean and the covariance matrix of cluster g. Let θ denote the set of all parameters including also those defining the $\pi_g(s;\beta)$. Thus, the model becomes

$$f(x|s;\theta) = \sum_{g=1}^{G} \pi_g(s;\beta) f_g(x;\theta_g). \tag{7.4}$$

In the finite dimensional setting (see for instance [23]), the multivariate probability density function is the main tool for estimating such a model using the Expectation–Maximization (EM) algorithm. For functional random variables, the notion of probability density in not well defined because of the infinite dimension of data. To overcome this difficulty, James and Sugar [13] and Bouveyron and Jacques [24] use the expansion coefficients of X into some finite basis of functions. This approach allows them to get a well-defined probability density function on the coefficients. In [25], the functional principal component analysis (PCA) is used to define a surrogate of the probability density for functional data. This approach is used in the context of model-based clustering in [10, 26]. In a spatial setting, [27] have proposed a mixed-effect model in which the fix effect can take into account the spatial dependencies. Moreover, assuming a spatial autoregressive dynamic for the random effect, they propose a functional classification criterion to detect local spatially homogeneous regions. In what follows, we assume that given $Z = g$, X is a Gaussian process. Then, within the cluster g, we consider a modified version of the pseudo-density defined in [25]:

$$f_g^{(q_g)}(x;\theta_k) = \prod_{j=1}^{q_g} f_{gj}(c_{gj}(x);\lambda_{gj}) \prod_{j'=q_g+1}^{d} f_{gj'}(c_{gj'}(x);\overline{\lambda}_g), \tag{7.5}$$

where f_{gj} is the probability density of the j-th principal component C_{gj} of X within the cluster g. The random variables C_{gj} $(j = 1, \ldots, q_g)$ are independent

Gaussian zero-mean with variance equal to the eigenvalues λ_{gj} of the covariance operator of X, and the random variables $C_{gj'}$ $(j' = q_g + 1, \dots, d)$ are independent Gaussian zero-mean with variance equal to the mean $\overline{\lambda}_g$ of the eigenvalues $\lambda_{gj'}$ $(j' = q_g + 1, \dots, d)$ of the covariance operator of X. Thus, the parameters $\theta_g = (\lambda_{g1}, \dots, \lambda_{gq_g}, \overline{\lambda}_g)$, q_g and d need to be defined. Notice that compared to the definition of [25], we have added the term $\prod_{j'=q_g+1}^{d} f_{gj'}(c_{gj'}(x); \overline{\lambda}_g)$.

In fact, the proposed surrogate density can be interpreted as a true density if the functional data belong to a finite dimensional space of functions spanned by some basis $\{\phi_1, \dots, \phi_d\}$, $d \geq 1$, i.e.

$$X(t) = \sum_{j=1}^{d} \alpha_j \phi_j(t), \quad t \in [0, T], T > 0.$$

Thus, we will take d as the dimension of the basis which has been used to perform the smoothing of the data. In this case, the principal components C_{kj} of the functional PCA can be obtained by performing PCA on the expansion coefficients of X in the metric M given by the inner product of the basis functions. Thus, if the learning data considered are now the expansion coefficients multiplied by $M^{1/2}$, then the proposed approach can simply be reinterpreted as learning a parsimonious high-dimensional model (see [28]) on these new data.

Note that it is also possible to consider sparse versions of the mixture model such as to consider the homoscedastic setting (equal covariance process by cluster).

7.2.1 The Expectation–Maximization (EM) Algorithm

We are now ready to describe the EM algorithm for estimating θ and therefore the clustering.

As in the finite setting, based on Eq. (7.5), we define a likelihood of the sample of curves $S = \{x_s,\ s \in \mathcal{I}\}$ by

$$l(\theta; S) = \prod_{s \in \mathcal{I}} \left(\sum_{g=1}^{G} \pi_g(s; \beta) f_g^{(q_g)}(x_s; \theta_g) \right). \tag{7.6}$$

A classical way to maximize the likelihood when data are missing (here the variable Z) is to use the iterative EM algorithm. We use this algorithm to maximize the likelihood (7.6) and adapt it for updating the principal component scores of each group as well as the parameters β defining the $\pi_g(s)$ in (7.3).

The algorithm consists in maximizing the approximated completed log-likelihood. Let $Z_g(s)$ be the indicator random variable for the cluster g at location s, then the completed log-likelihood is given by

$$L_c(\theta; S, Z) = \sum_{s \in \mathcal{I}} \sum_{g=1}^{G} Z_g(s) \left(\log \pi_g(s; \beta) + \log f_g^{(q_g)}(x_s; \theta_g) \right),$$

which is known to be easier to maximize than its incomplete version. Let $\theta^{(h)}$ be the estimated parameter value at iteration $h \geq 0$ of the algorithm.

7.2.1.1 E Step

As the groups belonging to $Z_g(s)$'s are unknown, the **E** step consists in computing the conditional expectation of the approximated completed log-likelihood:

$$\begin{aligned} Q(\theta; \theta^{(h)}) &= E_{\theta^{(h)}}[L_c(\theta; S, Z)|S] \\ &= \sum_{s \in I} \sum_{g=1}^{G} t_g^{(h+1)}(s) \left(\log \pi_g(s; \beta) + \log f_g^{(q_g)}(x_s; \theta_g) \right) \end{aligned}$$

where $t_g^{(h+1)}(s)$ is the probability for the curve X_s to belong to the group g conditionally to $C_{gj} = c_{gj}(x_s), j = 1, \ldots, q_g$:

$$t_g^{(h+1)}(s) = E_{\theta^{(h)}}[Z_g(s)|s] = \frac{\pi_g\left(s; \beta^{(h)}\right) f_g^{(q_g)}\left(x_s; \theta_g^{(h)}\right)}{\sum_{\ell=1}^{G} \pi_\ell\left(s; \beta^{(h)}\right) f_\ell^{(q_\ell)}\left(x_s; \theta_\ell^{(h)}\right)}. \tag{7.7}$$

7.2.1.2 M Step

The M-step consists in maximizing the conditional expectation of the completed log-likelihood with respect to θ:

$$\theta_g^{(h+1)} = \arg\max_{\theta_g} \sum_{s \in I} t_g^{(h+1)}(s) \log f_g^{(q_g)}(x_s; \theta_g),$$

and

$$\beta^{(h+1)} = \arg\max_{\beta} \sum_{s \in I} \sum_{g=1}^{G} t_g^{(h+1)}(s) \log \pi_g(s; \beta).$$

Note that $\beta^{(h+1)}$ is obtained as a solution of a weighted logistic regression.

The EM algorithm starts with an initial random partition of data S into G clusters.

If parsimonious models such as homoscedastic models are considered, this leads to a modification of the update of $\theta_g^{(h+1)}$, see [28] for more details.

7.2.2 Model Selection

In order to select the number of cluster G when q_g $(g = 1, \ldots, G)$ are known, we propose to maximize the Bayesian Information Criterion (BIC) criterion defined below:

$$BIC(G) = \log l(G) - \frac{\nu_G}{2} \log(n),$$

where $\nu_G = (N+1)(G-1) + Gd + \sum_{g=1}^{G}(q_g(d - (q_g - 1)/2) + 1)$ is the number of parameters of the model (spatial mixing proportions, center means, principal scores, and variances) and $n = |I|$.

When the values q_g $(g = 1, \dots, G)$ are unknown, they can be selected in order to maximize the BIC criterion by considering the following modified M-step which tries to maximize the conditional expectation of the BIC criterion:

$$\left(q_g, \theta_g^{(h+1)}\right) = \arg\max_{(q_g,\theta_g)} \sum_{s\in I} t_g^{(h+1)}(s) \log f_g^{(q_g)}(x_s;\theta_g) - \frac{\nu_{q_g}}{2}\log n,$$

where $\nu_{q_g} = q_g(d-(q_g-1)/2)$ is the additional number of parameters required for the model with q_g principal components.

Note that if we consider the homoscedastic setting, the value of the BIC criterion can be easily computed at each step of the EM algorithm for each possible value of q which does not depend on g since in this case this parameter is the same for each cluster. In this case, the expression of ν_G would be $\nu_G = (N+1)(G-1) + Gd + (q(d-(q-1)/2)+1)$.

In Section 7.4, we present the results of the application of this technique to air quality data.

7.3 Descendant Hierarchical Classification (HC) Based on Centrality Methods

Recent advances in nonparametric FDA allow to define centrality features for a sample of curves (see e.g. [29]). Dabo-Niang et al. [2] indicated that both the mean and the median curves are interesting when dealing with homogeneous data, while the modal curve would be more useful for detecting possible different structures in the data. Consequently, Dabo-Niang et al. [2] used a descendant HC method based on comparing the modal curve either with the mean or the median. Location measures (mean, mode, and median) summarize the data and aim to provide a representative element of the sample. The spatial mean used is the same as in the independent and identically distributed (i.i.d.) setting compared to the mode and median.

In our context, for the set of curves S, we define the mean curve as

$$X_{mean,S} = \frac{1}{n}\sum_{s\in I} X_s.$$

The notion of median curve for i.i.d. functional data can be extended to the spatial framework, see [30] and [2], for general definition in i.i.d. data and Dabo-Niang et al. [14], for a heterogeneity spatial index. Here, let the median curve be

$$X_{median,S} = \arg\min_{X_t\in S,\ t\in I} \sum_{s\in I} d_m(X_t, X_s),$$

with $d_m(X_t, X_s) = d(X_t, X_s)W_{s,t}$, where $W_{s,t}$ is a spatial weight. Indeed, the spatial dependency structure between the n spatial units is described by a nonstochastic

spatial weights $n \times n$ matrix W_n that depends on n. The elements $W_{s,t} = W_{s,t,n}$ of this matrix are usually considered as inversely proportional to distances between spatial units s and t with respect to some metric (physical distance, social networks, or economic distance, see for instance [31]) and Chapter 13 of this book. Here, the spatial weighted matrix W_n is constructed by taking k-neighbors of each spatial unit using k-nearest neighbor (kNN) method (k-Nearest Neighbors Algorithm). This k-neighbors matrix can be computed by, for instance the function *knn2nb* of the R package *spdep* [32] of the software R [33].

From a theoretical point of view, the mode, when it exists, is an observation whose probability is locally maximum. So the modal curve of the sample S can be estimated as follows:

$$X_{modal,S} = \arg \max_{X_i \in S,\ t \in I} \sum_{s \in I} K\left(\frac{d_m(X_t, X_s)}{h}\right)$$

where $K(\cdot)$ is a kernel function, $h = h_n$ is a sequence of positive numbers called bandwidth, considered as a smoothing parameter. The kernel K acts as a weight function: the larger is $d_m(x, X_s)$, and the smaller is $K\left(d_m(x, X_s)/h\right)$. This means that among all the curves X in S, the modal curve defines a spatial neighboring area, where the sample of curves is the most dense and dependent. The pertinence of this estimate of a modal curve and asymptotic properties are similar to that given in [14]. This last assumed a mixing condition on the spatial process and used it to measure the spatial heterogeneity of the data.

The elements K, $d(\cdot, \cdot)$, and h are essential in nonparametric estimation. In the functional context, a semimetric $d(\cdot, \cdot)$ is often used as a proximity measure. In particular, a semimetric based on the first q scores of a functional PCA, used in Section 7.4, is defined by

$$m_q^{PCA}(X_i, X_j)^2 = \int \left(X_i^{(q)}(t) - X_j^{(q)}(t)\right)^2 dt,$$

where $X^{(q)}$ denotes the vector of the first q-th scores components of X (see [29] for more details). A kernel K is a weighting function used in nonparametric estimation techniques. There exist a large variety of kernels in the FDA context, the most classical ones are the positive and symmetrical kernels, such as box, triangle, quadratic, and Gaussian (see [29]). In Section 7.4, we use the following kernel:

$$K(u) = \frac{3}{2}(1 - u^2)1_{(0,1)}(u).$$

We choose this kernel because it gives more relevant results (among several kernel functions investigated) from the classification point of view of the air quality data considered.

7.3.1 Methodology

The proposed methodology performs iteratively by splitting S into increasingly homogeneous classes. To measure the heterogeneity of a given sample S of curves, Dabo-Niang et al. [2] compared modal and mean curves by computing the Subsampling Heterogeneity Index (SHI). The median curve can also be used instead of the mean, e.g. when one wants to assign to the same group all curves that have the same shape but which are affected by some clearly horizontal shift (see [2]). The SHI is computed by using a large number L of randomly generated subsamples $S^{(l)} \subset S$ of the same size

$$\mathrm{SHI}_{mean}(S) = \frac{1}{L}\sum_{l=1}^{L} \frac{m(X_{modal,S^{(l)}}, X_{mean,S^{(l)}})}{m(X_{mean,S^{(l)}}, 0) + m(X_{modal,S^{(l)}}, 0)}, \tag{7.8}$$

where $m(X, 0)$ denotes the proximity measure between a function X and the constant null function 0. A large value of $m(X_{modal,S^{(l)}}, X_{mean,S^{(l)}})$ indicates that $X_{modal,S^{(l)}}$ and $X_{mean,S^{(l)}}$ have different behaviors according to $m(\cdot,\cdot)$. The larger $\mathrm{SHI}_{mean}(S)$ is, the more heterogeneous the sample S will be. However, since the goal is to decide if the set S should be split into G classes $S_1, \ldots, S_G$ another index is required. The splitting will be accepted if the heterogeneity in each class is smaller than before splitting. To this end, the Partitioning Heterogeneity Index (PHI) is considered. It is defined as a weighted average of the SHI over classes:

$$\mathrm{PHI}_{mean}(S; S_1, \ldots, S_G) = \frac{1}{\mathrm{Card}(S)}\sum_{g=1}^{G} \mathrm{Card}(S_g)\mathrm{SHI}_{mean}(S_g). \tag{7.9}$$

The larger PHI is, the more heterogeneous each class $S_1, \ldots, S_G$ is. Both $\mathrm{SHI}_{mean}(S)$ and $\mathrm{PHI}_{mean}(S; S_1, \ldots, S_G)$ are employed to define a score SC given by

$$\begin{aligned} \mathrm{SC} &= \mathrm{SC}_{mean}(S; S_1, \ldots, S_G) \\ &= \frac{\mathrm{SHI}_{mean}(S) - \mathrm{PHI}_{mean}(S; S_1, \ldots, S_G)}{\mathrm{SHI}_{mean}(S)} \end{aligned} \tag{7.10}$$

A positive score SC indicates a gain of homogeneity inside classes. The splitting is accepted if SC is greater than a fixed threshold τ. For instance, $\tau = 5\%$ indicates that a splitting is accepted if it brings more than 5% of homogeneity within classes. If the score is negative, then S does not require this splitting. A value of τ that is too small indicates that the considered splitting is not required and the gain in terms of homogeneity is not significant. The value of τ is chosen according to the type of the data and the purpose of the classification. It is analogous to the choice of the first-kind error in hypotheses testing. All the details concerning this methodology are given in [29].

Aside from the above splitting criteria, it is required to define classes of S. Ferraty and Vieu [29] proposed for independent data, a procedure to establish the subgroups $S_1, \ldots, S_G$ as well as their number G. The procedure is related

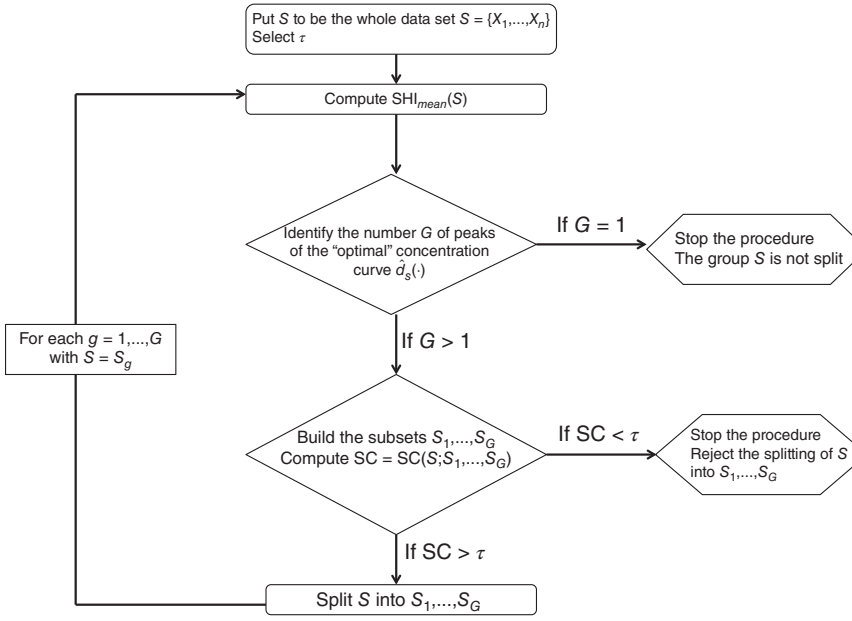


Figure 7.1 Algorithm of the descendant HC.

to the choice of the bandwidth parameters h. The choice of h is done by using the small ball probabilities. They play a key role in the theoretical properties of mode estimate (see [2]). A small ball probability is defined as $\mathbb{P}[X_i \in B(X,h)]$ for $X, X_i \in S$ which is the probability that a curve $X_i \in S$ belongs to the ball $B(X,h)$ with center X and radius h. For a given bandwidth h, one has at hand n probability points $\mathbb{P}[X_i \in B(X,h)], i = 1,\dots,n$ for which the corresponding density $d_{S,h}$ can be estimated by a kernel estimate $\hat{d}_{S,h}$. The estimated density can be computed using, for instance, the package *np* [34] of the R language [33]. The number of groups G will be determined by the number of peaks of $\widehat{d}_{S,\widehat{h}_S}$. In practice, the bandwidth is selected using the entropy such that $\widehat{h}_S = \arg\min_h \int \widehat{d}_{S,h}(t)\ \log \widehat{d}_{S,h}(t)\ dt$.

The main algorithm of this classification approach is illustrated in Figure 7.1. The reader is referred to [2, 14, 29] for more details. A *R* [33] code, in the context of i.i.d. data, is available at https://www.math.univ-toulouse.fr/ferraty/SOFTWARES/NPFDA/index.html.

7.4 Application

We illustrate the methodologies by using data of Ozone Concentration (OC) (units of measurement) collected in 106 monitoring stations of United States

Figure 7.2 Location of 106 monitoring stations (in the same number of cities) of ozone concentration in United States. Source: Environmental Protection Agency, https://www.epa.gov/outdoor-air-quality-data.

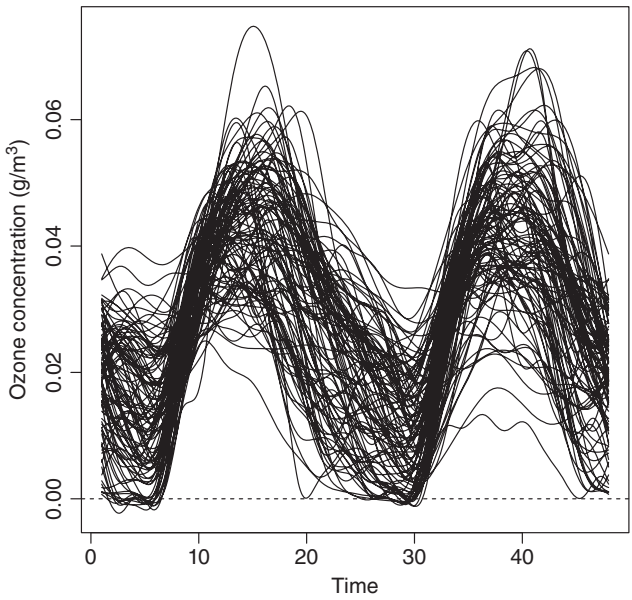


Figure 7.3 Ozone concentration curves (obtained after smoothing the data by using a Fourier basis) at 106 monitoring stations of the United States. Source: Environmental Protection Agency, https://www.epa.gov/outdoor-air-quality-data.

(see Figure 7.2) in 2015. The dataset is available at https://www.epa.gov/outdoor-air-quality-data.

Specifically, for each one of the 106 stations, we have data of OC recorded hourly from 19 July at 12 a.m. to 20 July at 11 p.m. (Figure 7.3). We use linear interpolation to estimate some missing values. We denote the OC at time t, $t \in [1, 48]$ as $X(t)$. In order to apply the methodologies described earlier, OC functions were obtained

(Figure 7.3) by expanding the discrete data (48 data at each station) in terms of 25 Fourier basis functions. The number of basis was chosen by using cross-validation.

7.4.1 Model-Based Clustering

We apply the EM algorithm for clustering the spatial functional data described above. A homoscedastic model has been applied since it gives more relevant results from the classification point of view, and the value of q was selected during the EM algorithm by maximizing the BIC computed at each step for each possible value of q.

Under this setting, the BIC indicates that two or three clusters could be appropriated (Figure 7.4).

In Figure 7.5, we show the classification of the monitoring stations in two and three groups, respectively. For the clustering in two clusters, $q = 18$ principal components have been retained. We see on the map that the obtained clustering well separates the East cities from the West cities. Moreover, we see on the curves that the clusters are also well separated from the curves point of view. On an average, we see in Figure 7.6 that West cities have higher pollution than East cities. For the clustering in three clusters, $q = 17$ principal components have been retained. We see on the map that the obtained clustering still separates well the East curves from the West curves, moreover it also separates the North from the South for the West side. When looking at these curves on Figure 7.6, we see that it is in the six first hours that cluster 1 (North) is the most different from cluster 3 (South).

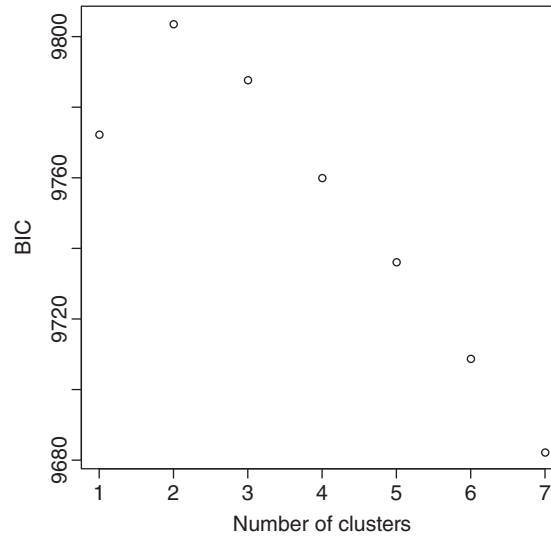


Figure 7.4 Value of BIC criterion according to the number of clusters.

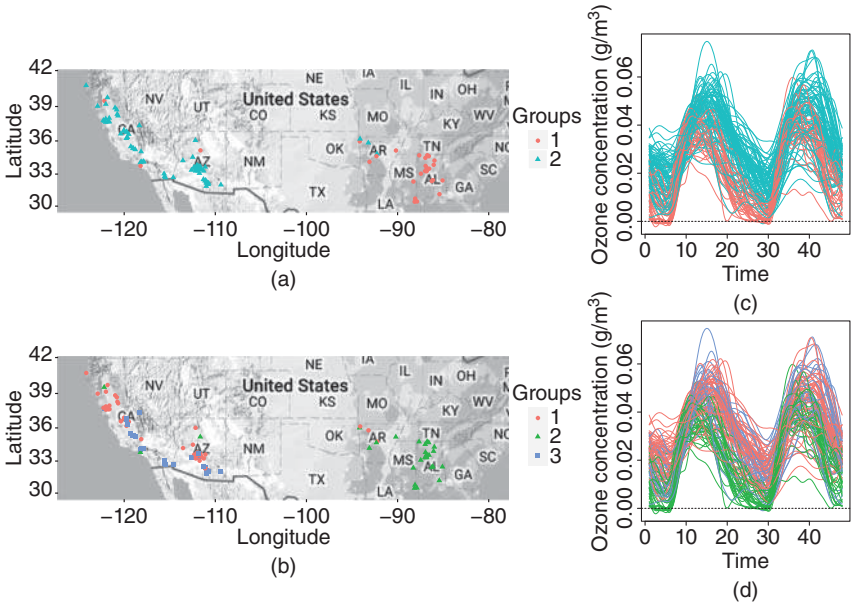


Figure 7.5 Locations of the stations are colored according to the cluster (a, b) and curves colored according to the cluster (c, d) for two clusters (c) and three clusters (d).

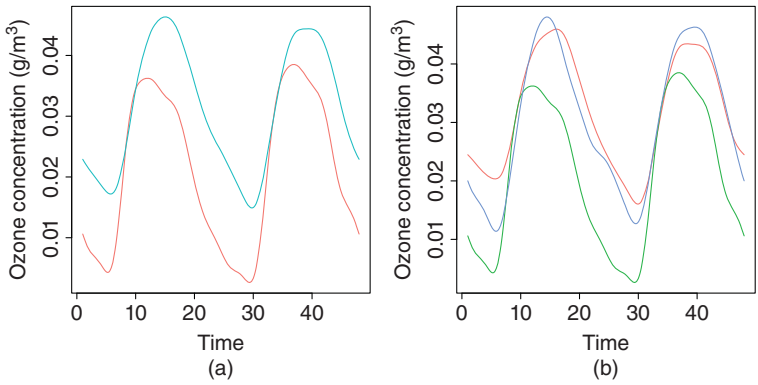


Figure 7.6 Average curves by cluster, respectively, for two clusters (a) and three clusters (b).

As a conclusion of this application, for the clusterings (with two or three clusters), let us observe that the method makes a trade-off between the geographical proximity and the common features of the curves, which allow to take into account spatial dependency while performing clustering. We see on the application that the obtained results are easily interpretable and give a relevant spatial segmentation.

7.4.2 Hierarchical Classification

The previous algorithm has been used on our set of 106 curves with the threshold parameter (τ) fixed at 25%, and $L = 0$ for reducing the computational cost (i.e. HI is used instead of SHI). Let us denote by S the whole sample of curves, given in Figure 7.3.

Recall that the spatial weighted matrix W_n is constructed by taking k neighbors of each unit using kNN method.

At the first iteration, the number of neighbors is equal to $k = 4$ and a number $q = 8$ of eigenvalues used in the semimetric $d(,\cdot,)$ (different values have been taken, this last gives better results in terms of homogeneity) and the data is split into two groups (CLASS 1 and CLASS 2 of, respective, sizes 5 and 101), since the corresponding concentration density $\hat{d}_S$ had two modes. This splitting is accepted since the gain of homogeneity (score SC) is larger (26%) than the threshold. Then, for the second iteration, the second (CLASS 2) group is split into two subgroups (CLASS 21, CLASS 22) with $k = 1$ and $q = 7$ and splitting score equal to 43%. At the third iteration, only CLASS 21 is split into two subgroups (CLASS 211 and CLASS 212) with $k = 1$ and $q = 6$ and splitting score equal to 55%. The procedure has been stopped with all the splitting scores smaller than 25% at the fourth iteration.

In Figure 7.7, the results of the different iterations of our procedure are presented. The sizes of the final groups CLASS 1, CLASS 22, CLASS 211, and CLASS 212 are 5, 51, 23, and 27, respectively. The mean and mode curves are given in Figure 7.8.

CLASS 1 (respectively CLASS 22) concerns essentially stations with higher main peak (around the middle of a day) ozone concentration (respectively smaller) than in other groups, particularly for the first day. The main difference

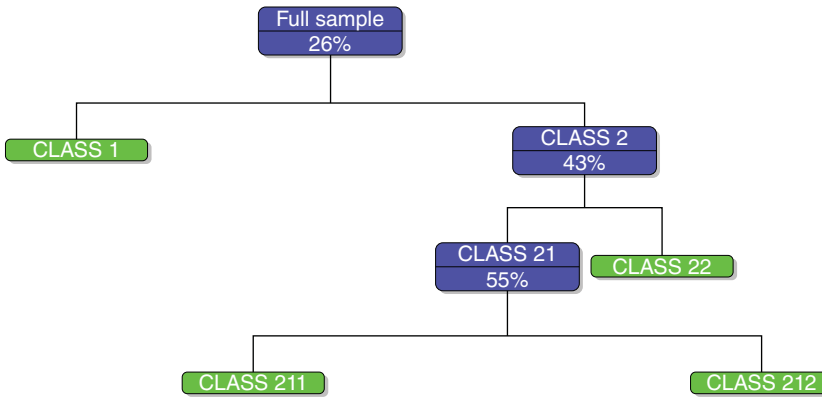


Figure 7.7 The classification results of the descendant HC.

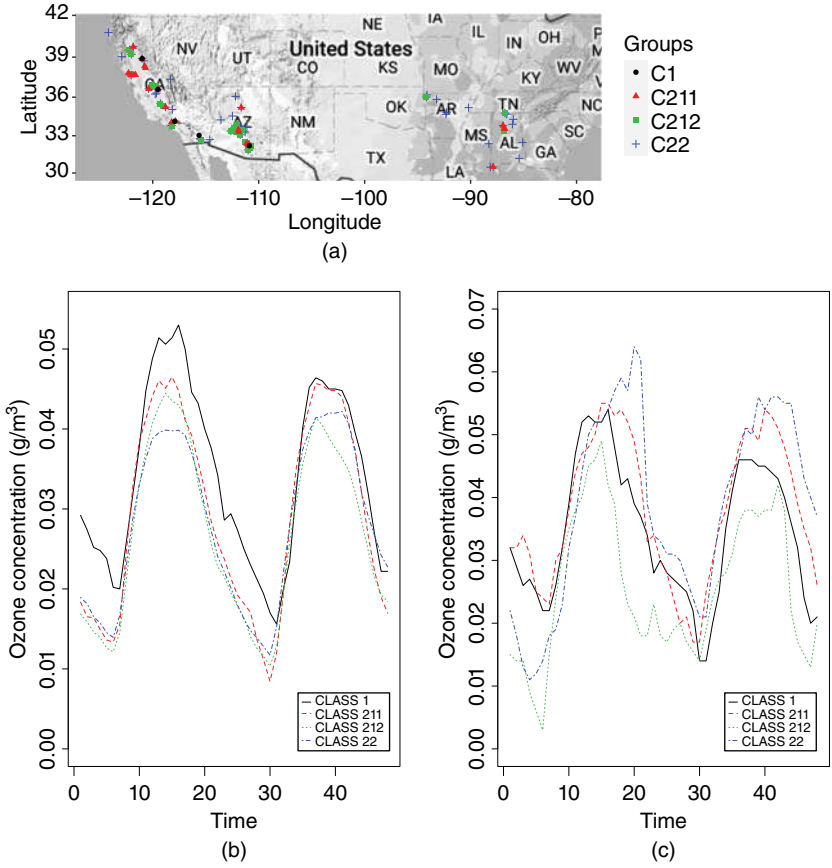


Figure 7.8 Locations of the stations colored according to the cluster (a), mean (b), and mode (c) curves colored according to the cluster (b, c).

between CLASS 211 and CLASS 212 comes also from the importance of the ozone concentration peak (smaller for the second group) and the width of their bases (larger for CLASS 211). It seems that at each iteration, the algorithm splits according to the maximum of the ozone concentration.

Regarding Figures 7.8 and 7.9, we may say that as in the first method, the clusters are separated from the curves point of view and geographical proximity. We see on the map of Figure 7.8 that there are mainly two groups of curves from the west (CLASS 1, CLASS 211, and CLASS 212) cities with higher pollution, while CLASS 22 (with the largest number of curves) is distributed in the west and east parts and has mainly smaller ozone concentration than the other groups. The CLASS 211 is mainly located in the west part.

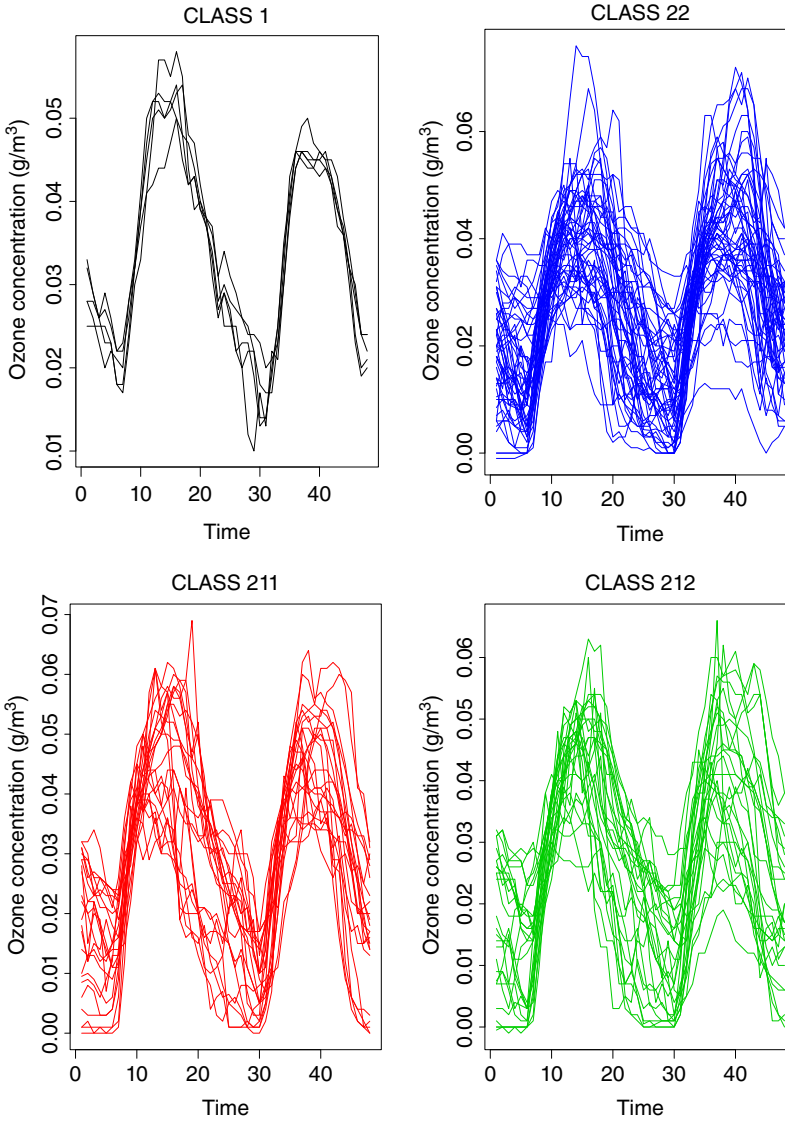


Figure 7.9 The curves of the different groups by the descendant HC.

7.5 Conclusion

The purpose of this chapter is the classification of functional spatial curves using the functional framework. Two functional classification methods are considered,

namely, descendant HC based on modal curve and the model-based clustering using a mixture model. These functional classification methods are presented and applied to ozone concentration data. We see on the application that the obtained results are interpretable and give a relevant spatial segmentation.

Although this work is mainly practical, consistency results may be easily obtained, see the references therein. An advantage of spatial functional approaches is that they allow the clustering to take in account some spatial dependency. For the two different methods considered, the different classification results allow to identify two main ozone concentration area. The first located in the west is characterized by large peak, and the second is characterized by high volume, it is located in the east and west parts. Two separate clustering could be tried in the west and east parts of the considered region. This could be adapted in order to account for regional spatial variability. Future efforts can focus on adapting the distance measure used in the hierarchical method to the classification objective (risk analysis, survival analysis, etc.) and to the data record length and quality.

References

1 Abraham, C., Biau, G., and Cadre, B. (2006). On the kernel rule for function classification. *Annals of the Institute of Statistical Mathematics* 58 (3): 619–633.

2 Dabo-Niang, S., Ferraty, F., and Vieu, P. (2007). On the using of modal curves for radar waveforms classification. *Computational Statistics and Data Analysis* 51 (10): 4878–4890.

3 Auder, B. and Fischer, A. (2012). Projection-based curve clustering. *Journal of Statistical Computation and Simulation* 82 (8): 1145–1168.

4 Abraham, C., Cornillon, P.A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics* 30 (3): 581–595.

5 Chiou, J.M. and Li, P.L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 (4): 679–699.

6 Cuevas, A., Febrero, M., and Fraiman, R. (2001). Cluster analysis: a further approach based on density estimation. *Computational Statistics and Data Analysis* 36 (4): 441–459.

7 García-Escudero, L.A. and Gordaliza, A. (2005). A proposal for robust curve clustering. *Journal of Classification* 22 (2): 185–201.

8 Romano, E., Mateu, J., and Giraldo, R. (2015). On the performance of two clustering methods for spatial functional data. *AStA Advances in Statistical Analysis* 99 (4): 467–492.

9 Tarpey, T. and Kinateder, K.K.J. (2003). Clustering functional data. *Journal of Classification* 20 (1): 93–114.

10 Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis* 71: 92–106.

11 Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification* 8 (3): 231–255.

12 Floriello, D. and Vitelli, V. (2017). Sparse clustering of functional data. *Journal of Multivariate Analysis* 154: 1–18.

13 James, G.M. and Sugar, C.A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* 98 (462): 397–408.

14 Dabo-Niang, S., Yao, A.F., Pischedda, L. et al. (2010). Spatial mode estimation for functional random fields with application to bioturbation problem. *Stochastic Environmental Research and Risk Assessment* 24 (4): 487–497.

15 Giraldo, R., Delicado, P., and Mateu, J. (2012). Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica* 66 (4): 403–421.

16 Giraldo, R., Delicado, P., and Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics* 18 (3): 411–426.

17 Romano, E., Balzanella, A., and Verde, R. (2017). Spatial variability clustering for spatially dependent functional data. *Statistics and Computing* 27 (3): 645–658.

18 Milligan, G.W. and Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50 (2): 159–179.

19 Krzanowski, W.J. and Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* 44 (1): 23–34.

20 Cuevas, A., Febrero, M., and Fraiman, R. (2000). Estimating the number of clusters. *Canadian Journal of Statistics* 28 (2): 367–382.

21 Klemelä, J. (2008). Density estimation with locally identically distributed data and with locally stationary data. *Journal of Time Series Analysis* 29 (1): 125–141.

22 Cheam, A.S.M., Marbac, M., and McNicholas, P.D. (2017). Model-based clustering for spatiotemporal data on air quality monitoring. *Environmetrics* 28 (3): e2437.

23 Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition* 28 (5): 781–793.

24 Bouveyron, C. and Jacques, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification* 5 (4): 281–300.

25 Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics* 38(2): 1171–1193.

26 Jacques, J. and Preda, C. (2013). Funclust: a curves clustering method using functional random variables density approximation. *Neurocomputing* 112: 164–171.

27 Ruiz-Medina, M.D., Espejo, R.M., and Romano, E. (2014). Spatial functional normal mixed effect approach for curve classification. *Advances in Data Analysis and Classification* 8 (3): 257–285.

28 Bouveyron, C., Girard, S., and Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics and Data Analysis* 52 (1): 502–519.

29 Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media.

30 Cadre, B. (2001). Convergent estimators for the l_1-median of banach valued random variable. *Statistics: A Journal of Theoretical and Applied Statistics* 35 (4): 509–521.

31 Pinkse, J. and Slade, M.E. (1998). Contracting in space: an application of spatial statistics to discrete-choice models. *Journal of Econometrics* 85 (1): 125–154.

32 Bivand, R., Altman, M., Anselin, L. et al. (2015). Package? spdep? https://cran.r-project.org/web/packages/spdep (accessed 9 December 2015).

33 R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

34 Hayfield, T., Racine, J.S. (2008). Nonparametric econometrics: the np package. *Journal of Statistical Software* 27 (5): 1–32.

8

Nonparametric Statistical Analysis of Spatially Distributed Functional Data

Sophie Dabo-Niang[1], Camille Ternynck[2], Baba Thiam[1], and Anne-Françoise Yao[3]

[1] *University of Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, INRIA-MODAL, F-59000 Lille, France*
[2] *University of Lille, URL 2694 -METRICS, F-59000 Lille, France*
[3] *Université Clermont Auvergne - UMR 6620 - CNRS, France*

8.1 Introduction

The spatial indexing, which provides geographical reference of data, is encountered in many subject areas such as oceanography, epidemiology, forestry survey, and economy. As a consequence, the scientific research community is increasingly interested in analyzing spatial data and then in developing more and more efficient spatial statistical tools. Early spatial models appeared at the beginning of the nineteenth century and are mainly related to parametric spatial statistics modeling (see [1–5] for more details on statistics for spatial data). The nonparametric methods are able to reveal structure in data that might be missed by classical parametric ones. Nowadays, a dynamic concerns the deployment of nonparametric methods to spatial statistics such as density estimation, regression, or prediction (e.g. [6–11]). However, most of nonparametric spatial contributions deal with univariate or multivariate data, whereas recent advances of real-time measurement instruments and data storage resources led to the emergence of functional data. The studied objects can then be curves, not variables or vectors of variables. This kind of data is more and more frequently involved in statistical problems since the 1990s. For an introduction to this field, the reader is directed to the books of [12–14].

Currently, the literature on spatial nonparametric statistics for functional data is not extensive (see for instance, [15–23], among others) compared to parametric models (see, e.g. Chapters 2–4, 13, and 15 of this book).

The baseline of this current chapter is nonparametric regression estimation for functional data presenting spatial dependence. The goal is also to predict

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

unsampled locations by taking into account neighborhood similarities imposed by both the geographic proximity and the values of available explanatory variables. To the best of our knowledge, very little research deals with this issue. Among the nonparametric methods, the usual kernel density estimator (see [24]) is often used in order to estimate the regression operator. In [11], a nonparametric kernel prediction is considered for spatial stochastic processes when a stochastic sampling design is assumed for selection of random locations. The particularity of this predictor is to be constructed with a kernel function on the locations. In the kernel-type estimator suggested in [25], the dependence structure is reduced to the estimation of one indicator variogram, as a nonparametric alternative to Matheron's indicator variogram. Wang et al. [26] proposed a local linear spatiotemporal prediction model, using a kernel weight function taking into account the distance between sites. The works of [22] and [23] proposed, respectively, a spatial density and regression estimators, for multivariate data, depending on two kernels, one of which controls the distance between observations and the other controls the spatial dependence structure. All these previous works concern real-valued data. The spatial kernel density estimator proposed in [18] for functional data does not directly take into account the spatial dependency in the form of the estimator, but the authors explained how this can be done by introducing a second kernel, based on distances between sites. Here, our interest lies in proposing spatial nonparametric regression and prediction approaches by combining these three last works using more general conditions adapted to some local identically distributed and spatially dependent variables. The originality of the suggested approach is to take advantage of each regression or prediction method introduced previously. In fact, our regression estimate uses some local identically distributed observations and permits to propose predictions based on local geographic proximity of the sites. A similar idea has been presented in [27] to deal with a functional regression problem for strictly stationary processes. The idea of incorporating explicitly a spatial correlation structure into a nonparametric regression estimator assuming that the error term is a second-order stationary error process with a parametric correlation model, has been used previously for real-valued data in [28] and [29]. These authors used local linear regression estimator and a generalized cross-validation criterion for the effect of spatial correlation. Despite these efforts, an explicit general spatial proximity structure into a kernel predictor is still not available. The present work goes along this direction and takes advantage of these previous works. Our model is different since there is no parametric correlation model on the error term, and the observations are not from strictly stationary processes.

Denote the integer lattice points in the N-dimensional Euclidean space by $\mathbb{Z}^N$, $N \geq 1$. We consider a spatial process $\{Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}), \mathbf{i} \in \mathbb{Z}^N\}$ defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

A point in bold $\mathbf{i} = (i_1, \dots, i_N) \in \mathbb{Z}^N$ will be referred as a site. Suppose $X_{\mathbf{i}}$ takes values in a separable semimetric space $(\mathcal{E}, d(\cdot,\cdot))$ of eventually infinite dimension (i.e. $X_{\mathbf{i}}$ is a functional random variable and d a semimetric) and $Y_{\mathbf{i}}$ takes values in $\mathbb{R}$. In the following, $\|\cdot\|$ will denote any norm in $\mathbb{R}^d$ or $\mathbb{R}^N$ (there will be no ambiguity since the vectors of $\mathbb{R}^N$ are in bold), C and C' will indicate some arbitrary positive constants that may vary from line to line, for each real u, $\lfloor u \rfloor$ will indicate the integer part of u. Moreover, we write $u_{\mathbf{n}} = O(v_{\mathbf{n}})$ means that $\exists C$ such that $\left|u_{\mathbf{n}}/v_{\mathbf{n}}\right| \leq C$ as $v_{\mathbf{n}} \to \infty$ and $u_{\mathbf{n}} = o(v_{\mathbf{n}})$ means that $\left|u_{\mathbf{n}}/v_{\mathbf{n}}\right| \to 0$ as $v_{\mathbf{n}} \to \infty$, where $\mathbf{n} \in \mathbb{R}^N$.

As it is classically assumed in the literature, the process under study $(Z_{\mathbf{i}})$ is observable over the rectangular domain $\mathcal{I}_{\mathbf{n}} = \left\{\mathbf{i} = (i_1, \dots, i_N), 1 \leq i_k \leq n_k,\ k = 1, \dots, N\right\}$, where a point $\mathbf{i} \in \mathbb{Z}^N$ refers to a site. We denote $\mathbf{n} = (n_1, \dots, n_N)$ and let $\widehat{\mathbf{n}} = n_1 \times \cdots \times n_N$ be the sample size. From now on, we assume for simplicity that $n_1 = n_2 = \cdots = n_N = n$ (e.g. [30, 31] and [32]), but the following results can be extended to a more general framework. We write $\mathbf{n} \to \infty$ if $n \to \infty$.

We do not suppose strict stationarity. We will assume that the variables $(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}}$ are locally identically distributed (see for instance [33] who considered density estimation for local identical time-series data): a sufficient number of $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ has a distribution close to that of a couple (X, Y). One may imagine that when $\mathbf{i}$ is close to some $\mathbf{i_0}$, and if there is enough sites $\mathbf{i}$ close to $\mathbf{i_0} \notin \mathcal{I}_{\mathbf{n}}$, then sequence $(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}}$ may be used to predict $Y_{\mathbf{i_0}}$.

We suppose that the spatial process satisfies the following nonparametric regression model: $Y_{\mathbf{i}} := r(X_{\mathbf{i}}) + \epsilon_{\mathbf{i}}$, where $r(x) = \mathbb{E}(Y_{\mathbf{i}} | X_{\mathbf{i}} = x)$ is assumed to be independent of $\mathbf{i}$, the noise $\epsilon_{\mathbf{i}}$ is centered, α-mixing, and independent of $X_{\mathbf{i}}$.

The regression estimate is defined by

$$r_{\mathbf{n}}(x) = \begin{cases} \frac{g_{\mathbf{n}}(x)}{f_{\mathbf{n}}(x)} \\ \frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} \quad \text{otherwise,} \end{cases} \tag{8.1}$$

where the functions $f_{\mathbf{n}}$ and $g_{\mathbf{n}}$ are defined by

$$f_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K\left(\frac{d(x, X_{\mathbf{i}})}{h_{\mathbf{n}}}\right) \quad \text{and} \quad g_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K\left(\frac{d(x, X_{\mathbf{i}})}{h_{\mathbf{n}}}\right),$$

with $a_{\mathbf{n}} = \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} \mathbb{E}\left[K\left(\frac{d(x, X_{\mathbf{i}})}{h_{\mathbf{n}}}\right)\right]$.

The model in Eq. (8.1) is used to predict $Y_{\mathbf{i_0}}$ at a location $\mathbf{i_0}$ in $\mathbb{Z}^N$ using as input the information $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ available at the neighboring locations.

The rest of the chapter is organized as follows. In Section 8.2, we provide the assumptions and large sample properties. An application to prediction is given in Section 8.3. Two different predictors are studied. To check the performance of the proposed methodology, numerical results are reported in Section 8.4. Conclusion is given in Section 8.5, while proofs and technical lemmas are postponed in Appendix 8.A.

8.2 Large Sample Properties

We first introduce some mixing assumptions. In fact, to take into account the spatial dependency, we assume that the process $\{Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}),\ \mathbf{i} \in \mathbb{Z}^N\}$ satisfies a mixing condition defined in [8] as follows: there exists a function $\chi(t) \searrow 0$ as $t \to \infty$, such that

$$\begin{aligned}\alpha(\sigma(S), \sigma(S')) &= \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|, A \in \sigma(S), B \in \sigma(S')\},\\ &\leq \psi(\mathrm{Card}(S), \mathrm{Card}(S'))\chi(\mathrm{dist}(S, S')),\end{aligned}$$

where $\mathrm{dist}(S, S')$ is the Euclidean distance between the two finite sets of sites S and S', $\mathrm{Card}(S)$ denotes the cardinality of the set S, $\sigma(S)$, and $\sigma(S')$ denote, respectively, the σ-fields generated by $\{Z_{\mathbf{i}}, \mathbf{i} \in S\}$, and $\{Z_{\mathbf{i}}, \mathbf{i} \in S'\}$, ψ is a positive symmetric function nondecreasing in each variable. We recall that the process is said to be strongly mixing if $\psi \equiv 1$. As usual, we will assume that one of both following conditions on $\chi(.)$ is verified. These conditions are defined by

$$\chi(i) \leq Ci^{-\theta}, \text{for some } \theta > 0, \tag{8.2}$$

i.e. that $\chi(i)$ tends to zero at a polynomial rate, or

$$\chi(i) \leq C\exp(-si), \quad \text{for some } s > 0,$$

i.e. that $\chi(i)$ tends to zero at an exponential rate. Concerning the function $\chi(\cdot)$, for the sake of simplicity, we will only study the case where $\chi(\cdot)$ tends to zero at a polynomial rate. However, similar asymptotic results could be obtained with $\chi(\cdot)$ tending to zero at an exponential rate (which implies the polynomial case). In what follows, it will be assumed that ψ satisfies either

$$\forall n, m \in \mathbb{N}, \quad \psi(n, m) \leq C\min(n, m), \tag{8.3}$$

or

$$\psi(n, m) \leq C(n + m + 1)^{\kappa}, \tag{8.4}$$

for some $C > 0$, and some $\kappa \geq 1$. Such functions $\psi(n, m)$ can be found, for instance, in [7–9, 21, 34].

Let $u_{\mathbf{n}} = \prod_{i=1}^{N}(\log n_i)(\log n_i)^{1+\epsilon}$ for $\epsilon > 0$, then $\sum_{\mathbf{n}\in\mathbb{N}^N}^{N} 1/\hat{\mathbf{n}}u_{\mathbf{n}} < \infty$.

We will denote by $p_{\mathbf{i}}$ the probability distribution of $X_{\mathbf{i}}$ and by $p_{\mathbf{i},\mathbf{j}}$, the joint probability distribution of $(X_{\mathbf{i}}, X_{\mathbf{j}})$, for all $\mathbf{i}$ and $\mathbf{j}$. The small ball probabilities are denoted by $\varphi_{\mathbf{i},x}(h) = \mathbb{P}[X_{\mathbf{i}} \in B(x, h)]$, recall that $\varphi_{\mathbf{i},x}(h)$ goes to zero when h goes to zero (see, e.g. [14] for more details).

For any random variable Z and $p \in \mathbb{N}^*$, $\|Z\|_p = (\mathbb{E}[|Z|^p])^{1/p}$.

The mean square consistency result of $r_{\mathbf{n}}$ is obtained under the following assumptions on r, the kernel, the bandwidth, and local dependence condition.

-H1: The kernel $K : \mathbb{R} \to \mathbb{R}^+$ is of integral 1 such that there exist two constants C_1 and C_2 with $0 < C_1 < C_2 < \infty$, such that

$$C_1 \mathbf{1}_{[0,1]}(t) \leq K(t) \leq C_2 \mathbf{1}_{[0,1]}(t).$$

-H2: r is a Lipschitz function, that is $r \in Lip_{\mathcal{E}}$, where

$$Lip_{\mathcal{E}} = \{f : \mathcal{E} \to \mathbb{R}, \exists C_3 \in \mathbb{R}_*^+, \forall x, x' \in \mathcal{E}, |f(x) - f(x')| < C_3 d(x, x')\}.$$

-H3: **(i) Local dependence condition:** For all $\mathbf{i} \neq \mathbf{j} \in \mathbb{N}^N$, the joint probability distribution $p_{\mathbf{i},\mathbf{j}}$ of $X_{\mathbf{i}}$ and $X_{\mathbf{j}}$ satisfies

$$\exists \varepsilon \in (0, 1], p_{\mathbf{i},\mathbf{j}}(B(x, h_{\mathbf{n}}) \times B(x, h_{\mathbf{n}})) \leq C_4 (\varphi_{\mathbf{i},x}(h_{\mathbf{n}}) \varphi_{\mathbf{j},x}(h_{\mathbf{n}}))^{\frac{1+\varepsilon}{2}}.$$

(ii) Small ball probabilities: For all $\mathbf{i}$ and x, there exist positive constants C_1' and C_2' and a function $\varphi_x(h)$ tending to zero as h goes to zero such that

$$0 < C_1' \varphi_x(h) \leq \varphi_{\mathbf{i},x}(h) \leq C_2' \varphi_x(h).$$

Remark 8.1 *These assumptions are very standard in the context of spatial nonparametric modeling. Indeed, Assumptions **H1** and **H2** allow to control the bias of the estimator. Assumption **H1** is satisfied, for instance, by several kernels with compact support such as triangular (Bartlett), biweight, triweight, Epanechnikov, and Parzen kernels. The Lipschitz condition **H2** allows the precise rate of convergence to be found, whereas a continuity-type model would give only convergence results. Local dependence condition **H3**(i) is a classical condition in kernel estimation based on dependent data nonnecessarily, strictly stationary (see, e.g. [8, 35] and [36]).*

In order to control the constraints on the bandwidth sequence due to the mixing coefficients with polynomial decreasing rate (8.2), we define

$$\gamma_1 = \frac{2N - \theta}{4N - \theta} \quad \text{and} \quad \gamma_1^* = \frac{N - \theta}{N(3 + 2\kappa) - \theta}.$$

The following result permits to have a bound of the mean squared error of $r_{\mathbf{n}}$.

Theorem 8.1 *Assume that assumptions **H1**–**H3** hold with $|Y_{\mathbf{i}}| \leq M$.*

(1) If (8.3) is satisfied and

$$\widehat{\mathbf{n}} \varphi_x(h_{\mathbf{n}})^{\gamma_1} \left(\log \widehat{\mathbf{n}}\right)^{-\gamma_1} \to \infty \text{ with } \theta > 4N,$$

or

(2) if (8.4) is satisfied and

$$\widehat{\mathbf{n}} \varphi_x(h_{\mathbf{n}})^{\gamma_1^*} \left(\log \widehat{\mathbf{n}}\right)^{-\gamma_1^*} \to \infty \text{ with } \theta > (3 + 2\kappa)N,$$

then

$$\|r_{\mathbf{n}}(x) - r(x)\|_2 = O\left(h_{\mathbf{n}} + \sqrt{\frac{1}{\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})}}\right).$$

Precisely, we have

$$\|r_{\mathbf{n}}(x) - r(x)\|_2 = C_3 \times h_{\mathbf{n}} + \left(2C(2MC_2 + 2M\sqrt{C_4} + C_0) + 4M\right) \times \sqrt{\frac{1}{\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})}},$$

where C_0 is a constant depending on the constant appearing in Lemma 8.1.

Remark 8.2 *The conditions on the bandwidth in Theorem 8.1 are classical technical assumptions, which appear (in the calculations when studying the asymptotic behavior of the estimator) in the particular case where the mixing coefficient is such that χ tends to zero at a polynomial rate (see [37] and [38] for some examples). Each of these conditions is related to a specific case of mixing in the spatial context and are used, respectively, in [37] and [39].*

8.2.1 Uniform Almost Complete Convergence

We consider a set $\mathcal{D}$ such that $\mathcal{D} \subset \bigcup_{k=1}^{\upsilon_{\mathbf{n}}} B_k$, where $B_k = B(x_k,\ \ell_{\mathbf{n}})$ (note that such set can always be built), $\upsilon_{\mathbf{n}} > 0$ is some integer, $x_k \in \mathcal{E},\ k = 1, \ldots, \upsilon_{\mathbf{n}}$, and $\ell_{\mathbf{n}} > 0$. We assume that

-H4 There exists a nonincreasing positive function Γ such that,

(i) $\lim_{\mathbf{n}\to\infty} \Gamma(h_{\mathbf{n}}) = 0$ and

$$0 < C_1'' \Gamma(h_{\mathbf{n}}) \leq \varphi_{\mathbf{i},x}(h_{\mathbf{n}}) \leq C_2'' \Gamma(h_{\mathbf{n}}), \text{for all } \mathbf{i}, x \in \mathcal{D},$$

where C_1'' and C_2'' are some constants.

(ii) $\lim_{\mathbf{n}\to\infty} \frac{\widehat{\mathbf{n}}\Gamma(h_{\mathbf{n}})}{\log \widehat{\mathbf{n}}} \to \infty$.

(iii) $\upsilon_{\mathbf{n}} = \widehat{\mathbf{n}}^{\beta}$ for some $\beta > 0$.

-H5 **Local dependence condition:** For any $\mathbf{i} \neq \mathbf{j} \in \mathbb{N}^N$, the joint probability distribution $p_{\mathbf{ij}}$ of $X_{\mathbf{i}}$ and $X_{\mathbf{j}}$ satisfies

$$\exists \varepsilon \in (0, 1], p_{\mathbf{ij}}(B(x, h_{\mathbf{n}}) \times B(x, h_{\mathbf{n}})) \leq C_3''(\Gamma(h_{\mathbf{n}}))^{1+\varepsilon},\ \text{for all } x \in \mathcal{D}.$$

-H6 There exists $s > 2$ and $C > 0$ such that

(i) $\sup_{\mathbf{i}} \mathbb{E}\left(\left|Y_{\mathbf{i}}\right|^s \middle| X_{\mathbf{i}}\right) < C$.

(ii) $\sup_{\mathbf{ij}} \mathbb{E}\left(\left|Y_{\mathbf{i}} Y_{\mathbf{j}}\right| \middle| X_{\mathbf{i}}, X_{\mathbf{j}}\right) < C$ for some constant $C > 0$.

Let us introduce the following functions of the mixing coefficient which is related to the conditions on the bandwidth and the moment of the functional covariate:

$$\theta_1 = \frac{2s(N-\theta)}{2Ns(\beta+2)+\theta(2-s)},\quad \theta_2 = \frac{(\theta-2N)s}{2Ns(\beta+2)+\theta(2-s)},$$
$$\theta_3 = \frac{2(Ns+\theta)}{2Ns(\beta+2)+\theta(2-s)},\quad \theta_1^* = \frac{s(-N-\theta)}{N(2s\beta+2s\kappa+s+2)+\theta(2-s)},$$
$$\theta_2^* = \frac{s(\theta-N)}{N(2s\beta+2s\kappa+s+2)+\theta(2-s)},\ \theta_3^* = \frac{2(N+\theta)}{N(2s\beta+2s\kappa+s+2)+\theta(2-s)}.$$

The following theorem gives an uniform almost sure convergence of the regression estimate.

Theorem 8.2 *Assume that assumptions **H1**–**H6** hold.*

(i) If (8.3) is satisfied and

$$\widehat{\mathbf{n}}\Gamma(h_{\mathbf{n}})^{\theta_1}\left(\log\widehat{\mathbf{n}}\right)^{\theta_2}u_{\mathbf{n}}^{\theta_3}\to\infty \text{ with } \theta > 2Ns(\beta+2)/(s-2), \tag{8.5}$$

(ii) or if (8.4) is satisfied and

$$\widehat{\mathbf{n}}\Gamma(h_{\mathbf{n}})^{\theta_1^*}\left(\log\widehat{\mathbf{n}}\right)^{\theta_2^*}u_{\mathbf{n}}^{\theta_3^*}\to\infty \text{ with } \theta > N(2s\beta+2s\kappa+s+2)/(s-2), \tag{8.6}$$

then

$$\sup_{x\in D}\left|r_{\mathbf{n}}(x)-r(x)\right| = O\left(h_{\mathbf{n}}+\sqrt{\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\,\Gamma(h_{\mathbf{n}})}}\right) \text{ a.s.}$$

Recall that [18] gave an uniform almost sure bound of their regression estimate on a specific set C that is $O\left(h_{\mathbf{n}}^{\bigstar}+\sqrt{\frac{\log\widehat{\mathbf{n}}}{\Gamma(h_{\mathbf{n}}^{\bigstar})\widehat{\mathbf{n}}}}\right)$ with $\Gamma(h_{\mathbf{n}}^{\bigstar}) = \sup_{x\in C}\varphi_x(h_{\mathbf{n}})^{\bigstar}$ when the considered process is strictly stationary.

8.3 Prediction

This section is concerned with the problem of predicting the process $\{Y_{\mathbf{i}},\ \mathbf{i}\in\mathbb{Z}^N\}$ at some unobserved locations, and more particularly, to predict the unobserved value $Y_{\mathbf{i}_0}$ at a location $\mathbf{i_0}\in\mathbb{Z}^N$, $\mathbf{i_0}\notin\mathcal{O}_{\mathbf{n}}$, where $\mathcal{O}_{\mathbf{n}}\subset\mathcal{I}_{\mathbf{n}}$ is the observed spatial set of finite cardinality tending to ∞ as $\mathbf{n}\to\infty$. The spatial dependence implies the need to determine which other units in $\mathcal{O}_{\mathbf{n}}$ have an influence on the considered location $\mathbf{i_0}$.

Let $(X_{\mathbf{i}_0}, Y_{\mathbf{i}_0})$ be of same distribution as (X, Y). As said in the introduction, one may imagine that when $\mathbf{i}$ is close to $\mathbf{i}_0$, and if there are enough sites $\mathbf{i}$ close to $\mathbf{i}_0$, then the sequence $(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}}$ may be used to predict $Y_{\mathbf{i}_0}$.

The predictor of $Y_{\mathbf{i}_0}$ derived from the regression estimate $r_{\mathbf{n}}$ is

$$\widehat{Y}_{\mathbf{i}_0} = r_{\mathbf{n}}(X_{\mathbf{i}_0}) = \frac{\sum_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}} Y_{\mathbf{i}} K\left(\frac{d(X_{\mathbf{i}_0}, X_{\mathbf{i}})}{h_{\mathbf{n}}}\right)}{\sum_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}} K\left(\frac{d(X_{\mathbf{i}_0}, X_{\mathbf{i}})}{h_{\mathbf{n}}}\right)}, \tag{8.7}$$

if the denominator is not null; otherwise, $\widehat{Y}_{\mathbf{i}_0}$ is the empirical mean of the observed $Y_{\mathbf{i}}$.

Note that this predictor does not take into account the spatial proximity. To take into account the spatial locations, we consider another predictor defined by

$$\tilde{Y}_{\mathbf{i}_0} = \frac{\sum_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{d(X_{\mathbf{i}_0}, X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|)}{\sum_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}} K_1\left(\frac{d(X_{\mathbf{i}_0}, X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|)}, \tag{8.8}$$

if the denominator is not null; otherwise, $\tilde{Y}_{\mathbf{i}_0}$ is the empirical mean of the observed $Y_{\mathbf{i}}$. The kernels K_1 and K_2 are defined on $\mathbb{R}$, $b_{\mathbf{n}}$ and $\rho_{\mathbf{n}}$ are sequences of bandwidths tending to zero as $\mathbf{n} \to \infty$, and we write $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}_0 - \mathbf{i}\|) = K_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}_0-\mathbf{i}}{\mathbf{n}}\right\|\right)$, $\left(\frac{\mathbf{i}}{\mathbf{n}} = \left(\frac{i_1}{n}, \frac{i_2}{n}, \ldots, \frac{i_N}{n}\right)\right)$ quantifying the proximity between sites. Then we have $\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}_0-\mathbf{i}}{\mathbf{n}}\right\| \leq 1$ means that $\|\mathbf{i}_0 - \mathbf{i}\| \leq n\rho_{\mathbf{n}}$.

Thus, for the site $\mathbf{i}_0$, let $k_{\mathbf{n}} = k_{\mathbf{n},\mathbf{i}_0} = \sum_{\mathbf{i}} \mathbb{1}_{\{\|\mathbf{i}-\mathbf{i}_0\|\leq d_{\mathbf{n}}\}}$ denote the number of neighbors $\mathbf{i}$ for which the distance between $\mathbf{i}$ and $\mathbf{i}_0$ is less than or equal to distance $d_{\mathbf{n}} > 0$ such that $d_{\mathbf{n}} \to \infty$ as $\mathbf{n} \to \infty$. This last assumes that the proximity between locations (eventually) increases as the sample size increases. Taking the Euclidean distance and if $N = 2$ (square grid), we have $k_{\mathbf{n}} \leq 4d_{\mathbf{n}}^2 - 4d_{\mathbf{n}} + 4$ which leads to $k_{\mathbf{n}} = O(d_{\mathbf{n}}^2)$ and $k_{\mathbf{n}} = o(d_{\mathbf{n}}^{\eta}), \eta > 2$. Moreover, if $d_{\mathbf{n}} = o(\widehat{\mathbf{n}}^{\epsilon})$, $0 < \epsilon < 1$, then $k_{\mathbf{n}} = o(\widehat{\mathbf{n}}^{2\epsilon})$, See, for instance [40]. Let $d_{\mathbf{n}} = n\rho_{\mathbf{n}}$; consequently, we have $d_{\mathbf{n}}^2 = \widehat{\mathbf{n}}\rho_{\mathbf{n}}^N$ and $k_{\mathbf{n}} = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)$ as well as $k_{\mathbf{n}} = o((\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)^{\eta/2})$, $\eta > 2$. Note that the role of the kernel K_2 here is to handle the nearness between locations. The corresponding weights on the sites are assumed to decline as a measure of distance between corresponding sites (that are normalized) increases. The predictor $\widetilde{Y}_{\mathbf{i}_0}$ is a function of the number $k_{\mathbf{n}}$ of neighbors $\mathbf{i}$ for which the distance $d_{\mathbf{n}}$ is chosen hereafter to be $n\rho_{\mathbf{n}}$, with $k_{\mathbf{n}} \to +\infty$, $k_{\mathbf{n}} = O(d_{\mathbf{n}}^N) = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)$. If one assumes that $d_{\mathbf{n}} = o(\widehat{\mathbf{n}}^{\epsilon})$, $\epsilon \in (0, 1)$, then $k_{\mathbf{n}}$ can be expressed in terms of $\widehat{\mathbf{n}}$. In what follows, we assume that $k_{\mathbf{n}} = C_N d_{\mathbf{n}}^N + O(d_{\mathbf{n}}^{\beta})$ as $d_{\mathbf{n}} \to +\infty$, $0 < \beta < N$ and C_N is a constant that depends on N.

Remark 8.3

- *To give some examples where the assumption on $k_{\mathbf{n}}$ is reasonable, consider $q_{\mathbf{n}}$ as the number of standard lattice (in $\mathbb{Z}^N$) points contained in a closed ball $\mathcal{B}(\mathbf{i}_0, d_{\mathbf{n}})$*

with center $\mathbf{i_0}$ *and radius* $d_\mathbf{n}$ *that is* $q_\mathbf{n} = Card\{\mathbf{i} \in \mathbb{R}^N,\ \|\mathbf{i} - \mathbf{i_0}\| \leq d_\mathbf{n}\}$, *where* $\mathbf{i_0}$ *is any vector of* $\mathbb{R}^N$. *It is well known that*

$$q_\mathbf{n} = \frac{\pi^{N/2}}{\Gamma(N/2+1)} d_\mathbf{n}^N + O(d_\mathbf{n}^{N-1}),$$

where $\Gamma(\cdot)$ *is the gamma function; see, for instance, [41–43] and [44], and notice that* $k_\mathbf{n} = C_N q_\mathbf{n}$. *In particular, if* $N = 2$, *we have* $q_\mathbf{n} = \frac{\pi}{\Gamma(2)} d_\mathbf{n}^2 + O(d_\mathbf{n})$, $q_\mathbf{n} = \frac{\pi}{\Gamma(2)} d_\mathbf{n}^2 + o(d_\mathbf{n}^{2/3})$.

- *Note that although the predictor* $\tilde{Y}_{\mathbf{i_0}}$ *takes into account the spatial proximity, it does not measure the spatial dependency. However, before using this predictor, one could evaluate the importance of the dependence, for instance, by fitting a variogram (e.g. [29, 45]) on the data to be processed.*

In particular and for simplicity, as proposed in the numerical Section 8.4, let us consider in the following that $\mathbf{i_0} \in \mathcal{I}_\mathbf{n}$, $\mathcal{O}_\mathbf{n} = \mathcal{I}_\mathbf{n} \backslash \{\mathbf{i_0}\}$, then the previous predictors become

$$\widehat{Y}_{\mathbf{i_0}}^{\sharp} = \frac{\sum_{\substack{\mathbf{i} \in \mathcal{I}_\mathbf{n} \\ \mathbf{i} \neq \mathbf{i_0}}} Y_\mathbf{i} K\left(\frac{d(x,X_\mathbf{i})}{h_\mathbf{n}}\right)}{\sum_{\substack{\mathbf{i} \in \mathcal{I}_\mathbf{n} \\ \mathbf{i} \neq \mathbf{i_0}}} K\left(\frac{d(x,X_\mathbf{i})}{h_\mathbf{n}}\right)},$$

and

$$\tilde{Y}_{\mathbf{i_0}}^{\star} = \frac{\sum_{\substack{\mathbf{i} \in \mathcal{I}_\mathbf{n} \\ \mathbf{i} \neq \mathbf{i_0}}} Y_\mathbf{i} K_1\left(\frac{d(x,X_\mathbf{i})}{b_\mathbf{n}}\right) K_{2,\rho_\mathbf{n}}(\|\mathbf{i_0} - \mathbf{i}\|)}{\sum_{\substack{\mathbf{i} \in \mathcal{I}_\mathbf{n} \\ \mathbf{i} \neq \mathbf{i_0}}} K_1\left(\frac{d(x,X_\mathbf{i})}{b_\mathbf{n}}\right) K_{2,\rho_\mathbf{n}}(\|\mathbf{i_0} - \mathbf{i}\|)}.$$

The following result gives an asymptotic property of $\widehat{Y}_{\mathbf{i_0}}^{\sharp}$ and is a consequence of Theorems 8.2. Its proof will be omitted.

Corollary 8.1 *Under conditions of Theorem 8.2,* $\widehat{Y}_{\mathbf{i_0}}^{\sharp}$ *converges almost surely to* $Y_{\mathbf{i_0}}$ *as* $\mathbf{n} \to \infty$.

Remark 8.4

- *Similar asymptotic results can be obtained for* $\widetilde{Y}_{\mathbf{i_0}}^{\star}$ *depending on the sample size,* $k_\mathbf{n}$ *and* $\rho_\mathbf{n}$.
- *Instead of* $\widetilde{Y}_{\mathbf{i_0}}^{\star}$, *one can consider a predictor using nonnormalized sites:*

$$\tilde{Y}_{\mathbf{i_0}}^{\star} = \frac{\sum_{\substack{\mathbf{i} \in \mathcal{I}_\mathbf{n} \\ \mathbf{i} \neq \mathbf{i_0}}} Y_\mathbf{i} K_1\left(\frac{d(x,X_\mathbf{i})}{b_\mathbf{n}}\right) K_2\left(\frac{\|\mathbf{i_0} - \mathbf{i}\|}{\rho_\mathbf{n}}\right)}{\sum_{\substack{\mathbf{i} \in \mathcal{I}_\mathbf{n} \\ \mathbf{i} \neq \mathbf{i_0}}} K_1\left(\frac{d(x,X_\mathbf{i})}{b_\mathbf{n}}\right) K_2\left(\frac{\|\mathbf{i_0} - \mathbf{i}\|}{\rho_\mathbf{n}}\right)}.$$

Such function allows the asymptotic result to remain valid with some minor changes in conditions on $k_{\mathbf{n}}$.

- *This current work is supported by a particular sampling scheme, which only includes deterministic designs for the spatial locations. One can generalize this work to random spatial sample such as in [11] (for real-valued regression) and in [40] (for spatial heteroskedasticity-and autocorrelation-consistent [HAC] estimation) and have a bound including* $\rho_{\mathbf{n}}^{\alpha}$.

Now that we have checked the theoretical behavior of our regression estimate, we are going to study its practical features through some numerical results. To this end, in Section 8.4, the regression estimate and the prediction procedure are illustrated by some simulations.

8.4 Numerical Results

In this section, we study the performance of the proposed regression estimator through some simulations which point out the importance of taking into account the spatial locations of the data. We remind that the theoretical results are obtained under a mixing condition whose role can be considered as that of the kernel function on the locations. We compare the two predictors, the basic one (see [18]) with the one that does take into account a spatial dependence in its structure. We consider a sample of dependent functional variables $X_{\mathbf{i}}$. That is, on each site $\mathbf{i}$, we have a curve $X_{\mathbf{i}}$ such that $X_{\mathbf{i}} = \{X_{\mathbf{i}}(t), t \in [0,1]\}$. Before studying the numerical results, we propose a useful procedure for estimating the spatial regression function.

8.4.1 Bandwidth Selection Procedure

(1) Specify sets of bandwidths $S(h)$, $S(b)$, and $S(\rho)$ for respectively K, K_1, and K_2.
(2) For each $h_{\mathbf{n}} \in S(h)$, $b_{\mathbf{n}} \in S(b)$, and $\rho_{\mathbf{n}} \in S(\rho)$ and each $\mathbf{j} \in \mathcal{I}_{\mathbf{n}}$, compute

$$\widehat{Y}_{\mathbf{j}}^{\sharp} = \frac{\sum_{\substack{\mathbf{i}\in I_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}} K\left(\frac{d(X_{\mathbf{i}},X_{\mathbf{j}})}{h_{\mathbf{n}}}\right)}{\sum_{\substack{\mathbf{i}\in I_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} K\left(\frac{d(X_{\mathbf{i}},X_{\mathbf{j}})}{h_{\mathbf{n}}}\right)},$$

and

$$\widetilde{Y}_{\mathbf{j}}^{\star} = \frac{\sum_{\substack{\mathbf{i}\in I_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}} K_1\left(\frac{d(X_{\mathbf{i}},X_{\mathbf{j}})}{b_{\mathbf{n}}}\right) K_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right)}{\sum_{\substack{\mathbf{i}\in I_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} K_1\left(\frac{d(X_{\mathbf{i}},X_{\mathbf{j}})}{b_{\mathbf{n}}}\right) K_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right)}.$$

(3) Compute $h_{\mathbf{n},opt}$, $b_{\mathbf{n},opt}$, and $\rho_{\mathbf{n},opt}$ by applying a cross-validation procedure over $S(h)$, $S(b)$, and $S(\rho)$. More precisely, consider the following minimization problem, i.e. determine $(b_{\mathbf{n},opt}, \rho_{\mathbf{n},opt})$ and $h_{\mathbf{n},opt}$ which minimize the mean squared errors over the $\widehat{\mathbf{n}}$ sites, respectively

$$\min_{b_{\mathbf{n}},\rho_{\mathbf{n}}} \quad \frac{1}{\widetilde{\mathbf{n}}} \sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}} (\tilde{Y}_{\mathbf{j}}^{\bigstar} - r(X_{\mathbf{j}}))^2, \quad \text{and} \min_{h_{\mathbf{n}}} \quad \frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}} (\widehat{Y}_{\mathbf{j}}^{\sharp} - r(X_{\mathbf{j}}))^2.$$

(4) For each $\mathbf{j}$, compute $\widehat{Y}_{\mathbf{j}}^{\sharp}$, $\tilde{Y}_{\mathbf{j}}^{\bigstar}$ corresponding to $h_{\mathbf{n},opt}$ and $(b_{\mathbf{n},opt}, \rho_{\mathbf{n},opt})$.

8.4.2 Simulation Study

This last procedure is used in the following simulation study dealing with $N = 2$. We consider observations $(X_{(i,j)}, Y_{(i,j)})$, $1 \le i,j \le n$ ($n = 30$ or 50), such that

$$Y_{(i,j)} = r(X_{(i,j)}) + \epsilon_{(i,j)} = 4 \times A_{(i,j)}^2 + \epsilon_{(i,j)},$$

and for $t \in [0,1]$, $X_{(i,j)}(t)$ is defined according to the following cases:

Case 1: $X_{(i,j)}(t) = A_{(i,j)}^2 \times (t - 0.5)^2 + A_{(i,j)} \times B_{(i,j)}$;
Case 2: $X_{(i,j)}(t) = A_{(i,j)} \times \cos(2\pi t)$,

where $A = (A_{(i,j)})$, $B = (B_{(i,j)})$, and $\epsilon = (\epsilon_{(i,j)})$ are random variables which will be specified according to the following considered model on $A = (A_{(i,j)})$. Several curve examples of $X_{(i,j)}(t)$, for each case, are drawn on Figure 8.1. More precisely, the figure on the left displays some curves simulated from Case 1, while that on the right concerns Case 2. In Case 1, an example of the function $r(\cdot)$ could be $r(X) = 2X''$ (where X'' denotes the second derivative of X with respect to t), whereas in Case 2, it could be $r(X) = \frac{A}{\pi^2} X''$ with $t = \frac{1}{2}$. We will denote by $GRF(m, \sigma^2, s)$ any stationary Gaussian Random Field with mean m and spatial exponential covariance function defined by

$$C(h) = \sigma^2 \exp\left(-\left(\frac{\|h\|}{s}\right)^2\right), \quad h \in \mathbb{R}^2 \text{ and } s > 0.$$

Then, we define the two considered models on $A = (A_{(i,j)})$ by

Model A: $A_{i,j} = D_{i,j} \times (\sin(2G_{i,j}) + 2\exp(-16G_{i,j}^2))$;
Model B: $A_{i,j} = D_{i,j} \times (2\cos(2G_{i,j}) + \exp(-4G_{i,j}^2))$.

Here, the number of observations $\widehat{\mathbf{n}}$ is equal to 30×30, i.e. 900 or 50×50, i.e. 2500. The several fields are defined by $D_{i,j} = \frac{1}{625}\sum_{1\le m,t\le 25} \exp\left(-\frac{\|(i,j)-(m,t)\|}{a}\right)$, $G_{i,j} = GRF(0,5,3)$, $B_{i,j} = GRF(2.5,5,3)$, and $\epsilon_{i,j} = GRF(0,0.1,5)$. We note that the local spatial dependence depends not only on the covariance function but also on a. In fact, the greater a is, the weaker the spatial dependency is. According to this fact, we provide simulation results obtained with different values of a which are $a = 5$, 20, and 50.

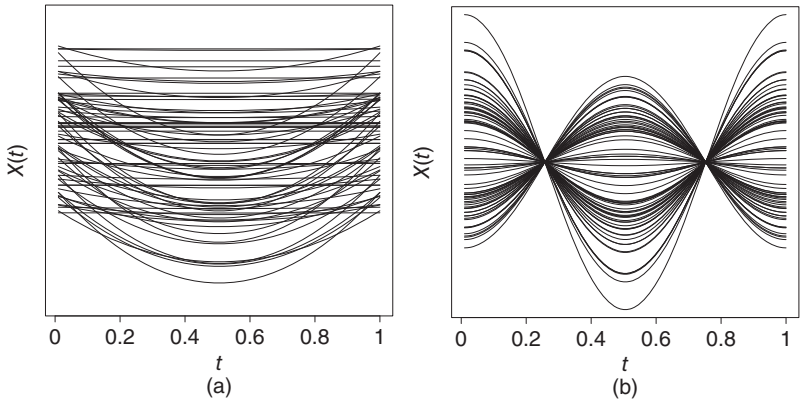


Figure 8.1 Some simulated curves of Case 1 (a) and Case 2 (b). In Case 1, X is simulated from $X_{(i,j)}(t) = A^2_{(i,j)} \times (t - 0.5)^2 + A_{(i,j)} \times B_{(i,j)}$ and in Case 2, X is simulated from $X_{(i,j)}(t) = A_{(i,j)} \times \cos(2\pi t)$, where A and B are random fields and $t \in [0, 1]$.

Along this part, the spatial regression is computed based on the kernels K, K_1 as the Epanechnikov kernel and K_2 as the Parzen kernel. The choice of the semimetric $d(\cdot, \cdot)$ is important and depends on the information one gets on the data. Ferraty and Vieu [14] present three families of semimetrics. The first is built from functional principal component analysis (FPCA) and is adapted to rough curves. The second is built from the partial least square (PLS) approach and is relevant when one considers multivariate response. The last, based on derivatives, is well adapted in the presence of smooth curves. Specifically, it approximates L_2 metric between derivatives of the curves based on their B-spline representation. Note that other semimetrics are encountered in the literature. However, according to [46], the theoretical justification of the usefulness of a particular semimetric is still an open problem. In this work, we consider a semimetric between curves based on their first $q = 2$ derivatives because of the smoothness of the curves. This semimetric (between X_i and X_j) is defined by

$$\sqrt{\int \left(X_i^{(q)}(t) - X_j^{(q)}(t)\right)^2 dt}, \qquad q = 0, 1, 2, \ldots$$

where, for any q-times differentiable real function X, $X^{(q)}$ denotes the q-th derivative of X (we refer, for example, to [14] for the theoretical setting about semimetrics used for functional nonparametric investigations). To confirm our semimetric choice, we tested, in addition to the semimetrics based on their first derivatives, two other semimetrics (based on PCA and on Fourier's decomposition) and different parameters such as the number of derivatives, principal components, basis. It turns out that the results are similar or worse than those obtained considering a

semimetric between curves based on their first $q = 2$ derivatives. We then present the results according to this last semimetric based on $q = 2$ derivatives.

To evaluate the performance of the proposed predictors, each studied model is replicated 350 times. At each replication k, we compute the mean squared error over the $\widehat{\mathbf{n}}$ sites. The bandwidths used, different at each replication, are those obtained using the previous procedure 8.4.1. For the kth replication, we define the mean squared error ($MSE^{(k)}$) by

$$MSE^{(k)} = \frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{j} \in \mathcal{I}_{\mathbf{n}}} (\hat{Y}^{\dagger}_{\mathbf{j},opt} - Y_{\mathbf{j}})^2, \text{with } \hat{Y}^{\dagger}_{\mathbf{j}} = \hat{Y}^{\sharp}_{\mathbf{j},opt} \text{or } \tilde{Y}^{\star}_{\mathbf{j},opt}. \tag{8.9}$$

The obtained results are summarized in Tables 8.1 and 8.2. For each situation (Model, Case, and value of a), the tables provide the average MSE over the 350 replications of Eq. (8.9) and the corresponding standard deviation. The AMSE* (average mean squared error) column makes reference to the proposed estimator $\hat{Y}^{\star}$, whereas the AMSE$^{\sharp}$'s column corresponds to the estimator $\widetilde{Y}^{\sharp}$ which does not takes into account the locations. Besides, we use a statistical hypothesis test rather than directly compare the average MSE accuracy. The column entitled "p-value" gives, for each considered situation, the p-value of Wilcoxon signed-rank test

Table 8.1 Simulation results for $\widehat{\mathbf{n}} = 900$ according to the models A and B, the cases 1 and 2 and the value of $a = 5$, 20, and 50: the table gives the average mean squared errors (AMSE) for each situation and in brackets, the corresponding standard deviation.

Model	Case	a	$AMSE^{\star}$	$AMSE^{\sharp}$	p-value	$AR^{2\star}$	$AR^{2\sharp}$
A	1	5	0.0035 (0.0011)	0.0091 (0.0024)	2.04×10^{-59}	0.6814	0.1846
		20	0.0067 (0.0014)	0.0104 (0.0026)	2.04×10^{-59}	0.9882	0.9818
		50	0.0105 (0.0037)	0.0126 (0.0045)	3.02×10^{-59}	0.9960	0.9953
	2	5	0.0009 (0.0003)	0.0091 (0.0025)	2.04×10^{-59}	0.9148	0.1820
		20	0.0063 (0.0011)	0.0098 (0.0025)	2.04×10^{-59}	0.9887	0.9825
		50	0.0091 (0.0019)	0.0106 (0.0027)	1.54×10^{-57}	0.9966	0.9960
B	1	5	0.0014 (0.0004)	0.0092 (0.0025)	2.04×10^{-59}	0.9480	0.6672
		20	0.0095 (0.0042)	0.0119 (0.0046)	3.66×10^{-59}	0.9982	0.9977
		50	0.0128 (0.0054)	0.0135 (0.0057)	1.86×10^{-33}	0.9995	0.9995
	2	5	0.0015 (0.0002)	0.0092 (0.0025)	2.04×10^{-59}	0.9446	0.6673
		20	0.0096 (0.0019)	0.0111 (0.0026)	2.79×10^{-57}	0.9982	0.9979
		50	0.0163 (0.0025)	0.0167 (0.0028)	1.22×10^{-15}	0.9994	0.9993

The column entitled "p-value" gives the p-value of a Wilcoxon signed-rank test performing in order to determine whether $AMSE^{\sharp}$ is significantly less than $AMSE^{\star}$. The two last columns display the average coefficients of determination (AR^2).

Table 8.2 Simulation results for $\hat{\mathbf{n}} = 2500$ according to the models *A* and *B* with cases 1 and 2, and the value of $a = 5$, 20, and 50: the table gives the average mean squared errors (AMSE) for each situation and in brackets, the corresponding standard deviation.

Model	Case	a	$AMSE^{\star}$	$AMSE^{\sharp}$	p-value	$AR^{2\star}$	$AR^{2\sharp}$
A	1	5	0.0039 (0.0009)	0.0096 (0.0017)	2.04×10^{-59}	0.5905	0.0113
		20	0.0047 (0.0008)	0.0097 (0.0017)	2.04×10^{-59}	0.9647	0.9272
		50	0.0077 (0.0013)	0.0102 (0.0020)	2.04×10^{-59}	0.9942	0.9924
	2	5	0.0011 (0.0002)	0.0096 (0.0017)	2.04×10^{-59}	0.8841	0.0119
		20	0.0040 (0.0005)	0.0096 (0.0017)	2.04×10^{-59}	0.9699	0.9273
		50	0.0074 (0.0009)	0.0099 (0.0017)	2.04×10^{-59}	0.9944	0.9926
B	1	5	0.0012 (0.0004)	0.0096 (0.0017)	2.04×10^{-59}	0.8764	0.0665
		20	0.0050 (0.0006)	0.0098 (0.0017)	2.04×10^{-59}	0.9958	0.9917
		50	0.0087 (0.0012)	0.0102 (0.0018)	4.40×10^{-59}	0.993	0.992
	2	5	0.0011 (0.0001)	0.0096 (0.0017)	2.04×10^{-59}	0.8884	0.0663
		20	0.0058 (0.0006)	0.0098 (0.0017)	2.04×10^{-59}	0.9951	0.9917
		50	0.0111 (0.0014)	0.0119 (0.0017)	2.77×10^{-59}	0.9991	0.9990

The column entitled "p-value" gives the p-value of a Wilcoxon signed-rank test performing in order to determine whether $AMSE^{\sharp}$ is significantly less than $AMSE^{\star}$. The two last columns display the average coefficients of determination (AR^2).

performing in order to determine if $\text{MSE}^{\star}$ is significantly less than $\text{MSE}^{\sharp}$ (the alternative hypothesis is then H_1: $\text{MSE}^{\star} < \text{MSE}^{\sharp}$). The two last columns give the average of the coefficients of determination, R^2, over the 350 replications. Recall that a value of R^2 close to 1 means that the quality of estimationis reliable. Here, we define R^2 as the square of the linear correlation coefficient between the vector of the $Y'_{\mathbf{i}}s$ and its estimated version.

The first general point to make about this study is that, when $a = 5$, regardless of the considered kind of model or case, the predictor $\hat{Y}^{\star}_{\mathbf{j}}$ leads to better results since the mean squared errors are significantly lower than with $\hat{Y}^{\sharp}_{\mathbf{j}}$. Moreover, it can be seen that the standard deviations are greater with $\hat{Y}^{\sharp}_{\mathbf{j}}$ than with $\hat{Y}^{\star}_{\mathbf{j}}$. Second, we note that when the value of a increases, $AMSE^{\sharp}$ is still higher than $AMSE^{\star}$, but the difference becomes narrower. Consequently, the higher the value of a (less spatial dependency), the lower the difference between the results of the two estimators is. In other words, our estimator $\hat{Y}^{\star}_{\mathbf{j}}$ outperforms $\hat{Y}^{\sharp}_{\mathbf{j}}$ when the spatial dependence is important. However, the two estimators tend to give similar performance in case of spatially independent fields. The low p-values obtained with Wilcoxon signed-rank test (less than 1.22×10^{-15}) confirm that $\hat{Y}^{\star}_{\mathbf{j}}$ produces less errors than $\hat{Y}^{\sharp}_{\mathbf{j}}$. Nevertheless, the probability of erroneously rejecting the null

hypothesis is highest when the value of a is equal to 20 or 50, rather than 5 (without one exception) since the p-value increases with the value of a. Finally, we may note that the R^2 criterion strengthens the previous comments. In fact, the values $AR^{2\star}$ are higher than $AR^{2\sharp}$ and the difference between them decreases as the value of a increases.

Insight into the performance of the two predictors can also be viewed from graphical outputs. In fact, Figures 8.2–8.4 illustrate different situations. The first deals with spatially dependent data ($a = 5$) simulated from Model A and Case 1 of which a representation of $\{Y_{\mathbf{j}},\ \mathbf{j} \in \mathcal{I}_{\mathbf{n}}\}$ is depicted in Figure 8.2a. Figures 8.2b, 8.2c show squared errors obtained by the two predictors, respectively.

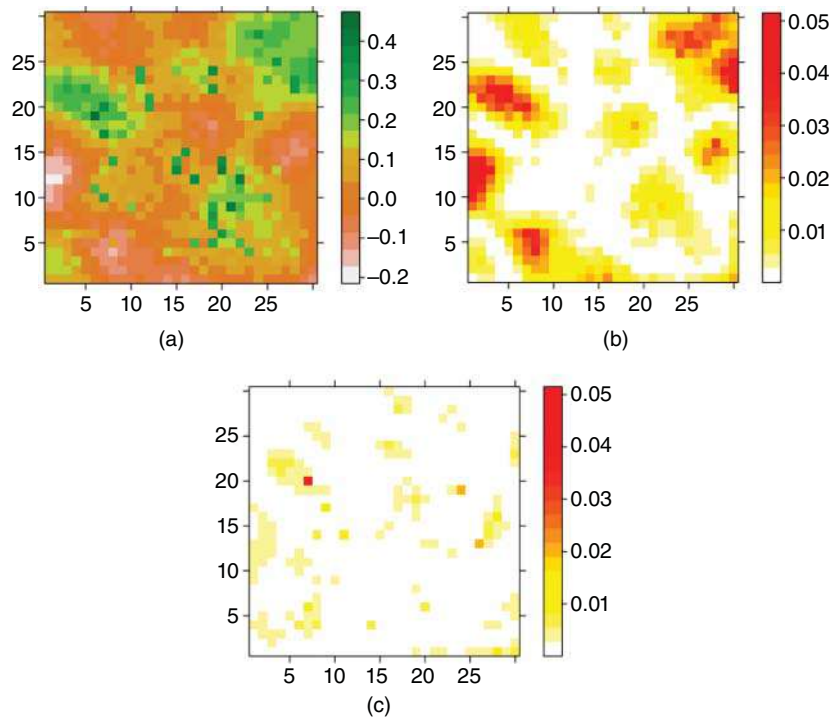


Figure 8.2 A simulated field considering Model A, Case 1 and $a = 5$ with (a) an image of the field Y; (b) the squared errors using $r^{\sharp}_{\mathbf{n}}$; (c) the squared errors using $r^{\star}_{\mathbf{n}}$. The sample size is 900. More precisely, the curves X are simulated from $X_{(i,j)}(t) = A^2_{(i,j)} \times (t - 0.5)^2 + A_{(i,j)} \times B_{(i,j)}$ (Case 1), and the field A is simulated from $A_{i,j} = D_{i,j} \times (\sin(2G_{i,j}) + 2\exp(-16G^2_{i,j}))$ (Model A). B and G are Gaussian random fields defined by $GRF(2.5, 5, 3)$ and $GRF(0, 5, 3)$, respectively. The parameter a acts on the local spatial dependence: the greater a is, the weaker the spatial dependency is.

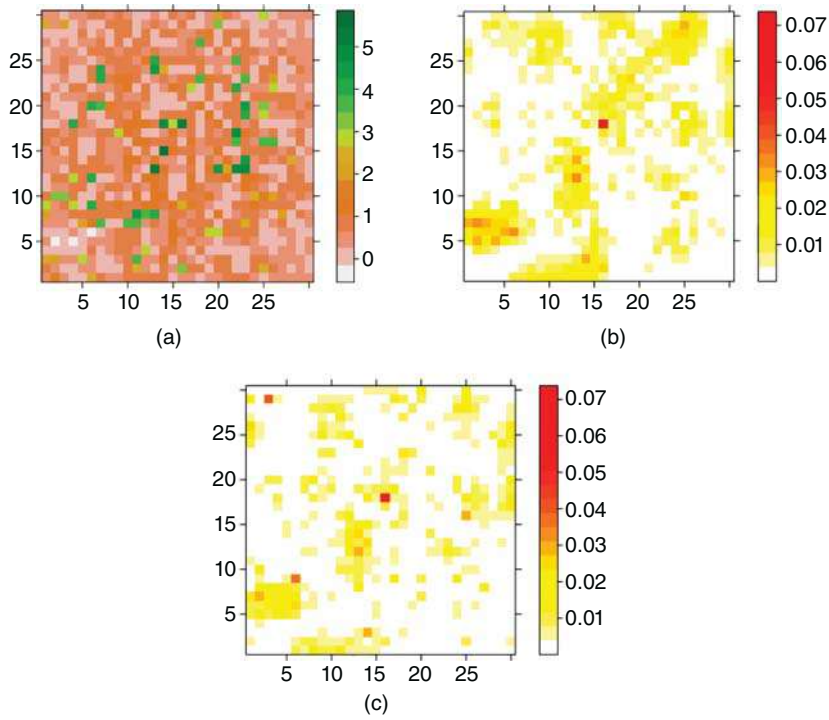


Figure 8.3 A simulated field considering Model *B*, Case 1, and $a = 20$ with (a) an image of the field Y; (b) the squared errors using $r_{\mathbf{n}}^{\sharp}$; (c) the squared errors using $r_{\mathbf{n}}^{\star}$. The sample size is 900. More precisely, the curves X are simulated from $X_{(i,j)}(t) = A_{(i,j)}^2 \times (t - 0.5)^2 + A_{(i,j)} \times B_{(i,j)}$ (Case 1), and the field A is simulated from $A_{i,j} = D_{i,j} \times (2\cos(2G_{i,j}) + \exp(-4G_{i,j}^2))$ (Model B). B and G are Gaussian random fields defined by $GRF(2.5, 5, 3)$ and $GRF(0, 5, 3)$, respectively. The parameter a acts on the local spatial dependence: the greater a is, the weaker the spatial dependency is.

These two figures confirm that $\hat{Y}_{\mathbf{j}}^{\star}$ generates less errors than using $\hat{Y}_{\mathbf{j}}^{\sharp}$ since the more colorful the representation is, the greater the error is. Figure 8.3 considers lower spatial dependence ($a = 20$) simulated from Model A and Case 1 for which the field Y is represented in Figure 8.3a. Figure 8.3b displays slightly less errors than in Figure 8.3c. Finally, Figure 8.4 gives summarized results of Model B and Case 2, with almost independent spatial data ($a = 50$). The two estimators seem to provide similar errors according to Figures 8.4b, 8.4c. It is not surprising to note that when a is high, the two estimators produce similar results. In fact, in this situation, the bandwidths $\rho_{\mathbf{n}}$ are large and could take the maximal distance between observations. In short, the two predictors work in an almost identical manner in the absence of spatial dependence.

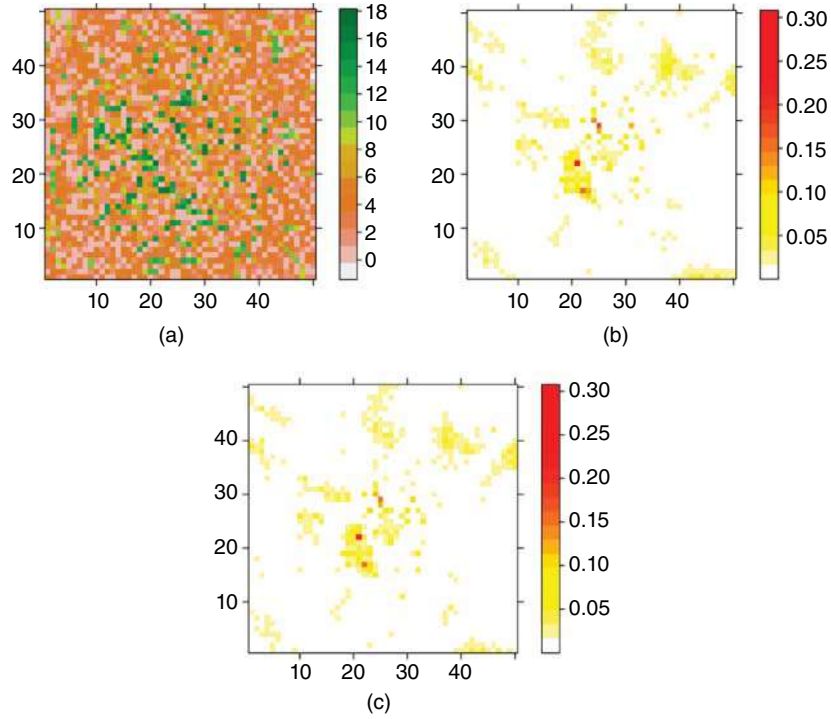


Figure 8.4 A simulated field considering Model *B*, Case 2, and $a = 50$ with (a) an image of the field Y; (b) the squared errors using $r_{\mathbf{n}}^{\sharp}$; (c) the squared errors using $r_{\mathbf{n}}^{\star}$. The sample size is 2500. More precisely, the curves X are simulated from $X_{(i,j)}(t) = A_{(i,j)} \times \cos(2\pi t)$ (Case 2), and the field A is simulated from $A_{i,j} = D_{i,j} \times (2\cos(2G_{i,j}) + \exp(-4G_{i,j}^2))$ (Model B). *B* and *G* are Gaussian random fields defined by *GRF*(2.5, 5, 3) and *GRF*(0, 5, 3) respectively. The parameter a acts on the local spatial dependence: the greater a is, the weaker the spatial dependency is.

Regarding the bandwidths selection, we carried out a cross-validation procedure. This selection is made differently, according to $\hat{Y}_{\mathbf{j}}^{\sharp}$ and $\hat{Y}_{\mathbf{j}}^{\star}$. First, with higher spatially dependent data ($a = 5$), the selected optimal bandwidths $\rho_{\mathbf{n},opt}^{\star}$ have the smallest values. This result was expected because when the spatial dependence is high, sites that are close together tend to be more related than sites that are far apart. For the bandwidth linked to the distance between the observations (according to K_1 and K), the selection differs with respect to the considered estimator. In fact, the optimal bandwidth are widely higher for $\hat{Y}_{\mathbf{j}}^{\star}$, rather than $\hat{Y}_{\mathbf{j}}^{\sharp}$. For more details on the values of the optimal bandwidths, through the replications, Figure 8.5 displays the corresponding boxplots. Second, when $a = 20$, considering Model A and Case 1, the bandwidths are slightly higher

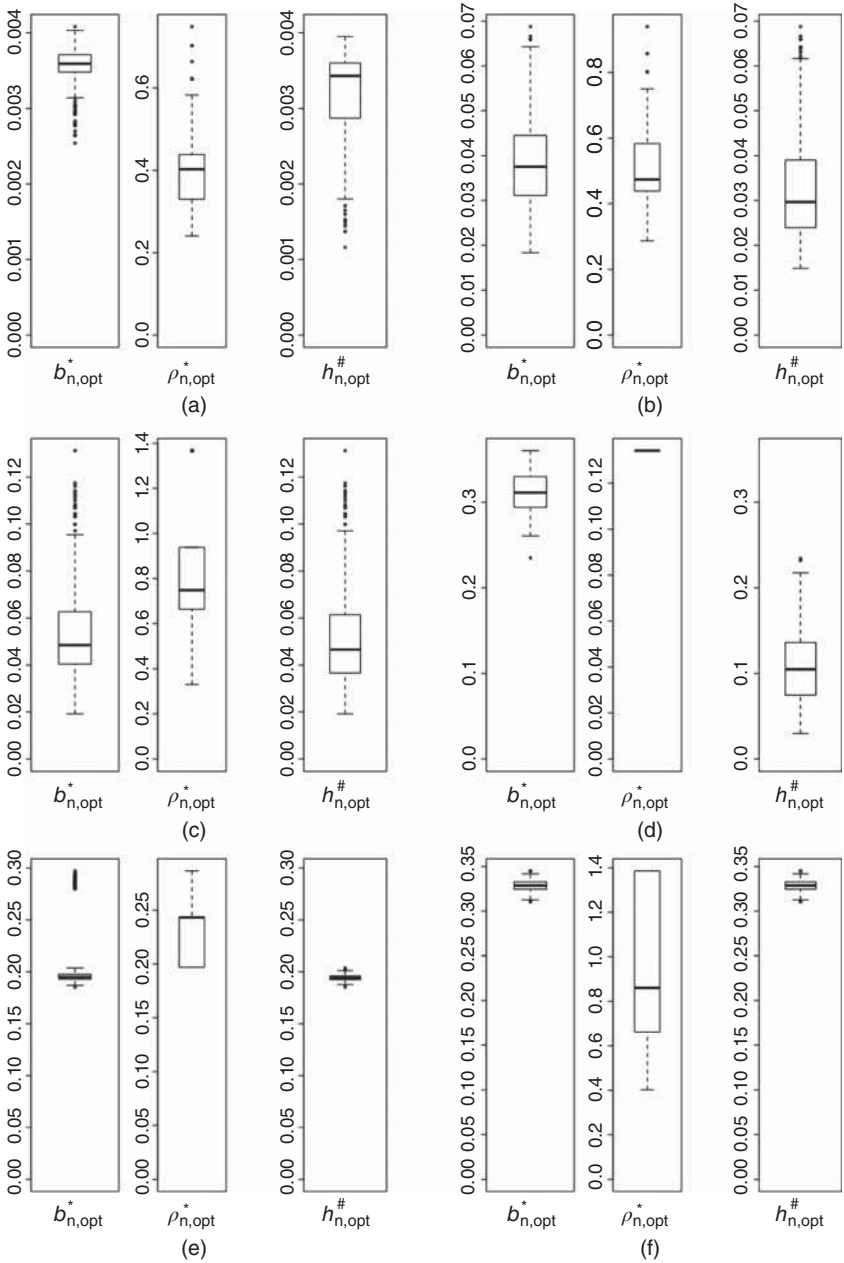


Figure 8.5 Boxplots of $b^{\star}_{\mathbf{n},opt}$, $\rho^{\star}_{\mathbf{n},opt}$ and $h^{\#}_{\mathbf{n},opt}$, respectively, over the 350 replications of the three following situations: (a) Model *A*, Case 1, $a = 5$ and $\widehat{\mathbf{n}} = 900$ (b) Model *A*, Case 1, $a = 20$ and $\widehat{\mathbf{n}} = 900$ (c) Model *A*, Case 1, $a = 50$ and $\widehat{\mathbf{n}} = 900$ (d) Model *B*, Case 1, $a = 5$ and $\widehat{\mathbf{n}} = 2500$ (e) Model *B*, Case 1, $a = 20$ and $\widehat{\mathbf{n}} = 2500$ (f) Model *B*, Case 2, $a = 50$ and $\widehat{\mathbf{n}} = 2500$. Cases 1 and 2 concern the simulation of *X*, models *A*, and *B* concern the simulation of the field *A*, the parameter *a* acts on the local spatial dependence and $\widehat{\mathbf{n}}$ is the number of sites.

than when $a = 5$ (see Figure 8.5b,8.5a). Finally, considering $a = 50$ with Model B and Case 2, the values of $\rho^{\star}_{\mathbf{n},opt}$ are more scattered and higher than with $a = 20$ (see Figure 8.5e). Moreover, for $a = 20$ and $a = 50$, the bandwidth selection is equivalent for $\hat{Y}^{\sharp}_{\mathbf{j}}$ and $\hat{Y}^{\star}_{\mathbf{j}}$ (see Figures 8.5b,8.5c,8.5e, and 8.5f). In these situations, the value of the bandwidth varies at each run while the locations do not change. In fact, contrary to the case $a = 5$, the values of $X_{i,j}(t)$ are more scattered and then imply a change in the value of $\rho^{\star}_{\mathbf{n},opt}$.

The previous study highlights the reliable performance of our estimator, particularly in the presence of spatial dependence. But a disadvantage may be that the cross-validation procedure on the two parameters $b_{\mathbf{n}}$ and $\rho_{\mathbf{n}}$ is very time-consuming. To this end, we tried to deal with simulations considering a fixed bandwidth $\rho_{\mathbf{n}}$ as in [40], where it is advised to take $d_{\mathbf{n}} = n\rho_{\mathbf{n}} = \lfloor \widehat{\mathbf{n}}^{1/4} \rfloor$ with $\lfloor \cdot \rfloor$ denotes the integer part. In our case, with $\widehat{\mathbf{n}} = 900$ sites, the corresponding bandwidths would be $\rho_{\mathbf{n}} \approx 0.18$. It allows to save time and obtain results that are quite satisfactory when the spatial dependence is high. More precisely, when $a = 5$, the results are similar or slightly worse than those obtained by the cross-validation procedure on the two parameters: it is explained by the fact that the cross-validation procedure chooses a value of $\rho_{\mathbf{n}}$ no always close to 0.18 (different at each replication). Nevertheless, the fixed bandwidth $\rho_{\mathbf{n}} = 0.18$ produces better results than using the estimator $\hat{Y}^{\sharp}_{\mathbf{j}}$. Note that the results depend largely on the spatial dependence structure considered. However, the results are worse with weaker spatial dependence ($a = 20$ or 50). In fact, in some cases (depending on the spatial dependency), the performance obtained by fixing $\rho_{\mathbf{n}}$ (according to the sample size $\widehat{\mathbf{n}}$ as above) is poorer than those obtained using the estimator $\hat{Y}^{\sharp}_{\mathbf{j}}$. In this case, the cross-validation procedure on the two parameters remains necessary.

8.5 Conclusion

This work proposes a new method to model spatial regression function for functional random fields providing an explicit general spatial proximity structure throughout a kernel estimator. This model requires no parametric correlation model on the error term, and the observations are supposed locally, identically distributed. Our main theoretical contribution was to derive the convergences in mean square and almost complete. One can see the proposed methodology as a good alternative to the classical kernel approach for functional spatial data. More precisely, it is apparent that the proposed approach is particularly well adapted to prediction with functional covariate, in the presence of spatial dependence. This good behavior is observed both from an asymptotic point of view and from a simulation study. However, in case of low spatial dependence, the two proposed predictors produce similar results.

In addition, this work offers very interesting perspectives of investigation. First of all, a future work will be tied up to the asymptotic normality of the regression estimator. Then, we could improve the choice of $h_{\mathbf{n}}$ and $\rho_{\mathbf{n}}$ which is outside the scope of this work. For further study, we could investigate this new approach using local linear spatial regression (see, for example, [34]). Also, an adaptation of this method to issues such as the spatial conditional mode or quantile regression estimation could be developed. Application of the proposed regression estimator to real data, and more particularly to data collected by the French Research Institute for Exploitation of the Sea (Ifremer) during IBTS campaign (International Bottom Trawl Survey), will be investigated. Moreover, another perspective is the study of regression estimation for continuous indexed spatial functional fields $\{Z_{\mathbf{i}}, \mathbf{i} \in \mathbb{R}^N\}$ that can be applied to spatial prediction.

8 Appendix

8.A.1 Some Preliminary Results for the Proofs

Lemma 8.A.1 *[8] Let the sets $S_1, S_2, \dots, S_k$ containing each m sites and such that, for all $i \neq j$, and for $1 \leq i,j \leq k$, $dist(S_i, S_j) \geq \delta_0$. Let $W_1, W_2, \dots, W_k$ a sequence of random variables with real values and measurable, respectively, with respect to $\mathcal{B}(S_1), \dots, \mathcal{B}(S_k)$. Let be W_l with values in $[a,b]$. There exists a sequence of independent random variables $W_1^*, W_2^*, \dots, W_k^*$ such that W_l^* has the same distribution as W_l and satisfies:*

$$\sum_{l=1}^{k} \mathbb{E}|W_l - W_l^*| \leq 2k(b-a)\psi((k-1)m, m)\chi(\delta_0).$$

Lemma 8.A.2 *[7] Denote by $\mathcal{L}_r(\mathcal{F})$ the class of $\mathcal{F}$-measurable random variables X which satisfy: $\|X\|_r = (\mathbb{E}|X|^r)^{1/r} < \infty$. Suppose that $X \in \mathcal{L}_r(\mathcal{B}(E))$, $Y \in \mathcal{L}_r(\mathcal{B}(E'))$, $1 \leq r,s,t < \infty$, and $\frac{1}{r} + \frac{1}{s} + \frac{1}{t} = 1$. Then,*

$$|\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y| \leq C\|X\|_r\|Y\|_s\{\psi(Card(E), Card(E'))\chi(dist(E,E'))\}^{1/t}.$$

For bounded random variables with probability 1, we have:

$$|\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y| \leq C\{\psi(Card(E), Card(E'))\chi(dist(E,E'))\}.$$

In the following, we will often use the notation $K_{\mathbf{i}}(x) = K\left(\frac{d(x,X_{\mathbf{i}})}{h_{\mathbf{n}}}\right)$ and $W_{\mathbf{ni}}(x) = \frac{K_{\mathbf{i}}(x)}{\sum_{\mathbf{j}\in I_{\mathbf{n}}} K_{\mathbf{j}}(x)}$ with the convention $0/0 = 0$, then $\sum_{\mathbf{i}\in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) = 0$ or 1. Thus, we have

$$r_{\mathbf{n}}(x) = \begin{cases} \sum_{\mathbf{i}\in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) Y_{\mathbf{i}} & \text{if} \quad \sum_{\mathbf{i}\in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) = 1; \\ \frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} Y_{\mathbf{i}} & \text{otherwise.} \end{cases}$$

Let us use the following decomposition:

$$r_{\mathbf{n}}(x) - r(x) = \frac{1}{f_{\mathbf{n}}(x)} \left[(g_{\mathbf{n}}(x) - \mathbb{E}(g_{\mathbf{n}}(x))) - (r(x) - \mathbb{E}(g_{\mathbf{n}}(x))) \right] - \frac{r(x)}{f_{\mathbf{n}}(x)} \left[f_{\mathbf{n}}(x) - 1 \right]. \tag{8.A.1}$$

Lemma 8.A.3 *Under hypotheses **H1–H2**, we have*

$$\mathbb{E}^{1/2} \left[\sum_{\mathbf{i}\in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}}) - r(x) \right]^2 = O(h_{\mathbf{n}}).$$

Lemma 8.A.4 *Under the conditions of Theorem 8.1, we have*

$$\mathbb{E}^{1/2} \left[\sum_{\mathbf{i}\in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) (Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})) \right]^2 = O\left(\frac{1}{\widehat{\mathbf{n}} \varphi_x(h_{\mathbf{n}})} \right)^{1/2}.$$

Lemma 8.A.5 *Under the conditions of Theorem 8.1, we have*

$$\mathbb{E}^{1/2} \left[\frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} Y_{\mathbf{i}} - r(x) \right]^2 = O\left(\frac{1}{\widehat{\mathbf{n}} \varphi_x(h_{\mathbf{n}})} \right)^{1/2}.$$

Define

$$\Lambda_{\mathbf{i}}(x) = \frac{1}{a_{\mathbf{n}}} \left[K_{\mathbf{i}}(x) - \mathbb{E}(K_{\mathbf{i}}(x)) \right],$$

$$I_{\mathbf{n}}(x) = \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[\left(\Lambda_{\mathbf{i}}(x)\right)^2 \right] \text{ and } R_{\mathbf{n}}(x) = \sum_{\mathbf{i}\neq\mathbf{k}} \left| \mathbb{E}\left[\Lambda_{\mathbf{i}}(x) \Lambda_{\mathbf{k}}(x) \right] \right|.$$

Lemma 8.A.6 *Under the conditions of Theorem 8.1, we have*

$$I_{\mathbf{n}}(x) + R_{\mathbf{n}}(x) = O\left(\frac{1}{\widehat{\mathbf{n}} \varphi_x(h_{\mathbf{n}})} \right).$$

8.A.2 Proofs

8.A.2.1 Proof of Theorem 8.1

Note that

$$\begin{aligned} r_{\mathbf{n}}(x) - r(x) = & \left(\sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) \mathbb{E}(Y_{\mathbf{i}} | X_{\mathbf{i}}) - r(x) \right) \mathbf{1}_{\left\{ \sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) = 1 \right\}} \\ & + \left(\sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) (Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}} | X_{\mathbf{i}})) \right) \mathbf{1}_{\left\{ \sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) = 1 \right\}} \\ & + \left(\frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{i} \in I_{\mathbf{n}}} Y_{\mathbf{i}} - r(x) \right) \mathbf{1}_{\left\{ \sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) = 0 \right\}} := \mathbf{A} + \mathbf{B} + \mathbf{C}. \end{aligned}$$

Applying Minkowski's inequality, we get

$$\| r_{\mathbf{n}}(x) - r(x) \|_2 \leq \mathbb{E}^{1/2}[\mathbf{A}]^2 + \mathbb{E}^{1/2}[\mathbf{B}]^2 + \mathbb{E}^{1/2}[\mathbf{C}]^2. \tag{8.A.2}$$

Therefore, Theorem 8.1 follows from (8.2) and Lemmas 8.3, 8.4 and 8.5. □

8.A.2.2 Proof of Lemma A.3

By the Lipschitz condition on Assumption **H2**, there exists a constant $C > 0$ such that

$$\begin{aligned} \mathbb{E}^{1/2}[\mathbf{A}]^2 & \leq \mathbb{E}^{1/2} \left[\left(\sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) |r(X_{\mathbf{i}}) - r(x)| \right) \mathbf{1}_{\left\{ \sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) = 1 \right\}} \right]^2 \\ & \leq C \mathbb{E}^{1/2} \left[\left(\sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) \| X_{\mathbf{i}} - x \| \right) \mathbf{1}_{\left\{ \sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) = 1 \right\}} \right]^2 \\ & \leq C \mathbb{E}^{1/2} \left[\times \sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) h_{\mathbf{n}} \right]^2 = O(h_{\mathbf{n}}). \end{aligned}$$

□

8.A.2.3 Proof of Lemma A.4

Define

$$\begin{aligned} G(x) & = \left(\sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) [Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}} | X_{\mathbf{i}})] \right) \mathbf{1}_{\left\{ \sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) = 1 \right\}} \\ & := \frac{e_{\mathbf{n}}(x)}{f_{\mathbf{n}}(x)} \mathbf{1}_{\left\{ \sum_{\mathbf{i} \in I_{\mathbf{n}}} W_{\mathbf{ni}}(x) = 1 \right\}}, \end{aligned}$$

where

$$e_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i} \in I_{\mathbf{n}}} K_{\mathbf{i}}(x) [Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}} | X_{\mathbf{i}})] \quad \text{and} \quad f_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i} \in I_{\mathbf{n}}} K_{\mathbf{i}}(x).$$

Note that, since Y is bounded, we have $\forall \mathbf{i}$, $0 \leq |Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})| \leq 2M$. It follows that $|G(x)| \leq 2M$ and

$$\begin{aligned}|G(x)| &= |G(x)|\mathbf{1}_{\left\{\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}(x)>c\right\}} + |G(x)|\mathbf{1}_{\left\{\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}(x)\leq c\right\}} \\ &\leq \frac{|e_{\mathbf{n}}(x)|}{f_{\mathbf{n}}(x)}\mathbf{1}_{\left\{\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}(x)>c\right\}} + 2M \times \mathbf{1}_{\left\{\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}(x)\leq c\right\}},\end{aligned}$$

where c is a given constant. Let us take $c = \frac{a_{\mathbf{n}}}{2}$ then by assumptions **H1** and **H3**, $\mathbb{E}[K_{\mathbf{i}}(x)] \leq C \times \varphi_x(h_{\mathbf{n}})$ since by **H1**, we have $C_1\varphi_{\mathbf{i},x}(h_{\mathbf{n}}) \leq \mathbb{E}\left[K_{\mathbf{i}}(x)\right] \leq C_2\varphi_{\mathbf{i},x}(h_{\mathbf{n}})$. If $\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}(x) > c = \frac{a_{\mathbf{n}}}{2}$ then $f_{\mathbf{n}}(x) > \dfrac{a_{\mathbf{n}}}{2a_{\mathbf{n}}} > \dfrac{1}{2}$. It follows that

$$\|G(x)\|_2 \leq 2\|e_{\mathbf{n}}(x)\|_2 + 2M\left(\mathbb{P}\left[\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}(x) \leq \frac{a_{\mathbf{n}}}{2}\right]\right)^{1/2},$$

and

$$\|e_{\mathbf{n}}(x)\|_2 = \frac{1}{a_{\mathbf{n}}}\left[\mathbb{E}\left(\sum_{\mathbf{i}\in I_{\mathbf{n}}} \xi_{\mathbf{i}}\right)^2\right]^{1/2},$$

where

$$\xi_{\mathbf{i}} = K_{\mathbf{i}}(x)\left[Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})\right].$$

To prove Lemma 8.4, we have to show that

$$\|e_{\mathbf{n}}(x)\|_2 = O(\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}}))^{-1/2}, \tag{8.A.3}$$

and

$$\mathbb{P}\left[\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}(x) \leq \frac{a_{\mathbf{n}}}{2}\right] \leq O\left(\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})\right)^{-1/2}. \tag{8.A.4}$$

Observe that, by Assumptions **H1** and **H3**, we have

$$\sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[\xi_{\mathbf{i}}^2\right] \leq \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[K_{\mathbf{i}}^2(x)\left[Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})\right]^2\right] = O(\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})).$$

Now, let $d_{\mathbf{n}}$ be a sequence of real numbers tending to ∞ as $\mathbf{n} \to \infty$ and set

$$S = \{(\mathbf{i},\mathbf{k}) \in \mathcal{I}_{\mathbf{n}}^2, \ \ \|\mathbf{i}-\mathbf{k}\| \leq d_{\mathbf{n}}\} \text{ and } \ S^c = \{(\mathbf{i},\mathbf{k}) \in \mathcal{I}_{\mathbf{n}}^2, \ \ \|\mathbf{i}-\mathbf{k}\| > d_{\mathbf{n}}\}.$$

Using Assumption **H3**, we have

$$\begin{aligned}\sum_{\mathbf{i},\mathbf{k}\in S} \mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] &\leq 4M^2 \sum_{\mathbf{i},\mathbf{k}\in S} \mathbb{E}\left[K_{\mathbf{i}}(x)K_{\mathbf{k}}(x)\right] \\ &\leq 4M^2 \sum_{\mathbf{i},\mathbf{k}\in S} \mathbb{P}\left[(X_{\mathbf{i}}, X_{\mathbf{k}}) \in B(x,h_{\mathbf{n}}) \times B(x,h_{\mathbf{n}})\right] \\ &\leq 4M^2C_4 \sum_{\mathbf{i},\mathbf{k}\in S} (\varphi_x(h_{\mathbf{n}}))^{1+\varepsilon} \leq 4M^2C_4\widehat{\mathbf{n}}d_{\mathbf{n}}^N(\varphi_x(h_{\mathbf{n}}))^{1+\varepsilon}.\end{aligned}$$

Since K is bounded, applying Lemma 8.2, we get

$$\sum_{\mathbf{i},\mathbf{k}\in S^c} \mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] \leq C \sum_{\mathbf{i},\mathbf{k}\in S^c} \{\psi(1,1)\chi(\|\mathbf{i}-\mathbf{k}\|)\} \leq C \sum_{\mathbf{i},\mathbf{k}\in S^c} \chi(\|\mathbf{i}-\mathbf{k}\|)$$
$$\leq C\widehat{\mathbf{n}} \sum_{\|\mathbf{i}\|>d_{\mathbf{n}}} \chi(\|\mathbf{i}\|).$$

Note that

$$\sum_{\|\mathbf{i}\|>d_{\mathbf{n}}} \|\mathbf{i}\|^{-\theta} = \sum_{\|\mathbf{i}\|>d_{\mathbf{n}}} \|\mathbf{i}\|^{-\theta}\|\mathbf{i}\|^{-N-\varepsilon}\|\mathbf{i}\|^{N+\varepsilon} \leq Cd_{\mathbf{n}}^{-N-\varepsilon} \sum_{\|\mathbf{i}\|>d_{\mathbf{n}}} \|\mathbf{i}\|^{N+\varepsilon-\theta}.$$

Then,

$$\sum_{\mathbf{i},\mathbf{k}\in S^c} \mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] \leq C\widehat{\mathbf{n}}d_{\mathbf{n}}^{-N-\varepsilon} \sum_{\|\mathbf{i}\|>d_{\mathbf{n}}} \|\mathbf{i}\|^{N+\varepsilon-\theta}.$$

Choosing $d_{\mathbf{n}} = (\varphi_x(h_{\mathbf{n}}))^{\frac{-\varepsilon}{N}+a}$ with $a>0$ such that $Na \leq \varepsilon - \frac{N}{N+\varepsilon}$ lead to

$$d_{\mathbf{n}}^{-(N+\varepsilon)} = \varphi_x(h_{\mathbf{n}})(\varphi_x(h_{\mathbf{n}}))^{\frac{-(N+\varepsilon)(Na-\varepsilon)-N}{N}} = O\left(\varphi_x(h_{\mathbf{n}})\right),$$

which implies that

$$\sum_{\mathbf{i},\mathbf{k}\in S} \mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] \leq 4M^2C_4\widehat{\mathbf{n}}d_{\mathbf{n}}^N(\varphi_x(h_{\mathbf{n}}))^{1+\varepsilon}$$
$$\leq 4M^2C_4\widehat{\mathbf{n}}(\varphi_x(h_{\mathbf{n}}))^{1+Na} = O(\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})).$$

Then, we deduce that

$$\mathbb{E}\left(\sum_{\mathbf{i}\in I_{\mathbf{n}}}\xi_{\mathbf{i}}\right)^2 = \sum_{\mathbf{i}\in I_{\mathbf{n}}}\mathbb{E}\left[\xi_{\mathbf{i}}^2\right] + \sum_{\mathbf{i},\mathbf{k}\in S}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] + \sum_{\mathbf{i},\mathbf{k}\in S^c}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] = O\left(\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})\right).$$

Consequently, $\left[\mathbb{E}\left(\sum_{\mathbf{i}\in I_{\mathbf{n}}}\xi_{\mathbf{i}}\right)^2\right]^{1/2} = O(\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}}))^{1/2}$ and $\|e_{\mathbf{n}}(x)\|_2 = O\big(\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})\big)^{-1/2}$ since by Assumption **H3**(ii), $a_{\mathbf{n}} \geq C_1'\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})$.

Next, for (8.4), define

$$S_{\mathbf{n}}(x) = \sum_{\mathbf{i}\in I_{\mathbf{n}}}\Lambda_{\mathbf{i}}(x) = f_{\mathbf{n}}(x) - \mathbb{E}(f_{\mathbf{n}}(x)).$$

Then, we have

$$\mathbb{P}\left[\sum_{\mathbf{i}\in I_{\mathbf{n}}}K_{\mathbf{i}}(x) \leq \frac{a_{\mathbf{n}}}{2}\right] = \mathbb{P}\left[\sum_{\mathbf{i}\in I_{\mathbf{n}}}\left(K_{\mathbf{i}}(x) - \mathbb{E}(K_{\mathbf{i}}(x))\right) \leq \frac{-a_{\mathbf{n}}}{2}\right]$$
$$\leq \mathbb{P}\left[\frac{1}{a_{\mathbf{n}}}\left|\sum_{\mathbf{i}\in I_{\mathbf{n}}}\left(K_{\mathbf{i}}(x) - \mathbb{E}(K_{\mathbf{i}}(x))\right)\right| \geq \frac{1}{2}\right]$$
$$\leq \mathbb{P}\left[\left|S_{\mathbf{n}}(x)\right| \geq \epsilon\right].$$

We will now introduce the spatial blocks decomposition introduced by Tran [7] which will be useful afterwards. Without loss of generality, we suppose that

$n_k = 2bq_k$, for $1 \le k \le N$. The random variables $\Lambda_{\mathbf{i}}(x)$ can be grouped into $2^N q_1 \dots q_N$ cubic blocks of side b. Let

$$U(1,\mathbf{n},x,\mathbf{j}) = \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,N.}}^{(2j_k+1)b} \Lambda_{\mathbf{i}}(x),$$

$$U(2,\mathbf{n},x,\mathbf{j}) = \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,N-1.}}^{(2j_k+1)b} \sum_{i_N=(2j_N+1)b+1}^{2(j_N+1)b} \Lambda_{\mathbf{i}}(x),$$

$$U(3,\mathbf{n},x,\mathbf{j}) = \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,N-2.}}^{(2j_k+1)b} \sum_{i_{N-1}=(2j_{N-1}+1)b+1}^{2(j_{N-1}+1)b} \sum_{i_N=2j_Nb+1}^{(2j_N+1)b} \Lambda_{\mathbf{i}}(x),$$

$$U(4,\mathbf{n},x,\mathbf{j}) = \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,N-2.}}^{(2j_k+1)b} \sum_{i_{N-1}=(2j_{N-1}+1)b+1}^{2(j_{N-1}+1)b} \sum_{i_N=(2j_N+1)b+1}^{(2j_N+1)b} \Lambda_{\mathbf{i}}(x),$$

and so on. Noting that

$$U(2^{N-1},\mathbf{n},x,\mathbf{j}) = \sum_{\substack{i_k=(2j_k+1)b+1,\\ k=1,\dots,N-1.}}^{2(j_k+1)b} \sum_{i_N=2j_Nb+1}^{(2j_N+1)b} \Lambda_{\mathbf{i}}(x)$$

$$U(2^N,\mathbf{n},x,\mathbf{j}) = \sum_{\substack{i_k=(2j_k+1)b+1,\\ k=1,\dots,N.}}^{2(j_k+1)b} \Lambda_{\mathbf{i}}(x),$$

for each integer $1 \le l \le 2^N$, we define $T(\mathbf{n},x,l) = \sum_{\substack{j_k=0\\ k=1,\dots,N.}}^{q_k-1} U(l,\mathbf{n},x,\mathbf{j})$. We obtain $S_{\mathbf{n}}(x) = \sum_{l=1}^{2^N} T(\mathbf{n},x,l)$. For $\epsilon > 0$, $P \le \mathbb{P}\left(\left|\sum_{l=1}^{2^N} T(\mathbf{n},x,l)\right| > \epsilon\right) \le 2^N \mathbb{P}\left(|T(\mathbf{n},x,1)| > \frac{\epsilon}{2^N}\right)$. We enumerate in arbitrary manner the $\widehat{q} = q_1 \times \dots \times q_N$ terms $U(1,\mathbf{n},x,\mathbf{j})$ of the sum $T(\mathbf{n},x,1)$, and refer to them as $W_1, \dots, W_{\widehat{q}}$. Note that $U(1,\mathbf{n},x,\mathbf{j})$ is a measurable σ-algebra generated by $X_{\mathbf{i}}$, with $\mathbf{i}$ such that $2j_kb + 1 \le i_k \le (2j_k+1)b$, $k = 1, \dots, N$. For all $l = 1, \dots, \widehat{q}$, the sets of the sites in W_l are separated by a distance of at least equal to b. In addition, since K_2 and K_1 are bounded, we can write $|W_l| \le C\frac{b^N}{a_{\mathbf{n}}}$ with $C = \|K\|_\infty$ (where $\|\cdot\|_\infty$ is the sup norm). Lemma 8.1 insures the existence of some random variables $W_1^*, W_2^*, \dots, W_{\widehat{q}}^*$ such that

$$\begin{aligned}\sum_{l=1}^{\widehat{q}} \mathbb{E}|W_l - W_l^*| &\le 2\widehat{q}C\frac{b^N}{a_{\mathbf{n}}}\psi((\widehat{q}-1)b^N, b^N)\chi(b)\\ &\le 2C\frac{\widehat{\mathbf{n}}}{2^N b^N}\frac{b^N}{a_{\mathbf{n}}}\psi(\widehat{\mathbf{n}}, b^N)\chi(b).\end{aligned}$$

Markov inequality allows us to write

$$\mathbb{P}\left(\sum_{l=1}^{\hat{q}}|W_l - W_l^*| > \frac{\epsilon}{2^{N+1}}\right) \le 2C\frac{\hat{\mathbf{n}}}{2^N b^N}\frac{b^N}{a_{\mathbf{n}}}\psi(\hat{\mathbf{n}}, b^N)\chi(b)2^{N+1}\epsilon^{-1},$$

and by Bernstein inequality, we have

$$\mathbb{P}\left(\sum_{l=1}^{\hat{q}}|W_l^*| > \frac{\epsilon}{2^{N+1}}\right) \le 2\exp\left\{\frac{-\epsilon^2/(2^{N+1})^2}{4\sum_{l=1}^{\hat{q}}\mathbb{E}(W_l^{*2}) + \frac{2\epsilon}{2^{N+1}}\frac{b^N}{a_{\mathbf{n}}}C}\right\},$$

which leads to

$$\begin{aligned}\mathbb{P}\left[\left|S_{\mathbf{n}}(x)\right| \ge \epsilon\right] &\le 2^{N+1}\exp\left\{\frac{-\epsilon^2/(2^{N+1})^2}{4\sum_{l=1}^{\hat{q}}\mathbb{E}(W_l^{*2}) + 2^{-N}C\epsilon\frac{b^N}{a_{\mathbf{n}}}}\right\}\\ &\quad + 2^{N+1}C\frac{\hat{\mathbf{n}}}{2^N b^N}\frac{b^N}{a_{\mathbf{n}}}\psi(\hat{\mathbf{n}}, b^N)\chi(b)2^{N+1}\epsilon^{-1}.\end{aligned}$$

Let $\delta > 0$, $\epsilon = \epsilon_{\mathbf{n}} = \delta\left(\frac{\log\hat{\mathbf{n}}}{\hat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})}\right)^{1/2}$ and $b = \left(\frac{\hat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})}{\log\hat{\mathbf{n}}}\right)^{\frac{1}{2N}}$. Since the variables W_l and W_l^* have the same distributions, we have $\sum_{l=1}^{\hat{q}}\mathbb{E}W_l^{*2} = \sum_{l=1}^{\hat{q}}\mathrm{var}(W_l^*) = \sum_{l=1}^{\hat{q}}\mathrm{var}(W_l) \le I_{\mathbf{n}}(x) + R_{\mathbf{n}}(x)$, and according to Lemma 8.6, we have $\sum_{l=1}^{\hat{q}}\mathbb{E}W_l^{*2} \le O\left([\hat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})]^{-1}\right)$. Then,

$$\begin{aligned}\mathbb{P}\left[\left|S_{\mathbf{n}}(x)\right| \ge \epsilon\right] &\le 2^{N+1}\exp\left\{\frac{-\epsilon^2}{2^{2N+2}\left(4\frac{C}{\hat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})} + C2^{-N}\epsilon\frac{b^N}{a_{\mathbf{n}}}\right)}\right\}\\ &\quad + 2^{N+2}C\frac{\hat{\mathbf{n}}}{a_{\mathbf{n}}}\psi(\hat{\mathbf{n}}, b^N)b^{-\theta}\epsilon^{-1}.\end{aligned}$$

Since $C_1'\hat{\mathbf{n}}\varphi_x(h_{\mathbf{n}}) \le a_{\mathbf{n}} \le C_2'\hat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})$, we have

$$\begin{aligned}\mathbb{P}\left[\left|S_{\mathbf{n}}(x)\right| \ge \epsilon_{\mathbf{n}}\right] &\le 2^{N+1}\exp\left\{\frac{-\delta^2\frac{\log\hat{\mathbf{n}}}{\hat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})}}{\frac{2^{2N+4}C}{\hat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})} + \frac{C2^{N+2}\delta}{\hat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})}}\right\}\\ &\quad + 2^{N+2}C\frac{\hat{\mathbf{n}}}{a_{\mathbf{n}}}\psi(\hat{\mathbf{n}}, b^N)b^{-\theta}\delta^{-1}\left(\frac{\hat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})}{\log\hat{\mathbf{n}}}\right)^{1/2}\\ &\le C2^{N+1}\exp\left\{\log\hat{\mathbf{n}}^{-a}\right\}\\ &\quad + 2^{N+2}C\delta^{-1}\frac{1}{\varphi_x(h_{\mathbf{n}})}\psi(\hat{\mathbf{n}}, b^N)\left(\frac{\hat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})}{\log\hat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}}\\ &:= C\hat{\mathbf{n}}^{-a} + C2^{N+1}\delta^{-1}D_{\mathbf{n}},\end{aligned}$$

with $a = \dfrac{\delta^2}{2^{2N+4}C + C2^{N+2}\delta} > 0$. Note that $\widehat{\mathbf{n}}^{1-a}\varphi_x(h_{\mathbf{n}})$ tends to 0 for $a > 1$ and then $C\widehat{\mathbf{n}}^{-a} = o\left([\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})]^{-1}\right)$. Moreover, $a > 1$ if and only if $\delta > 2^{N+1}C(1+\sqrt{4C}) > 2^{N+1}C$ (with $\delta > 0$). Now, we treat the second term.

When (8.3) is satisfied, i.e. $\psi(n,m) \leq C\min(n,m)$, $\forall n, m \in \mathbb{N}$, we have

$$\begin{aligned}\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})D_{\mathbf{n}} &\leq C\widehat{\mathbf{n}}\left(\frac{\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})}{\log\widehat{\mathbf{n}}}\right)^{\frac{2N-\theta}{2N}} \\ &\leq C\left[\widehat{\mathbf{n}}\left(\varphi_x(h_{\mathbf{n}})\right)^{\frac{2N-\theta}{4N-\theta}}\left(\log\widehat{\mathbf{n}}\right)^{\frac{\theta-2N}{4N-\theta}}\right]^{\frac{4N-\theta}{2N}},\end{aligned}$$

which tends to 0 as $\mathbf{n} \to 0$ since $\theta > 4N$.

When (8.4) is satisfied, i.e. $\psi(n,m) \leq C(n+m+1)^{\kappa}$, $\forall n, m \in \mathbb{N}$, and note that $\psi(\widehat{\mathbf{n}}, b^N) \leq C(\widehat{\mathbf{n}} + b^N + 1)^{\kappa} \leq C\widehat{\mathbf{n}}^{\kappa}$, we have

$$\begin{aligned}\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})D_{\mathbf{n}} &\leq C\widehat{\mathbf{n}}^{1+\kappa}\left(\frac{\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}} \\ &\leq C\left[\widehat{\mathbf{n}}\left(\varphi_x(h_{\mathbf{n}})\right)^{\frac{N-\theta}{N(3+2\kappa)-\theta}}\left(\log\widehat{\mathbf{n}}\right)^{\frac{\theta-N}{N(3+2\kappa)-\theta}}\right]^{\frac{N(3+2\kappa)-\theta}{2N}},\end{aligned}$$

which tends to 0 as $\mathbf{n} \to$ since $\theta > N(3+2\kappa)$. Therefore, (8.4) follows, which concludes the proof of Lemma 8.4. □

8.A.2.4 Proof of Lemma A.5

Since $Y_{\mathbf{i}}$ and r are bounded, we have

$$\begin{aligned}\mathbb{E}^{1/2}[\mathbf{C}] &\leq \mathbb{E}^{1/2}\left[\left|\frac{1}{\widehat{\mathbf{n}}}\sum_{\mathbf{i}\in I_{\mathbf{n}}} Y_{\mathbf{i}} - r(x)\right|\mathbf{1}_{\left\{\sum_{\mathbf{i}\in I_{\mathbf{n}}} W_{\mathbf{ni}}(x)=0\right\}}\right] \\ &\leq 2M\mathbb{E}^{1/2}\left[\mathbf{1}_{\left\{\sum_{\mathbf{i}\in I_{\mathbf{n}}} W_{\mathbf{ni}}(x)=0\right\}}\right] = 2M\left(\mathbb{P}\left[\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}(x) = 0\right]\right)^{1/2} \\ &\leq 2M\left(\mathbb{P}\left[\sum_{\mathbf{i}\in I_{\mathbf{n}}} K_{\mathbf{i}}(x) \leq \frac{a_{\mathbf{n}}}{2}\right]\right)^{1/2} = O\left(\frac{1}{\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})}\right)^{1/2},\end{aligned}$$

by Lemma 8.4. □

8.A.2.5 Proof of Lemma A.6

First, we deal with $I_{\mathbf{n}}(x) = \sum_{\mathbf{i}\in I_{\mathbf{n}}}\mathbb{E}\left[\left(\frac{1}{a_{\mathbf{n}}}K_{\mathbf{i}}(x)\right)^2\right] - \sum_{\mathbf{i}\in I_{\mathbf{n}}}\left(\frac{1}{a_{\mathbf{n}}}\mathbb{E}(K_{\mathbf{i}}(x))\right)^2$.

$$\sum_{\mathbf{i}\in I_{\mathbf{n}}}\mathbb{E}\left[\left(\frac{1}{a_{\mathbf{n}}}K_{\mathbf{i}}(x)\right)^2\right] = \frac{1}{a_{\mathbf{n}}^2}\sum_{\mathbf{i}\in I_{\mathbf{n}}}\mathbb{E}\left[K_{\mathbf{i}}^2(x)\right] = O\left([\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})]^{-1}\right),$$

for $\mathbf{n}$ sufficiently large.

Then, we have $I_{\mathbf{n}}(x) = O\left([\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})]^{-1}\right)$. We now treat the term $R_{\mathbf{n}}(x)$. Since K is bounded, applying Lemma 8.1, we get

$$|\mathbb{E}\left[\Lambda_{\mathbf{i}}(x)\Lambda_{\mathbf{k}}(x)\right]| \leq \frac{C}{a_{\mathbf{n}}^2}\psi(1,1)\chi(\|\mathbf{i}-\mathbf{k}\|).$$

Let $E_{\mathbf{n}}$ be a sequence of real numbers tending to ∞ as $\widehat{\mathbf{n}} \to \infty$. Set $T = \{\mathbf{i}, \mathbf{k} \in \mathcal{I}_{\mathbf{n}}, \|\mathbf{i}-\mathbf{k}\| \leq E_{\mathbf{n}}\}$ and denote by T^c the complementary of T. Let $R_{\mathbf{n}}^{(1)} = \sum_{\mathbf{i},\mathbf{k}\in T}\left|\mathbb{E}\left[\Lambda_{\mathbf{i}}(x)\Lambda_{\mathbf{k}}(x)\right]\right|$ and $R_{\mathbf{n}}^{(2)} = \sum_{\mathbf{i},\mathbf{k}\in T^c}\left|\mathbb{E}\left[\Lambda_{\mathbf{i}}(x)\Lambda_{\mathbf{k}}(x)\right]\right|$. Hence, $R_{\mathbf{n}}(x) \leq R_{\mathbf{n}}^{(1)} + R_{\mathbf{n}}^{(2)}$. Moreover, using the same arguments as in the proof of Lemma 8.4, we have $I_{\mathbf{n}}(x) + R_{\mathbf{n}}(x) = O\left([\widehat{\mathbf{n}}\varphi_x(h_{\mathbf{n}})]^{-1}\right)$. □

8.A.2.6 Proof of Theorem 8.2

Set $T_{\mathbf{n}} = \left(\widehat{\mathbf{n}}u_{\mathbf{n}}\right)^{1/s}$, where $u_{\mathbf{n}} = \prod_{i=1}^{N}(\log n_i)(\log n_i)^{1+\epsilon}$, and define

$$g_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}}K_{\mathbf{i}}(x), \quad f_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{\mathbf{i}}(x),$$
$$\widetilde{g}_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}}\mathbb{1}_{\{Y_{\mathbf{i}}\leq T_{\mathbf{n}}\}}K_{\mathbf{i}}(x).$$

Then, we can write

$$r_{\mathbf{n}}(x) - r(x) = -\frac{r(x)}{f_{\mathbf{n}}(x)}A_1(x) + \frac{1}{f_{\mathbf{n}}(x)}\left[A_2(x) + A_3(x) + A_4(x)\right], \tag{8.A.5}$$

where

$$A_1(x) = f_{\mathbf{n}}(x) - 1,$$
$$A_2(x) = \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right) - r(x),$$
$$A_3(x) = \widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right),$$
$$A_4(x) = g_{\mathbf{n}}(x) - \widetilde{g}_{\mathbf{n}}(x).$$

Therefore, Theorem 8.2 follows from (8.5) and Lemmas 8.7–8.9, 8.12. □

Lemma 8.7 *Under Assumptions **H1, H2** and **H6**,*

$$\sup_{x\in\mathcal{D}}\left|\mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right) - r(x)\right| = O\left(h_{\mathbf{n}} + \sqrt{\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\,\Gamma(h_{\mathbf{n}})}}\right).$$

Proof of Lemma A.7

Since

$$\begin{aligned}
&\mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right) - r(x) \\
&= \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[\left(Y_{\mathbf{i}} - Y_{\mathbf{i}} \mathbb{1}_{\{|Y_{\mathbf{i}}|>T_{\mathbf{n}}\}}\right) K_{\mathbf{i}}(x)\right] - r(x) \\
&= \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[\mathbb{E}\left(Y_{\mathbf{i}}|X_{\mathbf{i}}\right) K_{\mathbf{i}}(x)\right] - \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[Y_{\mathbf{i}} \mathbb{1}_{\{|Y_{\mathbf{i}}|>T_{\mathbf{n}}\}} K_{\mathbf{i}}(x)\right] - r(x) \\
&= \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[\left(r(X_{\mathbf{i}}) - r(x)\right) K_{\mathbf{i}}(x)\right] - \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[Y_{\mathbf{i}} \mathbb{1}_{\{|Y_{\mathbf{i}}|>T_{\mathbf{n}}\}} K_{\mathbf{i}}(x)\right],
\end{aligned}$$

we have

$$\begin{aligned}
\left|\mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right) - r(x)\right| &\leq \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[\left|r(X_{\mathbf{i}}) - r(x)\right| K_{\mathbf{i}}(x)\right] \\
&\quad + \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[\left|Y_{\mathbf{i}}\right| \mathbb{1}_{\{|Y_{\mathbf{i}}|>T_{\mathbf{n}}\}} K_{\mathbf{i}}(x)\right] := I + II.
\end{aligned}$$

Using assumptions **H1** and **H2**, we have

$$\left|r(X_{\mathbf{i}}) - r(x)\right| \leq \sup_{u\in B(x,h_{\mathbf{n}})} |r(x) - r(u)| = O(h_{\mathbf{n}}), \text{so that } I = O\left(h_{\mathbf{n}}\right).$$

For II, since $s > 2$, using Assumption **H6**, we can write

$$\begin{aligned}
II &\leq \frac{T_{\mathbf{n}}^{1-s}}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[\left|Y_{\mathbf{i}}\right|^s K_{\mathbf{i}}(x)\right] \leq \frac{T_{\mathbf{n}}^{1-s}}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \mathbb{E}\left[\mathbb{E}\left(\left|Y_{\mathbf{i}}\right|^s |X_{\mathbf{i}}\right) K_{\mathbf{i}}(x)\right] \\
&\leq C T_{\mathbf{n}}^{1-s} = o\left((\hat{\mathbf{n}} u_{\mathbf{n}})^{-1/2}\right) = o\left(\sqrt{\frac{\log \hat{\mathbf{n}}}{\hat{\mathbf{n}}\, \Gamma(h_{\mathbf{n}})}}\right),
\end{aligned}$$

which conclude the proof of Lemma 8.7. □

Lemma 8.8 *If Assumption **(H6)** (i) holds, then*

$$\sup_{x\in D} \left|g_{\mathbf{n}}(x) - \widetilde{g}_{\mathbf{n}}(x)\right| = 0$$

for sufficiently large $\mathbf{n}$.

Proof of Lemma A.8

Recall that $T_{\mathbf{n}} = \left(\hat{\mathbf{n}} u_{\mathbf{n}}\right)^{1/s}$ and note that

$$g_{\mathbf{n}}(x) - \widetilde{g}_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}} \sum_{\mathbf{i}\in I_{\mathbf{n}}} Y_{\mathbf{i}} \mathbb{1}_{\{|Y_{\mathbf{i}}|>T_{\mathbf{n}}\}} K_{\mathbf{i}}(x).$$

By the Markov inequality, $\mathbb{P}\left(|Y_{\mathbf{i}}| > T_{\mathbf{n}}\right) \leq T_{\mathbf{n}}^{-s} \mathbb{E}|Y_{\mathbf{i}}|^s$ for any $\mathbf{i} \in \mathbb{Z}^N$. Therefore,

$$\sum_{\mathbf{n}\in\mathbb{Z}^N} \mathbb{P}\left(|Y_{\mathbf{n}}| > T_{\mathbf{n}}\right) \leq C \sum_{\mathbf{n}\in\mathbb{Z}^N} \frac{1}{\hat{\mathbf{n}} u_{\mathbf{n}}} < \infty.$$

The Borel–Cantelli lemma ensures that almost surely $|Y_{\mathbf{i}}| \leq T_{\mathbf{n}}$ for sufficiently large $\mathbf{n}$. Since $T_{\mathbf{n}} \to \infty$ as $\mathbf{n} \to \infty$, we have almost surely $|Y_{\mathbf{i}}| < T_{\mathbf{n}}$ for all $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$ and for $\mathbf{n}$ sufficiently large enough, and thus the conclusion follows. □

Lemma 8.9 *Under the assumptions of Theorem 8.2,*

$$\sup_{x\in D}\left|\widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right)\right| = O\left(\left(\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\Gamma(h_{\mathbf{n}})}\right)^{1/2}\right) \ a.s.$$

Define

$$\widetilde{\Lambda}_{\mathbf{i}}(x) = Y_{\mathbf{i}}\mathbb{1}_{\{|Y_{\mathbf{i}}|\leq T_{\mathbf{n}}\}}K_{\mathbf{i}}(x) - \mathbb{E}(Y_{\mathbf{i}}\mathbb{1}_{\{|Y_{\mathbf{i}}|\leq T_{\mathbf{n}}\}}K_{\mathbf{i}}(x)),$$

$$\widetilde{I}_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}^2}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{E}\left(\widetilde{\Lambda}_{\mathbf{i}}(x)^2\right) \text{ and } \widetilde{R}_{\mathbf{n}}(x) = \frac{1}{a_{\mathbf{n}}^2}\sum_{\mathbf{i}\neq\mathbf{j}}\left|\mathbb{E}\left[\widetilde{\Lambda}_{\mathbf{i}}(x)\widetilde{\Lambda}_{\mathbf{j}}(x)\right]\right|. \tag{8.A.6}$$

Then, arguing as in the proof of Lemma 8.6 with $\varphi_x(h_{\mathbf{n}})$ replacing by $\Gamma(h_{\mathbf{n}})$, one can prove under assumptions **H1, H2**, **H4**–**H6** that,

$$\widetilde{I}_{\mathbf{n}}(x) + \widetilde{R}_{\mathbf{n}}(x) = O\left(\frac{1}{\widehat{\mathbf{n}}\Gamma(h_{\mathbf{n}})}\right) \quad \text{for any } x \in \mathcal{D}. \tag{8.A.7}$$

Let us define

$$\Omega_{\mathbf{n}} = \sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\Gamma(h_{\mathbf{n}})}} \text{and choose } \ell_{\mathbf{n}} \leq C\Omega_{\mathbf{n}}h_{\mathbf{n}}\Gamma(h_{\mathbf{n}})T_{\mathbf{n}}^{-1} \text{for some constant } C > 0.$$

We suppose that the compact set $\mathcal{D}$ is covered with $v_{\mathbf{n}}$ cubes B_k having sides of length $\ell_{\mathbf{n}}$ and centered at x_k. We have

$$\sup_{x\in D}\left|\widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right)\right| \leq Q_{1\mathbf{n}} + Q_{2\mathbf{n}} + Q_{3\mathbf{n}}, \tag{8.A.8}$$

where

$$Q_{1\mathbf{n}} = \max_{1\leq k\leq v_{\mathbf{n}}}\sup_{x\in B_k}\left|\widetilde{g}_{\mathbf{n}}(x) - \widetilde{g}_{\mathbf{n}}(x_k)\right|,$$

$$Q_{2\mathbf{n}} = \max_{1\leq k\leq v_{\mathbf{n}}}\sup_{x\in B_k}\left|\mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x_k)\right) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right)\right|,$$

$$Q_{3\mathbf{n}} = \max_{1\leq k\leq v_{\mathbf{n}}}\sup_{x\in B_k}\left|\widetilde{g}_{\mathbf{n}}(x_k) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x_k)\right)\right|.$$

Lemma 8.10 *Under Assumptions **H1**, **H2**, and **H4**, $Q_{1\mathbf{n}} = O\left(\Omega_{\mathbf{n}}\right)$ and $Q_{2\mathbf{n}} = O\left(\Omega_{\mathbf{n}}\right)$ a.s.*

Proof of Lemma A.10

By Assumptions **H1**, **H2**, and **H4**, for all $x \in B_k$,

$$\left|\widetilde{g}_{\mathbf{n}}(x) - \widetilde{g}_{\mathbf{n}}(x_k)\right| \leq a_{\mathbf{n}}^{-1}\widehat{\mathbf{n}}h_{\mathbf{n}}^{-1}T_{\mathbf{n}}\|x - x_k\| \leq Ch_{\mathbf{n}}^{-1}\Gamma(h_{\mathbf{n}})^{-1}T_{\mathbf{n}}\ell_{\mathbf{n}} = \left(\Omega_{\mathbf{n}}\right) a.s.$$

and Lemma 8.10 follows. □

Next, we have to show that

$$Q_{3\mathbf{n}} = O\left(\Omega_{\mathbf{n}}\right) \text{ a.s.} \tag{8.A.9}$$

Define

$$\widetilde{S}_{\mathbf{n}}(x) = a_{\mathbf{n}}^{-2} \sum_{\mathbf{i}\in I_{\mathbf{n}}} \widetilde{\Lambda}_{\mathbf{i}}(x) = \widetilde{g}_{\mathbf{n}}(x) - \mathbb{E}\left(\widetilde{g}_{\mathbf{n}}(x)\right).$$

Define also $\widetilde{U}(i,\mathbf{n},x,\mathbf{j})$ and $\widetilde{T}(\mathbf{n},x,i)$ to be the same as $U(i,\mathbf{n},\mathbf{j},x)$ and $T(\mathbf{n},i,x)$ in the proof of Lemma 8.4 except with $\Lambda_{\mathbf{j}}$ replacing by $\widetilde{\Lambda}_{\mathbf{j}}$. Arguing that $\widetilde{S}_{\mathbf{n}}$ is a finite sum of the $\widetilde{T}(\mathbf{n},x,i)$, then showing (8.9) is equivalent to show that

$$\max_{1\le k\le v_{\mathbf{n}}} \left|\widetilde{T}(\mathbf{n},x_k,1)\right| = O\left(\Omega_{\mathbf{n}}\right) \text{ a.s.} \tag{8.A.10}$$

By same arguments as in Lemma 8.4, $\widetilde{T}(\mathbf{n},1,x)$ is the sum of $\widehat{q} = q_1 \times \cdots \times q_N$ of the $\tilde{U}(i,\mathbf{n},\mathbf{j},x)$'s which are measurable with σ-field generated by $X_{\mathbf{i}}$, where $\mathbf{i}$ belong to the set of sites which are separated by a distance at least p. Enumerate these random variables as $Z_1,\ldots,Z_{\widehat{q}}$ and approximate them by the independent random variables $Z_1^*,\ldots,Z_{\widehat{q}}^*$ as was done in Lemma 8.1. Define

$$p \sim \Omega_{\mathbf{n}}^{-1/N} T_{\mathbf{n}}^{-1/N},$$

and

$$\widetilde{\beta}_{\mathbf{n}} = T_{\mathbf{n}}\Gamma(h_{\mathbf{n}})^{-1}\psi(\widehat{\mathbf{n}},p^N)p^{-\theta}\Omega_{\mathbf{n}}^{-1}.$$

Lemma 8.11 *Under assumptions of Theorem 8.2, there exist two positive constants A and C such that, for any $\lambda > 0$,*

$$\mathbb{P}\left(\max_{1\le k\le v_{\mathbf{n}}} \left|\widetilde{T}(\mathbf{n},x_k,i)\right| > \lambda\Omega_{\mathbf{n}}\right) \le C\widehat{\mathbf{n}}^{\beta}\left[\widehat{\mathbf{n}}^{-A} + \widetilde{\beta}_{\mathbf{n}}\right].$$

Proof of Lemma A.11

Since $\widetilde{T}(\mathbf{n},x,i) = \sum_{i=1}^{\widehat{q}} Z_i$, we have, for any $\lambda > 0$,

$$\mathbb{P}\left(\left|\widetilde{T}(\mathbf{n},x,i)\right| > \lambda\Omega_{\mathbf{n}}\right) \le \mathbb{P}\left(\sum_{i=1}^{\widehat{q}} \left|Z_i - Z_i^*\right| > \lambda\Omega_{\mathbf{n}}/2\right) + \mathbb{P}\left(\left|\sum_{i=1}^{\widehat{q}} Z_i^*\right| > \lambda\Omega_{\mathbf{n}}/2\right).$$

By the boundedness of K, we have

$$\left|Z_{\mathbf{i}}\right| \le Cp^N T_{\mathbf{n}} a_{\mathbf{n}}^{-1} \le CT_{\mathbf{n}} p^N \left(\widehat{\mathbf{n}}\Gamma(h_{\mathbf{n}})\right)^{-1}.$$

Note that $\widehat{\mathbf{n}} = 2^N p^N \widehat{q}$. Therefore, Markov inequality gives: for any $\lambda > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{\widehat{q}} \left|Z_i - Z_i^*\right| > \lambda\Omega_{\mathbf{n}}\right) \le 2\widehat{q}p^N T_{\mathbf{n}}\left(\widehat{\mathbf{n}}\Gamma(h_{\mathbf{n}})\right)^{-1}\psi(\widehat{\mathbf{n}},p^N)\chi(p)\lambda^{-1}\Omega_{\mathbf{n}}^{-1} \le C\widetilde{\beta}_{\mathbf{n}}.$$

By Lemma 8.7, we get, for any $\lambda > 0$, there exists a constant $C > 0$ such that

$$\mathbb{P}\left(\left|\sum_{i=1}^{\hat{q}} Z_i^*\right| > \lambda\Omega_{\mathbf{n}}\right) \leq C\hat{\mathbf{n}}^{-A},$$

and the conclusion follows. □

Proof of Lemma A.9 Note that by the Fubini's theorem, it can be seen that $\sum_{\mathbf{n}\in\mathbb{Z}^N} 1/(\hat{\mathbf{n}}u_{\mathbf{n}}) < \infty$. By (8.8), Lemma 8.10, and Lemma 8.11, proving Lemma 8.9 is equivalent to show that

$$\hat{\mathbf{n}}u_{\mathbf{n}}\hat{\mathbf{n}}^{\beta-A} \to 0 \text{ and } \hat{\mathbf{n}}u_{\mathbf{n}}\hat{\mathbf{n}}^{\beta}\widetilde{\beta}_{\mathbf{n}} \to 0 \text{ as } \mathbf{n} \to \infty. \tag{8.A.11}$$

Note that, the first part of (8.11) holds by choosing A such that $A > \beta + 2$. For its second part, when (8.3) is satisfied, $\psi(\hat{\mathbf{n}}, p^N) = p^N$ for $\mathbf{n}$ large enough. Then

$$\begin{aligned}
\hat{\mathbf{n}}^{\beta+1}u_{\mathbf{n}}\widetilde{\beta}_{\mathbf{n}} &\leq C\hat{\mathbf{n}}^{\beta}\left(\hat{\mathbf{n}}u_{\mathbf{n}}\right)^{1/s+1}\Gamma(h_{\mathbf{n}})^{-1}\Omega_{\mathbf{n}}^{(\theta-2N)/N}\left(\hat{\mathbf{n}}u_{\mathbf{n}}\right)^{(\theta-N)/sN} \\
&= C\hat{\mathbf{n}}^{\beta+1/s+1+(\theta-N)/(sN)+(2N-\theta)/(2N)}\Gamma(h_{\mathbf{n}})^{\frac{-\theta}{2N}}\left(\log\hat{\mathbf{n}}\right)^{\frac{\theta-2N}{2N}}u_{\mathbf{n}}^{\frac{sN+\theta}{sN}} \\
&= C\left[\hat{\mathbf{n}}\Gamma(h_{\mathbf{n}})^{\theta_1}\left(\log\hat{\mathbf{n}}\right)^{\theta_2}u_{\mathbf{n}}^{\theta_3}\right]^{\frac{2sN(\beta+2)+\theta(2-s)}{2sN}},
\end{aligned}$$

which goes to zero when $\theta > (2Ns(\beta+2))/(s-2)$.

Similarly, when (8.4) is satisfied, we have $\psi(\hat{\mathbf{n}}, p^N) \leq C\hat{\mathbf{n}}^{\kappa}$ for $\mathbf{n}$ large enough. Then,

$$\begin{aligned}
\hat{\mathbf{n}}^{\beta+1}u_{\mathbf{n}}\widetilde{\beta}_{\mathbf{n}} &\leq C\hat{\mathbf{n}}^{\beta+\kappa}\Gamma(h_{\mathbf{n}})^{-1}T_{\mathbf{n}}^{1+\theta/N}\Omega_{\mathbf{n}}^{\frac{\theta-N}{N}} \\
= C\hat{\mathbf{n}}^{\beta+\kappa+(N+\theta)/(sN)+(N-\theta)/(2N)}&\left(\Gamma(h_{\mathbf{n}})\right)^{\frac{-N-\theta}{2N}}\left(\log\hat{\mathbf{n}}\right)^{\frac{\theta-N}{2N}}u_{\mathbf{n}}^{\frac{N+\theta}{sN}} \\
&= C\left[\hat{\mathbf{n}}\Gamma(h_{\mathbf{n}})^{\theta_1^*}\left(\log\hat{\mathbf{n}}\right)^{\theta_2^*}u_{\mathbf{n}}^{\theta_3^*}\right]^{\frac{N(2s\beta+2s\kappa+s+2)+\theta(2-s)}{2sN}},
\end{aligned}$$

which goes to zero when $\theta > (N(2s\beta+2s\kappa+s+2))/(s-2)$ and Lemma 8.9 follows. □

Lemma 8.12 *Under Assumptions **H1**, **H2**, **H4**, and **H5**,*

(1) if (8.3) is satisfied and

$$\hat{\mathbf{n}}\Gamma(h_{\mathbf{n}})^{\theta_4}\left(\log\hat{\mathbf{n}}\right)^{\theta_5}u_{\mathbf{n}}^{\theta_6} \to \infty \text{ with } \theta > 2N(\beta+2),$$

(2) or if (8.4) is satisfied and

$$\hat{\mathbf{n}}\Gamma(h_{\mathbf{n}})^{\theta_4^*}\left(\log\hat{\mathbf{n}}\right)^{\theta_5^*}u_{\mathbf{n}}^{\theta_6^*} \to \infty \text{ with } \theta > N(2\beta+2\kappa+3),$$

then,

$$\sup_{x\in D}|f_{\mathbf{n}}(x)-1| = O\left(\left(\frac{\log\hat{\mathbf{n}}}{\hat{\mathbf{n}}\Gamma(b_{\mathbf{n}})}\right)^{1/2}\right) \text{ a.s,}$$

where

$$\theta_4 = \frac{\theta}{\theta - 2N(\beta+2)} \quad \theta_5 = \frac{\theta - 2N}{2N(\beta+2) - \theta} \quad \theta_6 = \frac{2N}{2N(\beta+2) - \theta},$$
$$\theta_4^* = \frac{-N - \theta}{N(2\beta + 2\kappa + 3) - \theta} \quad \theta_5^* = \frac{\theta - N}{N(2\beta + 2\kappa + 3) - \theta}$$
$$\theta_6^* = \frac{2N}{N(2\beta + 2\kappa + 3) - \theta}.$$

Proof of Lemma A.12

To prove Lemma 8.12, just adapt the arguments considered in the proof of Lemma 8.9 to the case where $Y_{\mathbf{i}} \equiv 1$ and $T_{\mathbf{n}} = 1$. □

References

1 Ripley, B.D. (1981). *Spatial Statistics*, *Wiley Series in Probability and Statistics*. Wiley.

2 Cressie, N.A.C. (1993). *Statistics for Spatial Data*, *Wiley Series in Probability and Statistics*, vol. 110. Wiley-Interscience, revised edn.

3 Guyon, X. (1995). *Random Fields on a Network: Modeling, Statistics, and Applications*, *Probability and its Applications*. Springer-Verlag.

4 Anselin, L. and Florax, R.J.G.M. (1995). *New Directions in Spatial Econometrics*, *Advances in Spatial Science*. Springer.

5 Chilés, J.P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*, *Wiley Series in Applied Probability and Statistics*. Wiley.

6 Journel, A.G. (1983). Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology* 15 (3): 445–468.

7 Tran, L.T. (1990). Kernel density estimation on random fields. *Journal of Multivariate Analysis* 34 (1): 37–53.

8 Carbon, M., Tran, L.T., and Wu, B. (1997). Kernel density estimation for random fields. *Statistics & Probability Letters* 36 (2): 115–125.

9 Biau, G. and Cadre, B. (2004). Nonparametric spatial prediction. *Statistical Inference for Stochastic Processes* 7 (3): 327–349.

10 Hallin, M., Lu, Z., and Yu, K. (2009). Local linear spatial quantile regression. *Bernoulli* 15 (3): 659–686.

11 Menezes, R., García-Soidán, P., and Ferreira, C. (2010). Nonparametric spatial prediction under stochastic sampling design. *Journal of Nonparametric Statistics* 22 (3): 363–377.

12 Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*, *Lecture Notes in Statistics*, vol. 149. New York: Springer-Verlag.

13 Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, *Springer Series in Statistics*, 2e. Springer.

14 Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*, *Springer Series in Statistics*. Springer.

15 Laksaci, A. and Maref, F. (2009). Estimation non paramétrique de quantiles conditionnels pour des variables fonctionnelles spatialement dépendantes. *Comptes Rendus Mathematique* 347 (17–18): 1075–1080.

16 Dabo-Niang, S., Yao, A.F., Pischedda, L. et al. (2010). Spatial mode estimation for functional random fields with application to bioturbation problem. *Stochastic Environmental Research and Risk Assessment* 24 (4): 487–497.

17 Laksaci, A. and Mechab, B. (2010). Estimation non paramétrique de la fonction de hasard avec variable explicative fonctionnelle: cas des données spatiales. *Revue Roumaine de Mathématiques Pures et Appliquées* 55 (1): 35–51.

18 Dabo-Niang, S., Rachdi, M., and Yao, A.F. (2011). Kernel regression estimation for spatial functional random variables. *Far East Journal of Theoretical Statistics* 37 (2): 77–113.

19 Attouch, M.K., Gheriballah, A., and Laksaci, A. (2011). Robust nonparametric estimation for functional spatial regression. In: *Recent Advances in Functional Data Analysis and Related Topics*, *Contributions to Statistics* (ed. F. Ferraty), 27–31. Physica-Verlag HD.

20 Dabo-Niang, S., Kaid, Z., and Laksaci, A. (2012). On spatial conditional mode estimation for a functional regressor. *Statistics & Probability Letters* 82 (7): 1413–1421.

21 Dabo-Niang, S. and Yao, A.F. (2013). Kernel spatial density estimation in infinite dimension space. *Metrika* 76 (1): 19–52.

22 Dabo-Niang, S., Hamdad, L., Ternynck, C., and Yao, A.F. (2014). A kernel spatial density estimation allowing for the analysis of spatial clustering: application to Monsoon Asia Drought Atlas data. *Stochastic Environmental Research and Risk Assessment* (Accepted, Online), 28: 2075–2099. http://dx.doi.org/10.1007/s00477-014-0903-6.

23 Dabo-Niang, S., Ternynck, C., and Yao, A.F. (2016). Nonparametric prediction of spatial multivariate. *Nonparametric Statistics* 2.: 428–458.

24 Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics* 27 (3): 832–837.

25 García-Soidán, P. and Menezes, R. (2012). Estimation of the spatial distribution through the kernel indicator variogram. *Environmetrics* 23 (6): 535–548.

26 Wang, H., Wang, J., and Huang, B. (2012). Prediction for spatio-temporal models with autoregression in errors. *Journal of Nonparametric Statistics* 24 (1): 217–244.

27 Ternynck, C. (2014). Spatial Regression Estimation for Functional Data with Spatial Dependency. *SFDS,155,2*.

28 Francisco-Fernández, M. and Opsomer, J.D. (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated

errors. *Canadian Journal of Statistics* 33 (2): 279–295. http://dx.doi.org/10.1002/cjs.5550330208.

29 Francisco-Fernández, M., Quintela-del Río, A., and Fernández-Casal, R. (2012). Nonparametric methods for spatial regression. An application to seismic events. *Environmetrics* 23 (1): 85–93.

30 El Machkouri, M. (2007). Nonparametric regression estimation for random fields in a fixed-design. *Statistical Inference for Stochastic Processes* 10 (1): 29–47.

31 El Machkouri, M. and Stoica, R. (2010). Asymptotic normality of kernel estimates in a regression model for random fields. *Journal of Nonparametric Statistics* 22 (8): 955–971.

32 El Machkouri, M. (2011). Asymptotic normality of the Parzen–Rosenblatt density estimator for strongly mixing random fields. *Statistical Inference for Stochastic Processes* 14 (1): 73–84.

33 Klemelä, J. (2008). Density estimation with locally identically distributed data and with locally stationary data. *Journal of Time Series Analysis* 29 (1): 125–141. http://dx.doi.org/10.1111/j.1467-9892.2007.00547.x.

34 Hallin, M., Lu, Z., and Tran, L.T. (2004). Local linear spatial regression. *The Annals of Statistics* 32 (6): 2469–2500.

35 Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*, *Lecture Notes in Statistics*, 2e, vol. 110. New York: Springer-Verlag.

36 Masry, E. (2005). Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic Processes and their Applications* 115 (1): 155–177.

37 Neaderhouser, C.C. (1980). Convergence of block spins defined by a random field. *Journal of Statistical Physics* 22 (6): 673–684.

38 Rosenblatt, M. (1985). *Stationary Sequences and Random Fields*. Boston, MA: Birkhauser.

39 Takahata, H. (1983). On the rates in the central limit theorem for weakly dependent random fields. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 64 (4): 445–456.

40 Kelejian, H.H. and Prucha, I.R. (2007). HAC estimation in a spatial framework. *Journal of Econometrics* 140 (1): 131–154.

41 Mitchell, W. (1966). The number of lattice points in a *k*-dimensional hypersphere. *Mathematics of Computation* 20 (94): 300–310.

42 Chamizo, L.F. and Iwaniec, H. (1995). On the sphere problem. *Revista Matemática Iberoamericana* 11 (2): 417–430.

43 Tsang, K.M. (2000). Counting lattice points in the sphere. *Bulletin of the London Mathematical Society* 32 (6): 679–688.

44 Meyer, A. (2011). On the number of lattice points in a small sphere. In: *WCC 2011 - Workshop on Coding and Cryptography*, pp. 463–472.

45 Gaetan, C. and Guyon, X. (2008). *Modélisation et statistique spatiales*, vol. 63. Springer.

46 Delsol, L. (2008). Régression sur variable fonctionnelle: Estimation, tests de structure et Applications. PhD thesis. Université Paul Sabatier-Toulouse III.

9

A Nonparametric Algorithm for Spatially Dependent Functional Data: Bagging Voronoi for Clustering, Dimensional Reduction, and Regression

Valeria Vitelli[1], Federica Passamonti[2], Simone Vantini[2], and Piercesare Secchi[2,3]

[1] *Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Domus Medica, Sognsvannsveien 9, 0372, Oslo, Norway*
[2] *Department of Mathematics, MOX, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy*
[3] *CADS - Center for Analysis Decisions and Society, Human Technopole, Via Cristina Belgioioso, 171, 20157, Milano, Italy*

9.1 Introduction

We aim at considering the problem of the nonparametric treatment of spatially dependent functional data, where the curves are indexed by the sites of a spatial finite lattice $S_0 \subset S$, being S the region of interest for the analysis. The literature on spatially dependent functional data is growing at a rapid pace [1, 2], and nonparametric methods play an important role (see [3–8], and Chapters 8, 10, and 11). Some functional regression techniques have recently been proposed for functional data indexed by a lattice [9–11], even though most of the available methods tackle the problem via spatiotemporal autoregressive models [12, 13]. Also clustering methods suited for functional lattice data are being developed (see [14–17], and Chapter 7).

We describe here a quite flexible nonparametric approach to the exploitation of spatial dependence on a lattice, which can be targeted to any kind of functional data analysis: e.g. classification, dimensional reduction, and regression. The approach is based on bagging [18], and it consists of two phases: a bootstrap phase, where many weak randomized analyses are repeatedly carried out, and an aggregation phase, where the many weak results are merged into a final strong result. Precisely, in the bootstrap phase, random connectivity maps are repeatedly generated and used to compute local representatives of neighboring functional data. A weak form of the target analysis (classification, dimensional reduction, and regression) is then performed on the sample of functional local

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

representatives at each repetition of the algorithm. The final aggregation phase has the purpose of combining all weak results together, and its structure strongly depends on the specific target of the analysis performed.

The algorithm is completely nonparametric, since no explicit assumption is made neither on the distribution generating functional data, nor on their mutual dependence induced by the spatial localization. A great advantage of this approach is its flexibility in the exploitation of further information on the considered region, which is not paid off by an excessive increment of the computational cost. The fact that our data is functional is not irrelevant to the computational cost of standard procedures for the analysis of lattice data, thus the use of a method which implicitly performs a reduction in the problem dimension (by analyzing a reduced number n of functional local representatives) has to be strongly preferred.

This chapter is structured as follows: our motivating application, which we use to test the different specifications of the method, is described in Section 9.2. In Section 9.3, the Bagging Voronoi (BV) strategy for spatially dependent functional data is introduced and described in its most general structure. In Section 9.4, its adaptation to functional clustering (Bagging Voronoi Clustering [BVClu]) is detailed. Bagging Voronoi Dimensional Reduction (BVDim) is described in Section 9.5, while a possible use of Bagging Voronoi for Regression (BVReg) is sketched in Section 9.6. Concluding remarks are discussed in Section 9.7. All analyses are performed in R [19].

9.2 The Motivating Application

In this chapter, we describe a case study in city management[1] that stimulated our research in the analysis of functional data spatially distributed on a lattice. Data are measures along time (every 15 minutes for 2 weeks) of the use of the Telecom mobile phone network across a lattice covering the area of Milan (Italy). These measures, named *Erlang*, refer to the average number of mobile phones connected to the network in each particular site of the lattice, in each considered time interval.

The great advantage in the exploitation of Erlang measures is that they are costless and freely available to any mobile phone network provider. Nevertheless, the analysis of such a kind of data can give insight on different aspects of the urban area they are referred to and can be developed with various scopes: the segmentation of the area into districts characterized by homogeneous usage patterns; the

1 Data are courtesy of Convenzione di ricerca DiAP – Telecom Italia, Politecnico di Milano (Italy).

identification of a set of "reference signals" able to describe the different patterns of utilization of the mobile phone network; the description of the influence of each telephonic pattern in each site of the lattice, via a linear model describing its link to a set of geographic covariates, when available.

Erlang measurements show a great variability across locations and time: this fact seems natural, since in the same lattice are included both sites referred to the center of Milan, and sites referred to the countryside. More precisely,

- The metropolitan area of Milan is partitioned as a uniform lattice S_0 of $|S_0| = N = 10\,573$ rectangular sites of 232 m × 309 m, distributed on a grid of 97 × 109 and covering an area of 757 km^2 included between latitude 45.37° and 45.57° North and longitude 9.05° and 95.35° East (see Figure 9.1).
- The data are provided every 15 minute for 14 days, from 18 March 2009, 00:15, until 31 March 2009, 23:45, as the average number of mobile phones simultaneously using the network for calling for 15 minutes. Given the presence of nonadmissible or missing values, we have a nonuniform time grid of $p = 1308$ elements, each one referring to a 15-minutes interval for which a measurement is available in at least a lattice site (see Figure 9.2).

The Erlang $E_{\mathbf{x}j}$, associated with the lattice site $\mathbf{x} \in S_0$ in the j-th quarter of an hour, is defined as

$$E_{\mathbf{x}j} = \frac{1}{15} \sum_{q=1}^{Q} \left| T^q_{\mathbf{x}j} \right|, \tag{9.1}$$

Figure 9.1 Map of the region around Milan (metropolitan area) covered by the Telecom Italia database.

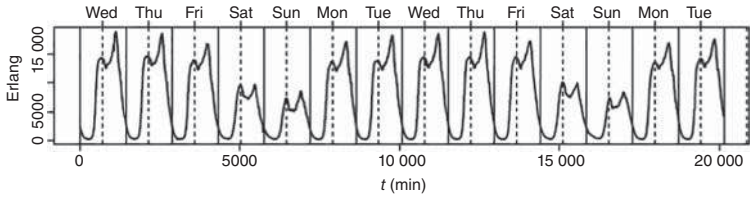


Figure 9.2 The total Erlang data as a function of time. Continuous vertical line separate weekdays, while dotted lines are drawn at noon.

where Q is the number of phones having at least one connection event to the Telecom network within site $\mathbf{x}$ during the i-th quarter, and $T^q_{\mathbf{x}j}$ represents the time interval in which phone q is calling while being within site $\mathbf{x}$ and during the j-th quarter of an hour; $|T^q_{\mathbf{x}j}|$ is its duration (minutes). The definition of the Erlang data given in (9.1) is the one actually used by the company for the computation, but it can be mathematically clarified with the following expression:

$$E_{\mathbf{x}j} = \frac{1}{15}\int_{15(j-1)}^{15j} N_{\mathbf{x}}(t)dt. \tag{9.2}$$

If we indicate with $N_{\mathbf{x}}(t)$ the number of mobile phones using the network within site $\mathbf{x}$ at time t, the expression in (9.2) clearly shows that $E_{\mathbf{x}j}$ is an average over the j-th quarter of an hour. The two expressions (9.1) and (9.2) are in fact equivalent, as it can be deduced from the following identities [20]:

$$\frac{1}{15}\sum_{q=1}^{Q}\left|T^q_{\mathbf{x}j}\right| = \frac{1}{15}\sum_{q=1}^{Q}\int_{15(j-1)}^{15j}\mathbf{1}_{\left\{T^q_{\mathbf{x}j}\right\}}(t)dt = \frac{1}{15}\int_{15(j-1)}^{15j}N_{\mathbf{x}}(t)dt.$$

The whole Erlang data set as just described includes 13 829 484 records, among which 110 475 are missing. Erlang data can be considered as functional data with spatial dependence, due to the high frequency of measurements in time and, on the other hand, to the georeferentiation of observations on the spatial lattice. An overview of the data as a function of time is given in Figure 9.2, where the sum of the Erlang data on the whole region of interest $\sum_{\mathbf{x}\in S_0}E_{\mathbf{x}j}$ is shown. We can immediately notice differences between night/day and weekdays/weekends, and this general trend will be investigated in the analysis together with more local behaviors, both restricted to specific time windows and/or associated to particular subregions.

9.2.1 Data Preprocessing

The discrete sequence of Erlang values in each given site can be considered as a sampling of a continuous process in time [21], describing the average

number of mobile phones using the network in that site, as expressed in Eq. (9.2). Indeed, in each site of the lattice, we observe a discrete version of the Erlang continuous process, recorded approximately every quarter of an hour: due to discontinuities in the information provided by the network antennas, the Erlang measure is missing at some time instances, and hence, the time grid of Erlang measurements is nonuniform. Moreover, some Erlang recordings are negative due to measurement errors and should be treated as missing values. We thus need to choose a proper basis expansion to reconstruct the functional form of the time-varying Erlang data and to evaluate them on a common uniform grid of time values, before applying the methodologies presented in the rest of the chapter.

We perform a sitewise smoothing of the Erlang data via a Fourier basis expansion due to the evident seasonality in the Erlang profiles. We set the period of the Fourier basis equal to one week: hence, the reconstructed functional form of the Erlang profile for site $\mathbf{x} \in S_0$ is a function $E_{\mathbf{x}}(t)$ such that

$$E_{\mathbf{x}}(t) = \frac{c_0^{\mathbf{x}}}{2} + \sum_{h=1}^{H} \left[a_h^{\mathbf{x}} \cos(h\omega t) + b_h^{\mathbf{x}} \sin(h\omega t) \right], \tag{9.3}$$

where $t \in [0; T]$, $\omega = 2\pi/T$, and $T = 60 \times 24 \times 7$ are the period expressed in minutes. The coefficients, $c_0^{\mathbf{x}}$, $a_h^{\mathbf{x}}$, and $b_h^{\mathbf{x}}$, are estimated by means of least squares. To choose the basis dimension H, we analyze the power spectrum associated with the sitewise smoothing of the Erlang data with a Fourier basis of large dimension ($H = 200$). The power spectrum of the Fourier expansion of a signal represents the amplitude of the signal as a function of the frequency, and at the h-th frequency, it is given by

$$P_{\mathbf{x}}(h) = \sqrt{\left(a_h^{\mathbf{x}}\right)^2 + \left(b_h^{\mathbf{x}}\right)^2}. \tag{9.4}$$

Hence, a local maximum in the power spectrum detects a frequency explaining relevant features in the data, while when it vanishes toward zero, there is no need to include higher frequency terms. We choose the most proper value of H by inspecting the shape of the average power spectrum over all sites of the lattice, i.e. $\overline{P}(h) = \frac{1}{N} \sum_{\mathbf{x} \in S_0} P_{\mathbf{x}}(h)$, plotted as a function of h in Figure 9.3: the frequencies significantly contributing to the Erlang time variation are the smaller ones (all less than 7), capturing differences among days or blocks of days (e.g. the working and weekend days variation), and the multiples of 7, capturing the recurring daily dynamics. Due to the huge dimension of the Telecom Italia database, we choose a basis of very high dimension, in order to be reasonably sure to catch all relevant localized features: we thus set $H = 100$ for subsequent analyses, which ensures a rich enough basis. Other approaches for tuning H are of course conceivable (MSE minimization, cross-validation, …).

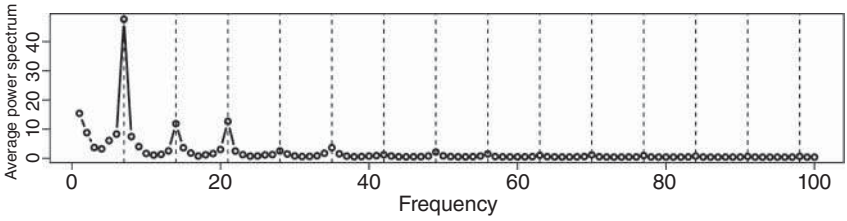


Figure 9.3 Average power spectrum $\overline{P}(h)$ obtained via sitewise smoothing of the Erlang measures with a Fourier basis of dimension $H = 200$. Only the values of $\overline{P}(h)$ for $h = 1, \dots, 100$ are shown in the plot. Dotted vertical lines are drawn for multiples of 7.

9.3 The Bagging Voronoi Strategy

Let $E_{\mathbf{x}}(t)$ be the value of the functional random variable at site $\mathbf{x} \in S_0$ and time t. It is evident that a proper handling of spatial dependence is crucial for treating functional data indexed by a lattice: the dependency along time observed in neighboring sites can be strong, both for physical and for technical reasons. We treat spatial dependence by following a Bagging Voronoi strategy first introduced in [22], and here sketched via a pseudo-code scheme in the next page.

The rationale beyond this strategy is simple, but effective: the algorithm is divided in two main parts, a *Bootstrap Phase* and an *Aggregation Phase*, the former consisting in the bootstrap repetition of many "weak versions" of the target analysis, and the latter having the scope of bagging together the weak analyses results to obtain a conclusive strong result. In the Bootstrap Phase, the initial randomization is given by a random system of neighborhoods, differently capturing local behaviors. The original data set is then replaced with a reduced one, composed by functional local representatives of subsets of data belonging to the same neighborhood, which is then analyzed. The repetition of this weak analysis many times for different reduced data sets associated with different randomly generated systems of neighborhoods makes the final aggregated result more accurate, and it substantially reduces the overall computational costs associated with the procedure.

More precisely, at the b-th bootstrap repetition, $b = 1, \dots, B$: first, a n-dimensional partition of the considered region S in random neighborhoods is obtained by generating n random locations, or nuclei, $\{\mathbf{v}_1^b, \dots, \mathbf{v}_n^b\} \in S$, and then computing the Voronoi tessellation of S_0 induced by those nuclei, i.e. $V_i^b := \left\{ \mathbf{x} \in S_0 : d\left(\mathbf{x}, \mathbf{v}_i^b\right) = \min_{j=1,\dots,n} d\left(\mathbf{x}, \mathbf{v}_j^b\right) \right\}$ for $i = 1, \dots, n$. Hence, the distance $d(\cdot, \cdot)$ defines the notion of closeness on the lattice, and its choice depends on the application (it can simply be an Euclidean distance on $\mathbb{R}^2$, or a geodesic if the lattice is defined on a manifold). A functional local representative

Algorithm. Bagging Voronoi.

Initialize B, n, and choose a metric $d(\cdot, \cdot)$.

`Bootstrap Phase`
`for` $b := 1$ `to` B `do`

- `step 1.` generate a n-dimensional random Voronoi tessellation of the lattice, i.e. group together neighboring sites of the lattice, where the notion of "closeness" is defined by $d(\cdot, \cdot)$;
- `step 2.` identify a functional local representative for each element of the tessellation to sum up local information: due to spatial dependence, neighboring data are most likely drawn from the same functional distribution;
- `step 3.` analyze the sample of functional local representatives with a suitable statistical technique, depending on the scopes of the whole analysis: functional clustering, dimensional reduction, and functional regression.

`end for`

`Aggregation Phase`

- perform a `matching` of the results along the B bootstrap repetitions, to deal with possible identifiability issues;
- `aggregate` the results of each single bootstrap repetition into a stronger final result.

$g_i^b(t)$ is then computed as a summary of the functional data $\{E_{\mathbf{x}}(t)\}_{\mathbf{x}\in V_i^b}$, via a weighted mean with Gaussian isotropic weights (also called spatial smoothing in [23]). The set of functional local representatives $\left\{g_1^b(t), \ldots, g_n^b(t)\right\}$ exploits a specific structure of spatial dependence, and it is expected to be less noisy and less spatially dependent [22]. Finally, any analysis suited to treat functional data (functional clustering, dimensional reduction, and functional regression) can be used on the set of functional local representatives, to obtain a coarse (or weak) estimate of the final target analysis. The coarse estimate (a cluster label or a set of basis scores) is then assigned to all sites of the lattice belonging to the element of the partition associated with the considered representative.

Note that one has to fix some parameters in advance: n, the dimension of the random partition and the size of the sample of functional local representatives; B, the number of bootstrap replicates; $d(\cdot, \cdot)$, the most proper metric to measure distances in the considered region. While $d(\cdot, \cdot)$ has to be chosen according to the particular application at hand, both the choice of B and of n require further

attention. In general, larger values of B imply a higher accuracy of the final estimate, so this parameter can be tuned in order to achieve the desired accuracy; however, we also investigate a stability analysis to guide the specification (see Section 9.5.1.1). Also n, which sets the tessellation dimension and thus the number of local representatives to be computed, deserves some attention: the tuning of this parameter has been extensively studied in the light of simulations [20, 22], and for the specific purpose of maximizing analysis-specific performance indicators. Nevertheless, these studies pointed out a quite general conclusion. The optimal choice of n is the one that finds a good compromise between variance and bias of the local representatives: as n decreases, noise is reduced in the local representatives sample, since local representatives are weighted sample means calculated on subsamples that are larger on average (minimal variance), but at the same time, the associated Voronoi tessellation becomes coarser, thus including data with different characteristics in the computation of local representatives (maximal bias). On the other hand, as n increases, the resulting Voronoi tessellation becomes more and more refined, being able to catch very localized spatial features (minimal bias), but at the same time, the variability reduction due to spatial smoothing is smaller (maximal variance). The optimal value of n determined by this trade-off depends both on the strength of spatial dependence and on the distribution of the spatial signal generating the functional data. In [22], a tuning criterion for this parameter based on the *total entropy* had been proposed for the purposes of clustering, while in [20], the total *average variance* of the scores is considered, since it is more suited to the purposes of dimensional reduction. Details of these tuning criteria will be given for each model specification in the coming sessions.

9.4 Bagging Voronoi Clustering (BVClu)

Suppose a latent field of labels $\Lambda_0 : S_0 \to \{1, \dots, L\}$ is associated with each site of the lattice S_0, i.e. $\Lambda_0(\mathbf{x})$ is the true unknown label associated with the site $\mathbf{x} \in S_0$: the label sums up some characteristics of the considered area which are interesting for the scopes of the analysis, and L is the unknown number of labels present in the area. Moreover, suppose that, given Λ_0, the Erlang $E_{\mathbf{x}}(t)$ are independently generated in each site $\mathbf{x} \in S_0$ from a distribution indexed by $\Lambda_0(\mathbf{x})$: this means that, given the characteristics of the area in site $\mathbf{x}$ as summarized by $\Lambda_0(\mathbf{x})$, the Erlang profile $E_{\mathbf{x}}(t)$ will be drawn from a different distribution. This hypothesis is the basis for models like Hidden Markov Random Fields (HMRF) are a typical setup for spatially dependent multivariate data on a lattice. However, most algorithms for image analysis based on HMRF models (see [24, 25] for details on these procedures) heavily depend on hypotheses on the multivariate distribution of the observed signal, which are often too restrictive or anyhow unrealistic in

a functional data context. Hence, our nonparametric treatment of spatial dependence might help in solving the functional unsupervised classification problem when functions are indexed on the sites of a lattice S_0.

Aim of the classification procedure is to reconstruct the unknown field Λ_0 of labels. Hence, the final result of the procedure is a label assignment for each site of the lattice. The general methodology described in Section 9.3 for handling spatially dependent functional data can be adapted to the present clustering purpose by specifying both the Bootstrap and the Aggregation Phases of the Bagging Voronoi (BV) algorithm. This entails:

(1) performing functional clustering on the set of local representatives at each bootstrap repetition;
(2) matching the cluster labels in each site along bootstrap repetitions, to ensure identifiability, and then computing the frequencies of assignment of the site to each one of the K clusters.

Point (1) refers to the Bootstrap Phase of the BV algorithm, while point (2) refers to the Aggregation Phase. Clarifying these two points is the scope of the present section, which will give the details of the BVClu strategy. Note that BVClu is a refined version of the method first introduced in [22], where a different case study was shown, since BVClu does not necessarily rely on the use of a functional basis to project the functional lattice data.

Concerning point (1), many strategies to functional clustering are conceivable, many of which are based on performing dimensional reduction first [26]. Given that we already reduced the problem dimension by computing a sample of functional local representatives at each bootstrap repetition, we here choose a functional clustering method which directly handles the curves. Due to the data characteristics described in Section 9.2, curves misalignment is not an issue in this case. However, the strong localized features of the data in time have to be properly taken into account. We thus use K-medoid functional clustering [27], based on the L^1 distance among the curves, to be reasonably robust to localized features. The Bootstrap Phase of the BV algorithm is thus specified such that, at each bootstrap repetition $b = 1, \dots, B$, K-medoid functional clustering is used on the set of local representatives $\{g_1^b, \dots, g_n^b\}$ to obtain a set of labels $\{\Gamma_1^b, \dots, \Gamma_n^b\} \in \{1, \dots, K\}$ assigning each local representative to one of the K clusters. Then, all sites $\mathbf{x} \in V_i^b$ get the same label Γ_i^b at the b-th bootstrap iteration: for $k = 1, \dots, K$, and $b = 1, \dots, B$, we indicate with C_k^b the set of $\mathbf{x} \in S_0$ whose label is equal to k.

Concerning point (2), the B coarse results obtained after the Bootstrap Phase $\{C_k^b\}_{k=1,\dots,K}^{b=1,\dots,B}$ must be aggregated in the Aggregation Phase, which consists of two tasks: cluster matching across bootstrap repetitions, to ensure identifiability, and the actual aggregation of clustering results in a final frequency of assignment,

which is indeed based on the assumption that cluster labels are coherent along bootstrap repetitions. Cluster matching is structured as follows: for $b \geq 2$, look for the label permutation $\{l_1, \dots, l_K\}$ in the set $\{1, \dots, K\}$ that minimizes the total sum of the off-diagonal frequencies in the contingency table describing the joint distribution of sites along the two classifications $C_1^{b-1}, \dots, C_K^{b-1}$ and $C_{l_1}^b, \dots, C_{l_K}^b$. Then, the labels identifying the clusters $C_1^b, \dots, C_K^b$ are renamed by permuting them according to $\{l_1, \dots, l_K\}$.

Once cluster matching has been performed, we can finally aggregate the B coarse results of the Bootstrap Phase. The frequency distribution of assignment of each site to each of the K clusters along the B repetitions is thus computed: for each site $\mathbf{x} \in S_0$, one can compute $\pi_{\mathbf{x}}^k = \#\left\{b \in \{1, \dots, B\} : \mathbf{x} \in C_k^b\right\}/B, \ \forall\ k = 1, \dots, K$. A final assignment of site $\mathbf{x}$ to *one* of the K clusters can be obtained by selecting that label corresponding to a mode of the distribution $\boldsymbol{\pi}_{\mathbf{x}} = \left(\pi_{\mathbf{x}}^1, \dots, \pi_{\mathbf{x}}^K\right)$.

In order to perform BVClu, some parameters need to be properly tuned. In order to set n, the number of elements of each random partition, and K, the number of clusters, we can vary them in reasonable ranges and examine the behavior of two performance measures: the average normalized entropy, designed to assess the uncertainty associated with cluster assignments along bootstrap replicates, and the Wilks's λ, which evaluates the quality of the final classification.

The spatial entropy criterion for assessing the uncertainty associated with cluster assignments was first introduced in [22]. Consider the frequency distribution of the assignments $\boldsymbol{\pi}_{\mathbf{x}} = \left(\pi_{\mathbf{x}}^1, \dots, \pi_{\mathbf{x}}^K\right)$ of each site $\mathbf{x} \in S_0$ to each of the K clusters. The entropy associated with the final classification in the site $\mathbf{x} \in S_0$ is obtained as $\eta_{\mathbf{x}}^K = -\sum_{k=1}^K \pi_{\mathbf{x}}^k \cdot \log\left(\pi_{\mathbf{x}}^k\right)$, which is close to 0 for peaked distributions of assignments and close to the maximum $\log(K)$ for quite uniform ones. The more the frequency distribution $\boldsymbol{\pi}_{\mathbf{x}}$ is concentrated on one particular label, the more the classification is precise and stable along replicates. Hence, minimizing the entropy is a good strategy to assess the uncertainty of the clustering result. A global measure, involving all sites of the lattice, can be computed as

$$\eta^K = \frac{\sum_{\mathbf{x} \in S_0} \eta_{\mathbf{x}}^K}{\log(K) \cdot |S_0|}, \tag{9.5}$$

and we name this global measure *average normalized entropy*. Since the criterion expressed in (9.5) is a measure of the uncertainty associated with the cluster assignments, we expect η^K to be minimal if n properly accounts for the (unknown) spatial dependence in the latent field of labels.

One could guess that the average normalized entropy would be a good criterion also for selecting K. Simulation studies performed in [22] indeed showed that the entropy criterion generally leads to a choice for K more parsimonious than necessary, and an alternative approach was there proposed, directly targeted to instead check the goodness/quality of the final classification. This measure is close to a

Wilks's λ [28], a classical quality measure in cluster analysis, and has the following form:

$$\theta = \frac{\text{tr}(B)}{\text{tr}(B+W)}, \tag{9.6}$$

where B and W are the final *between* and *within* cluster sum of squares matrix, respectively.

9.4.1 BVClu of the Telecom Data

9.4.1.1 Setting the BVClu Parameters

To set n and K, we examined the behavior of the previously introduced performance measures, while for what concerns B, the number of bootstrap replicates, we decided to fix it to a "reasonable" value ($B = 50$ for BVClu), meaning that it is large enough to give meaningful results, but small enough not to make computations too intense.

We varied $n \in \{500, 750, 850, 1000, 1250\}$, and $K \in \{1, \dots, 10\}$. We used the BVClu strategy with functional K-medoid clustering on the Telecom data. Results are shown in Figure 9.4a, we plot the average normalized spatial entropy η as a function of n for different choices of K, and we seek to find a minimum with respect to n for reasonable choices of K, taking also into account that the entropy criterion might be parsimonious for lower K's. We see that $n = 750$ is a good candidate choice, showing a minimum for various reasonable K's and showing stabilized results for $K = 2$. In Figure 9.4b, we can then inspect the behavior

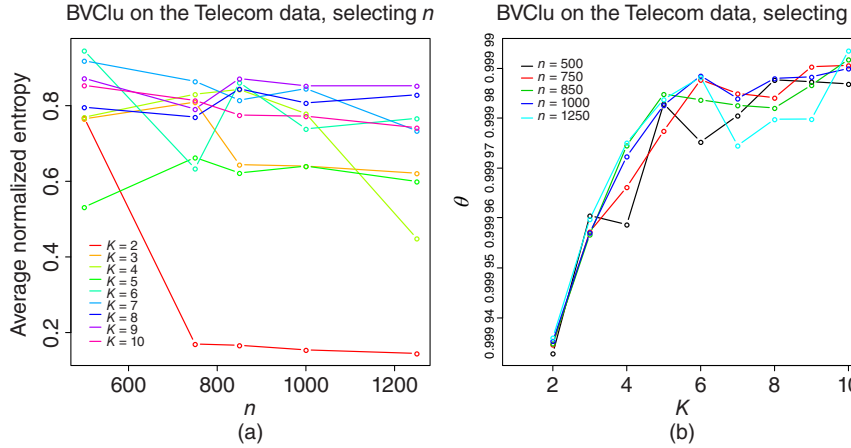


Figure 9.4 Results of BVClu on the Telecom data. Average normalized entropy η along 50 bootstrap replicates (a), and θ values for evaluating the quality of the final classification (b), obtained for various choices of n and K.

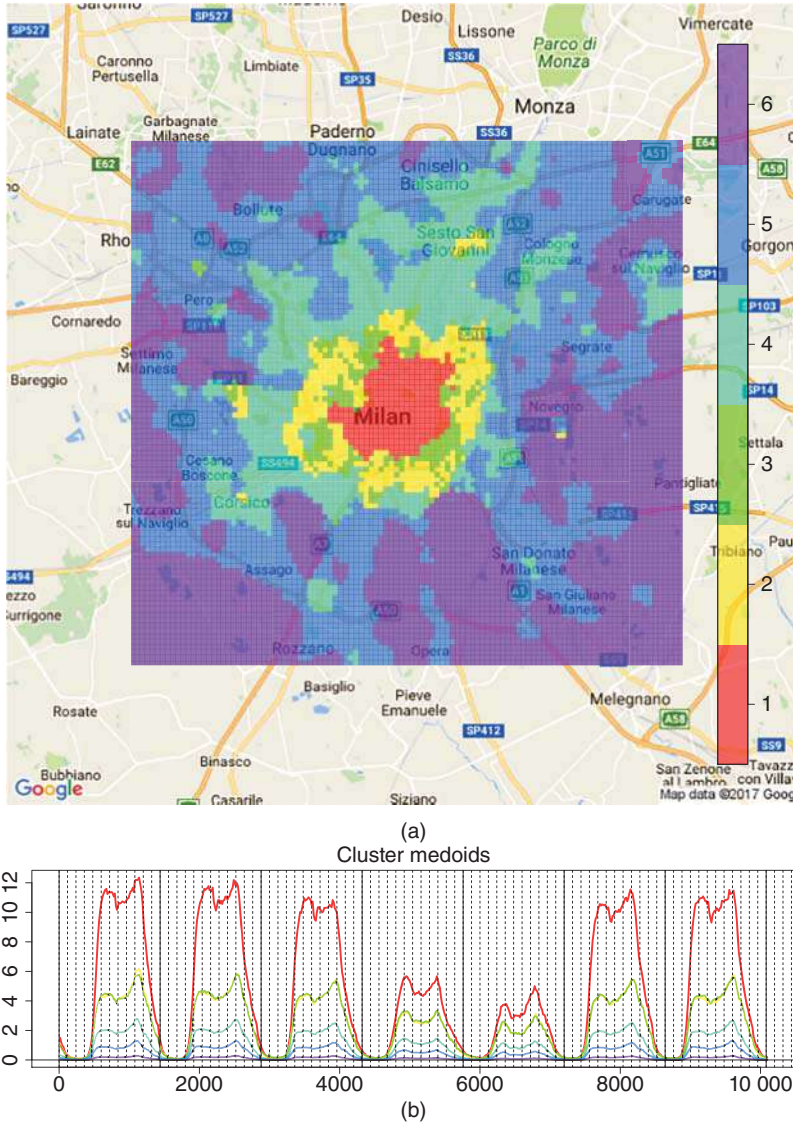


Figure 9.5 Results of BVClu on the Telecom data, with $n = 750$ and $K = 6$. (a) Map of the final cluster assignments superimposed to the map of the metropolitan area of Milan. (b) The six final cluster medoids. The shade of gray identifying a cluster on the map in (a) is associated with that of the cluster medoid in (b).

of θ with respect to K for different values of n: here we look for an "elbow," showing that a further increase in K does not imply a reasonable increase in the proportion of data variability explained by the clusters. For $n = 750$, the elbow is clearly visible at $K = 6$, but even equally reasonable values for n point to the same conclusion. Looking back at Figure 9.4a, we recognize a clear minimum in $n = 750$ in the entropy curve for $K = 6$. These two choices thus seem the most reasonable ones.

9.4.1.2 Results

The results obtained performing BVClu as detailed in Section 9.4, run using $B = 50$ bootstrap iterations with Voronoi tessellations of dimension $n = 750$, and when estimating six clusters, are shown in Figure 9.5. In Figure 9.5a, the final cluster assignments in each site of the lattice $\mathbf{x} \in S_0$ are superimposed to the considered geographical area, i.e. the metropolitan area of Milan (see Figure 9.1). In Figure 9.5b, the six final cluster medoids estimated by the procedure are shown. As it is evident from the medoids, the main feature distinguishing among the clusters is the scale of the Erlang measurements: the medoid associated with cluster 1, which corresponds to the map to sites in the very center of Milan, has peaks up to 12 Erlangs; clusters 2 and 3, localized in the close neighborhoods and periferic areas of the city, peak at around 6 and are very similar in shape; cluster 4, representing the suburbs, peaks at 2; cluster 5, which points at highways and at the city ringroads, peaks at 1.2; and finally cluster 6, mainly countryside, has maximum value 0.3. Apart from this very neat distinction, the medoids do not show any other relevant difference: the shape is approximately the same, with the same daily and weekly periodicity already evident in the total Erlang data in Figure 9.2. This result is totally realistic if we think that functional clustering with respect to the L^1 metric does not imply a distinction among the data in terms of shape, but rather of their respective scale, especially if the scale varies consistently within the dataset. For capturing repeated localized behaviors, either the functional distance has to be changed, or a dimensional reduction method has to be employed, to capture via few time profiles the average behavior of many neighboring sites. This will be the focus of Section 9.5.

9.5 Bagging Voronoi Dimensional Reduction (BVDim)

When the purpose of the analysis is dimensional reduction, what we have in mind is an additive model where $E_{\mathbf{x}}(t)$ is represented by the value at t taken by a limited number K of time-varying functions $\{\psi_1, \dots, \psi_K\}$, common to the entire lattice

S_0, and coupled with the values at $\mathbf{x}$ taken by K latent random fields $\{D_1, \dots, D_K\}$ indexed by the sites of S_0

$$E_{\mathbf{x}}(t) = \sum_{k=1}^{K} D_k(\mathbf{x})\psi_k(t) + \epsilon. \tag{9.7}$$

The random error term ϵ is assumed to be independent on $\{D_1, \dots, D_K\}$, with zero mean and bounded variance. Model (9.7) is often assumed in the context of reduced basis representation methods, like Functional Principal Component Analysis (FPCA) and Functional Independent Component Analysis. The important difference here stands in the fact that we take into account spatial dependence, through the random fields $\{D_1, \dots, D_K\}$ generating the scores of the basis expansion.

Model (9.7) implies the following regression model for the collection $\{E_{\mathbf{x}}(t)\}_{\mathbf{x}\in S_0}$ of time profiles belonging to the Erlang dataset

$$\mathbb{E}[E_{\mathbf{x}}(t)|D_1 = d_1, \dots, D_K = d_K] = \sum_{k=1}^{K} d_k(\mathbf{x})\psi_k(t), \tag{9.8}$$

for $\mathbf{x} \in S_0$ and $t \in [0, T]$. The time-varying functions $\{\psi_1, \dots, \psi_K\}$ are unknown, each function describing a time profile for mobile phone activity. The surfaces $\{d_1, \dots, d_K\}$ are the *unobserved* realizations of the random fields $\{D_1, \dots, D_K\}$. The K values $\{d_1(\mathbf{x}), \dots, d_K(\mathbf{x})\}$ represent the contributions to the Erlang time profile $E_{\mathbf{x}}$ of their coupled time-varying functions. The other way round, the K time-varying functions $\{\psi_1, \dots, \psi_K\}$ express the evolution in time of the coupled surfaces.

The general methodology described in Section 9.3 for handling spatially dependent functional data can be adapted to the present purpose of performing dimensional reduction, i.e. estimating both sets of functions $\{\psi_1, \dots, \psi_K\}$ and $\{d_1, \dots, d_K\}$. This entails:

(1) performing dimensional reduction on the set of local representatives at each bootstrap repetition;
(2) matching the bases obtained across bootstrap repetitions and then combining them into a final basis.

Note that, as in case of BVClu, point (1) refers to the Bootstrap Phase of the BV algorithm, while point (2) refers to the Aggregation Phase, and we give here the details of both. The strategy and results here reported are a refined version, where the tuning of the parameters has been better explored, of the analysis described in [20] under the name of *Bagging-Voronoi Treelet Analysis*. We named the procedure BVDim to stress its generality for what concerns the employed dimensional reduction strategy.

Concerning point (1), the most common approach [29] when dealing with functions is FPCA, which allows finding optimal subspaces for representing data. However, this kind of method usually builds the new basis as a linear combination of all (or most of) the original variables, thus providing a global representation, without being able to capture localized behaviors. Following Secchi et al. [20, 30], we instead use *treelets*, which are a multiscale data-driven orthonormal basis forming a hierarchical tree. Treelets are built on wavelets: the wavelet approach is combined with a PCA performed hierarchically on the couple of most correlated variables at each level (see [30] for details on this dimensional reduction procedure).

The Bootstrap Phase of the BV algorithm is thus modified so that, at each bootstrap repetition $b = 1, \dots, B$, a treelet analysis is used on the set of local representatives $\left\{g_1^b, \dots, g_n^b\right\}$, evaluated on a fine grid of equally spaced abscissa values (see Section 9.2.1), to obtain an orthogonal treelet basis $\left\{\varphi_1^b, \dots, \varphi_J^b\right\}$. Then, each functional local representative g_i^b, $i = 1, \dots, n$ is orthogonally projected onto each basis element φ_j^b, $j = 1, \dots, J$ and the corresponding score $d_j^b(\mathbf{x})$ is assigned to all $\mathbf{x} \in V_i^b$.

Concerning point (2), it is important to recall what is the output of the Bootstrap Phase: a collection of reference basis functions $\left\{\varphi_1^b(t), \dots, \varphi_J^b(t)\right\}_{b=1}^B$, and of their coupled surfaces $\left\{d_1^b(\mathbf{x}), \dots, d_J^b(\mathbf{x})\right\}_{b=1}^B$. These B coarse results must then be aggregated in the Aggregation Phase, which consists of two tasks: a matching of the bases along the B bootstrap repetitions, to ensure their comparability, and their actual aggregation in a final reference basis. We handle both tasks simultaneously via 1-median basis alignment, which jointly computes the reference basis from the B coarse bases, while also reordering their elements. This procedure is inspired by the 1-medoid alignment method [31], but in the context of bases matching each datum is a multivariate function (one of the coarse bases), and we look for the unique prototype (the reference basis) which best describes the set of functional objects, while also aligning their components, by permutations in the order of basis functions. Note that the chosen aggregation strategy is a discrete variation of a Procrustes alignment procedure [32–35]. For more details on the Alignment Phase in BVDim see [20].

9.5.1 BVDim of the Telecom Data

9.5.1.1 Setting the BVDim Parameters

In order to perform BVDim, we need to fix some parameters in advance. Specifically, we have to set n, the number of elements of each random partition; B, the number of bootstrap replicates; and K, the number of selected relevant elements of the final reference basis. The choice of n, B, and K is somehow related, thus making their tuning even more tricky. We proceed in the following way: we first fix B to a "reasonable" value, meaning that it is large enough to give meaningful results, but small enough not to make computations too intense. We then perform

BVDim for large J (i.e. without selecting K) with the scope of tuning n. Once n is selected, we focus on a good choice for K. Finally, we perform a stability analysis of the so obtained basis functions (for fixed n and K) to find the optimal B.

Hence, for fixed $B = 50$, we focus on n and K. The right value for n is based on the optimization of a measure of stability across bootstrap replications: for each site $\mathbf{x} \in S_0$ and $J = 1, \dots, J$, we compute the bootstrap variance of $\tilde{d}_j^b(\mathbf{x})$ (i.e. the aligned $d_j^b(\mathbf{x})$) over $b = 1, \dots, B$; we then compute the average over $\mathbf{x} \in S_0$ and sum over j. This quantity is called *Total Average Variance* (TAV) and has to be minimized, varying the possible number of elements in the tessellation, in order to obtain the optimal n. In [20], the optimal value turned out to be $n = 850$, indicating that 1 km is the relevant practical spatial range of the Erlang data. The number K of relevant basis elements, as in most dimensional reduction techniques, is instead chosen considering the fraction of total variance explained by each component. In particular, for each element j, given the final collection of surfaces $\left\{\tilde{d}_j^b(\mathbf{x})\right\}_{b=1}^B$, we compute the variance over b and sum over $\mathbf{x} \in S_0$ to obtain $\tilde{s}_j^2$. The selected elements are those associated to the K-th significantly largest variances. Following this strategy, six basis elements turned out to have higher variance, even though only four of them were fully commented and described in [20]. Of course, a lower threshold on the proportion of explained total variance could lead to the inclusion of more treelets, up to a couple dozens.

Once n and K have been selected, the number of bootstrap replicates B can be carefully tuned so to avoid possible instabilities in the results due to the random components of BVDim. This is what we call *stability analysis*: the concept of stability relates to the variation of the basis elements with respect to a reference result, which we set as the basis associated to $B_{max} = 200$ (the largest number of boostrap replicates here considered): $\Psi_{ref} = \{\psi_1, \dots, \psi_6\}_{B=200}$. In each basis, for $B \in \mathcal{B} = \{50, 80, 110, 140, 170, 200\}$, we consider the elements in decreasing order according to their total variance $\tilde{s}_j^2$, meaning that the first or second element in a basis is the one with the largest or second-largest value of $\tilde{s}_j^2$, and so on. We recall that for a proper comparison of the bases $\{\psi_1, \dots, \psi_6\}_{B \in \mathcal{B}}$ an alignment of the elements is also needed, since for different values of B positions in the variance-based order might vary. After the alignment, for each element, we measure its distance from the corresponding element of the reference basis Ψ_{ref}. A graphical representation of the stability analysis is given in Figure 9.6, where the Euclidean distance from Ψ_{ref} is plotted vs. the number of bootstrap iterations (x-axis), and the different basis elements $j = 1, \dots, 6$ are drawn in shades of gray; we see that for the first, second, and fifth elements, 50 bootstrap iterations are sufficient for a good matching with the corresponding components of the reference basis. However, the other elements are more unstable, and the stability of the complete basis is achieved with a number of iterations $B \geq 140$. We thus chose this threshold as the value for B to be used in BVDim.

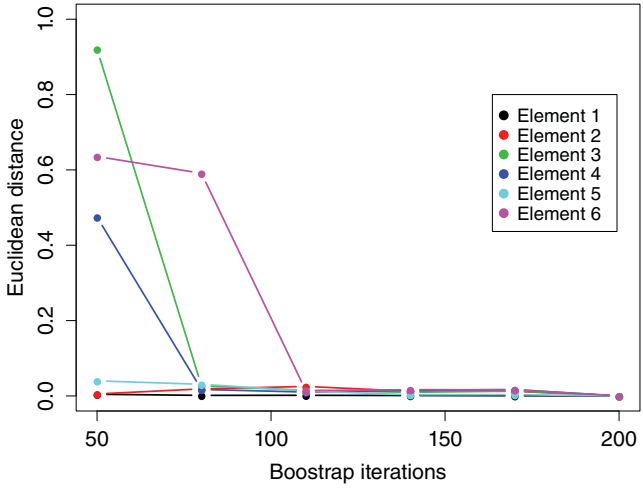


Figure 9.6 Results of BVDim on the Telecom data. Euclidean distance from the reference basis Ψ_{ref} for the first six aligned basis elements, varying the number of bootstrap iterations $B \in \mathcal{B}$.

9.5.1.2 Results

In this section, we report results obtained performing BVDim as detailed in Section 9.5, run using $B = 140$ bootstrap iterations with Voronoi tessellations of dimension $n = 850$. The strength of our analysis stands in finding the most exhaustive, and at the same time synthetic, description of the signal components, and of their action both in time and space. We discuss here only the first six time-varying functions and the coupled surfaces, focusing on their interpretation in the framework of our analysis of mobile phones data:

- time function ψ_1: *general mobile phone activity*;
- time function ψ_2: *working/nonworking hours*;
- time function ψ_3: *after-work activity*;
- time function ψ_4 and ψ_5: *rush-hour*;
- time function ψ_6: *commuting vs. long-distance traveling dynamics.*

The time profiles are shown in Figure 9.7, reported over a week period starting from Wednesday 00:00 and ending with Tuesday 24:00, with days separated by full vertical lines. The associated surfaces are represented in Figure 9.8, with the map of the metropolitan area of Milan in the background, for a better geographical understanding of the spatial dynamics. In the maps, a value close to 0 in a particular area means that the corresponding basis function does not give a relevantly contribution to the mobile phone activity (Erlang measurement) in that site. The 0-levels contour lines are traced in bold.

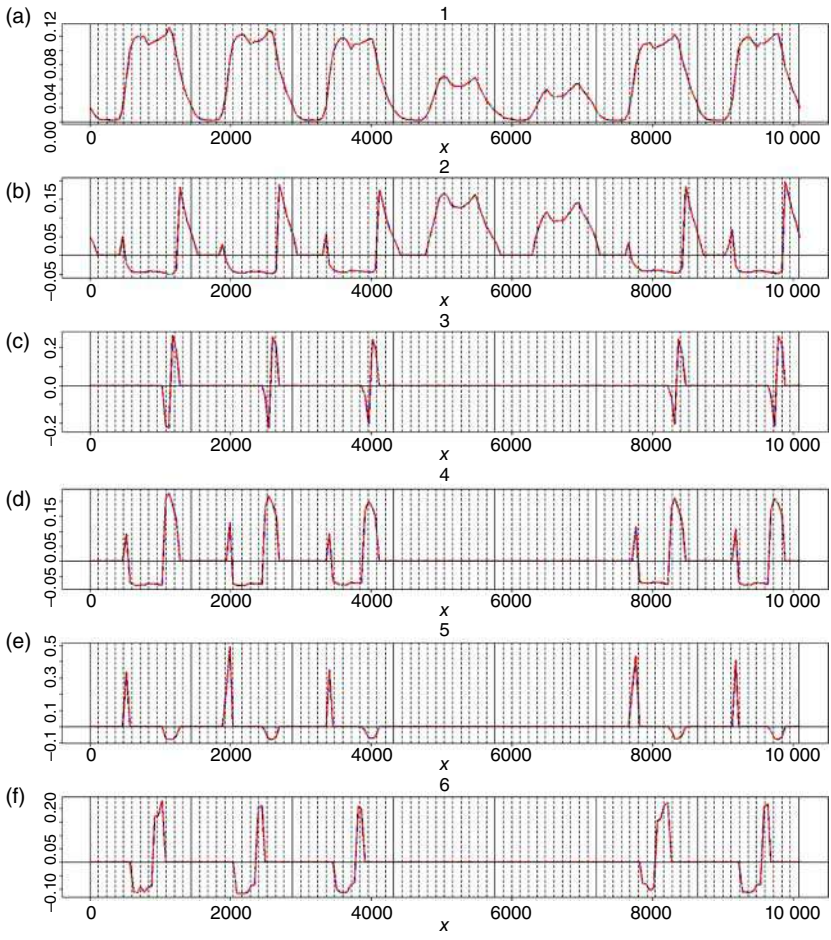


Figure 9.7 Results of BVDim on the Telecom data: the first six elements of the reference basis (from (a) to (f), ψ_1 to ψ_6), in decreasing order based on relative variance. Continuous vertical lines separate the weekdays (WED-THU-FRI-SAT-SUN-MON-TUE), while dotted lines are drawn every two hours starting from midnight.

- ψ_1 - **General mobile phone activity**. The fist time profile ψ_1, given in Figure 9.7a, is associated with the surface $d_1(\mathbf{x})$ (Figure 9.8a), which catches the urbanization of the area and can be related to the average population density, distinguishing day-time low-density population areas and high-density population areas. It can, therefore, be considered an indicator of the general mobile phone activity. This profile brings the largest contribution to the Erlang signal (values as large as 120), and it is clearly identified as the most relevant one even with a low number of bootstrap iterations, indicating that it may

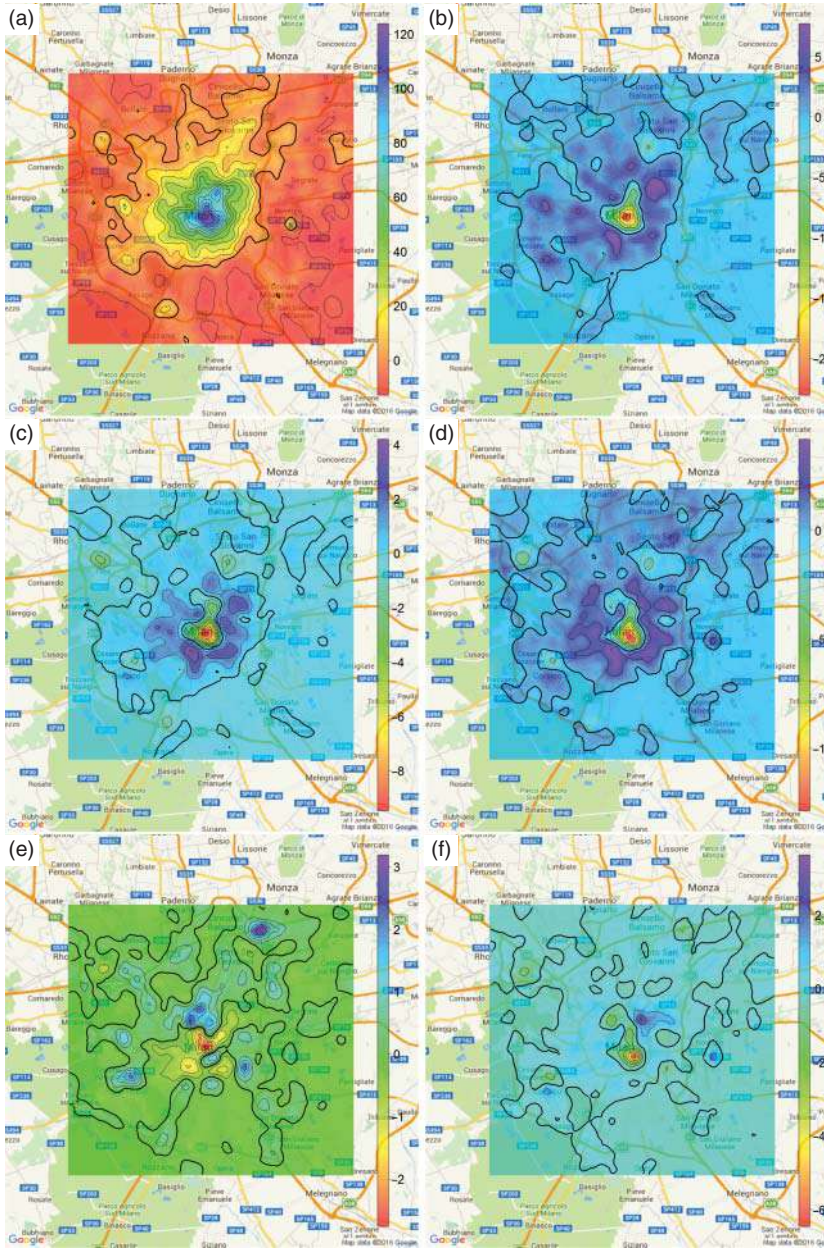


Figure 9.8 Results of BVDim on the Telecom data: maps of the estimated surfaces $\{d_1(\mathbf{x}), d_2(\mathbf{x}), d_3(\mathbf{x}), d_4(\mathbf{x}), d_5(\mathbf{x}), d_6(\mathbf{x})\}$, in lexicographic order, coupled to the elements in Figure 9.7.

indeed be detected through much simpler analyses. The time function describes daily and weekly periodicity, with higher peaks during the day (vs. night) and in working days (vs. weekends) and, considering the associated map, it confirms the city of Milan to be an attractor during the daytime of working days.

- ψ_2 - **Working/nonworking time**. This basis element (Figure 9.7b) contrasts the working time (negative, from 08:00 to 20:00) and nonworking time (positive peaks in early morning, evening, and weekends). In the associated surface, given in Figure 9.8b, we can observe negative values (indicating high activity during working hours) in the city center with an extension northeast toward the Central Railway Station. Indeed, these areas are mostly devoted to tertiary activities and are contrasted with a wider positive-valued (residential and recreational) area around the center. The map clearly points out that the municipality Milan is an attractor for work-related reasons, while the satellite towns as attractors during nonworking hours, due to the flows of commuters.
- ψ_3 - **After-work activity**. The third time profile ψ_3 (Figure 9.7c) focuses on late afternoon and evening, opposing the time slot 5:00–7:00 p.m. (negative) to the slot 7:00–9:00 p.m. (positive), and can be therefore related to the after-work activities. The associated surface (Figure 9.8c) shows a negative concentration in the historical city center and in the area around the central station, while the positive peaks are in the residential areas of the city, especially in the eastern part. While the previous maps concern the general trends during nonworking hours, and it thus takes into account dynamics involving the suburbs, this map is more focused on the city itself, and it considers more specific features: sites with positive value seem to be the residential and leisure areas of the city where people spend their evening, while negative ones might be those where people commonly go right after work, remaining close to the working site for the late afternoon shopping and/or relax, or heading to the train station to go back home.
- ψ_4 **and** ψ_5 - **Rush-hour**. We describe here two different functions pointing out the same phenomenon from two different points of view. Indeed, both ψ_4 and ψ_5 show peaks between 8:00 and 10:00 a.m. and between 5:00 and 9:00 p.m., which correspond to the morning and evening rush hours. However, while the fourth profile (Figure 9.7d) is positive in both time slots and negative in the central hours of the day, ψ_5 (Figure 9.7e) contrasts the morning rush hours (positive) with the evening rush hours (negative). The corresponding surfaces, $d_4(\mathbf{x})$ and $d_5(\mathbf{x})$ (Figure 9.8d,e, respectively) accordingly show two different spatial distributions. The former shows that areas particularly active during the rush hours are concentrated along the city external ring-roads, the roads connecting satellite towns to Milan, at the central station, and in Linate Airport (eastern blue spot). The latter surface shows instead a nonuniform pattern connected to differences between the morning and evening activities: there are positive spots (morning) in the northwest part of the city (Central Railway Station, Garibaldi

Railway Station, new financial district) and in several crossroads in the road network surrounding Milan, while the negative areas (evening) are mainly in the center and in the southeastern part of the city.

- ψ_6 - **Commuting vs. long-distance traveling dynamics**. The last function focuses on the central hours of the day, contrasting the activities carried out from 10:00 a.m. to 3:00 p.m. (negative) and those between 3:00 and 6:00 p.m. (positive), highlighting some particular daytime dynamics. The map clearly spots the Central Railway Station and Linate Airport as positive, while the negative areas correspond to the historical center, the southeast part of the city (within the ring-road), and the new financial district, including Garibaldi Railway Station. Hence, this function seems to suggest contrast between some kind of travel within or nearby the city in the first part of the day (most of the trains in Garibaldi Station are local trains), and long-distance (Linate Airport) or regional trips (Central Station) in the second part of the day.

9.6 Bagging Voronoi Regression (BVReg)

When dealing with the problem of modeling the relationship between a response and a set of explanatory variables, linear regression represents the simplest and most common approach. The matter complicates when the response is a spatially dependent functional variable, and the aim is to find its relationship with a set of p spatial covariates, summing up relevant characteristics of the area under study. Note that, if we forget the spatial component, the sitewise version of model (9.8) translates into a functional linear model with scalar predictor [29]. Aim of the present section is to show how the Bagging Voronoi strategy described in Section 9.3 can be tailored to handle functional regression with spatial dependence, originating the BVReg procedure. A more detailed explanation of the strategy and results here reported can be found in [36] under the name of *Bagging Voronoi Lasso Regression* (BVLR).

Precisely, we need to modify both the Bootstrap and the Aggregation Phases, in the following directions:

(1) estimating a functional linear model at each bootstrap repetition, where each functional local representative has to be related to the spatial covariate values, when referred to the same Voronoi element;
(2) combining the model coefficients obtained along bootstrap repetitions into a final regression model.

Point (1) deserves some clarifications: first, given the Voronoi element V_i^b of the tessellation, we have to find the local representatives of the spatial covariates $\{Z_1(\mathbf{x}), \dots, Z_p(\mathbf{x})\}$ by computing an average (with a Gaussian isotropic kernel) of the covariate values $Z_j(\mathbf{x})$ on $\mathbf{x} \in V_i^b$, for $j = 1, \dots, p$ and $i = 1, \dots, n$. This

leaves us with the set of covariate local representatives $\{\mathbf{z}_i^b\}_{i=1}^n$ at the b-th bootstrap repetition. Note that the vector $\mathbf{z}_i^b \in \mathbb{R}^p$ is the local representative of all covariates on the i-th Voronoi element. Then, we estimate the functional linear model $g^b(t) = \mathbf{Z}^b \beta^b(t)$, where $\mathbf{Z}^b$ is a $n \times (p+1)$ design matrix combining the set of covariate local representatives and including the intercept. Since the spatial covariates are fixed, point (2) is straightforward: no matching is needed, and the set of coefficients obtained along bootstrap repetitions $\{\beta^1(t), \dots, \beta^B(t)\}$ can be combined by simply taking an average.

9.6.1 Covariate Information: The DUSAF Data

We consider a set of covariates $\{Z_1(\mathbf{x}), \dots, Z_p(\mathbf{x})\}$, including information about soil use and land cover, available for the region of Milan (Lombardia) under the name of DUSAF data: indeed, soil use and its dynamics are a strategic element for urban and territorial planning, allowing to understand the current state of territory as the result of past modifications, and to monitor its changes and future opportunities.

DUSAF is a geographical data bank born in 2001–2002, promoted by different regional institutions focused on territory, agriculture, environment, and urban planning. It is made up of categories describing the use of the area, georeferenced via shapefiles, and each category is divided in five hierarchical levels. The first level, giving a first rough classification of the land use, includes five general land cover categories: "Anthropic areas," "Agricultural areas," "Woods and seminatural areas," "Humid areas," and "Water bodies." Each of these categories is then detailed in the levels from the second to the fifth. The resulting structure is very complex, and it includes some very specific local categories (e.g. cemeteries, rice fields, olive cultivations). We thus first need to select only the categories that could be relevant for our analysis based on mobile phone data. The selection procedure was performed manually and resulted in $p = 18$ DUSAF surface covariates, which are reported in the first column of Table 9.1. In Table 9.1, the five macro categories of level 1 are highlighted in red and, when the selection is made at a deeper level, level 2 classes are pointed out in bold. In particular, among the five macro categories, we decided to analyze in depth only "Anthropic areas," while we stopped at level 1 for the remaining four: this is motivated by the assumption that any further specification for this kind of low population density classes would not have a relevant influence in explaining mobile phone use dynamics in the metropolitan area we are examining. Note that the levels of the selected categories can be deduced also from the digits included in the label, reported in the third column of Table 9.1. Finally, the last column of Table 9.1 shows the variable number, which represents the covariates order in the analysis reported in Section 9.6.2 (alphabetical order according to the new reference names in the first column). Four of the

Table 9.1 DUSAF data: soil use categories and corresponding explanations

Selected	Explanation	Label	Variable No.
	Anthropic areas (1)		
	Urban areas (11)		
urban_cont	Continuous urban tissue	111	17
urban_disc	Sparse urban tissue	112	18
	Industrial plants and communication networks (12)		
	Production areas, public and private services (121)		
	Industrial, commercial, agricultural settlements (1211)		
prod_indcomm	Industrial and commercial settlements	12111	9
prod_agri	Agricultural plants	12112	8
	Public and private services (1212)		
serv_osp	Hospitals	12121	12
serv_gen	Public and private services settlements	12122	11
serv_tech	Technological plants	12123	13
	Road and rails networks (122)		
road	Road networks	1221	10
train	Rail networks	1222	15
aero	Airports and heliports (124)	124	1
	Dumps, abandoned lands, construction sites (13)		
dumps	Dumps and abandoned lands	131,132,134	5
constr	Construction sites	133	4
	Nonagricultural green areas (14)		
green	Urban green areas (141)	141	6
sport	Sports and recreational areas (142)	1421	14
	Agricultural areas (2)		
agri	Agricultural areas	2	2
	Woods and seminatural areas (3)		
bosc	Woods and seminatural areas	3	3
	Humid areas (4)		
umid	Humid areas	4	16
	Water bodies (5)		
idro	Water bodies	5	7

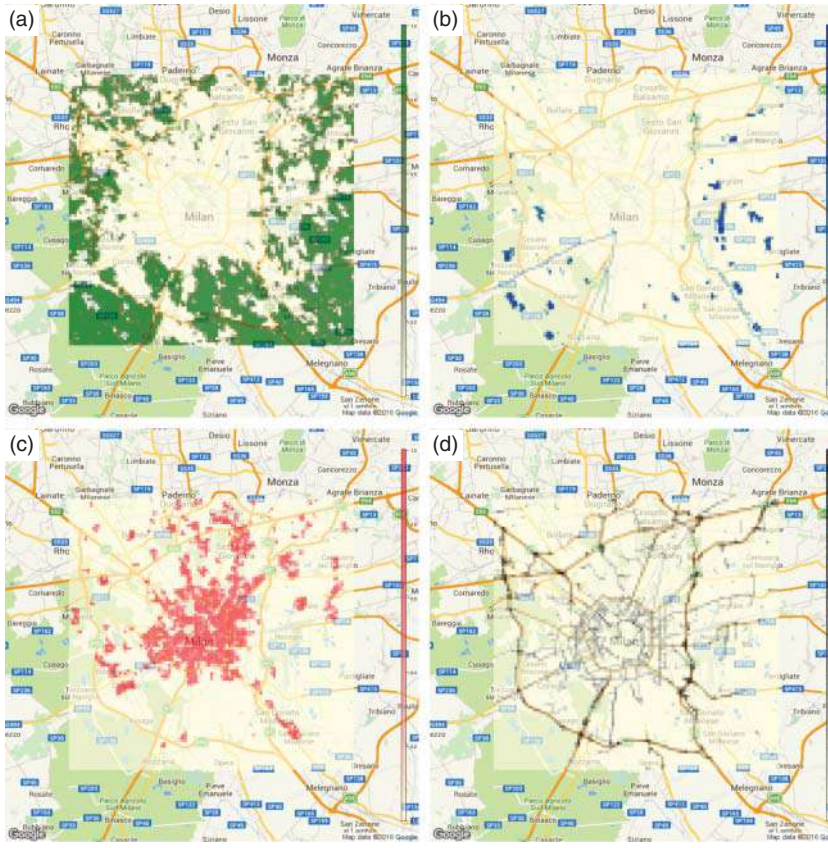


Figure 9.9 Four of the $p = 18$ selected DUSAF covariates superimposed to the metropolitan area of Milan: agricultural areas – *agri* (a); water bodies – *idro* (b); continuous urban tissue – *urban_cont* (c); road networks – *road* (d).

p selected covariates, superimposed on the metropolitan area of Milan, are shown in Figure 9.9. For more details on the preprocessing of covariate data, see [36].

9.6.2 BVReg of the Telecom Data

9.6.2.1 Setting the BVReg Parameters

Also in the case of BVReg, before performing the actual analysis, we need to specify some further choices.

One of the possible approaches to estimating a functional linear model is via basis function expansion [29]. Somehow mimicking this approach, we exploit the BVDim strategy described in Section 9.5: instead of directly considering the Erlang data $E_{\mathbf{x}}(t)$, we consider as response a set of selected spatial surfaces

obtained after the Aggregation Phase of BVDim, and in particular, the first four treelet maps $\left\{d^1_{\mathbf{x}}, d^2_{\mathbf{x}}, d^3_{\mathbf{x}}, d^4_{\mathbf{x}}\right\}, \mathbf{x} \in S_0$, as shown in Figure 9.8. This choice allows us to estimate a multivariate regression, instead of a functional linear model, at each bootstrap repetition, leaving us with a set of multivariate coefficients at the end of the Bootstrap Phase.

Another important aspect concerns the choice of the most suited linear model: since we lack any prior information about the relevance of the many regressors, we also aim at investigating which variables, i.e. which cover land types, are actually related to our response surfaces, thus avoiding overfitting. Hence, the proposed BVReg approach for the Telecom data is in fact a spatial version of lasso regression, a regularization approach to variable selection [37]. This implies that the set of multivariate coefficients obtained at the end of the Bootstrap Phase are also dependent on the choice of λ, the lasso regularization parameter: assuming to explore L different values of the regularization parameter $\{\lambda_1, \dots, \lambda_L\}$, we obtain the coarse estimates $\{\mathcal{B}^b_l\}^L_{l=1} = \{\boldsymbol{\beta}^{(1),b}_l, \dots, \boldsymbol{\beta}^{(4),b}_l\}^L_{l=1}$ at the b-th bootstrap repetition, where $\boldsymbol{\beta}^{(k),b}_l \in \mathbb{R}^p$, for $k = 1, \dots, 4$, is the estimated coefficient vector of the lasso regression with $\lambda = \lambda_l$ for predicting $d^k_{\mathbf{x}}$ given the covariates $\mathbf{Z}^b$.

9.6.2.2 Results

For the present application, we chose $L = 20$ values for the lasso regularization parameter, selected uniformly in log-scale in the interval $[-5, -0.5]$. The final choice of the optimal λ^* was obtained by analyzing the 10-fold cross-validation

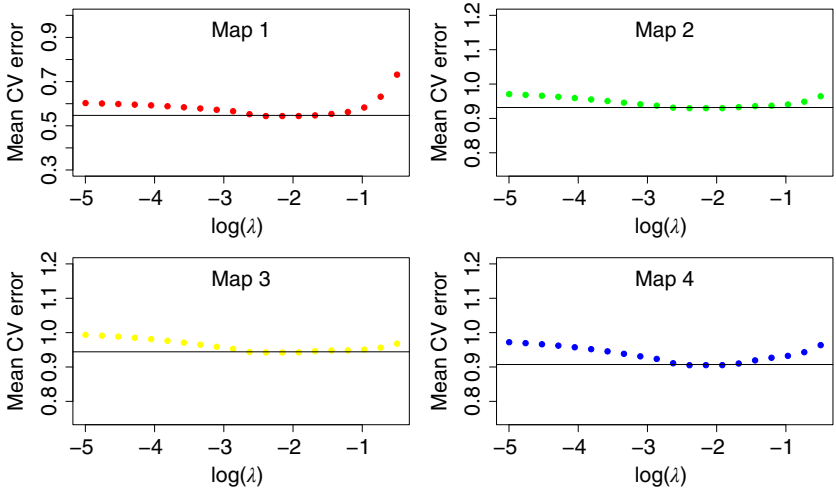


Figure 9.10 Results of BVReg on the Telecom data. Mean cross-validation error (10-fold) associated to the lasso, when varying the regularization parameter $\lambda = \lambda_1, \dots, \lambda_L$ (x-axis).

error of BVReg with lasso regression, as shown in Figure 9.10: in the four panels, we represent the average cross-validation errors associated with the four treelet maps, when varying λ in the x-axis. The error tends to increase when we use a small regularization parameter, meaning that a weak penalization implies overfitting along bootstrap iterations, possibly leading to a nonconsistent aggregation. The model associated with the minimum cross-validation error, with $\lambda^* \cong 0.09$ (i.e. $\log(\lambda^*) \cong -2.4$), is the best compromise between a complete and highly explanatory model capable of accurate prediction, and a parsimonious and easily interpretable model including only important features. The aggregated output $\{\boldsymbol{\beta}^{(1),*}, \dots, \boldsymbol{\beta}^{(4),*}\}$, averaged over $b = 1, \dots, B$ and obtained for $\lambda = \lambda^*$ consists of four coefficient vectors associated with the four treelet maps responses, shown in the four panels of Figure 9.11 ($k = 1, 2, 3, 4$ in lexicographic order): in each panel, intervals are built for each coefficient i with semiwidths corresponding to one bootstrap standard deviation around the coefficient mean, in order to assess the variability across bootstrap replicates. The striped bars are those of coefficients whose associated interval does not include zero, meaning that the corresponding covariate resulted in stable results (similar estimates) across bootstrap repetitions, and it can thus undoubtedly be considered relevant in the model.

The interpretation of the coefficients shown in Figure 9.11 confirms our qualitative interpretation given to the first four estimated treelet surfaces in Section 9.5.1.2: the first surface was related to the general mobile phone activity, and indeed, coefficients showing large positive values are those of covariates concentrated in the urban areas (including road networks and services settlements), opposed to low-density population areas (agricultural and natural areas, industrial sites). The second and third surfaces show nearly the same selected relevant coefficients, related to residential or recreational areas (the presence of roads could be related to travel between home and work), in contrast with areas devoted to work activities, especially services settlements and industrial production sites. Finally, the fourth surface, according to our interpretation, emphasizes phenomena associated with rush hours, and indeed, the importance of road networks is clearly identified, as well as an opposition between residential areas (continuous and sparse urban tissue) and tertiary-devoted areas (services). All in all, we can conclude that the BVReg results describing the dependence of treelet-associated surfaces on land cover use allowed us to validate the heuristic interpretation of the treelet surfaces, given in Section 9.5.1.2.

9.7 Conclusions and Discussion

In this chapter, we presented a unified view on Bagging Voronoi, a nonparametric approach to the treatment of spatial dependence when dealing with functional

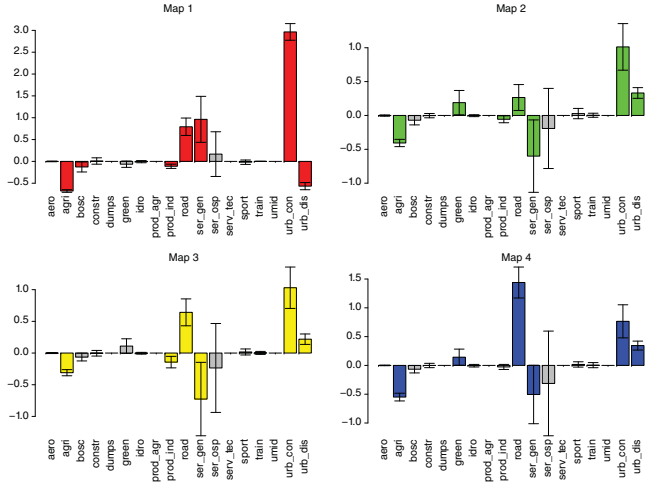


Figure 9.11 Results of the BVReg lasso regression on the first four estimated surfaces $\{d_1(\mathbf{x}), d_2(\mathbf{x}), d_3(\mathbf{x}), d_4(\mathbf{x})\}$ shown in Figure 9.8: barplots of the aggregated lasso coefficients, with whiskers showing $\pm$ a bootstrap standard deviation around each coefficient value.

data indexed on the sites of a lattice. The methodology was first introduced in [22] for classification purposes, and then extended in [20] to dimensional reduction and in [36] to regression. The present chapter introduces the main concepts in a pretty simple and schematic way, and then specifies it further to the different purposes (classification, dimensional reduction, and regression), thus showing the overall flexibility of the approach. We also show the potentialities of the method on a case study concerning the mobile phone use in the metropolitan area of Milan, which guides us in the method explanation and interpretation along the whole chapter. Both the clustering and the dimensional reduction approaches lead to interesting results, whose interpretation was validated thanks to the use of covariate data related to soil use and land cover.

Indeed, the Bagging Voronoi strategy can be investigated for further developments: the overall structure of the method (alignment and aggregation procedure), the treatment of spatial dependence (Voronoi tessellation, chosen distances, and Gaussian kernels), and the dimensional reduction technique (orthogonality constraints, strategy for the selection of a number of relevant basis elements, cross-validation to assess the performance) could require further thinking. A first step in this direction was introduced in [38], where the Bagging Voronoi strategy was combined to a joint clustering and alignment procedure [39], to allow for joint clustering and alignment of spatially dependent functional data. Other statistical methods for the exploration of spatial dependence, such as Local Indicators of Spatial Association (LISA) [40], could also be investigated.

From the point of view of the considered case study, we believe that the proposed analysis can find widespread applications in several fields related to urban planning. This work provides a structured methodology for the understanding and explanation of population dynamics in urban areas based on mobile phone data, and it could be extended in several directions, possibly representing a strategic tool to boost the trends associated to the development of smart cities. Furthermore, we can also think of other practical situations where our nonparametric approach to the treatment of functional data indexed by a lattice might be more natural than other model-based methods, e.g. geostatistical space–time models: for example, when the lattice is defined on a non-Euclidean space, e.g. a manifold, our approach is easily adaptable via a different choice of the distance $d(\cdot, \cdot)$, while model-based approaches might be harder to adapt. Indeed, this is not a rare situation: we already successfully used the Bagging Voronoi strategy on satellite data indexed by sites on a nonhomgeneous lattice covering the Earth [22], and to this scope, we defined $d(\cdot, \cdot)$ to be the geodesic, but, of course, the same could be done for other kinds of non-Euclidean data such as functional magnetic resonance images (fMRIs).

References

1 Menafoglio, A. and Secchi, P. (2017). Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. *European Journal of Operational Research* 258 (2): 401–410.

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

3 Laksaci, A. and Maref, F. (2009). Estimation non paramétrique de quantiles conditionnels pour des variables fonctionnelles spatialement dépendantes. *Comptes Rendus Mathematique* 347 (17–18): 1075–1080.

4 Dabo-Niang, S., Yao, A.F., Pischedda, L. et al. (2010). Spatial mode estimation for functional random fields with application to bioturbation problem. *Stochastic Environmental Research and Risk Assessment* 24 (4): 487–497.

5 Attouch, M.K., Gheriballah, A., and Laksaci, A. (2011). Robust nonparametric estimation for functional spatial regression. In: (ed. F. Ferraty) *Recent Advances in Functional Data Analysis and Related Topics*, 27–31. Springer.

6 Dabo-Niang, S., Kaid, Z., and Laksaci, A. (2012). On spatial conditional mode estimation for a functional regressor. *Statistics & Probability Letters* 82 (7): 1413–1421.

7 Dabo-Niang, S. and Yao, A.F. (2013). Kernel spatial density estimation in infinite dimension space. *Metrika* 76 (1): 19–52.

8 Dabo-Niang, S., Hamdad, L., Ternynck, C., and Yao, A.F. (2014). A kernel spatial density estimation allowing for the analysis of spatial clustering. Application to Monsoon Asia Drought Atlas Data. *Stochastic Environmental Research and Risk Assessment* 28 (8): 2075–2099.

9 Ruiz-Medina, M.D. (2011). Spatial autoregressive and moving average Hilbertian processes. *Journal of Multivariate Analysis* 102 (2): 292–305.

10 Ruiz-Medina, M. (2012). Spatial functional prediction from spatial autoregressive Hilbertian processes. *Environmetrics* 23 (1): 119–128.

11 Ternynck, C. (2014). Spatial regression estimation for functional data with spatial dependency. *Journal de la Société Française de Statistique* 155 (2): 138–160.

12 Wang, H., Wang, J., and Huang, B. (2012). Prediction for spatio-temporal models with autoregression in errors. *Journal of Nonparametric Statistics* 24 (1): 217–244.

13 De Iaco, S., Palma, M., and Posa, D. (2005). Modeling and prediction of multivariate space–time random fields. *Computational Statistics and Data Analysis* 48 (3): 525–547.

14 Giraldo, R., Delicado, P., and Mateu, J. (2012). Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica* 66 (4): 403–421.

15 Romano, E., Mateu, J., and Giraldo, R. (2015). On the performance of two clustering methods for spatial functional data. *AStA Advances in Statistical Analysis* 99 (4): 467–492.

16 Romano, E., Balzanella, A., and Verde, R. (2017). Spatial variability clustering for spatially dependent functional data. *Statistics and Computing* 27 (3): 645–658.

17 Ruiz-Medina, M.D., Espejo, R.M., and Romano, E. (2014). Spatial functional normal mixed effect approach for curve classification. *Advances in Data Analysis and Classification* 8 (3): 257–285.

18 Breiman, L. (1996). Bagging predictors. *Machine Learning* 24: 123–140.

19 R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/. ISBN:3-900051-07-0.

20 Secchi, P., Vantini, S., and Vitelli, V. (2015). Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. *Statistical Methods and Applications* 24 (2): 279–300.

21 Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications. *Journal of the American Statistical Association* 96: 1272–1298.

22 Secchi, P., Vantini, S., and Vitelli, V. (2013). Bagging Voronoi classifiers for clustering spatial functional data. *International Journal of Applied Earth Observation and Geoinformation* 22: 53–64.

23 Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, *Monographs on Statistics and Applied Probability*. Chapman & Hall.

24 Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological)* 48 (3): 259–302.

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

26 Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer.

27 Tarpey, T. and Kinateder, K.K.J. (2003). Clustering functional data. *Journal of Classification* 20: 93–114.

28 Rao, C.R. (1951). An asymptotic expansion of the distribution of Wilks' criterion. *Bulletin of the International Statistical Institute* 33 (2): 177–180.

29 Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*. Springer.

30 Lee, A.B., Nadler, B., and Wasserman, L. (2008). Treelets – an adaptive multi-scale basis for sparse unordered data. *The Annals of Applied Statistics* 2 (2): 435–471.

31 Sangalli, L.M., Secchi, P., Vantini, S., and Vitelli, V. (2010). Functional clustering and alignment methods with applications. *Communications in Applied and Industrial Mathematics* 1 (1): 205–224.

32 Ramsay, J.O. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60: 351–363.

33 James, G.M. (2007). Curve alignment by moments. *The Annals of Applied Statistics* 1: 480–501.

34 Kaziska, D. and Srivastava, A. (2007). Gait-based human recognition by classification of cyclostationary processes on nonlinear shape manifolds. *Journal of the American Statistical Association* 102: 1114–1128.

35 Sangalli, L.M., Secchi, P., Vantini, S., and Veneziani, A. (2009). A case study in exploratory functional data analysis: geometrical features of the internal carotid artery. *Journal of the American Statistical Association* 104: 37–48.

36 Passamonti, F. (2016). Spatio-temporal mobile phone data in Milan: Bagging-Voronoi exploration and modeling through soil use and land cover data. Master thesis. Politecnico di Milano.

37 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58: 267–288.

38 Abramowicz, K., Arnqvist, P., Secchi, P. et al. (2016). Clustering misaligned dependent curves applied to varved lake sediment for climate reconstruction. *Stochastic Environmental Research and Risk Assessment* 1 (31): 71–85.

39 Sangalli, L.M., Secchi, P., Vantini, S., and Vitelli, V. (2010). K-mean alignment for curve clustering. *Computational Statistics and Data Analysis* 54: 1219–1233.

40 Anselin, L. (1995). Local indicators of spatial association – lisa. *Geographical Analysis* 27: 93–115.

10

Nonparametric Inference for Spatiotemporal Data Based on Local Null Hypothesis Testing for Functional Data

Alessia Pini[1] *and Simone Vantini*[2]

[1] *Department of Statistical Sciences, Università Cattolica del Sacro Cuore, largo A. Gemelli 1, Milan, 20123, Italy*
[2] *MOX – Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, 20133, Italy*

10.1 Introduction

Spatiotemporal data are treated by the current literature of functional data analysis in two different ways. They can indeed be modeled as a collection of time-varying data possibly showing spatial dependence, or alternatively, a collection of space-varying data possibly showing temporal dependence. In this chapter, we discuss a general inferential technique for functional data that can be applied to both cases, providing different local information depending on the chosen representation of the functional data. For a complete overview of functional inference methods for spatiotemporal data see Chapters 14, 15, and 16 and references therein.

Statistical inference for functional data is tackled by the literature from two different perspectives: parametric inference relying on distributional assumptions on functional data (e.g. Gaussianity) and/or on asymptotic results (e.g. [1–6]), and nonparametric inference relying instead on very computational intensive techniques based on permutations or bootstrapping (e.g. [7–14]). Bootstrap techniques are only asymptotically valid, i.e. exact when the sample size goes to infinity. Conversely, at the expense of minimal assumptions on the distribution of the data, the permutation-based approach can generate exact statistical inference even for small sample sizes (e.g. [15]).

Although being based on very different modeling assumptions, all the mentioned techniques can be classified as *global* inferential techniques: they look at functions as data points and share the common aim of performing a global test focusing on the distribution of the functional data considered as random objects in

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

a functional space. Consistently, the conclusion of these statistical tests is a unique *p*-value that is used to decide whether the null hypothesis should be accepted or rejected over the whole domain.

While being mathematically appealing, this global approach to inference may not be not fully satisfactory in many functional data applications: whenever a functional null hypothesis is rejected, scientists are often interested in also imputing such a rejection to specific parts of the domain of the functional data. For instance, whenever there is strong statistical evidence to reject a null hypothesis of equality in distribution between two functional populations, one is often interested in selecting the specific parts of the domain where the significant differences are observed. This is why, in the recent literature, there has been growing interest in *local* inferential techniques, i.e. techniques that aim at testing the null hypothesis locally, and that provide a set of local *p*-values associated to specific points of the domain. In the case of spatiotemporal data, a local *p*-value can be associated to specific time instants or space locations, depending on the chosen representation of the data.

A first approach to local inference for functional data was basically based on finite dimensional approximations of the problem above. In detail, the inferential procedure proposed by Vsevolozhskaya et al. [16] is based on a discretization of the domain. In that work, a finite partition of the domain in subintervals is a priori defined, and a global test is performed on each subinterval. Since several tests are jointly performed on the same data set, the results of the tests are finally adjusted for multiplicity by means of a closed testing procedure [17] on the subintervals. Thanks to the multiplicity correction, this procedure controls the probability of wrongly selecting any set of subintervals. The disadvantage of using such a procedure is that the conclusions of the test and the achieved control depend on the preliminary-chosen partition of the domain, and that the control of type-I error probability is lost within each subinterval, as discussed by Pini and Vantini [18]. Pini and Vantini [18] proposed instead a procedure for locally testing functional data based on a finite dimensional approximation of the functional data by means of a B-spline basis expansion. Functional data are described by means of the coefficients of an a priori-defined basis expansion. A statistical test is performed on each coefficient of the basis expansion, and the results of the tests are finally adjusted for multiplicity. The disadvantage in this case is that conclusions depend on the preliminary-chosen basis expansion.

The extension of the latter techniques to the truly infinite-dimensional case, avoiding both discretization of the domain and final dimensional approximation the functional data poses at least two major challenges. The first issue is that pointwise *p*-values are in general not trivially defined in functional spaces. When functional data are, for instance, embedded in the L^2 space (being such space the natural extension of the Euclidean geometry to the functional data analysis

framework), pointwise evaluations of functions are meaningless. The second issue is that any possible multiplicity correction would involve a continuous infinity of univariate tests which Bonferroni–Holm, Benjamini–Hochberg, and closed testing procedures are not able to deal with it. The challenge in this framework is thus twofold: (i) defining a continuous infinity of tests in bijective correspondence with the points of the domain without relying on pointwise evaluations of functional data, and then (ii) developing a multiplicity correction technique for that continuous infinity of tests.

The first issue has been first addressed by Abramovich and Heller [19], assuming that the functional data follow a Wiener process. The second issue has been instead addressed in a nonparametric framework by Cox and Lee [20] and Vsevolozhskaya et al. [21], assuming instead the continuity of functional data and proposing two different strategies to adjust pointwise p-values. Finally, Pini and Vantini [18] present a domain-selective inferential strategy – i.e. the intervalwise testing (IWT) procedure – that focus on both issues, and that neither requires distributional assumptions on functional data, nor assumes their continuity.

The aim of this chapter is to present the IWT procedure for local inferential analysis of functional data and discuss its inferential properties. As an illustration, we report the application of the IWT on a well-known benchmark functional dataset, i.e. Canadian daily temperatures measured in Canadian weather stations presented in [22].

10.2 Methodology

In this section, the IWT procedure is first described for the case of comparing the means of two functional populations. The second part of the section reports an extension of the described procedure to other more complex null hypothesis testing problems.

10.2.1 Comparing Means of Two Functional Populations

Assume observing two samples of functional data taking values in the $L^2(T)$ space of squared-integrable functions on the domain $T = (a, b) \subset \mathbb{R}$. Let χ_{ji} $i = 1, \dots, n_j$, $j = 1, 2$ denote the i-th function of sample j. Finally, assume $\forall i = 1, \dots, n_j$, $\forall j = 1, 2$ $\chi_{ji} = \mu_j + \varepsilon_{ji}$, where $\mu_j \in L^2(T)$ is the fixed mean function of population j, and ε_{ji} are random i.i.d. error functions. We aim at testing – in a local perspective – the following hypotheses:

$$H_0 : \mu_1 = \mu_2 \text{ against } H_1 : \mu_1 \neq \mu_2, \tag{10.1}$$

where the equality is defined in the usual L^2 sense. Specifically, we aim at identifying a local criterion to pointwise reject (or not to reject) the null hypothesis along

the domain T such as to be able to select the portions of T, where the two means are significantly different. The method that we describe is based on the concepts of unadjusted and adjusted p-value functions.

Unadjusted *p*-value function: A function defining a p-value for a continuous infinity of tests in bijective correspondence with the points of the domain, without relying on pointwise evaluations of functional data which are not properly defined in the L^2 setting.

Adjusted *p*-value function: A function defining a p-value for a continuous infinity of tests in bijective correspondence with the points of the domain adjusted by means of a suitable multiplicity correction technique for that continuous infinity of tests.

Let $\mathcal{I} \subseteq T$ be a generic interval of the form (t_1, t_2) or complementary interval of the form $T\backslash(t_1, t_2)$, where $a \leq t_1 < t_2 \leq b$. For every $\mathcal{I} \subseteq T$, consider the following functional test:

$$H_0^{\mathcal{I}} : \mu_1^{\mathcal{I}} = \mu_2^{\mathcal{I}} \text{against } H_1^{\mathcal{I}} : \mu_1^{\mathcal{I}} \neq \mu_2^{\mathcal{I}}, \tag{10.2}$$

being $\mu_j^{\mathcal{I}}$ the restriction of μ_j over $\mathcal{I}$, $j = 1, 2$. Let $p^{\mathcal{I}}$ be the global p-value of test (10.2). Different methods can be employed to compute $p^{\mathcal{I}}$. We here describe two different approaches, coming from the recent literature of global testing for functional data. Both approaches are based on the same test statistic $T^{\mathcal{I}}$, that is the squared $L^2(\mathcal{I})$ distance between the two restricted sample means are divided by the measure of $\mathcal{I}$:

$$T^{\mathcal{I}}(\chi_{11}, \dots, \chi_{1n_1}, \chi_{21}, \dots, \chi_{2n_2}) = \frac{\|\overline{\chi}_1 - \overline{\chi}_2\|^2_{L^2(\mathcal{I})}}{|\mathcal{I}|} \tag{10.3}$$

$$= \frac{1}{|\mathcal{I}|} \int_{\mathcal{I}} \left(\overline{\chi}_1(t) - \overline{\chi}_2(t)\right)^2 \mathrm{d}t, \tag{10.4}$$

where the integration is intended in the Lebesgue sense, $\overline{\chi}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \chi_{ji}$, $j = 1, 2$, and $|\mathcal{I}|$ is the Lebesgue measure of interval $\mathcal{I}$. The two methods described here for testing (10.2) are the following:

Parametric/asymptotic test: When assuming functional normality of the data (i.e. $\forall g \in L^2(T)$, $\forall j, i$ $\int_T g\chi_{ji}$ is a normal random variable) and the covariance structure is known, it is possible to perform parametric tests on hypotheses (10.2). Horváth and Kokoszka [4] provide – under some regularity conditions – the distribution of the test statistic $T^{\mathcal{I}}$ under the null hypothesis. Specifically, assuming that $\forall i = 1, \dots, n_j$; $j = 1, 2$: $\mathbb{E}[\|\varepsilon_{ij}\|^4] < \infty$, we have

$$H_0^{\mathcal{I}} \;\Rightarrow\; T^{\mathcal{I}} \frac{n_1 n_2}{n_1 + n_2} \sim \sum_{k=1}^{\infty} \tau_k N_k^2,$$

$$H_1^{\mathcal{I}} \;\Rightarrow\; T^{\mathcal{I}} \frac{n_1 n_2}{n_1 + n_2} \xrightarrow[n_1, n_2 \to \infty]{P} \infty.$$

where N_k are independent standard normal random variables and τ_k are the eigenvalues of the operator with kernel $(1-\theta)\Sigma_1(t,s)+\theta\Sigma_2(t,s)$, $\Sigma_j(t,s)$ being the covariance function of functional data χ_{ji}, and $\theta = n_1/(n_1+n_2)$. As suggested by Horváth and Kokoszka [4], such properties can be used to compute the p-value p^I, obtaining an exact and consistent test of hypotheses (10.2). If functional normality cannot be assumed, it is possible to use an analogous asymptotic test. In detail, when the sample sizes tend to infinity, regardless of the distribution of the χ_{ji} with $n_1/(n_1+n_2) \to \theta$ with some $0 \le \theta \le 1$, we have

$$H_0^I \;\Rightarrow\; T^I \frac{n_1 n_2}{n_1+n_2} \xrightarrow[n_1,n_2\to\infty]{d} \sum_{k=1}^{\infty} \tau_k N_k^2.$$

Such a property suggests how to compute – for high sample sizes – the p-value p^I, obtaining an asymptotically exact and consistent test of hypotheses (10.2).

Permutation test: When the sample sizes are low and functional normality cannot be assumed, one can rely on nonparametric permutation methods, which have also the advantage of not assuming any particular structure of the covariance operator. For instance, Hall and Tajvidi [8] propose a permutation test based on the test statistic T^I. Under the null hypothesis, the two random samples have the same distribution and are independent. All functions are therefore exchangeable. Thus, equally likely samples under H_0 are obtained by all possible rearrangement of the values χ_{ji} across the units:

$$(\chi_{11},\dots,\chi_{1n_1},\chi_{21},\dots,\chi_{2n_2}) \mapsto (\chi^*_{11},\dots,\chi^*_{1n_1},\chi^*_{21},\dots,\chi^*_{2n_2}). \tag{10.5}$$

The number of possible rearrangements (or permutations) leading to a different allocation in the two groups is $\binom{n_1+n_2}{n_1}$. The p-value can be computed as the proportion of permuted samples (among the $\binom{n_1+n_2}{n_1}$ possible ones) associated with a value of the test statistic T^I that exceeds the value associated with the original sample, which reads

$$p^I = \frac{\sum_{b=1}^{\binom{n_1+n_2}{n_1}} \mathbb{I}\left[T^I\left(\chi^*_{11_b},\dots,\chi^*_{2n_{2_b}}\right) \ge T^I\left(\chi_{11},\dots,\chi_{2n_2}\right)\right]}{\binom{n_1+n_2}{n_1}} \tag{10.6}$$

with b indexing the $\binom{n_1+n_2}{n_1}$ permuted samples obtained from (10.5). The resulting test is exact for every finite sample size and consistent.

No matter the way they are computed, the p-values p^I of tests (10.2) are used to define the two p-value functions.

Definition 10.1 The unadjusted and adjusted p-value functions are defined as follows:

- The **unadjusted p-value function** $p(t): T \to [0,1]$ is

$$p(t) = \limsup_{\mathcal{I} \to t} p^{\mathcal{I}} \quad \forall t \in T\,, \tag{10.7}$$

  i.e. its value at point $t \in T$ is the superior limit of the p-value $p^{\mathcal{I}}$ as both extremes of $\mathcal{I}$ converge to t.
- The **adjusted p-value function** $\tilde{p}(t)$ is

$$\tilde{p}(t) = \sup_{\mathcal{I} \ni t} p^{\mathcal{I}} \quad \forall t \in T\,, \tag{10.8}$$

  i.e. its value at point t is the supremum over all p-values $p^{\mathcal{I}}$ pertaining to intervals and complementary intervals containing t.

Note that although the pointwise evaluations of the functions χ_{ji} in $L^2(T)$ are generally not properly defined, both the unadjusted and the adjusted p-value functions are instead univocally defined $\forall t \in T$. This directly derives from the boundedness of the p-values $p^{\mathcal{I}}$. Furthermore, in the special case of data embedded in $L^2(T) \cap C^0(T)$, the identity between the superior limit in (10.7) and its corresponding limit, and the integral mean value theorem jointly guarantee that $p(t)$ coincides with the p-value of the permutation test based on the test statistic $T(t) := (\overline{\chi}_1(t) - \overline{\chi}_2(t))^2$, $\forall t \in T$ (i.e. the pointwise evaluations of the integrand in (10.3)). In this case, we also have that both the unadjusted and the adjusted p-value functions are continuous functions of t. Indeed, the pointwise test statistic $T(t)$ is continuous in T so the unadjusted p-value is continuous. In addition, the interval test statistic $T^{\mathcal{I}}$ defined in (10.3) is continuous with respect to both extremes of the interval $\mathcal{I}$, making the adjusted p-value function continuous as well.

The unadjusted p-value function $p(t)$ and the adjusted p-value function $\tilde{p}(t)$ present different inferential properties with respect to both type I error control and consistency. In detail:

- If all the tests used to compute the p-values $p^{\mathcal{I}}$ are exact, the unadjusted p-value function $p(t)$ provides a control of the **pointwise error rate** (see Theorem A.1 of [18]), that is, $\forall \alpha \in (0,1)$:

$$\forall t \in T \text{ s.t. } \exists \mathcal{I} \ni t\colon H_0^{\mathcal{I}} \text{ is true} \Rightarrow \mathbb{P}[p(t) \leq \alpha] \leq \alpha. \tag{10.9}$$

  If all the tests used to compute the p-values $p^{\mathcal{I}}$ are consistent, the unadjusted p-value function $p(t)$ is **pointwise consistent** (see Theorem A.4 of [18]), that is, $\forall \alpha \in (0,1)$:

$$\forall t \in T \text{ s.t. } \nexists \mathcal{I} \ni t\colon H_0^{\mathcal{I}} \text{ is true} \Rightarrow \mathbb{P}\left[p(t) \leq \alpha\right] \xrightarrow[n_1, n_2 \to \infty]{} 1. \tag{10.10}$$

- If all the tests used to compute the p-values $p^{\mathcal{I}}$ are exact, the adjusted p-value function $\tilde{p}(t)$ provides a control of the **intervalwise error rate** (see Theorem A.3 of [18]), that is, $\forall \alpha \in (0,1)$:

$$\forall \mathcal{I} \subseteq T\colon H_0^{\mathcal{I}} \text{ is true} \Rightarrow \mathbb{P}\left[\forall t \in \mathcal{I}, \tilde{p}(t) \leq \alpha\right] \leq \alpha. \tag{10.11}$$

  If all the tests used to compute the p-values $p^{\mathcal{I}}$ are consistent, the adjusted p-value function $\tilde{p}(t)$ is **intervalwise consistent** (see Theorem A.4 of [18]), that is, $\forall \alpha \in (0,1)$:

$$\forall \mathcal{I} \subseteq T \text{ s.t. } \nexists \mathcal{J} \subseteq \mathcal{I}\colon H_0^{\mathcal{J}} \text{ is true} \Rightarrow \mathbb{P}\left[\forall t \in \mathcal{I},\ \tilde{p}(t) \leq \alpha\right] \xrightarrow[n_1, n_2 \to \infty]{} 1 \tag{10.12}$$

The results hold for any sample size n_1 and n_2, given the exactness and consistency of the tests used to compute the p-values $p^{\mathcal{I}}$. Note that – if an asymptotic test is used to compute p-values $p^{\mathcal{I}}$ – the control of the pointwise and intervalwise error rates are only guaranteed asymptotically.

Heuristically speaking, property (10.9) states that when a thresholding of $p(t)$ is performed at level α, for each point of the domain "where H_0 is true," the probability that H_0 is rejected in that point is less or equal to α. Property (10.11) states instead that, when a thresholding of $\tilde{p}(t)$ is performed at level α, for each interval of the domain "where H_0 is true," the probability that H_0 is rejected on the interval is less or equal to α. Thus, if one is interested in controlling the pointwise error rate at level $\alpha \in (0,1)$, the points $t \in T$ such that $p(t) \leq \alpha$ should be selected. Instead, if one is interested in controlling the intervalwise error rate at level α, the points $t \in T$ such that $\tilde{p}(t) \leq \alpha$ should be selected.

In terms of consistency, property (10.10) states that when a thresholding of the $p(t)$ is performed at level α, the consequent domain selection criterion is such that, for each point of the domain "where H_0 is false," the probability of that H_0 is rejected in that point converges to one as the sample size increases. Property (10.12) states instead that, when a thresholding of $\tilde{p}(t)$ is performed at level α, for each interval of the domain "where H_0 is false," the probability that H_0 is rejected on the entire interval converges to one as the sample size increases.

10.2.2 Extensions

The procedure described in Section 10.2.1 can be modified for dealing with more complex functional null hypothesis testing problems. Indeed, the definition of the unadjusted and adjusted p-value functions in Eqs. (10.7) and (10.8) directly descends from the p-values $p^{\mathcal{I}}$. Hence, by suitably changing the procedure used to compute the p-values of tests on intervals, IWT can be used to tackle other null hypothesis significance testing problems. What needs to be defined is a suitable parametric or nonparametric test that can be used to compute the p-values $p^{\mathcal{I}}$. As an example, we describe here the extension to multiway functional analysis of

variance (FANOVA). Extensions to functional-on-scalar linear models and to tests for comparing different distributional properties of the samples (e.g. the variance) can be found, respectively, in [23, 24].

10.2.2.1 Multiway FANOVA

We report here a brief description of the extension of IWT to the case of a two-way FANOVA for testing the effects of two factors A and B on a functional data set. The procedure is described in detail in [25]. The procedure can be straightforwardly extended to the general case of one-way and multiway FANOVA.

Let $\chi_{ijl}:(a,b)\mapsto\mathbb{R}$ be the functional data, where $i=1,\dots,I$ denotes the level of the first factor, $j=1,\dots,J$ denotes the level of the second factor, and $l=1,\dots,n_{ijl}$ denotes the replicate. We assume functional data to be $L^2(T)$ random functions. The FANOVA model that we want to test is the following:

$$\chi_{ijl}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\epsilon_{ijl}, \tag{10.13}$$

with $i=1,\dots,I,\ j=1,\dots,J,\ l=1,\dots,n_{ij},\ n_{ij}$ being the number of replicates for ith level of the first factor and jth level of the second factor. In model (10.13), $\mu\in L^2(T)$ is the functional grand mean, $\alpha_i\in L^2(T)$ and $\beta_j\in L^2(T)$ are functional main effects, and $\gamma_{ij}\in L^2(T)$ is the functional interaction effect. The functional errors ϵ_{ijl} are assumed to be independent and identically distributed zero-mean random functions of $L^2(T)$. For sake of identifiability, we require the classical constraints on the effects, i.e. $\sum_{i=1}^{I} n_{i\cdot}\alpha_i=0$; $\sum_{j=1}^{J} n_{\cdot j}\beta_j=0$; $\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\gamma_{ij}=0$, where $n_{i\cdot}=\sum_{j=1}^{J} n_{ij}$ denotes the number of units at ith level of the first factor and $n_{\cdot j}=\sum_{i=1}^{I} n_{ij}$ denotes the number of units at jth level of the second factor.

The aim of the analysis is to test the significance of all functional coefficients of model (10.13). In particular, we want to perform – in a local perspective – the functional counterparts of three classical ANOVA tests, i.e. three functional tests for the effects of each factor and interaction:

$$H_{0,A}\colon\ \alpha_i=0\ \forall i=1,\dots,I;\quad H_{1,A}:(H_{0,A})^C \tag{10.14}$$

$$H_{0,B}\colon\ \beta_j=0\ \forall j=1,\dots,J;\quad H_{1,B}:(H_{0,B})^C \tag{10.15}$$

$$H_{0,AB}\colon\ \gamma_{ij}=0\ \forall i=1,\dots,I; j=1,\dots,J;\quad H_{1,AB}:(H_{0,AB})^C. \tag{10.16}$$

For every interval $\mathcal{I}$, denote with $H_{0,A}^{\mathcal{I}}$, $H_{1,A}^{\mathcal{I}}$, $H_{0,B}^{\mathcal{I}}$, $H_{1,B}^{\mathcal{I}}$, $H_{0,AB}^{\mathcal{I}}$, and $H_{1,AB}^{\mathcal{I}}$ the restriction of null and alternative hypotheses to $\mathcal{I}$. Pini et al. [25] propose to perform all tests in a permutation framework by applying permutations of the residuals of the reduced model according to the Freedman and Lane permutation scheme [26], and test statistics based on the integral over the interval $\mathcal{I}$ of the two-way ANOVA statistics of the corresponding classical F-tests.

In detail, the proposed test statistics are

$$F_A^{\mathcal{I}} = \int_{\mathcal{I}} \frac{\sum_{i=1}^{I} n_{i\cdot} \left(\overline{\chi}_{i\cdot}(t) - \overline{\chi}(t)\right)^2 / (I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} (\chi_{ijl}(t) - \overline{\chi}_{ij}(t))^2 / df_{Error}} dt$$

$$F_B^{\mathcal{I}} = \int_{\mathcal{I}} \frac{\sum_{j=1}^{J} n_{\cdot j} \left(\overline{\chi}_{\cdot j}(t) - \overline{\chi}(t)\right)^2 / (J-1)}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} (\chi_{ijl} - \overline{\chi}_{ij}(t))^2 / df_{Error}} dt$$

$$F_{AB}^{\mathcal{I}} = \int_{\mathcal{I}} \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \left(\overline{\chi}_{ij}(t) - \overline{\chi}_{i\cdot}(t) - \overline{\chi}_{\cdot j}(t) + \overline{\chi}(t)\right)^2 / (I-1)(J-1)}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} (\chi_{ijl}(t) - \overline{\chi}_{ij}(t))^2 / df_{Error}} dt$$

where, with the common ANOVA notation, $\overline{\chi}_{ij}(t) = \sum_{l=1}^{n_{ij}} \chi_{ijl}(t)/n_{ij}$, $\overline{\chi}(t)$ is the grand mean, $df_{Error} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} - (I-1) - (J-1) - (I-1)(J-1) - 1$, $\overline{\chi}_{i\cdot}(t) = \sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} \chi_{ijl}(t)/(n_{i\cdot})$ and $\overline{\chi}_{\cdot j}(t) = \sum_{i=1}^{I}\sum_{l=1}^{n_{ij}} \chi_{ijl}(t)/(n_{\cdot j})$.

This provides approximated (asymptotically exact) tests for $H_{0,A}^{\mathcal{I}}$, $H_{0,B}^{\mathcal{I}}$, and $H_{0,AB}^{\mathcal{I}}$. In addition, in the particular case of a one-way FANOVA, the corresponding test is shown to be exact. In the following, we denote with $p_A^{\mathcal{I}}$, $p_B^{\mathcal{I}}$, and $p_{AB}^{\mathcal{I}}$ as the p-values of the corresponding tests.

The adjusted p-value functions of each test ((10.14)–(10.16)) are defined as follows:

$$\tilde{p}_A(t) = \sup_{\mathcal{I} \ni t} p_A^{\mathcal{I}}, \quad \tilde{p}_B(t) = \sup_{\mathcal{I} \ni t} p_B^{\mathcal{I}}, \quad \tilde{p}_{AB}(t) = \sup_{\mathcal{I} \ni t} p_{AB}^{\mathcal{I}}.$$

Since the tests based on permutations of the residuals are in this case only asymptotically exact, the adjusted p-value functions $\tilde{p}_A(t)$, $\tilde{p}_B(t)$, and $\tilde{p}_{AB}(t)$ are provided with an asymptotic control of the intervalwise error rate.

10.3 Data Analysis

The aim of this section is to illustrate the potential of the functional IWT approach described in Section 10.2 by applying the procedure to a well-known benchmark functional data set derived from spatiotemporal weather records, i.e. the Canadian daily temperatures data set [22].

The data set contains the daily temperatures along the year (averaged over 30 years) recorded by 35 weather stations in Canada (Figure 10.1a). The weather stations are divided into four climate zones: Atlantic, Pacific, Continental, and Arctic. The locations of the stations and the functional data are reported in Figure 10.1a–c, respectively. Figure 10.1a shows the locations of the weather stations on the map, and Figure 10.1b,c report the functional data and their first derivatives. The four different shades of gray are associated with the different climatic regions. As done by Hall and Tajvidi [8], we test the equality of the mean

temperatures of the four climatic zones. Specifically, we first test differences between the four zones by applying a one-way FANOVA test with one factor with four levels. Then we test differences between each couple of zones in a pairwise perspective. Note that the geographical information about the locations of the 35 weather stations is only exploited to assign them to the different climatic regions, and functional data are assumed to be exchangeable, since a permutation test is employed. Such an assumption is reasonable for the data analyzed here since the climatic zones can capture the spatial information of data, and the functional data are averaged temperatures in 30 years.

Since in this example, sample sizes are quite low (Atlantic region: 15 curves, Pacific region: 5 curves; Continental region: 12 curves; Arctic region: 3 curves), and the functional normality assumption cannot be verified, in order to make exact inference, we apply to each test an IWT procedure based on nonparametric permutation tests. Due to the nonparametric nature of the resulting procedure, it is possible to apply the procedure to different quantities of interest of the data set. In detail, the comparison between the different climatic regions is performed here on the curves and on their first derivatives. Specifically:

- the test on the **curves** provides an information about which climatic regions are characterized by higher/lower mean temperatures in the different periods of the year;
- the test on the **first derivatives** provides an information about which climatic regions are characterized by faster/slower transitions of the mean temperatures in the different periods of the year.

Testing both the curves and the derivatives can provide a highly informative characterization of data. For instance, most models of energy consumption take as inputs the temperature level and the temperature derivative. In particular, the derivatives have a considerable impact on the perceived temperature, and hence, to the heating and cooling energy consumption. In general, low temperatures and negative derivatives are associated with high expenses for heating, while high temperatures and positive derivatives are associated with high expenses for cooling.

In order to obtain accurate estimates of the first derivatives of the functional data, we performed a Fourier smoothing of the raw data based on a reduced number of harmonics as suggested in [22]. In detail, the results shown here were obtained with seven harmonics, and the results of the inferential procedure are robust with respect to this choice. The smoothed temperature curves and their estimated first derivatives are shown in Figure 10.1b,c. In the following, we report the results of the analysis on the curves and on their first derivatives.

The results of the FANOVA tests are summarized in Figure 10.2. The results of the tests on temperature curves are reported in Figure 10.2a,b, and the ones on temperature derivatives are reported in Figure 10.2c,d. For each test, Figure 10.2a,c the

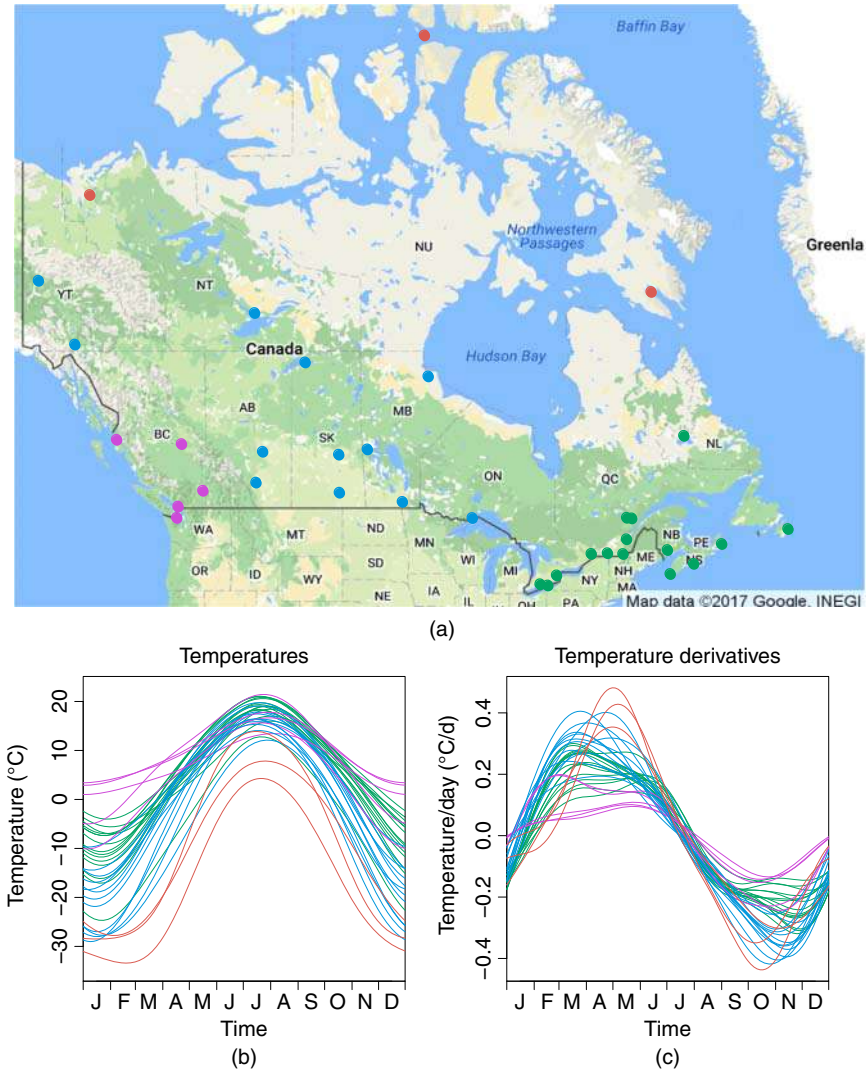


Figure 10.1 (a) Map of Canada with locations of the 35 weather stations. Smoothed functional data (b) and first derivatives (c). Different gray colors correspond to different climatic regions.

sample means of the four groups, and Figure 10.2b,d reports the adjusted (full line) and unadjusted (dashed line) p-value functions. Note that in this example, the difference between these two curves is only perceivable in a very short time interval during summer for the test on first derivatives. The bands in the lower parts of the plots highlight the intervals presenting significant differences at a 5% (light) and

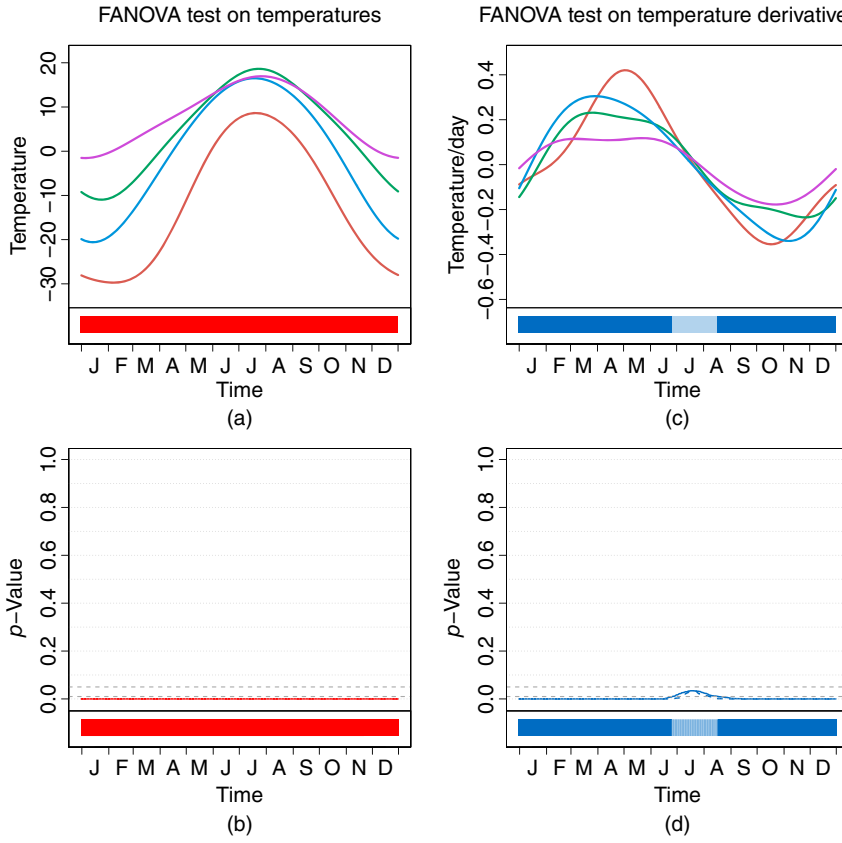


Figure 10.2 FANOVA test on curves (a, b) and first derivatives (c, d) of Canadian temperature data. (a, c) Functional sample means of the four regions. (b, d) Unadjusted (dashed line) and adjusted (full line) p-value functions of the FANOVA test. The bands in the lower parts of the plots highlight the intervals presenting significant differences at a 5% (light) and 1% (dark) significance level.

1% (dark) significance level, i.e. the gray dashed lines reported in Figure 10.2b,d. The FANOVA test highlights significant differences between the four groups along the whole year, both for temperatures and temperature derivatives.

As a comparison, both p-value of the test proposed by Cuevas et al. [1] and the p-value of a global permutation test based on the F-test statistic result in a p-value equal to zero for curves and first derivatives. The difference between the latter two approaches and the IWT is that the IWT provides an adjusted p-value function that is able to locate the areas of the domain imputable for the rejection of the null hypothesis. To better describe the differences between the regions, we then apply the IWT to test differences between each couple of groups.

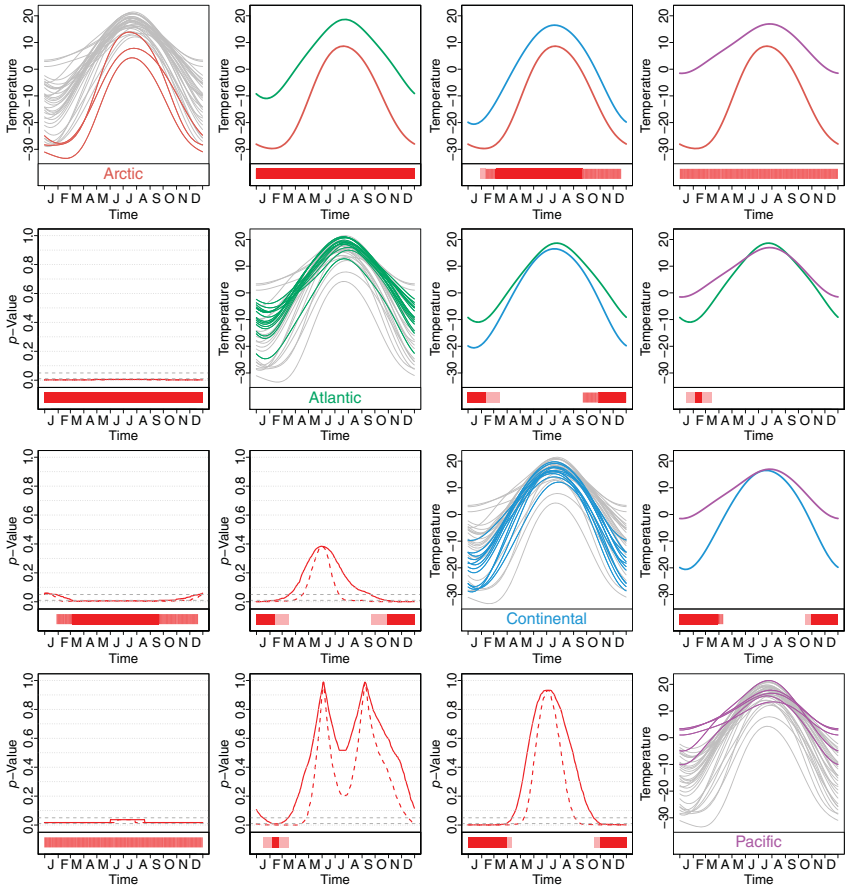


Figure 10.3 Pairwise comparisons between curves. Diagonal panels: all temperature curves (gray) and in each panel curves of each of the five climatic regions (darker gray). Upper-diagonal panels: two functional sample means of each comparison. Lower diagonal panels: unadjusted (dashed line) and adjusted (full line) p-value functions of each comparison. The bands in the lower parts of each extra-diagonal panel highlight the intervals presenting significant differences at a 5% (light) and 1% (dark) significance level.

Figure 10.3 reports the results of the IWT applied to the curves in a pairwise perspective. In detail, the IWT is applied to each comparison between two climatic regions, using a nonparametric permutation test based on the L^2 distance between the two sample means. The diagonal panels of Figure 10.3 show the curves of each of the five climatic regions. For each comparison of a couple of groups identified in the diagonal, the upper-diagonal panels of Figure 10.3 report the two functional sample means. The lower diagonal panels report instead the unadjusted and adjusted p-value functions evaluated according to the procedure described

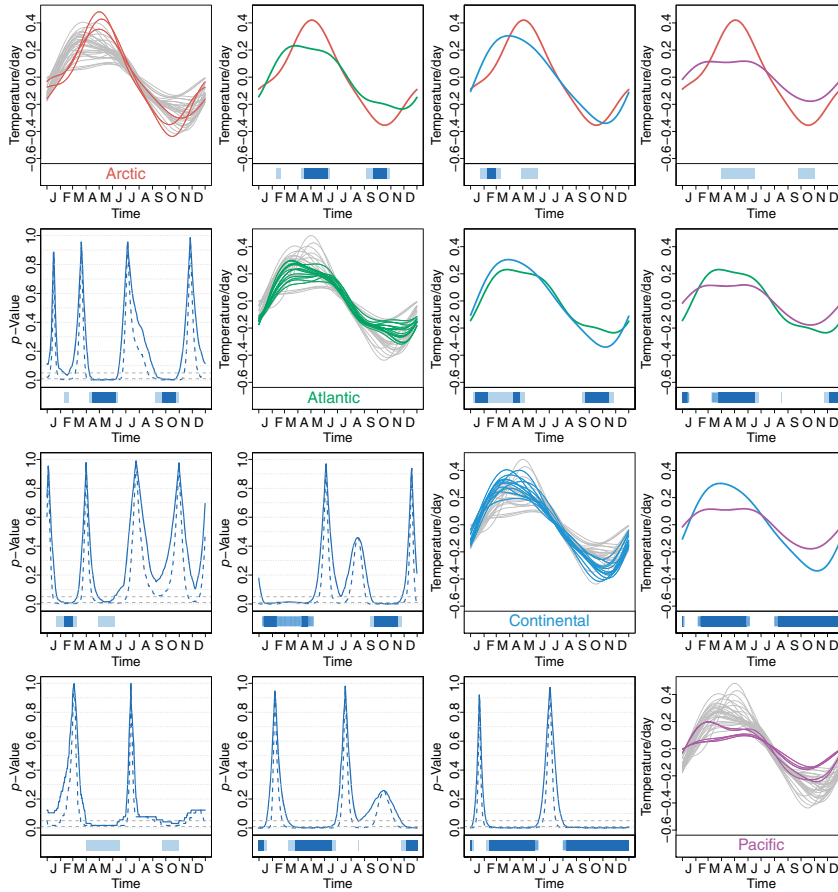


Figure 10.4 Pairwise comparisons between first derivatives. Diagonal panels: all temperature derivatives (gray) and in each panel derivatives of each of the five climatic regions (darker gray). Upper-diagonal panels: two functional sample means of each comparison. Lower diagonal panels: unadjusted (dashed line) and adjusted (full line) *p*-value functions of each comparison. The bands in the lower parts of each extra-diagonal panel highlight the intervals presenting significant differences at a 5% (light) and 1% (dark) significance level.

in Section 10.2. The unadjusted and the adjusted *p*-value functions are reported with dashed and full lines, respectively. Finally, the bands reported in the lower parts of each extra-diagonal panel highlight the intervals of the domain presenting significant differences between each couple of regions at a 5% (light) and 1% (dark) significance level. The displayed intervals are obtained by applying a threshold to the adjusted *p*-value function at levels 5% and 1% (displayed in the lower diagonal panels of Figure 10.3). Hence, such a domain selection is provided with a

control of the intervalwise error rate, and it is intervalwise consistent, as shown in Section 10.2. Analogously, Figure 10.4 reports – with the same scheme – the results of the comparison between the derivatives of the functional data.

The IWT-based comparison of the curves and the first derivatives result in a very clear interpretation of the data differences. Focusing on the curves, we notice that both Atlantic and Pacific Zones differ from the Arctic Zone over the entire year. Temperatures of these two zones also differ from the Continental ones during winter. The Continental and Arctic Zones are significantly different during the whole year from February to December. So the Continental climate is similar to the Arctic one during winter, and similar to the Pacific and Atlantic ones during summer. The Atlantic and Pacific zones are pointed out as significantly different only during the first months of the year, from January to March.

The pairwise tests of derivatives (Figure 10.4) add several information about data differences completing the characterization of the different climatic zones. For instance, let us focus on Atlantic and Pacific zones. The temperature of these two zones was not detected as statistically different throughout the great majority of the year. Looking at data derivatives – instead – we notice how the climate of these two zones present statistically significant differences in the spring transition from winter to summer. Indeed, these two zones differ significantly in early winter and spring, when the latter is characterized by smaller absolute derivatives. The temperature derivatives of both Continental and Arctic zones is instead significantly different with respect to the one of both Atlantic and Pacific zones during Spring and Autumn. In general, Atlantic and Pacific zones are characterized by a slower transition in the temperature profiles in such periods.

10.4 Conclusion and Future Works

In this chapter, we presented a null hypothesis testing technique for performing local inference on functional data embedded in the $L^2(T)$ space of squared-integrable functions on the domain T. The technique – namely IWT procedure – is based on the definition of an unadjusted and an adjusted p-value function that can be used to locally test a functional null hypothesis over the domain of functional data. The IWT for testing differences between two populations is described in detail [18], and its extension to multiway FANOVA and functional on scalar linear models can be found in [23–25].

When applied to spatiotemporal data, this technique can identify intervals of time or regions of space imputable for the rejection of a functional null hypothesis.

The IWT is a very general procedure, since it can be plugged-in with either parametric or nonparametric exact and consistent tests for the functional hypothesis at hand. The procedure is first described for testing the equality between the means of

two functional populations. Then, an extension to the case of testing hypotheses in a functional multiway ANOVA is also detailed. IWT can be extended to more complex inferential problems such as functional linear models and tests for comparing variances or higher-order moments.

In all such cases, it is possible to characterize the inferential properties of the unadjusted and adjusted p-value functions. In detail, the unadjusted p-value function is provided with a control of the pointwise error rate and it is pointwise consistent. The pointwise control implies that the probability of wrongly rejecting the null hypothesis on a point of the domain where it is not violated – in an L^2-sense suitably defined – is controlled. The pointwise consistency instead implies that the probability that the null hypothesis is rejected on a point of the domain where it is violated – in an L^2-sense suitably defined – converges to one as the sample size increases. The adjusted p-value function is instead provided with a control of the intervalwise error rate and it is intervalwise consistent. Intervalwise control implies that the probability of wrongly rejecting the null hypothesis on an interval where it is not violated is controlled, and intervalwise consistency implies that the probability that the null hypothesis is rejected on an interval where it is violated converges to one as the sample size increases.

These properties can be of help in deciding to base inference in the unadjusted or in the adjusted p-value function. In most applications – when a selection of intervals imputed to the rejection of the null hypothesis is desired – inference should be based on the adjusted p-value function that is based on a sound control of the probability of falsely detecting intervals. The unadjusted p-value function indeed only pointwise controls the type I error rate, and it does not provide any global control of the probability of type I errors.

The IWT is here applied to a spatiotemporal data set of Canadian daily temperatures. The data are modeled as a collection of functional data on the time domain pertaining to different groups – identified by climatic zones. IWT is applied to test differences between the climatic zones. The climatic zones are compared in terms of the level of average daily temperatures (IWT performed on curves) and the velocity of transition in time of the average daily temperatures (IWT performed on derivatives). Since the domain of functional data is here time, the procedure results in a selection of the periods of the year presenting significant differences between each pair of regions in terms of data and first derivatives. The general nature of the IWT and the possibility of plugging it in with nonparametric tests based on very few modeling assumptions makes its application either to curves or to differential quantities such as derivatives straightforward. In addition, IWT-based comparison between the climatic zones in terms of the two quantities results in a very clear and informative characterization of data differences.

References

1 Cuevas, A., Febrero, M., and Fraiman, R. (2004). An ANOVA test for functional data. *Computational Statistics and Data Analysis* 47 (1): 111–122.
2 Abramovich, F. and Angelini, C. (2006). Testing in mixed-effects FANOVA models. *Journal of Statistical Planning and Inference* 136 (12): 4326–4348.
3 Antoniadis, A. and Sapatinas, T. (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics and Data Analysis* 51 (10): 4793–4813.
4 Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, vol. 200. Springer.
5 Staicu, A., Li, Y., Crainiceanu, C.M., and Ruppert, D. (2014). Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian Journal of Statistics* 41 (4): 932–949.
6 Zhang, J. and Liang, X. (2014). One-way ANOVA for functional data via globalizing the pointwise F-test. *Scandinavian Journal of Statistics* 41 (1): 51–71.
7 Cuesta-Albertos, J.A. and Febrero-Bande, M. (2010). A simple multiway ANOVA for functional data. *Test* 19 (3): 537–557.
8 Hall, P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika* 89 (2): 359–374.
9 Cardot, H., Goia, A., and Sarda, P. (2004). Testing for no effect in functional linear regression models, some computational approaches. *Communications in Statistics - Simulation and Computation* 33 (1): 179–199.
10 Cardot, H., Prchal, L., and Sarda, P. (2007). No effect and lack-of-fit permutation tests for functional regression. *Computational Statistics* 22 (3): 371–390.
11 Hall, P. and Van Keilegom, I. (2007). Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica* 17 (4): 1511.
12 Wang, H. and Akritas, M.G. (2010). Inference from heteroscedastic functional data. *Journal of Nonparametric Statistics* 22 (2): 149–168.
13 Zhang, J.T. (2013). *Analysis of Variance for Functional Data*. CRC Press.
14 Corain, L., Melas, V.B., Pepelyshev, A., and Salmaso, L. (2014). New insights on permutation approach for hypothesis testing on functional data. *Advances in Data Analysis and Classification* 8 (3): 339–356.
15 Pesarin, F. and Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Applications and Software*. Wiley.
16 Vsevolozhskaya, O., Greenwood, M., and Holodov, D. (2014). Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. *Annals of Applied Statistics* 8 (2): 905–925.

17 Marcus, R., Peritz, E., and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63 (3): 655–660.

18 Pini, A. and Vantini, S. (2017). Interval-wise testing for functional data. *Journal of Nonparametric Statistics* 29(2): 407–424.

19 Abramovich, F. and Heller, R. (2005). Local functional hypothesis testing. *Mathematical Methods of Statistics* 14 (3): 253.

20 Cox, D.D. and Lee, J.S. (2008). Pointwise testing with functional data using the Westfall–Young randomization method. *Biometrika* 95 (3): 621–634.

21 Vsevolozhskaya, O.A., Greenwood, M.C., Powell, S.L., and Zaykin, D.V. (2015). Resampling-based multiple comparison procedure with application to point-wise testing with functional data. *Environmental and Ecological Statistics* 22 (1): 45–59.

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

23 Abramowicz, K., Häger, C., Pini, A., Schelin, L., Sjöstedt de Luna, S., and Vantini, S. (2018). Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scandinavian Journal of Statistics* 45(4): 1036–1061.

24 Pini, A., Spreafico, L., Vantini, S., and Vietti, A. (2018). Multi-Aspect Local Inference for Functional Data: Analysis of Ultrasound Tongue Profiles. *Journal of Multivariate Analysis* 170: 162–185.

25 Pini, A., Vantini, S., Colosimo, B.M., and Grasso, M. (2017). Domain-selective functional analysis of variance for supervised statistical profile monitoring of signal data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67(1): 55–81.

26 Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics* 1 (4): 292–298.

11

Modeling Spatially Dependent Functional Data by Spatial Regression with Differential Regularization

Mara S. Bernardi and Laura M. Sangalli

MOX – Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, MI, 20133, Italy

11.1 Introduction

In this chapter, we describe the modeling of spatially dependent functional data by regression with differential regularization [1]. The chapter is based on [2].

Spatial regression with differential regularization defines the unknown spatial field f as deterministic and models the spatial (or spatiotemporal) variation of the phenomenon under study via a regularizing term involving a partial differential equation (PDE). This contrasts with the main approach of geostatistical modeling, where the unknown field is modeled as stochastic and the spatial variation of the phenomenon is controlled via the definition of the covariance structure of the random field.

The different modeling of the spatial variation considered by spatial regression with differential regularization, combined with the use of advanced numerical analysis techniques, such as finite element methods and isogeometric analysis based on splines and extensions, leads to important advantages. One main advantage, that we illustrate here, is the ability to efficiently deal with data distributed over a spatial domain featuring peninsulas, islands, holes, and other complex geometries that influence the phenomenon under study. Moreover, the method can comply with specific conditions at the boundaries of the problem domain, which is fundamental in many applications to obtain meaningful estimates.

As an illustrative example, consider the estimation of the temporal evolution of the amount of per capita municipal waste produced in the towns of Venice province. Figure 11.1 shows the Venice province, with dots indicating town centers, including municipalities and other tourist localities of particular relevance.

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

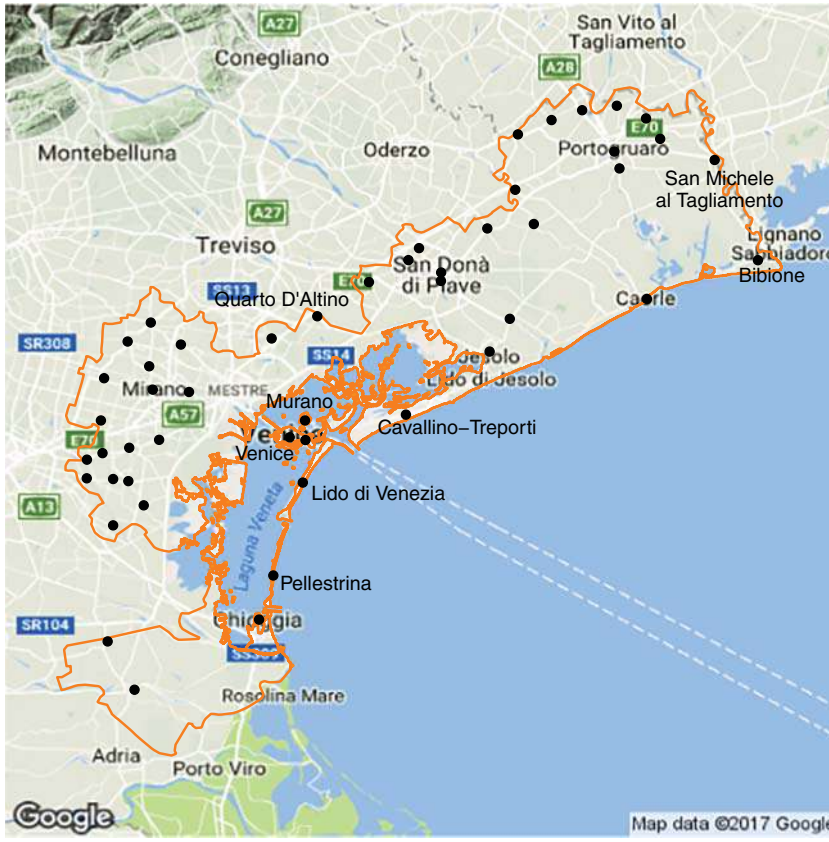


Figure 11.1 Spatial domain of the Venice waste data, with a line highlighting the province boundary and dots indicating the towns centers.

The province boundary is shown by a line, highlighting the irregular shape of the province administrative borders and its complex coastlines, with the Venice lagoon partly enclosed by elongated peninsulas and small islands.

The data are measurements from 1997 to 2011 of the yearly amount of per capita municipal waste (total kilograms divided by the number of municipality residents) and are provided by the Arpav, the Agenzia regionale per la prevenzione e protezione ambientale del Veneto. Figure 11.2 shows the temporal evolution of the production of per capita waste in the towns of Venice province; Figure 11.3 is a bubble plot of the data at a fixed year, 2006. The phenomenon portrayed by these data is expressed differently in different parts of the domain. Consider, for instance, the two towns of Cavallino-Treporti (in the peninsula at the northeast of Venice) and Quarto d'Altino (north of Venice), indicated by

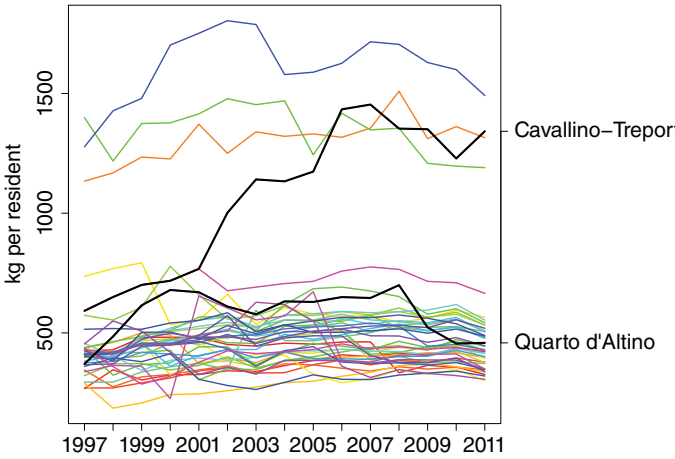


Figure 11.2 Temporal evolution of the yearly per capita production (kilogram per resident) of municipal waste in the towns of Venice province. Source: Adapted from Bernardi et al. [2].

black dots in Figure 11.3. The temporal evolution of the production of per capita municipal waste in the two towns, highlighted in Figure 11.2, is rather different, with strongly increasing and high values in the seaside and tourist town of Cavallino-Treporti, opposed to the not increasing and lower values measured in the hinterland town of Quarto d'Altino. These two towns are close in terms of their geodesic distance, but they are much further apart in terms of land connections, as they are separated by the Venice Lagoon. Appropriately accounting for the shape of the domain, characterized by the strong concavity formed by the lagoon, is crucial to accurately handle these data.

When analyzing the temporal evolutions of the amount of per capita municipal waste, we shall make a strong simplification of the nature of these data and consider them in the framework of geostatistical functional data [3], where the datum is observable in principle in any point of the domain, instead of in the framework of functional area data. As detailed in Section 11.4, this is due to the fact that we miss the information concerning the urbanized areas of the municipalities, where the type of waste considered here (that does not include agricultural, industrial, construction/demolition, and hazardous waste) is produced.

This book reviews in detail many of recently proposed methods for the analysis of spatially dependent functional data, mostly in the framework of kriging for functional data (see, e.g. [3–10]). On the other hand, these methods, as well as the extensive literature in the more classical spatial–time data framework (see, e.g. [11] and references therein), are not well suited to handle data distributed over irregular domains, as they do not take into account the shape of the

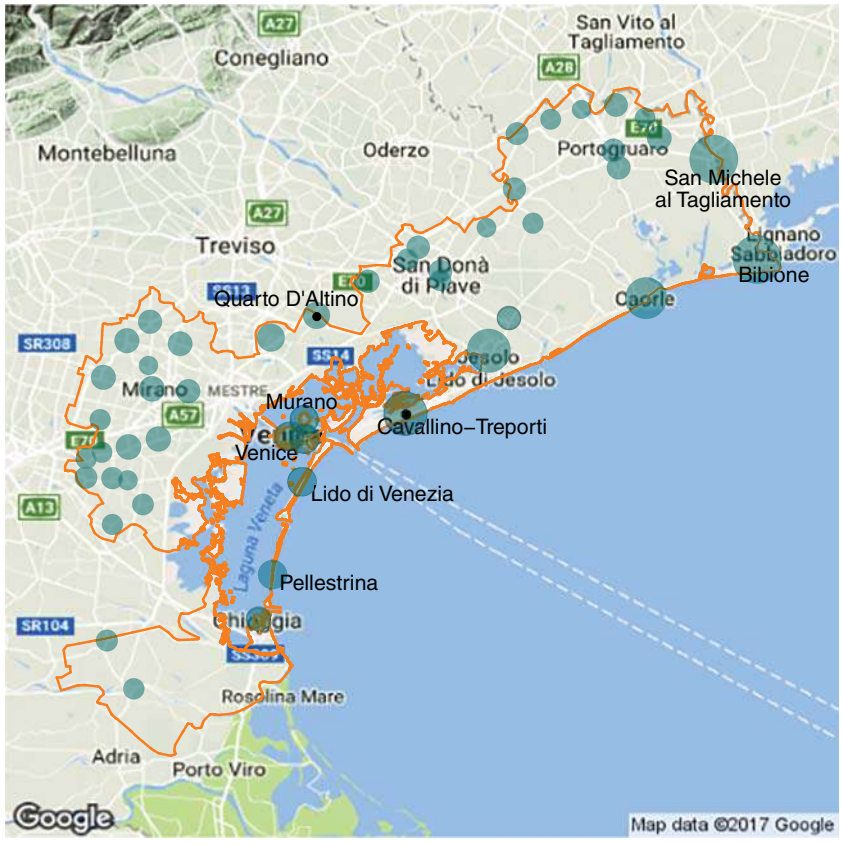


Figure 11.3 Per capita production (kilogram per resident) of municipal waste in the towns of Venice province in 2006. The data include all municipalities of Venice province and additional four localities (Bibione, Murano, Lido di Venezia, and Pellestrina), that do not constitute a municipality on their own, but have been included due to their tourist relevance and their location on the domain. For these additional four localities, the considered datum is a replicate of the datum of their corresponding municipalities (see Section 11.4).

domain; they would, for instance, smooth across concave boundary regions, thus closely linking data points that are in fact far apart by land connections.

Recent methods for the analysis of spatiotemporal data that instead specifically account for the geometry of the domain of interest are described in [12, 13]. These models are based on the spatial smoother proposed by Wood et al. [14]. Here, we describe the method proposed in [2], that extends the spatial regression models with differential regularization described in [1, 15, 16]. The model is implemented in R [17], based on the package fdaPDE [18].

11.2 Spatial Regression with Differential Regularization for Geostatistical Functional Data

Let $\Omega \subset \mathbb{R}^2$ be a bounded spatial domain, possibly with an irregular geometry, and let $\{\mathbf{p}_i = (x_i, y_i) \in \Omega; i = 1, \dots, n\}$ be a set of n spatial locations within this domain. Moreover, consider m time instants $\{t_j \in T; j = 1, \dots, m\}$ over the temporal interval $T = [t_{start}, t_{end}] \subset \mathbb{R}$. Let z_{ij} be the value of a real-valued variable observed at location $\mathbf{p}_i$ and time instant t_j. Additionally, let $\mathbf{w}_{ij} \in \mathbb{R}^q$ be a vector of q space–time varying covariates associated with the observation z_{ij} at the spatiotemporal location $(\mathbf{p}_i, t_j)$. In our illustrative application, the spatial domain Ω is the province of Venice, the spatial locations $\mathbf{p}_i$ are the centers of the towns, the time instants t_j are the years between 1997 and 2011, the variable of interest z_{ij} is the amount of per capita municipal waste produced in the town i and year t_j; furthermore, since intuition suggest that the tourism may play an important role in the production of waste, we consider as covariate w_{ij} the number of beds in accommodation facilities in the town i and year t_j.

Assume the following semiparametric generalized additive model:

$$z_{ij} = \mathbf{w}_{ij}^T \boldsymbol{\beta} + f(\mathbf{p}_i, t_j) + \varepsilon_{ij} \quad i = 1, \dots, n, \quad j = 1, \dots, m, \tag{11.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is a vector of q regression coefficients, $f(\mathbf{p}, t) : \Omega \times T \to \mathbb{R}$ is an unknown smooth spatiotemporal function, and $\{\varepsilon_{ij}; i = 1, \dots, n; j = 1, \dots, m\}$ are independently distributed residuals with mean zero and constant variance σ^2. As detailed in Section 11.2.4, one may as well consider a model without covariates. In this case, the values z_{ij} can be directly seen as discrete and noisy observations of dependent functional data, either spatially dependent curves, or time-dependent surfaces.

The vector of regression coefficients $\boldsymbol{\beta}$ and the spatiotemporal field f can be jointly estimated minimizing a penalized sum of square error functional. In particular, in [2], we propose to consider two roughness penalties that account separately for the regularity of the field in space and in time. Let

$$J(f, \boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(z_{ij} - \mathbf{w}_{ij}^T \boldsymbol{\beta} - f(\mathbf{p}_i, t_j) \right)^2 + \lambda_\Omega \int_T \int_\Omega (\Delta f(\mathbf{p}, t))^2 d\mathbf{p} dt + \lambda_T \int_\Omega \int_T \left(\frac{\partial^2 f(\mathbf{p}, t)}{\partial t^2} \right)^2 dt d\mathbf{p} , \tag{11.2}$$

where the two smoothing parameters $\lambda_\Omega > 0$ and $\lambda_T > 0$ weight the penalizations, respectively, in space and time; the choice of these parameters will be discussed in Section 11.2.3. The partial differential operator $\Delta f(\mathbf{p}, t) = \frac{\partial^2 f}{\partial x^2}(\mathbf{p}, t) + \frac{\partial^2 f}{\partial y^2}(\mathbf{p}, t)$ is the Laplacian of the spatial component of the field; it provides the local curvature of the spatial field, at a given time t. The Laplacian is invariant to rigid transformations of the spatial coordinates, thus ensuring that the concept of smoothness

does not depend on the arbitrary choice of the coordinate system. The smoothness penalties in (11.2) are isotropic and stationary. As detailed in Section 11.2.3, they induce the spatiotemporal mean and covariance structures of the estimator. Different regularizations may be considered, as briefly discussed in Section 11.5, implying different mean and covariance structures and modeling anisotropic and nonstationary effects.

11.2.1 A Separable Spatiotemporal Basis System

We represent the spatiotemporal field $f(\mathbf{p}, t)$ as an expansion on a separable spatiotemporal basis system. Specifically, let $\{\varphi_k(t);\ k = 1, \dots, M\}$ be a set of M basis functions defined on T and $\{\psi_l(\mathbf{p});\ l = 1, \dots, N\}$ a set of N basis functions defined on Ω. Then, f is represented by the following basis expansion:

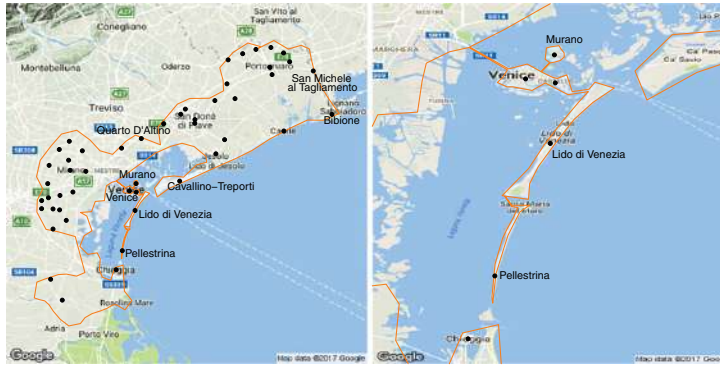
$$f(\mathbf{p}, t) = \sum_{l=1}^{N} \sum_{k=1}^{M} c_{lk}\ \psi_l(\mathbf{p})\ \varphi_k(t), \tag{11.3}$$

where $\{c_{lk};\ l = 1, \dots, N;\ k = 1, ldots, M\}$ are the coefficients of the expansion on the separable spatiotemporal basis.

Various possible bases can be used for the expansions in the spatial and temporal domains. In this chapter, we describe the use, for the spatial domain, of a finite element basis on a triangulation Ω_τ of the domain Ω. This choice leads to an efficient discretization of the functional J and allows to accurately take into account the shape of the spatial domain.

We illustrate the construction of such basis on Venice domain. Before building the basis, we simplify the original spatial domain represented in Figure 11.1, excluding the coastal uninhabited regions and the smaller islands, and keeping in the domain of study only the four major islands: Venice, Murano (at the northeast of Venice), Lido di Venezia (at the southeast of Venice), and Pellestrina (at the south of Lido). We then smooth the boundary of the domain with regression splines. Finally, we obtain a piecewise linear boundary, subsampling from this smooth curve so that the features characterizing the domain are preserved. Figure 11.4a shows the simplified boundary of Venice province, while Figure 11.4b shows the detail around the city of Venice. This region is particularly interesting since it shows the four islands retained in the domain. Here the domain includes four bridges: one linking Venice to the continent and the others linking some of the islands between themselves; the first one is an actual bridge with a road and a railway, while the other bridges represent regular and frequent ferries among the islands.

A triangulation of the resulting simplified domain is then obtained using the R package fdaPDE [18]. In particular, we start from a Delaunay triangulation, constrained within the simplified boundary, where each of the town locations and



(a) (b)

Figure 11.4 Simplified boundary of the Venice province (a) and detail of the Venice lagoon (b).

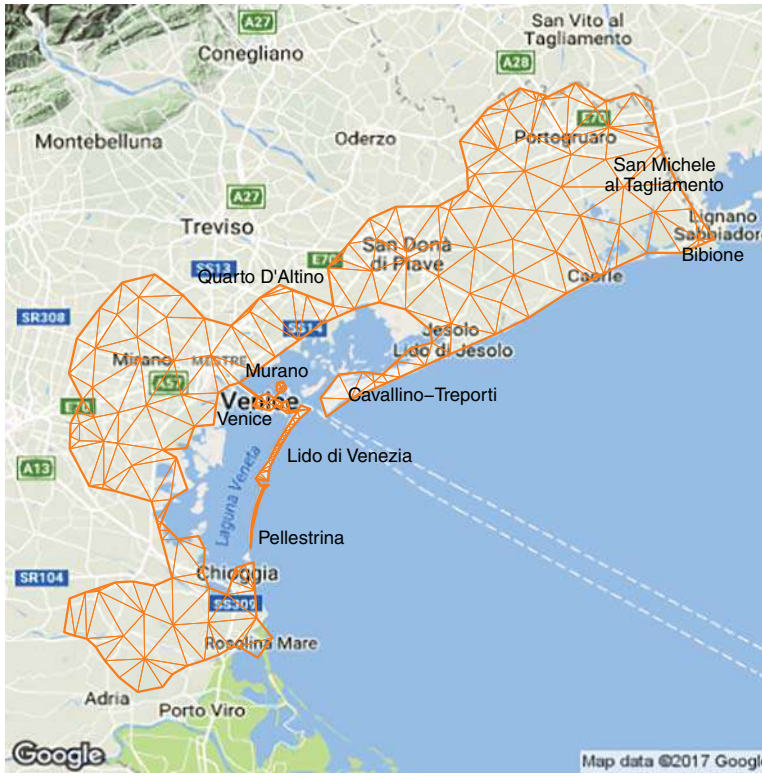


Figure 11.5 Triangulation of the Venice province.

each point defining the simplified boundary becomes a triangle vertex. A more regular mesh is then obtained imposing a maximum value to the triangle areas. Figure 11.5 displays the resulting triangulation of Venice province. For this application, here and in Section 11.4, instead of using as coordinates the latitude and longitude, we employ the universal transverse mercator (UTM) coordinate system, which allows to compute the distance between two points on the Earth's surface by means of the Euclidean distance instead of the geodesic distance.

The finite element basis is composed by globally continuous functions that coincide with a polynomial of a certain degree on each element of the domain triangulation. In particular, we use here linear finite element basis, that are piecewise linear functions. The dimension of the spatial basis is strictly related to the triangulation of the spatial domain: there is one basis function for each knot of the triangulation; for linear finite elements, each basis is associated with a vertex of the triangulation and has value 1 at that vertex and 0 at all other vertices. Figure 11.6 shows an example of linear basis function.

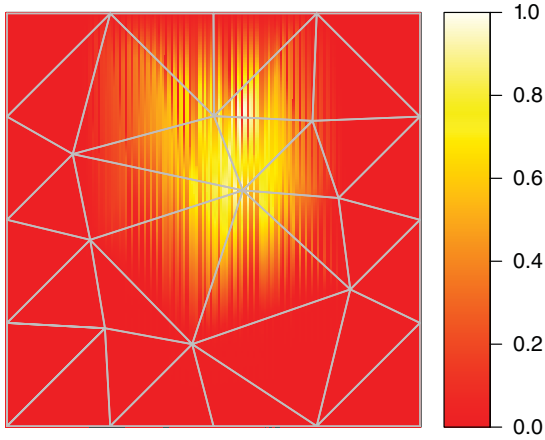


Figure 11.6 Example of linear finite element basis function.

For the temporal dimension, we use here a cubic B-spline basis with penalization of the second derivative, with knots coinciding with the sampling time instants of the dataset. Other basis systems may turn out to be more appropriate in other applicative contexts. For instance, Fourier basis are well suited to the case of cyclic data, possibly with penalization of the harmonic acceleration operator, instead of the order 2 derivative considered in (11.2).

The chosen basis system should be rich enough to enable an accurate representation of the spatiotemporal evolution of the phenomenon. In general, the number of bases, and thus of coefficients to be estimated, $M \times N$, can be larger than the sample size, $m \times n$. This is for instance the case of the application to Venice waste data, as well as of the simulation studies reported in Section 11.3. In these examples, in space, we start from a constrained Delaunay triangulation of the spatial locations, that is further refined in the application to Venice waste data, and then consider the associated linear finite element basis, whose dimension N (equal to number of internal and boundary nodes) is thus larger than n. In time, we use a cubic B-spline basis having knots at the m time instants of observation, resulting in a basis dimension M larger than m. This fact does not create any problem from the estimation point of view, thanks to the presence of the regularizing terms. We, indeed, never experienced any numerical instability of the method. Of course, in presence of dense sampling schemes, in space, or time, coarser spatial or temporal grids may be preferred for computational saving.

11.2.2 Discretization of the Penalized Sum-of-Square Error Functional

Let $\mathbf{z} \in \mathbb{R}^{nm}$ be the vector of observed data values at the $n \times m$ spatiotemporal locations, $\mathbf{f} \in \mathbb{R}^{nm}$ the vector of evaluations of the spatiotemporal function f at the

$n \times m$ spatiotemporal locations, and $\mathbf{c} \in \mathbb{R}^{NM}$ the vector of coefficients of the basis expansion (11.3) of the spatiotemporal field f, with entries ordered as follows:

$$\mathbf{z} = \begin{bmatrix} z_{11} \\ \vdots \\ z_{1m} \\ z_{21} \\ \vdots \\ z_{2m} \\ \vdots \\ z_{nm} \end{bmatrix} \qquad \mathbf{f} = \begin{bmatrix} f(\mathbf{p}_1, t_1) \\ \vdots \\ f(\mathbf{p}_1, t_m) \\ f(\mathbf{p}_2, t_1) \\ \vdots \\ f(\mathbf{p}_2, t_m) \\ \vdots \\ f(\mathbf{p}_n, t_m) \end{bmatrix} \qquad \mathbf{c} = \begin{bmatrix} c_{11} \\ \vdots \\ c_{1M} \\ c_{21} \\ \vdots \\ c_{2M} \\ \vdots \\ c_{NM} \end{bmatrix}.$$

Coherently, let $W \in \mathbb{R}^{nm \times q}$ be the design matrix containing the covariates $\{\mathbf{w}_{ij};\ i = 1, \dots, n;\ j = 1, \dots, m\}$:

$$W = \begin{bmatrix} \mathbf{w}_{11}^T \\ \vdots \\ \mathbf{w}_{1m}^T \\ \mathbf{w}_{21}^T \\ \vdots \\ \mathbf{w}_{2m}^T \\ \vdots \\ \mathbf{w}_{nm}^T \end{bmatrix}.$$

Set $H_W = W(W^T W)^{-1} W^T$ as the matrix that projects orthogonally into the subspace of $\mathbb{R}^{nm}$ generated by the columns of W, and set $Q = I_{nm} - H_W$ as the matrix that projects into the orthogonal complement. We denote by $I_d \in \mathbb{R}^{d \times d}$ the identity matrix. Let $\Psi \in \mathbb{R}^{n \times N}$ be the matrix of the evaluations of the N spatial basis functions in the n spatial locations $\{\mathbf{p}_i;\ i = 1, \dots, n\}$,

$$\Psi = \begin{bmatrix} \psi_1(\mathbf{p}_1) & \psi_2(\mathbf{p}_1) & \cdots & \psi_N(\mathbf{p}_1) \\ \psi_1(\mathbf{p}_2) & \psi_2(\mathbf{p}_2) & \cdots & \psi_N(\mathbf{p}_2) \\ \vdots & \vdots & \cdots & \vdots \\ \psi_1(\mathbf{p}_n) & \psi_2(\mathbf{p}_n) & \cdots & \psi_N(\mathbf{p}_n) \end{bmatrix}.$$

Moreover, define the vectors $\boldsymbol{\psi}, \boldsymbol{\psi}_x, \boldsymbol{\psi}_y \in \mathbb{R}^N$ of the spatial basis functions and of their first-order partial derivatives:

$$\boldsymbol{\psi} = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_N \end{bmatrix} \qquad \boldsymbol{\psi}_x = \begin{bmatrix} \partial\psi_1/\partial x \\ \partial\psi_2/\partial x \\ \vdots \\ \partial\psi_N/\partial x \end{bmatrix} \qquad \boldsymbol{\psi}_y = \begin{bmatrix} \partial\psi_1/\partial y \\ \partial\psi_2/\partial y \\ \vdots \\ \partial\psi_N/\partial y \end{bmatrix}.$$

Finally, let $R_0, R_1 \in \mathbb{R}^{N \times N}$ be two matrices, respectively, containing the integrals over Ω_τ of the cross products of the N spatial basis, and the integrals over Ω_τ of

the cross products of the first derivatives of the N spatial basis, i.e.

$$R_0 = \int_{\Omega_s} \boldsymbol{\psi}\boldsymbol{\psi}^T \qquad R_1 = \int_{\Omega_s} \left(\boldsymbol{\psi}_x\boldsymbol{\psi}_x^T + \boldsymbol{\psi}_y\boldsymbol{\psi}_y^T\right) .$$

Analogously, let $\Phi \in \mathbb{R}^{m\times M}$ be the matrix of the evaluations of the M temporal basis functions in the m time instants $\{t_j;\ j = 1, \dots, m\}$:

$$\Phi = \begin{bmatrix} \varphi_1(t_1) & \varphi_2(t_1) & \cdots & \varphi_M(t_1) \\ \varphi_1(t_2) & \varphi_2(t_2) & \cdots & \varphi_M(t_2) \\ \vdots & \vdots & \cdots & \vdots \\ \varphi_1(t_m) & \varphi_2(t_m) & \cdots & \varphi_M(t_m) \end{bmatrix}.$$

Moreover, define the vectors $\boldsymbol{\varphi}, \boldsymbol{\varphi}_{tt} \in \mathbb{R}^M$ of the temporal basis functions and of their second-order derivatives:

$$\boldsymbol{\varphi} = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_M \end{bmatrix} \quad \boldsymbol{\varphi}_{tt} = \begin{bmatrix} d^2\varphi_1/dt^2 \\ d^2\varphi_2/dt^2 \\ \vdots \\ d^2\varphi_M/dt^2 \end{bmatrix}.$$

Finally, let $K_0 \in \mathbb{R}^{M\times M}$ be the matrix of the integrals over T of the cross products of the M temporal basis, i.e.

$$K_0 = \int_T \boldsymbol{\varphi}\boldsymbol{\varphi}^T . \tag{11.4}$$

Consider now the matrix $B = \Psi \otimes \Phi \in \mathbb{R}^{nm\times NM}$, where $\otimes$ denotes the Kronecker product. Then $\mathbf{f} = B\mathbf{c}$.

We may then rewrite the sum of square error functional J in (11.2) as follows:

$$\begin{aligned} J &= (\mathbf{z} - W\boldsymbol{\beta} - B\mathbf{c})^T(\mathbf{z} - W\boldsymbol{\beta} - B\mathbf{c}) + \lambda_\Omega \mathbf{c}^T(P_S \otimes K_0)\mathbf{c} + \lambda_T \mathbf{c}^T(R_0 \otimes P_T)\mathbf{c} \\ &= (\mathbf{z} - W\boldsymbol{\beta} - B\mathbf{c})^T(\mathbf{z} - W\boldsymbol{\beta} - B\mathbf{c}) + \mathbf{c}^T P\mathbf{c} , \end{aligned} \tag{11.5}$$

where P_S and P_T are the matrix discretizations of the spatial and temporal penalization terms, and P is the overall penalty $P = \lambda_\Omega\ (P_S \otimes K_0) + \lambda_T\ (R_0 \otimes P_T)$. Specifically, the matrix P_T is obtained by direct discretization of the temporal penalty term in (11.2):

$$P_T = \int_T \boldsymbol{\varphi}_{tt}\boldsymbol{\varphi}_{tt}^T;$$

see [19] for details. For the matrix P_S, following [1, 15], we consider a computationally efficient discretization of the spatial penalty term in (11.2), that does not involve the computation of second-order derivatives of the basis functions, but only of first-order derivatives. This discretization is given by $P_S = R_1 R_0^{-1} R_1$, and it is based on a variational characterization of the estimation problem; see [15] for details. As shown in [16], in the finite element space used to discretize the problem,

the matrix P_S is in fact equivalent to the penalty matrix that would be obtained as direct discretization of the penalty term in (11.2) and involving the computation of second-order derivatives.

This formulation uses the homogeneous Neumann condition at the boundary of the domain of interest implying zero flow across the boundary $\partial\Omega$. Boundary conditions are a way to control the behavior of the estimated function at the boundaries of the domain. Various boundary conditions are possible: Dirichlet conditions control the value of the function, that is $f|_{\partial\Omega} = \gamma_D$, Neumann conditions control the value of the normal derivative of the function, that is $\partial_{\mathbf{n}} f|_{\partial\Omega} = \gamma_N$, and Robin conditions are linear combinations of the previous two. Homogeneous conditions correspond to the case when γ_D or γ_N are null functions. Moreover, different types of boundary conditions can be imposed on different parts of the boundary, forming a partition of it. In the simulations reported in Section 11.3 and in the application to Venice waste data, we impose homogeneous Neumann boundary conditions, i.e. null flow across the boundary; we are thus considering closed systems with respect to the phenomenon considered. In the context of Venice data, this means that we assume no exchange of waste between Venice province and the sea or between Venice province and other neighboring provinces.

To compute the estimates of the vector of regression coefficients $\boldsymbol{\beta}$ and of the vector $\mathbf{c}$ of coefficients of the basis expansion of the spatiotemporal field f, we compute the first partial derivatives of J with respect to $\boldsymbol{\beta}$ and $\mathbf{c}$, and set them equal to zero, getting the following explicit solution to the estimation problem:

$$\hat{\boldsymbol{\beta}} = (W^T W)^{-1} W^T (\mathbf{z} - B\hat{\mathbf{c}}),$$
$$\hat{\mathbf{c}} = (B^T Q B + P)^{-1} B^T Q \mathbf{z}.$$

The estimators are linear in the observed data values $\mathbf{z}$; the estimator $\hat{\mathbf{c}}$ has a penalized least-square form, and given $\hat{\mathbf{c}}$, the estimator $\hat{\boldsymbol{\beta}}$ has a least-square form.

11.2.3 Properties of the Estimators

Let $S_{\mathbf{f}} = B(B^T Q B + P)^{-1} B^T Q$, so that $\hat{\boldsymbol{\beta}} = (W^T W)^{-1} W^T (I_{nm} - S_{\mathbf{f}})\mathbf{z}$.

Since $E[\mathbf{z}] = W\boldsymbol{\beta} + \mathbf{f}$ and $\mathrm{Var}[\mathbf{z}] = \sigma^2 I_{nm}$, and exploiting the fact that the matrix Q is idempotent and $QW = 0$ (where 0 is the $nm \times q$ zero matrix), we obtain

$$E[\hat{\mathbf{c}}] = (B^T Q B + P)^{-1} B^T Q \mathbf{f},$$
$$\mathrm{Var}[\hat{\mathbf{c}}] = \sigma^2 (B^T Q B + P)^{-1} B^T Q B (B^T Q B + P)^{-1}$$

and

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (W^T W)^{-1} W^T (I_{nm} - S_{\mathbf{f}})\mathbf{f},$$
$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2 (W^T W)^{-1} + \sigma^2 (W^T W)^{-1} W^T S_{\mathbf{f}} S_{\mathbf{f}}^T W (W^T W)^{-1}. \tag{11.6}$$

Consider the vector $\mathbf{B}(\mathbf{p}, t) = \boldsymbol{\psi}(\mathbf{p})^T \otimes \boldsymbol{\varphi}(t)^T$ of evaluations of the separable basis system at the spatiotemporal location $(\mathbf{p}, t) \in \Omega \times T$. The estimate of the field f at this generic location is thus given by

$$\hat{f}(\mathbf{p}, t) = \mathbf{B}(\mathbf{p}, t)\hat{\mathbf{c}} = \mathbf{B}(\mathbf{p}, t)(B^T Q B + P)^{-1} B^T Q \mathbf{z}$$

and its mean and variance are given by

$$E[\hat{f}(\mathbf{p}, t)] = \mathbf{B}(\mathbf{p}, t)(B^T Q B + P)^{-1} B^T Q \mathbf{f}$$
$$\mathrm{Var}[\hat{f}(\mathbf{p}, t)] = \sigma^2 \mathbf{B}(\mathbf{p}, t)(B^T Q B + P)^{-1} B^T Q B (B^T Q B + P)^{-1} \mathbf{B}(\mathbf{p}, t)^T, \tag{11.7}$$

with covariance at any two spatiotemporal locations $(\mathbf{p}_1, t_1), (\mathbf{p}_2, t_2) \in \Omega \times T$ given by

$$\mathrm{Cov}[\hat{f}(\mathbf{p}_1, t_1), \hat{f}(\mathbf{p}_2, t_2)]$$
$$= \sigma^2 \mathbf{B}(\mathbf{p}_1, t_1)(B^T Q B + P)^{-1} B^T Q B (B^T Q B + P)^{-1} \mathbf{B}(\mathbf{p}_2, t_2)^T. \tag{11.8}$$

It should be noticed that the regularizing terms in (11.2), and their corresponding discretization P, induce both the first-order structure (i.e. the mean) and the second-order structure (i.e. the spatiotemporal covariance) of the estimator $\hat{f}$. Different regularizations would imply different mean and covariance structures. For instance, Azzimonti et al. [16] consider a regularized spatial regression model and show that by changing the regularizing term and considering more complex differential operators it is possible to include in the model a priori information about the spatial variation of the phenomenon, and to model also anisotropies and nonstationarities.

The smoothing matrix S, which maps the vector of observed values $\mathbf{z}$ to the vector of fitted values $\hat{\mathbf{z}} = S\mathbf{z}$, is given by

$$S = H_W + Q S_{\mathbf{f}}.$$

The trace of the smoothing matrix constitutes a commonly used measure of the equivalent degrees of freedom for linear estimators (this notion was first introduced by Buja et al. [20]). For the model considered, this is given by $\mathrm{tr}(S) = q + \mathrm{tr}(S_{\mathbf{f}})$, thus coinciding with the sum of the q degrees of freedom corresponding to the parametric part of the model and the $\mathrm{tr}(S_{\mathbf{f}})$ degrees of freedom

corresponding to the nonparametric part of the model. A robust estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{nm - \text{tr}(S)}(\mathbf{z} - \hat{\mathbf{z}})^T(\mathbf{z} - \hat{\mathbf{z}}). \tag{11.9}$$

This estimate of the error variance, plugged into (11.7), can be used to compute approximate Gaussian pointwise confidence intervals for f; similarly, plugged into (11.6), it can be used to compute approximate Gaussian confidence intervals for $\boldsymbol{\beta}$. Moreover, the value of a new observation at the spatial location $\mathbf{p}_{n+1} \in \Omega$ and time instant $t_{m+1} \in T$, and with associated covariates $\mathbf{w}_{n+1\ m+1}$, can be predicted by $\hat{z}_{n+1\ m+1} = \mathbf{w}_{n+1\ m+1}^T\ \hat{\boldsymbol{\beta}} + \hat{f}(\mathbf{p}_{n+1}, t_{m+1})$, and approximate prediction intervals may as well be constructed.

Finally, the values of the smoothing parameters λ_Ω and λ_T may be chosen via generalized cross-validation (GCV), searching for the values of λ_Ω and λ_T that minimize

$$\text{GCV}(\lambda_\Omega, \lambda_T) = \frac{nm}{(nm - \text{tr}(S))^2}(\mathbf{z} - \hat{\mathbf{z}})^T(\mathbf{z} - \hat{\mathbf{z}}).$$

11.2.4 Model Without Covariates

If covariates are not included in the model, then (11.1) is replaced by

$$z_{ij} = f(\mathbf{p}_i, t_j) + \varepsilon_{ij} \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,$$

a classical functional data analysis model. The spatiotemporal field f can thus be estimated minimizing the functional:

$$\begin{aligned} J(f) = &\sum_{i=1}^{n}\sum_{j=1}^{m}\left(z_{ij} - f(\mathbf{p}_i, t_j)\right)^2 \\ &+\lambda_\Omega \int_T\int_\Omega (\Delta f(\mathbf{p}, t))^2 d\mathbf{p}dt + \lambda_T \int_\Omega\int_T \left(\frac{\partial^2 f(\mathbf{p}, t)}{\partial t^2}\right)^2 dtd\mathbf{p}. \end{aligned} \tag{11.10}$$

The numerical discretization of the functional follows as in Section 11.2.1, leading to

$$J = (\mathbf{z} - B\mathbf{c})^T(\mathbf{z} - B\mathbf{c}) + \mathbf{c}^T P\mathbf{c}\ ,$$

and hence, to the following estimator of the vector of coefficients for the spatiotemporal field:

$$\hat{\mathbf{c}} = (B^TB + P)^{-1}B^T\mathbf{z}.$$

The mean and variance of this estimator are given by

$$\begin{aligned} E[\hat{\mathbf{c}}] &= (B^TB + P)^{-1}B^T\mathbf{f}, \\ \text{Var}[\hat{\mathbf{c}}] &= \sigma^2(B^TB + P)^{-1}B^TB(B^TB + P)^{-1}. \end{aligned}$$

The estimator of the field f at a generic location $(\mathbf{p}, t)$ is thus given by

$$\hat{f}(\mathbf{p}, t) = \mathbf{B}(\mathbf{p}, t)\hat{\mathbf{c}} = \mathbf{B}(\mathbf{p}, t)(B^T B + P)^{-1} B^T \mathbf{z}$$

and has the following mean, variance, and covariance structures:

$$\begin{aligned} E[\hat{f}(\mathbf{p}, t)] &= \mathbf{B}(\mathbf{p}, t)(B^T B + P)^{-1} B^T \mathbf{f} \\ \mathrm{Var}[\hat{f}(\mathbf{p}, t)] &= \sigma^2 \mathbf{B}(\mathbf{p}, t)(B^T B + P)^{-1} B^T B (B^T B + P)^{-1} \mathbf{B}(\mathbf{p}, t)^T \\ \mathrm{Cov}[\hat{f}(\mathbf{p}_1, t_1), \hat{f}(\mathbf{p}_2, t_2)] &= \sigma^2 \mathbf{B}(\mathbf{p}_1, t_1)(B^T B + P)^{-1} B^T B (B^T B + P)^{-1} \mathbf{B}(\mathbf{p}_2, t_2)^T. \end{aligned}$$

These above expressions coincide with those derived in Section 11.2.3, setting $Q = I$. As noted earlier, the mean and covariance structure of the estimator are characterized by the chosen regularizing terms, through their discretization P. Finally, the smoothing matrix is in this case given by $S = B(B^T B + P)^{-1} B^T$. The computation of the degrees of freedom of the estimator, the estimate of the error variance, the optimal selection of the smoothing parameters λ_Ω and λ_T, and the computation of confidence/prediction intervals follows along the same lines outlined in the case of the model with covariates.

11.2.5 An Alternative Formulation of the Model

Instead of considering the functionals (11.2) or (11.10), respectively, in the case with or without covariates, it is possible to consider alternative functionals, that regularize directly the coefficients of the basis expansion of the spatiotemporal field, in analogy with the models proposed by [12, 13]. This alternative formulation is detailed in [2], Section 5.

11.3 Simulation Studies

In [2], the performances of the proposed spatiotemporal regression with partial differential equation regularization (ST-PDE) are tested via extensive simulation studies under various settings, with different sampling designs in space and time, with and without covariates, with correlated and uncorrelated noise. Spatial regression with differential regularization is compared to separable spatiotemporal kriging, to the space–time models proposed by [12, 13] and based on soap film smoothing [14], and to an analogous space–time model based on thin-plate splines (TPS). We report here the results from two simulation studies, referring the interested reader to [2] for details.

We consider a test function defined on a C-shaped spatial domain, shown at three different time instants in the first row of Figure 11.7. The test function displays similar features as Venice waste data: its domain is characterized by a strong

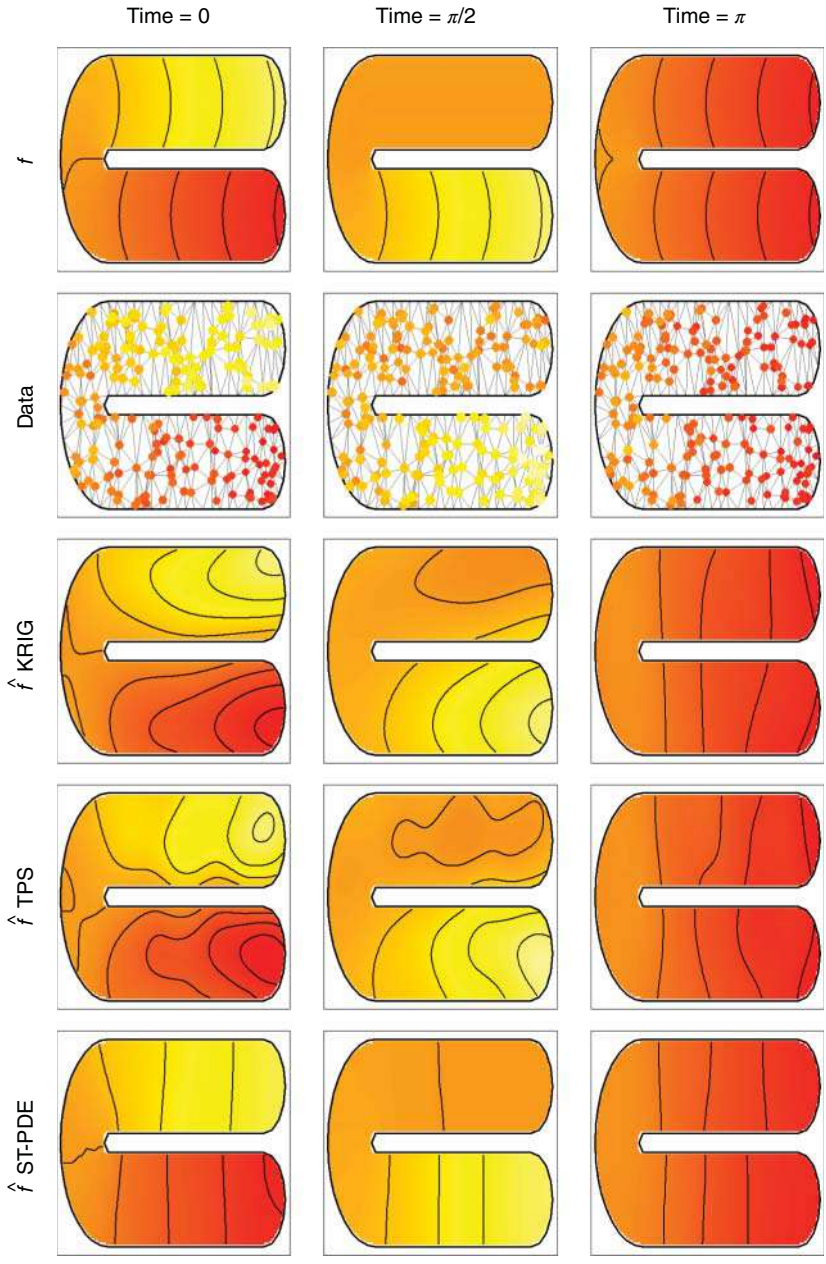


Figure 11.7 Simulation without covariates: test function (first row), sampled data (second row), field estimates provided by spatiotemporal kriging (third row), by spatiotemporal smoothing using thin-plate splines (fourth row), and by ST-PDE (fifth row).

concavity, and different values of the field are observed in the two arms of the domain, across the concavity, with different behaviors over time. The second row in Figure 11.7 shows the data sampled at the three time instants, for the first simulation replicate.

The implementation of ST-PDE is based on the R package fdaPDE [18]. In space, we use a linear finite element basis defined on the mesh shown in the second row of Figure 11.7, a constrained Delaunay triangulation of the sampling spatial locations. In time, we use a cubic B-spline basis, with knots coinciding with the sampling time instants. The values of the smoothing parameters λ_Ω and λ_T are chosen via GCV. The field estimate obtained in the first simulation replicate, at the three considered time instants, is shown in the last row of Figure 11.7. The third and fourth rows of Figure 11.7 illustrate instead the field estimates obtained by spatiotemporal kriging (KRIG), implemented using the R package gstat [21], and by spatiotemporal smoothing with a TPS basis in space and a B-spline basis in time, implemented using the R package mgcv [22]. For the kriging, we use a separable variogram marginally Gaussian in space and exponential in time, with parameters estimated from the empirical variogram. For the spatiotemporal model based on TPS, we select the smoothing parameters via GCV. Figure 11.7 shows that KRIG and TPS return poor estimates of the true spatiotemporal field, especially in those time instants where the true field is characterized by different values in the two arms of the C-shaped domain (see the first and second columns in the figure): the different values have in fact been smoothed across the concavity in the domain. ST-PDE instead accurately estimates the spatiotemporal field, being able to comply with the shape of the domain. Figure 11.9a shows the boxplots of the root mean square errors (RMSE), over 50 replicates of the noise generation, of the space–time field estimates yielded by the three methods. These boxplots confirm the comparative advantage of ST-PDE over KRIG and TPS.

Figures 11.8 and 11.9b show the results from a second simulation study detailed in [2], where we as well include a space–time varying covariate. The second row in Figure 11.8 shows the added contributions of covariates and true function. In this simulation setting, we do not compare to spatiotemporal kriging, as the function `krigeST` of the R package gstat does not allow for the inclusion of covariates. The results are otherwise similar to those obtained in the simulation without covariates, with a superiority of ST-PDE over TPS. This superiority also reflects in the estimation of the β coefficient: the corresponding RMSE over the 50 replicates is 0.14 for TPS and 0.09 for ST-PDE. Also in this simulation setting, the main reason of the comparative advantage shown by ST-PDE consists in its ability to comply with the shape of the domain.

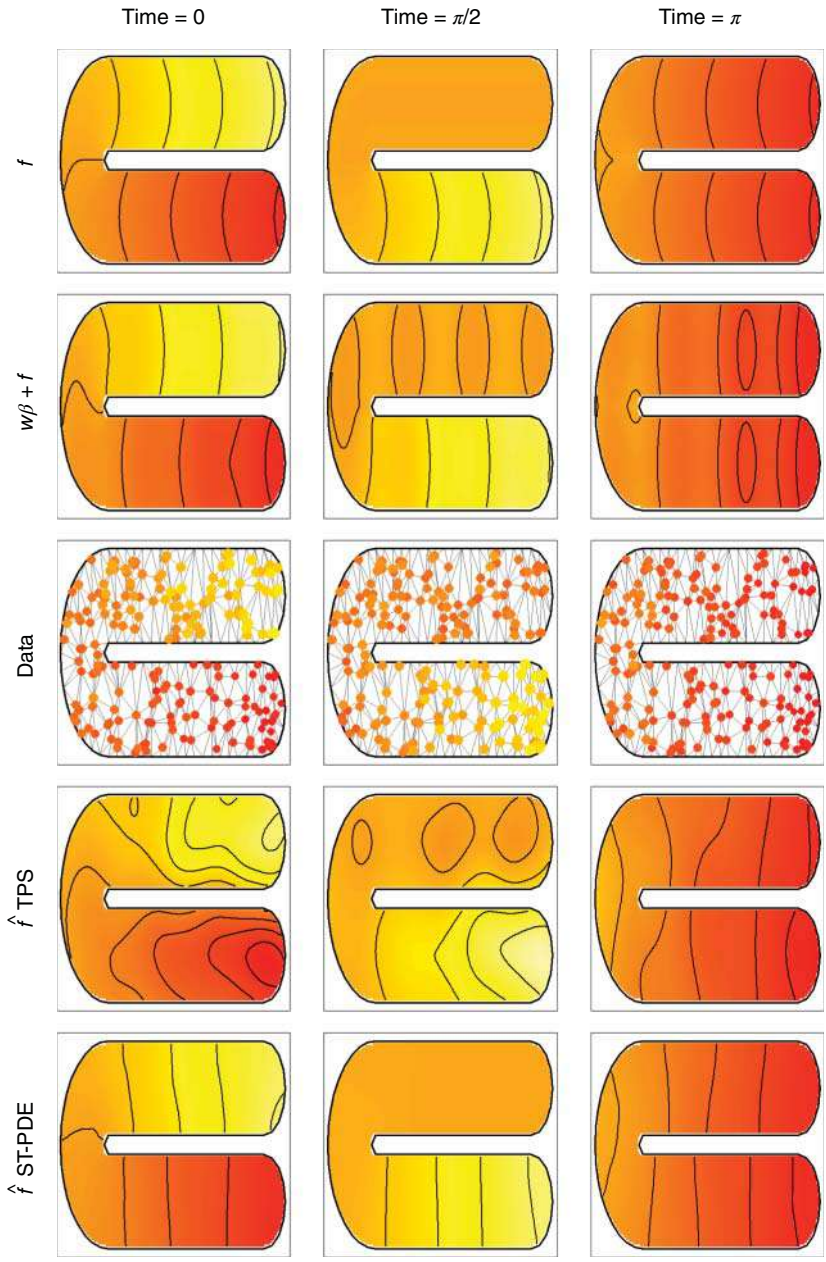


Figure 11.8 Simulation with covariates: test function (first row), added contributions of the spatiotemporal covariate field and of the test function (second row), sampled data (third row), and field estimates provided by spatiotemporal smoothing using thin-plate splines (fourth row), and by ST-PDE (fifth row).

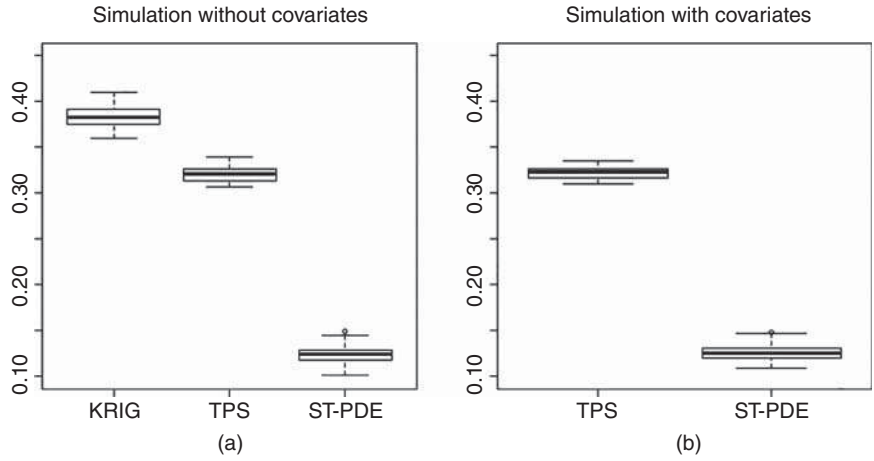


Figure 11.9 (a) Simulation without covariates: boxplots of the RMSE, over 50 simulation replicates, of the field estimates provided by spatiotemporal kriging (KRIG), by spatiotemporal smoothing using thin-plate splines (TPS) and by spatiotemporal regression with PDE regularization (ST-PDE). (b) Simulation with covariates: boxplots of the RMSE, over 50 simulation replicates, of the field estimates provided by spatiotemporal smoothing using thin-plate splines (TPS) and by spatiotemporal regression with PDE regularization (ST-PDE).

11.4 An Illustrative Example: Study of the Waste Production in Venice Province

We now illustrate the described method via an application to the study of the annual amount of per capita municipal waste produced in the Venice province.

11.4.1 The Venice Waste Dataset

Open Data Veneto[1] provides the gross and per capita annual amount of municipal waste produced in each municipality of the Venice province in the period from 1997 to 2011. We consider here for the analysis the annual per capita municipal waste, in kilogram per municipality resident.

Municipal waste includes waste that is produced in houses and public areas, but does not include special waste, i.e. industrial, agricultural, construction and demolition waste, or hazardous waste, for which there are special disposal programs. Therefore, the data refer only to the urbanized areas of the municipality, while they do not refer to the agricultural or industrial areas in the municipality territories.

1 http://dati.veneto.it/dataset/produzione-annua-di-rifiuti-urbani-totale-e-pro-capite-1997-2011

Since no data identifying the urbanized areas of the municipalities is available, we face here two possible simplifications of the problem. We can either partition the Venice province in the municipality territories and attribute each datum to the whole territory of its municipality, or assign each datum to a point representing the center of the municipality. We adopt here the second simplification. The spatial coordinates of the town centers are available online.[2] As mentioned in Section 11.2.1, latitude and longitude are converted into UTM coordinate system.

In some cases, there are localities that do not constitute a municipality on their own, but are under the jurisdiction of another town. In this case, there are two or more main urbanized areas in the municipality territory. Some of these localities are not negligible for the problem under analysis due to their tourist relevance and their location on the domain; for this reason, we add them to the data. Specifically, we include the seaside town of Bibione, the easternmost village indicated in Figure 11.1. This popular vacation destination falls under the jurisdiction of the municipality of San Michele al Tagliamento, northwest of Bibione; the waste data considered for Bibione are a replicate of the data of San Michele al Tagliamento. Moreover, we replicate the data of Venice in the islands of Murano, Lido di Venezia, and Pellestrina because of their tourist relevance and the particular shape of the domain.

We include as a covariate the number of beds in accommodation facilities (such as hotels, bed and breakfast, guesthouses, campings) divided by the number of residents. This ratio may be as large as 7 in some tourist towns by the sea. The number of beds in accommodation facilities is provided by Istat,[3] the Italian national institute for statistics.

11.4.2 Analysis of Venice Waste Data by Spatial Regression with Differential Regularization

Figure 11.10 shows the estimated spatiotemporal field at fixed time instants. The estimate for the coefficient β is 39.7 meaning that one more unit in the ratio between the number of beds in accommodation facilities and the number of residents is estimated to increase the yearly per capita production of waste by residents by about 40 kg. The estimated spatial field f shows the highest values, across the years, in correspondence of the coastline, around the towns of Bibione, Lido di Jesolo, and Cavallino-Treporti. These higher values may be due to a type of tourism that is not captured by the available covariate, such as daily tourists who do not stay overnight, and vacationers who either own or rent vacation houses. The higher values of the field are also probably due to the presence of many seasonal workers, who are not residents of these towns, and are employed in the numerous accommodation facilities, restaurants, cafés, shops, beach resorts, and other services.

2 http://www.dossier.net/utilities/coordinate-geografiche/
3 http://www.istat.it/it/archivio/113712

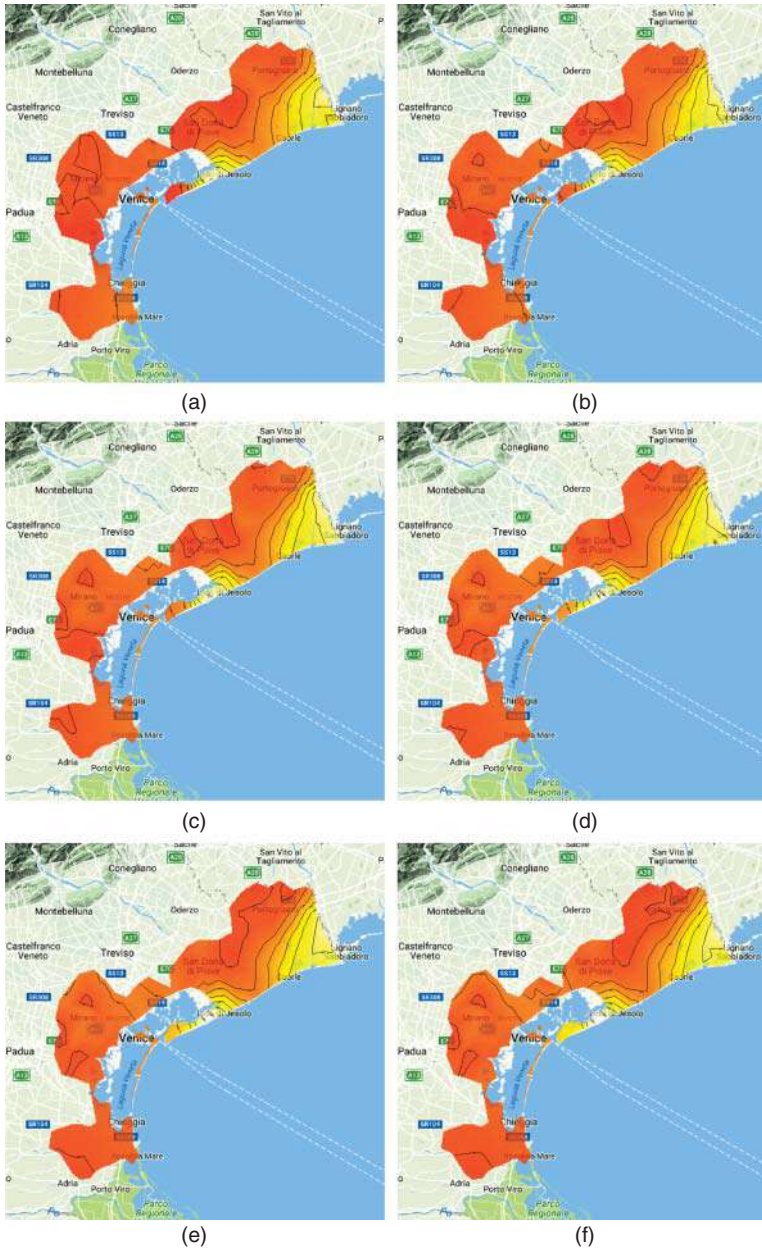


Figure 11.10 Estimated spatiotemporal field for the Venice waste data (yearly per capita production) at fixed time instants. (a) 1997, (b) 2000, (c) 2003, (d) 2006, (e) 2009, and (f) 2011.

Although Venice is one of the most visited cities in Italy, and this tourism is active all year-round, the production of per capita waste in Venice appears to be lower than in other nearby tourist localities by the seaside. This might be partly explained by the fact that the tourist activities in Venice are not so highly characterized by seasonality as in the smaller seaside villages, and people working in tourist activities in Venice are more likely to be themselves residents of this large city.

It is significant to notice how the estimated function does not smooth across concave boundaries. For example, the area of the city of Quarto d'Altino and the one around the city of Cavallino-Treporti show different ranges of values. Indeed, even though the two towns are geographically close, they are separated by the Venetian lagoon. This difference is evident also from the first two panels of Figure 11.11, which shows the estimated spatiotemporal field at fixed localities: Quarto d'Altino,

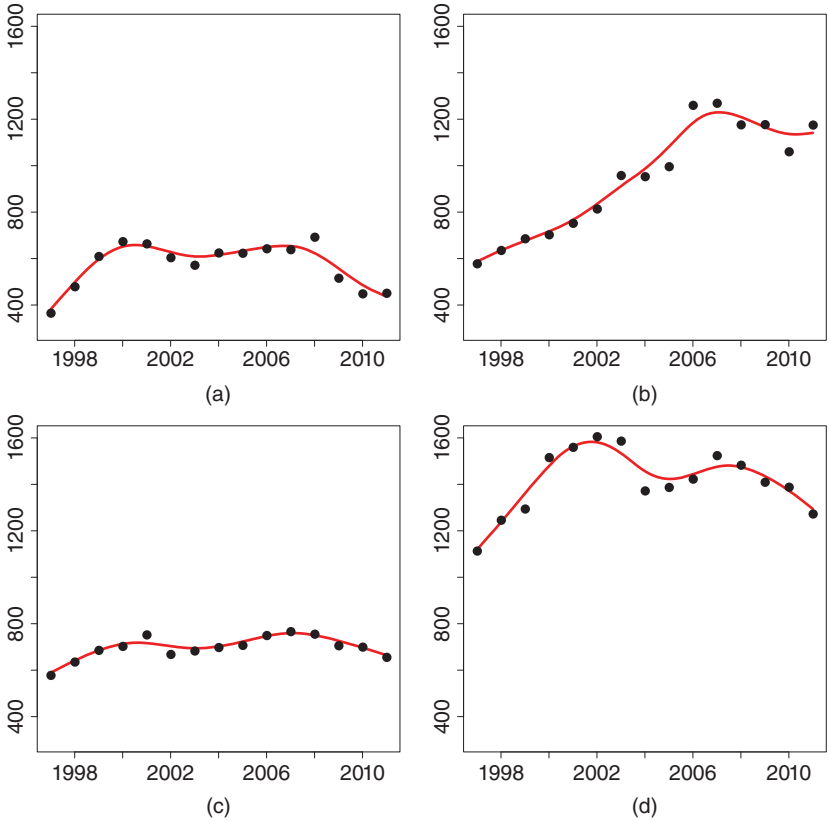


Figure 11.11 Temporal evolution of the estimated spatiotemporal field for the Venice waste data (yearly per capita production) at fixed spatial locations. (a) Quarto d'Altino, (b) Cavallino-Treporti, (c) Venezia, and (d) Bibione. Source: Adapted from Bernardi et al. [2].

Cavallino-Treporti, Venice, and Bibione. In these plots, the dots are obtained subtracting from the data the estimated contribution by the covariate, i.e. $\hat{\beta} w_{ij}$.

The temporal evolution plots in Figure 11.11 show the ability of the method to capture the temporal trend of the phenomenon. The method provides good estimates also for the municipality of Cavallino-Treporti, which presents a strong variation of per capita waste over the year. The large increase of the per capita waste of Cavallino-Treporti is partly explained by the fact that, during the first years of this study, this town was under the jurisdiction of Venice, while the data for this new municipality are available only from 2002. In particular, the data for Cavallino-Treporti for years 1997–2001 are a replicate of the data of the municipality of Venice. Nevertheless, the strong variation in the data is well captured by the estimated function.

11.5 Model Extensions

Various extensions of the model described in this chapter are possible. A first generalization consists in modeling data that are areal in space and integral in time, and estimating an underlying spatiotemporal intensity function. In the application to Venice waste data, if information about the urbanized areas of each municipality would become available, such a model extension would for instance allows to appropriately refer the waste datum to the area and year where it is produced, estimating a spatiotemporal intensity of waste production.

Extending the work of [16], it is also possible to include a priori information available on the phenomenon under study, using more complex differential regularizations modeling the spatial and/or temporal behavior of the phenomenon. This also allows to account for nonstationarities and anisotropies in space and/or time. Along the same lines, if a priori information about the interaction between space and time was available, then it would make sense to consider a unique space/time regularizing term based on a time-dependent PDE that governs the phenomenon behavior. Azzimonti et al. [16], for instance, analyze the blood flow velocity in a section of the carotid artery at a fixed time instant corresponding to the systolic peak, starting from Echo-Color Doppler data, and including a priori information on the problem under study. By introducing the time dimension, we could study how the blood flow velocity field varies during the time of the heartbeat. PDEs are commonly used to describe complex phenomena behavior in many fields of engineering and sciences, including biosciences, geosciences, and physical sciences. Potential applications of particular interest of this space–time technique in the environmental sciences would, for example concern the study of the dispersion of pollutant released in water or in air and transported by streams or winds, and the study of the propagation of earthquakes, tsunamis, and

other wave phenomena. If one wishes instead to consider simpler isotropic and stationary regularizations, then a possibility to allow for stronger interactions in space/time, with respect to the model here presented, would consist in defining a unique regularizing term based on a heat equation.

Finally, data distributed over curved domains, instead of over planar domains, could be handled by extending the model proposed in [23]. Considering the same application presented by Ettinger et al. [23], this would, for instance enable the study of time-dependent hemodynamic forces exerted by blood-flow over the wall of inner carotid arteries affected by aneurysms, taking into account the complex morphology of these vessels. Another fascinating field of application of this modeling extension would be in the neurosciences [24, 25], studying signals associated to neuronal activity over the cortical surface, a highly convoluted thin sheet of neural tissue that constitutes the outermost part of the brain. In the geosciences, this would permit the study of data distributed over regions with complex orographies. Moreover, generalizations to time-dependent data of the spatial regression model introduced by Wilhelm et al. [26] would be particularly well suited for important engineering applications, especially in the automotive, naval, aircraft, and space sectors, where space–time varying quantities of interest are observed over the surface of a designed 3D object, such as the pressure over the surface of a shuttle winglet.

References

1 Sangalli, L.M., Ramsay, J.O., and Ramsay, T.O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (4): 681–703.

2 Bernardi, M.S., Sangalli, L.M., Mazza, G., and Ramsay, J.O. (2017). A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. *Stochastic Environmental Research and Risk Assessment* 31 (1): 23–38.

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

4 Goulard, M. and Voltz, M. (1993). Geostatistical interpolation of curves: a case study in soil science. In: (ed. A. Soares) *Geostatistics Tróia'92. Quantitative Geology and Geostatistics* 5, 805–816. Springer. 10.1007/978-94-011-1739-5_64

5 Nerini, D., Monestiez, P., and Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis* 101 (2): 409–418.

6 Giraldo, R., Delicado, P., and Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics* 18 (3): 411–426.

7 Caballero, W., Giraldo, R., and Mateu, J. (2013). A universal kriging approach for spatial functional data. *Stochastic Environmental Research and Risk Assessment* 27 (7): 1553–1563.

8 Menafoglio, A., Secchi, P., and Dalla Rosa, M. (2013). A universal kriging predictor for spatially dependent functional data of a Hilbert space. *Electronic Journal of Statistics* 7: 2209–2240.

9 Menafoglio, A., Guadagnini, A., and Secchi, P. (2014). A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment* 28 (7): 1835–1851.

10 Ignaccolo, R., Mateu, J., and Giraldo, R. (2014). Kriging with external drift for functional data for air quality monitoring. *Stochastic Environmental Research and Risk Assessment* 28 (5): 1171–1186.

11 Cressie, N. and Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data*. Wiley.

12 Augustin, N.H., Trenkel, V.M., Wood, S.N., and Lorance, P. (2013). Space-time modelling of blue ling for fisheries stock management. *Environmetrics* 24 (2): 109–119.

13 Marra, G., Miller, D.L., and Zanin, L. (2012). Modelling the spatiotemporal distribution of the incidence of resident foreign population. *Statistica Neerlandica* 66 (2): 133–160.

14 Wood, S.N., Bravington, M.V., and Hedley, S.L. (2008). Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (5): 931–955.

15 Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (2): 307–319.

16 Azzimonti, L., Sangalli, L.M., Secchi, P. et al. (2015). Blood flow velocity field estimation via spatial regression with PDE penalization. *Journal of the American Statistical Association* 110 (511): 1057–1071.

17 R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

18 Lila, E., Sangalli, L.M., Ramsay, J., and Formaggia, L. (2016). fdaPDE: Regression with partial differential regularizations, using the finite element method, R package version 0.1-1. https://CRAN.R-project.org/package=fdaPDE (accessed 01 February 2021).

19 Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*. Springer.

20 Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics* 17 (2): 453–510. http://www.jstor.org/stable/2241560 (Retrieved 15 March 2021).

21 Pebesma, E.J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences* 30: 683–691.

22 Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. CRC Press.

23 Ettinger, B., Perotto, S., and Sangalli, L.M. (2016). Spatial regression models over two-dimensional manifolds. *Biometrika* 103 (1): 71–88.

24 Dassi, F., Ettinger, B., Perotto, S., and Sangalli, L.M. (2015). A mesh simplification strategy for a spatial regression analysis over the cortical surface of the brain. *Applied Numerical Mathematics* 90: 111–131.

25 Lila, E., Aston, J.A., and Sangalli, L.M. (2016). Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *The Annals of Applied Statistics* 10 (4): 1854–1879.

26 Wilhelm, M., Dedè, L., Sangalli, L.M., and Wilhelm, P. (2016). IGS: an IsoGeometric approach for Smoothing on surfaces. *Computer Methods in Applied Mechanics and Engineering*. https://doi.org/10.1016/j.cma.2015.12.028.

12

Quasi-maximum Likelihood Estimators for Functional Linear Spatial Autoregressive Models

Mohamed-Salem Ahmed[1], Laurence Broze[3], Sophie Dabo-Niang[2], and Zied Gharbi[4]

[1] *University of Lille, CHU Lille,ULR 2694 - METRICS, Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France*
[2] *University of Lille, CNRS,UMR 8524 - Laboratoire Paul Painlevé, INRIA-MODAL, F-59000 Lille, France*
[3] *University of Lille, CNRS,UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France*
[4] *University of Lille, CNRS,UMR 9221 - Laboratoire LEM, F-59000 Lille, France*

12.1 Introduction

In this chapter, we address two research areas: spatial statistics and functional data analysis. Spatial functional random variables are becoming more common in statistical analyses due to the availability of high-frequency spatial data and new mathematical strategies to address such statistical objects.

Many fields, such as urban systems, agriculture, environmental science, and economics, often consider spatially dependent data. Therefore, modeling spatial dependency in statistical inferences (estimation of the spatial distribution, regression, prediction, among others) is a significant feature of spatial data analysis. Spatial statistics provides tools to solve such modeling. Various spatial models and methods have been proposed, particularly within the scope of geostatistics or lattice data. Most of the spatial modeling methods are parametric and concern nonfunctional data.

Several types of functional linear models (FLMs) for independent data have been developed for different purposes. The most studied model is perhaps the FLM for scalar response, originally introduced by Hastie and Mallows [1]. Estimation and prediction problems for this model and some of its generalizations have been reported mainly for independent data (see, e.g. [2–5]). Some research exists on functional spatial linear prediction using kriging methods (see, e.g. [6–12],…), highlighting the interest in considering spatial linear functional models.

Complex issues arise in spatial econometrics (statistical techniques to address economic modeling), many of which are neither clearly defined nor completely

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

resolved but form the basis for current research. Among the practical considerations that influence the available techniques used in spatial data and geostatistics modeling, particularly in econometrics, are data dependency. This is the case, for instance, in images analysis, remote sensing from satellites, agriculture, climatology, environmental monitoring, or geology, where data are often dependent, and a spatial model must be able to account for this characteristic. Linear spatial models, which are common in geostatistical modeling, generally impose a dependency structure model based on linear covariance relationships between spatial locations. However, under many circumstances, the spatial index does not vary continuously over a subset of $\mathbb{R}^N$, $N \geq 2$ and may be of the lattice type, the baseline of this current work. This is, for instance, the case in a number of problems. In images analysis, remote sensing from satellites, agriculture, and so on, data are often received as regular lattice and identified as the centroids of square pixels, whereas a mapping forms often an irregular lattice. Basically, statistical models for lattice data are linked to nearest neighbors to express the fact that data are nearby.

We are concerned here about functional (functional covariates) spatial models for lattice data. One of the well-known spatial lattice models is the spatial autoregressive (SAR) model of [13], which extends regression in time series to spatial data. This model has been extensively studied and extended in several ways in the case of real-valued data, compared to the functional framework.

SAR models for real-valued data proposed in the literature, closest to models with functional variables are those with functional autoregressive coefficients. Namely, these models are based on nonparametric spatial interactive structures. Sun [14] proposed a nonparametric approach to estimate a functional coefficient SAR model with the spatial dependence described by an unknown smooth function of geographic distances. In this model, the relation between the covariate and the response is also nonparametric. In the same spirit, Koroglu and Sun [15] provided an two-stage least squares (2SLS) estimation method for an upgraded version of [14]'s model with a spatial dependence in the explanatory variable. Sun and Malikov [16] extended the SAR fixed effect panel data model of [17] to a functional coefficient SAR model for panel data with fixed effects.

So far the literature on autoregressive spatial models with functional variables is very limited. Ruiz-Medina [18, 19] considered a spatial autoregressive Hilbertian(SARH(1)) processes, where the autoregressive part is given in terms of three functional random components located in three points defining the boundary between some notions of past and future. Recently, Pineda-Ríos and Giraldo [20] studied FLMs with real-valued response and a functional covariate, with SAR disturbances.

The structure of SAR models for real-valued data, its identification, and estimation, among others, 2SLS [21, 22], maximum likelihood (ML) [23], and generalized

method of moments (GMM) [24] have been developed. The identification and estimation of SAR models by quasi-maximum likelihood (QML) are limited. Lee [25] and more recently [26], proposed QML estimators for a SAR model with a spatial dependency structure based on a spatial weights matrix. The quasi-maximum likelihood estimator (QMLE) is appropriate when the disturbances in the considered model are not normally distributed. In the literature on SAR models for real-valued data, the QMLE and maximum likelihood estimator (MLE) are proved to be computationally challenging, consistent with rates of convergence depending on the spatial weights matrix of the considered model [25, 26].

The present work considers an estimation of a functional spatial linear model with a random functional covariate and a real-valued response using spatial autoregression on the response based on a weight matrix. We investigate parameter identification and asymptotic properties of the QMLE estimator using the so-called *increasing domain asymptotics*. We provide identification conditions combining identification in the classical SAR model and identification in the FLM. Monte Carlo experiments illustrate the performance of the QML estimation.

The rest of this chapter is organized as follows. In Section 12.2, we provide the functional spatial autoregressive (FSAR) model and its QMLE. In Section 12.3, we state the consistency and asymptotic normality of the estimator. To check the performance of the estimator, numerical results are reported in Section 12.3 using different spatial scenarios, where each unit is influenced by neighboring units. Proofs and technical lemmas are given in the Appendix.

12.2 Model

We consider that at n spatial units located on $\mathcal{I}_n$, a finite subset of cardinal n of a regular or irregularly spaced, countable lattice $\mathcal{I} \subset \mathbb{R}^N$, we observe a real-valued random variable Y considered as the *response variable* and a functional covariate $\{X(t), t \in \mathcal{T}\}$, a square-integrable stochastic process on the interval $\mathcal{T} \subset \mathbb{R}$. Assume that the process $\{X(t), t \in \mathcal{T}\}$ takes values in space $\mathcal{X} \subset L^2(\mathcal{T})$, where $L^2(\mathcal{T})$ is the space of square-integrable functions in $\mathcal{T}$. The spatial dependency structure between these n spatial units is described by an $n \times n$ nonstochastic spatial weights matrix W_n that depends on n. The elements $w_{ij} = w_{ij,n}$ of this matrix are usually considered as inversely proportional to the distance between spatial units i and j with respect to some metric (physical distance, social network, or economic distance, see for instance [27]). Since the weight matrix changes with n, we consider these observations as triangular array observations. This is required to conduct an asymptotic study of the following model that describes the relationship between the response variable Y and the covariate function $X(\cdot)$ [28].

There are mainly three different types of interaction effects that may explain why an observation associated with a specific location may be dependent on observations at other locations:

- Endogenous interaction effects, where the variable Y at some spatial unit depends on values of Y taken by other spatial units.
- Exogenous interaction effects, where the variable Y at some spatial unit depends on independent explanatory variables at other spatial units.
- Correlated effects, where similar unobserved characteristics result in similar behavior.

In this work, we assume that the relationship between Y and X follows a FSAR model with endogenous interactions:

$$Y_i = \lambda_0 \sum_{j=1}^{n} w_{ij} Y_j + \int_{\mathcal{T}} X_i(t)\beta^*(t)dt + U_i, \quad i = 1, \dots, n, \quad n = 1, 2, \dots, \tag{12.1}$$

where the autoregressive parameter λ_0 is in compact space Λ, $\beta^*(\cdot)$ is a functional parameter assumed to belong to the space of functions $L^2(\mathcal{T})$, and $(w_{ij})_{j=1,\dots,n}$ is the i-th row of W_n.

Note that [20] introduced a FSAR model with correlated effects, meaning that the spatial dependence is in the disturbance term. They proposed the following SAR process:

$$Y_i = \int_{\mathcal{T}} X_i(t)\beta^*(t)dt + U_i, \quad i = 1, \dots, n, \quad n = 1, 2, \dots, \tag{12.2}$$

where the vector of disturbances $U = \{U_i,\ i = 1, \dots, n,\ n = 1, 2, \dots\}$ verifies $U = \rho W_n U + \varepsilon$, the vector $\varepsilon = \{\varepsilon_i,\ i = 1, \dots, n,\ n = 1, 2, \dots\}$ is composed of independent Gaussian random variables.

Assume that $w_{ij} = O(h_n^{-1})$ uniformly in all i, j, where the rate sequence h_n can be bounded or divergent, such as $h_n = o(n)$. This kind of matrix can be obtained by Nearest Neighbor weights.

In practice, it is common, but not necessary, to *row normalize* the spatial weight matrix so in each row-normalized weight, $0 \leqslant w_{ij} \leqslant 1$ can be interpreted as the *fraction* of all spatial influence on unit i attributable to unit j. In general, these matrices W_n can be classified into two groups: *Weights Based on Distance* and *Weights Based on Boundaries*.

For Weights Based on Distance, one way to construct spatial weight matrices is to use the distance d_{ij} between each pair of spatial units (regions, cities, centroids,…) i and j.

- *k-Nearest Neighbor weights*

$$w_{ij} = \begin{cases} 1 & \texttt{if } j \in N_k(i) \\ 0 & \texttt{Otherwise} \end{cases},$$

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

- *Radial Distance weights*

$$w_{ij} = \begin{cases} 1 & \texttt{if } 0 \le d_{ij} \le \delta \\ 0 & \texttt{if } d_{ij} > \delta \end{cases},$$

where δ is a critical distance (*threshold distance* or *bandwidth*) cut-off after which spatial effects are considered to be negligible, it should be able to guarantee that each region has at least one neighbor.

- *Power Distance Decay weights*

$$w_{ij} = \begin{cases} d_{ij}^{-\alpha} & \texttt{if } 0 \le d_{ij} \le \delta \\ 0 & \texttt{if } d_{ij} > \delta \end{cases}$$

where α is any positive exponent, typically $\alpha = 1$ or $\alpha = 2$.

- *Exponential Distance Decay weights*

$$w_{ij} = \begin{cases} \exp(-\alpha d_{ij}) & \texttt{if } 0 \le d_{ij} \le \delta \\ 0 & \texttt{if } d_{ij} > \delta \end{cases}$$

- *Double-Power Distance weights*

$$w_{ij} = \begin{cases} \left[1 - (d_{ij}/\delta)\right]^k & \texttt{if } 0 \le d_{ij} \le \delta \\ 0 & \texttt{if } d_{ij} > \delta \end{cases},$$

with k is a positive integer, typically $k = 2$, $k = 3$, or $k = 4$.

- *Cliff–Ord weights*
  Cliff and Ord [13] suggested to use the length of the common border between contiguous regions, weighted by a distance function:

$$w_{ij} = d_{ij}^{-a} D_{ij}^{b}$$

where D_{ij} is the share of common boundary between i and j, a and b are parameters estimated from data or chosen a priori.

- *Block structure*
  In this case, $w_{ij} = 1$ for all i and j in the same block, and the blocks are defined according to some specific criterion.

For weights based on boundaries, spatial contiguity is often used to specify neighboring location in the sense of sharing a common border. There are different type of spatial contiguity, but the classical cases are those referred to *Rook contiguity* (with only common boundaries), *Bishop contiguity* (with only common vertices), and *Queen contiguity* (with both Rook and Bishop contiguity).

$$w_{ij} = \begin{cases} 1 & \texttt{if } i \texttt{ and } j \texttt{ are contiguity} \\ 0 & \texttt{Otherwise} \end{cases}$$

In general, we can rewrite the last equation as follows:

$$w_{ij} = \begin{cases} 1 & \ell_{ij} > 0 \\ 0 & \ell_{ij} = 0 \end{cases},$$

with ℓ_{ij} denotes the length of shared boundary.

The disturbances $\{U_i,\ i = 1, \dots, n,\ n = 1, 2, \dots\}$ are assumed to be independent Gaussian random variables such that $E(U_i) = 0$, $E(U_i^2) = \sigma_0^2$. They are also independent of $\{X_i(t), t \in \mathcal{T},\ i = 1, \dots, n,\ n = 1, 2, \dots\}$.

We are interested in estimating the unknown true parameters λ_0, $\beta^*(\cdot)$, and σ_0^2. Let $\mathbf{X}_n(\beta^*(\cdot))$ be the $n \times 1$ vector of i-th element $\int_{\mathcal{T}} X_i(t)\beta^*(t)dt$; then, one can rewrite (12.2) as follows:

$$S_n \mathbf{Y}_n = \mathbf{X}_n(\beta^*(\cdot)) + \mathbf{U}_n\ , \quad n = 1, 2, \dots \tag{12.3}$$

where $S_n = (I_n - \lambda_0 W_n)$, $\mathbf{Y}_n$, and $\mathbf{U}_n$ are two $n \times 1$ vectors of elements Y_i and U_i, $i = 1, \dots, n$, respectively, and I_n denotes the $n \times n$ identity matrix.

Let $S_n(\lambda) = I_n - \lambda W_n$, so the conditional log-likelihood function of the vector $\mathbf{Y}_n$, given $\{X_i(t), t \in \mathcal{T},\ i = 1, \dots, n,\ n = 1, 2, \dots\}$ is

$$\begin{aligned} L_n(\lambda, \beta(\cdot), \sigma^2) = &-\frac{n}{2}\ \ln\ \sigma^2 - \frac{n}{2}\ \ln(2\pi) + \ln|S_n(\lambda)| \\ &- \frac{1}{2\sigma^2}\left[S_n(\lambda)\mathbf{Y}_n - \mathbf{X}_n(\beta(\cdot))\right]'\ \left[S_n(\lambda)\mathbf{Y}_n - \mathbf{X}_n(\beta(\cdot))\right], \end{aligned} \tag{12.4}$$

where A' denotes the transpose of matrix A.

Maximizing (12.4) with respect to λ, $\beta(\cdot)$, and σ^2 will give us the QMLE estimation of λ_0, $\beta^*(\cdot)$, and σ_0^2. But this likelihood cannot be maximized without addressing the difficulty produced by the infinite dimensionality of the explanatory random function. To solve this problem, we project, as usual, the functional explanatory variable and parameter function into the space of the functions generated by a basis of functions with dimensions that increase asymptotically as the sample size tends to infinity. Several truncation techniques exist. Cardot et al. [29] proposed the estimated eigenbasis of the sample; Cardot and Sarda [30] considered a Spline basis, adding a penalty that controls the degree of smoothness of the parameter function. Müller and Stadtmüller [31] proposed the use of any basis of functions that verifies some truncation criterion. We adapt the alternative proposed by Müller and Stadtmüller [31] to solve the infinite dimension problem of the functional space. This method is denoted *truncated conditional likelihood method*.

12.2.1 Truncated Conditional Likelihood Method

Let $\{\varphi_j,\ j = 1, 2, \dots\}$ be an orthonormal basis of the functional space $L^2(\mathcal{T})$, usually a Fourier or a Spline basis or a basis constructed by the eigenfunctions of the covariance operator Γ, defined by

$$\Gamma x(t) = \int_{\mathcal{T}} E(X(t)X(s))x(s)ds\ , \quad x \in \mathcal{X},\ t \in \mathcal{T}. \tag{12.5}$$

Using an expansion on this orthonormal basis, we can write $X(\cdot)$ and $\beta^*(\cdot)$ as follows:

$$X(t) = \sum_{j\geq 1} \varepsilon_j \varphi_j(t) \quad \text{and} \quad \beta^*(t) = \sum_{j\geq 1} \beta_j^* \varphi_j(t) \quad \text{for all } t \in \mathcal{T},$$

where the real random variables ε_j and the coefficients β_j^* are given by

$$\varepsilon_j = \int_{\mathcal{T}} X(t)\varphi_j(t)dt \quad \text{and} \quad \beta_j^* = \int_{\mathcal{T}} \beta^*(t)\varphi_j(t)dt.$$

Let p_n be a positive sequence of integers that increase asymptotically as $n \to \infty$; by the orthonormality of the basis $\{\varphi_j,\ j = 1, 2, \dots\}$, we can consider the following decomposition:

$$\int_{\mathcal{T}} X(t)\beta^*(t)dt = \sum_{j=1}^{\infty} \beta_j^* \varepsilon_j = \sum_{j=1}^{p_n} \beta_j^* \varepsilon_j + \sum_{j=p_n+1}^{\infty} \beta_j^* \varepsilon_j. \tag{12.6}$$

The truncation strategy introduced by Müller and Stadtmüller [31] consists of approximating the left-hand side in (12.6) using only the first term of the right-hand side. This is possible when the approximation error vanishes asymptotically, where this error is controlled by a square expectation of the second term on the right-hand side of (12.6). In particular, the approximation error vanishes asymptotically when one considers the eigenbasis of the variance-covariance operator Γ by remarking that

$$E\left(\sum_{j=p_n+1}^{\infty} \beta_j^* \varepsilon_j\right)^2 = \sum_{j=p_n+1}^{\infty} \beta_j^{*^2} E\left(\varepsilon_j^2\right) = \sum_{j=p_n+1}^{\infty} \beta_j^{*^2} \delta_j$$

where $\delta_j, j = 1, 2, \dots$ are the eigenvalues. Under this truncation strategy, $\mathbf{X}_n(\beta^*(\cdot))$ may be approximated by $\xi_{p_n}\beta^*$, where $\beta^* = (\beta_1^*, \dots, \beta_{p_n}^*)'$ and ξ_{p_n} is an $n \times p_n$ matrix of the (i,j)-th element given by

$$\varepsilon_j^{(i)} = \int_{\mathcal{T}} X_i(t)\varphi_j(t)dt\,, \qquad i = 1, \dots, n\ \ j = 1, \dots, p_n.$$

Now, the truncated conditional log-likelihood function can be obtained by replacing in (12.4), $\mathbf{X}_n(\beta(\cdot))$ with $\xi_{p_n}\beta$ for all $\beta(\cdot) \in L^2(\mathcal{T})$ and $\beta \in \mathbb{R}^{p_n}$. The corresponding and feasible log conditional likelihood is

$$\begin{aligned}\tilde{L}_n(\lambda, \beta, \sigma^2) = &-\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ \ln(2\pi) + \ln|S_n(\lambda)| \\ &-\frac{1}{2\sigma^2}\left[S_n(\lambda)\mathbf{Y}_n - \xi_{p_n}\beta\right]'\left[S_n(\lambda)\mathbf{Y}_n - \xi_{p_n}\beta\right].\end{aligned} \tag{12.7}$$

For a fixed λ, (12.7) is maximized at

$$\hat{\beta}_{n,\lambda} = (\xi'_{p_n}\xi_{p_n})^{-1}\xi'_{p_n}S_n(\lambda)\mathbf{Y}_n = (\hat{\beta}_{nj,\lambda})_{j=1,\dots,p_n} \tag{12.8}$$

and

$$\hat{\sigma}^2_{n,\lambda} = \frac{1}{n}\left(S_n(\lambda)\mathbf{Y}_n - \xi_{p_n}\hat{\beta}_{n,\lambda}\right)'\left(S_n(\lambda)\mathbf{Y}_n - \xi_{p_n}\hat{\beta}_{n,\lambda}\right)$$
$$= \frac{1}{n}\mathbf{Y}'_n S'_n(\lambda) M_n S_n(\lambda)\mathbf{Y}_n, \tag{12.9}$$

where $M_n = I_n - \xi_{p_n}(\xi'_{p_n}\xi_{p_n})^{-1}\xi'_{p_n}$.

The concentrated truncated conditional log-likelihood function of λ is

$$\tilde{L}_n(\lambda) = -\frac{n}{2}(\ln(2\pi) + 1) - \frac{n}{2}\ \ln\hat{\sigma}^2_{n,\lambda} + \ln|S_n(\lambda)|. \tag{12.10}$$

Then the estimator of λ_0 is $\hat{\lambda}_n$, which maximizes $\tilde{L}_n(\lambda)$, and those of the vector β^* and variance σ_0^2 are, respectively, $\hat{\beta}_{n,\hat{\lambda}_n}$, $\hat{\sigma}^2_{n,\hat{\lambda}_n}$. The corresponding estimator of the function parameter $\beta^*(\cdot)$ is

$$\hat{\beta}_n(t) = \sum_{j=1}^{p_n}\hat{\beta}_{nj,\hat{\lambda}_n}\varphi_j(t).$$

The estimation of the model is given, we focus on the asymptotics results in Section 12.3.

For that purpose, we need to define some asymptotic method. There are two main asymptotic methods in the spatial literature: increasing domain and infill asymptotics (see [32], p. 480). The growth of the sample in increasing domain asymptotics is a consequence of an unbounded expansion of the sample region, whereas under infill asymptotics, the sample region is fixed and the growth of the sample size is due to dense sampling in the considered region. In the following, we consider increasing domain asymptotics.

12.3 Results and Assumptions

Let us first state some combining condition assumptions related to the spatial dependency structure and assumptions on the functional nature of the data.

Let $I_n + \lambda_0 G_n = S_n^{-1}$, where $G_n = W_n S_n^{-1}$, $B_n(\lambda) = S_n(\lambda)S_n^{-1} = I_n + (\lambda_0 - \lambda)G_n$ for all $\lambda \in \Lambda$ and $A_n(\lambda) = B_n^{'}(\lambda)B_n(\lambda)$.

We assume that

Assumption 12.1

i. The matrix S_n is nonsingular.
ii. The sequences of matrices $\{W_n\}$ and $\{S_n^{-1}\}$ are uniformly bounded in both row and column sums.
iii. The matrices $\{S_n^{-1}(\lambda)\}$ are uniformly bounded in either row or column sums and uniformly bounded in λ, in compact parameter space Λ. The true λ_0 is in the interior of Λ.

Assumption 12.2 The sequence p_n satisfies $p_n \to \infty$ and $p_n n^{-1/4} \to 0$ as $n \to \infty$, and

i. $p_n \sum_{r_1, r_2 > p_n} E(\varepsilon_{r_1} \varepsilon_{r_2}) = o(1)$
ii. $\sum_{r_1, \ldots, r_4 > p_n} E(\varepsilon_{r_1} \ldots \varepsilon_{r_4}) = o(1)$
iii. $\sqrt{n} \sum_{s=1}^{p_n} \sum_{r_1, r_2 > p_n} E(\varepsilon_s \varepsilon_{r_1}) E(\varepsilon_s \varepsilon_{r_2}) = o(1)$.

Remark 12.1 ***Assumption 12.1****-i ensures that* $\mathbf{Y}_n$ *has mean* $S_n^{-1}\mathbf{X}_n(\beta^*(\cdot))$ *and variance* $\sigma_0^2 S_n^{-1} S_n^{'-1}$. *The uniform boundedness of* W_n *and* S_n^{-1} *in* ***Assumption 12.1****-ii enables the control of the degree of spatial correlation and plays an important role in the asymptotic properties of the estimators. By easy computation, one can show under this assumption that the matrix* $G_n = W_n S_n^{-1}$ *is uniformly bounded in both row and column sums together with elements of order* h_n^{-1}. *Consequently, the matrix* $A_n(\lambda) = B_n^{'}(\lambda) B_n(\lambda)$ *has a trace of order n uniformly in* $\lambda \in \Lambda$ *by the compactness condition of* Λ *in* ***Assumption 12.1****-iii.* ***Assumption 12.1****-iii makes it possible to address the nonlinearity of* $\ln |S_n(\lambda)|$ *as a function of* λ *in (*12.7*). For more detail and a discussion of* ***Assumption 12.1****, see [25].* ***Assumption 12.3*** *considers the rate of convergence of* p_n *with respect to n. Condition iii of Assumption 12.2 is satisfied when one consider the eigenbasis, since in this case* $E(\varepsilon_r \varepsilon_s) = 0$, *for* $s \neq r$.

To obtain the identifiability of λ_0, β^*, and σ_0^2 in the truncated model, remark that

$$E\left(\tilde{L}_n(\lambda, \beta, \sigma^2)\right) = -\frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi) + \ln |S_n(\lambda)| - \frac{1}{2\sigma^2} E\left(\left[S_n(\lambda)\mathbf{Y}_n - \xi_{p_n}\beta \right]' \left[S_n(\lambda)\mathbf{Y}_n - \xi_{p_n}\beta \right] \right).$$

We have

$$E\left(\left[S_n(\lambda)\mathbf{Y}_n - \xi_{p_n}\beta \right]' \left[S_n(\lambda)\mathbf{Y}_n - \xi_{p_n}\beta \right] \right) = E\left(\left[B_n(\lambda)\xi_{p_n}\beta^* - \xi_{p_n}\beta \right]' \left[B_n(\lambda)\xi_{p_n}\beta^* - \xi_{p_n}\beta \right] \right) + E\left(R_n' A_n(\lambda) R_n \right) + \sigma_0^2 \mathrm{tr}\left(A_n(\lambda) \right) + 2E\left(\left[B_n(\lambda)\xi_{p_n}\beta^* - \xi_{p_n}\beta \right]' B_n(\lambda)\mathbf{R}_n \right),$$

where $\mathbf{R}_n = (R_1, \ldots, R_n)'$ with $R_i = \sum_{j > p_n} \beta_j^* \varepsilon_j^{(i)}$. Let R denote the generic copy of R_i, $i = 1, \ldots, n$, where $E(R) = 0$.

We then have

$$E\left(\beta^{*\prime} \xi_{p_n}' B_n(\lambda)\mathbf{R}_n \right) = \mathrm{tr}\left(B_n(\lambda) \right) \epsilon_{n1}, \quad \text{where} \quad \epsilon_{n1} = \sum_{r=1}^{p_n} \sum_{s > p_n} \beta_r \beta_s^* E(\varepsilon_r \varepsilon_s),$$

$$E\left(\beta'\xi'_{p_n}A_n(\lambda)\mathbf{R}_n\right) = \operatorname{tr}\left(A_n(\lambda)\right)\epsilon_{n2}, \quad \text{where } \epsilon_{n2} = \sum_{r=1}^{p_n}\sum_{s>p_n}\beta_r^*\beta_s^*E(\varepsilon_r\varepsilon_s),$$

$$E\left(\mathbf{R}'_nA_n(\lambda)\mathbf{R}_n\right) = \operatorname{tr}(A_n(\lambda))\epsilon_{n3}, \quad \text{where } \epsilon_{n3} = E(R^2).$$

Note that $\epsilon_{n1}, \epsilon_{n2}$, and ϵ_{n3} are of order $o(1)$ by **Assumption 12.2**, and they are independent of λ. In addition, ϵ_{n1} and ϵ_{n2} are null if one uses the eigenbasis.

Consequently,

$$\begin{aligned}
&E\left(\tilde{L}_n(\lambda,\beta,\sigma^2)\right)\\
&\quad = -\frac{1}{2\sigma^2}\,E\left(\left(B_n(\lambda)\xi_{p_n}\beta^* - \xi_{p_n}\beta\right)'\left(B_n(\lambda)\xi_{p_n}\beta^* - \xi_{p_n}\beta\right)\right)\\
&\qquad + \ln|S_n(\lambda)| - \frac{n}{2}\left(\ln\,\sigma^2 + \ln 2\pi\right) - \frac{\sigma_0^2}{2\sigma^2}\,\operatorname{tr}\left(A_n(\lambda)\right)\\
&\qquad + \epsilon_{n1}\,\operatorname{tr}\left(B_n(\lambda)\right) + \epsilon_{n4}\,\operatorname{tr}\left(A_n(\lambda)\right),
\end{aligned} \tag{12.11}$$

with $\epsilon_{n4} := \epsilon_{n2} + \epsilon_{n3}$. Note that the terms that contain ϵ_{n1} and ϵ_{n4} are negligible with respect to the others.

For fixed λ, the expectation $E\left(\tilde{L}_n(\lambda,\beta,\sigma^2)\right)$ is maximum with respect to β and σ^2 at

$$\beta^*_{n,\lambda} = \frac{1}{n}\Gamma^{-1}_{p_n}E\left(\xi'_{p_n}B_n(\lambda)\xi_{p_n}\right)\beta^* = \beta^* + (\lambda_0-\lambda)\beta^*\frac{1}{n}\,\operatorname{tr}\left(G_n\right)$$

and

$$\begin{aligned}
\sigma^{*2}_{n,\lambda} &= \frac{1}{n}E\left(\left[B_n(\lambda)\xi_{p_n}\beta^* - \xi_{p_n}\beta^*_{n,\lambda}\right]'\left[B_n(\lambda)\xi_{p_n}\beta^* - \xi_{p_n}\beta^*_{n,\lambda}\right]\right)\\
&\qquad + \frac{\sigma_0^2}{n}\,\operatorname{tr}\left(A_n(\lambda)\right)\\
&= (\lambda_0-\lambda)^2\frac{1}{n}\Delta_n + \frac{\sigma_0^2}{n}\,\operatorname{tr}\left(A_n(\lambda)\right),
\end{aligned} \tag{12.12}$$

with $\Delta_n = n\left(\operatorname{tr}\left(\frac{G'_nG_n}{n}\right) - \operatorname{tr}^2\left(\frac{G_n}{n}\right)\right)\beta^{*'}\Gamma_{p_n}\beta^*$ since

$$E\left(\xi'_{p_n}G'_nG_n\xi_{p_n}\right) = \operatorname{tr}(G'_nG_n)\Gamma_{p_n} \text{ and } E\left(\xi'_{p_n}G_n\xi_{p_n}\right) = \operatorname{tr}(G_n)\Gamma_{p_n},$$

where $\Gamma_{p_n} = E\left(\frac{1}{n}\xi'_{p_n}\xi_{p_n}\right)$ is assumed to be positive definite. This is the case when the eigenbasis is considered in the truncation strategy.

Based on these results, it is clear that $\beta^*_{n,\lambda_0} = \beta^*$ and $\sigma^{*2}_{n,\lambda_0} = \sigma_0^2$. Hence, the identifiability of β^* and σ_0^2 depends on that of λ_0. Note that

$$\begin{aligned}
Q_n(\lambda) &= E\left(\tilde{L}_n\left(\lambda,\beta^*_\lambda,\sigma^{*2}_{n,\lambda}\right)\right)\\
&= \ln|S_n(\lambda)| - \frac{n}{2}\,\ln\sigma^{*2}_{n,\lambda} - \frac{n}{2}\left(1+\ln(2\pi)\right)\\
&\quad + \epsilon_{n1}\,\operatorname{tr}\left(B_n(\lambda)\right) + \epsilon_{n4}\,\operatorname{tr}\left(A_n(\lambda)\right).
\end{aligned}$$

Therefore, proving the identifiability of λ_0 is equivalent to showing that λ_0 maximizes $Q_n(\lambda)$. This will be proved before addressing the consistency of the estimators.

We will need to compose an additional assumption

Assumption 12.3 Let $\lim_{n\to\infty}\frac{1}{n}\Delta_n = c$, where (a) $c > 0$; (b) $c = 0$. Under the later condition,

$$\lim_{n\to\infty}\frac{h_n}{n}\left\{\ln\left|\sigma_0^2 S_n^{-1}S_n'^{-1}\right| - \ln\left|\sigma_{n,\lambda}^2 S_n^{-1}(\lambda)S_n'^{-1}(\lambda)\right|\right\} \neq 0,$$

whenever $\lambda \neq \lambda_0$, with $\sigma_{n,\lambda}^2 = \frac{\sigma_0^2}{n}\,\mathrm{tr}(A_n(\lambda))$.

Assumption 12.4 U_i, $i = 1, \ldots, n$ in $\mathbf{U}_n = (U_1, \ldots, U_n)'$ are independent and identically distributed (i.i.d.) with mean zero and variance σ_0^2. The moment $E\left(|U_i|^{4+\delta}\right)$ exists for some $\delta > 0$. Let $\mu_4 = E(U_i^4)$.

Remark 12.2 ***Assumption 12.3*** *enables the identification of λ_0 according to the boundless of h_n. It is similar to that used in [25] in the case of multivariate deterministic covariates. This assumption ensures that* $\mathrm{tr}^2(G_n/n)$ *is dominated by* $\mathrm{tr}(G_n'G_n/n)$*, which is the case when $h_n \to \infty$, as under* ***Assumption 12.1****,* $\mathrm{tr}(G_n'G_n)$ *and* $\mathrm{tr}(G_n)$ *are of order $O(n/h_n)$. Situation (b) is related to the existence of a unique variance of $\mathbf{Y}_n$.* ***Assumption 12.4*** *characterizes the properties of the disturbance term.*

Under assumptions similar to those used in [25] but adapted to the functional context, we show that the proposed QMLE estimator has the same asymptotic properties as those in the context of independent data (see e.g. [31]) and the spatial model with real-valued covariates (see e.g. [25]). The following theorems give the identification, consistency, and asymptotic normality results of the autoregressive, functional, and variance parameters estimates.

Theorem 12.1 *Under* ***Assumptions 12.1–12.4*** *and $h_n^4 = O(n)$ for divergent h_n, the QMLE $\hat{\lambda}_n$ derived from the maximization of $\tilde{L}_n(\lambda)$ is consistent and satisfies*

$$\sqrt{\frac{n}{h_n}}(\hat{\lambda}_n - \lambda_0) \to \mathcal{N}(0, s_\lambda^2),$$

with $s_\lambda^2 = \lim_{n\to\infty}\frac{s_n^2 h_n}{n}\left\{\frac{h_n}{n}\left[\Delta_n + \sigma_0^2\mathrm{tr}(G_n(G_n' + G_n))\right]\right\}^{-2}$, *where*

$$\begin{aligned} s_n^2 = {} & \sigma_0^2\left[\beta^{*\prime}\Gamma_{p_n}\beta^* + \sigma_0^2\right]\ \mathrm{tr}\left(G_n(G_n' + G_n)\right) \\ & + \left[3\sigma_0^2\beta^{*\prime}\Gamma_{p_n}\beta^* + \sigma_0^4 - \mu_4\right]\frac{1}{n}\mathrm{tr}^2(G_n) \\ & + \left[\mu_4 - 3\sigma_0^4 - \sigma_0^2\beta^{*\prime}\Gamma_{p_n}\beta^*\right]\sum_{i=1}^{n} G_{ii}^2. \end{aligned} \tag{12.13}$$

Note that when h_n is divergent, the last two terms in (12.13) are of order $o(1)$.

Theorem 12.2 *Under assumptions of Theorem 12.1, $\hat{\sigma}_n^2$ is a consistent estimator of σ_0^2 and satisfies*

$$\sqrt{n}\left(\hat{\sigma}^2_{n,\hat{\lambda}_n} - \sigma_0^2\right) \to \mathcal{N}\left(0, s_\sigma^2\right)$$

$$\text{with} \qquad s_\sigma^2 = \mu_4 - \sigma_0^4 + 4s_\lambda^2 \lim_{n\to\infty} h_n \left[\frac{\mathrm{tr}(G_n)}{n}\right]^2.$$

When h_n is divergent, s_σ^2 will be reduced to $\mu_4 - \sigma_0^4$.

The following assumptions are needed to ensure the asymptotic property of the parameter function estimator. They are similar to ones used in [31].

Assumption 12.5 We have

$$\sum_{r_1,r_2,r_3,r_4=0}^{p_n} E\left(\varepsilon_{r_1}\varepsilon_{r_2}\varepsilon_{r_3}\varepsilon_{r_4}\right) v_{r_1r_2}v_{r_3r_4} = o(n/p_n^2),$$

where the v_{kl}, $k,l = 1,\ldots,p_n$, are the elements of $\Gamma_{p_n}^{-1}$.

Assumption 12.6 We assume that

$$\sum_{r_1,\ldots,r_8=0}^{p_n} E\left(\varepsilon_{r_1}\varepsilon_{r_3}\varepsilon_{r_5}\varepsilon_{r_7}\right) E\left(\varepsilon_{r_2}\varepsilon_{r_4}\varepsilon_{r_6}\varepsilon_{r_8}\right) v_{r_1r_2}v_{r_3r_4}v_{r_5r_6}v_{r_7r_8} = o(n^2p_n^2).$$

The asymptotic normality of the parameter function estimator is given in the following theorem:

Theorem 12.3 *Under **Assumptions 12.1–12.6**, we have*

$$\frac{n\left(\hat{\beta}_{n,\hat{\lambda}_n} - \beta^*\right)'\Gamma_{p_n}\left(\hat{\beta}_{n,\hat{\lambda}_n} - \beta^*\right) - p_n}{\sqrt{2p_n}} \to \mathcal{N}(0,\sigma_0^4).$$

$$\text{Moreover, if} \qquad \sum_{j>p_n} E\left((\varepsilon_j)^2\right)\left(\int \beta^*(t)\varphi_j(t)dt\right)^2 = o(\sqrt{p_n}/n), \tag{12.14}$$

where here $\{\varphi_j,\ j = 1,2,\ldots\}$ is the eigenbasis associated with the variance-covariance operator Γ, we have

$$\frac{nd^2\left(\hat{\beta}_n(\cdot),\beta^*(\cdot)\right) - p_n}{\sqrt{2p_n}} \to \mathcal{N}(0,\sigma_0^4), \tag{12.15}$$

where $d^2(\cdot,\cdot)$ denotes the metric defined in $L^2(\mathcal{T})$ through operator Γ, and defined by

$$d^2(f,g) = \int_{\mathcal{T}}\int_{\mathcal{T}} (f(t) - g(t))\, E(X(t)X(s))\, (f(s) - g(s))\, dtds,$$

for all $f, g \in L^2(\mathcal{T})$.

Now that we have checked the theoretical behavior of the estimator, we study its practical features through numerical results. We investigate the numerical performance of the proposed methodology based on some simulations and an application to ozone concentrations.

12.4 Numerical Experiments

In this section, we study the performance of the proposed model based on numerical results that highlight the importance of truncation of the functional covariate and the spatial nature of the data. We first describe the estimation procedure for the considered model.

Recall that the truncation strategy requires an appropriate selection of orthonormal basis. This one can be chosen to be a fixed orthonormal basis, such as the Fourier basis, or it can be constructed by estimating the eigenfunctions of the covariance kernel (12.5) and applying functional principal component analysis (FPCA) to the explanatory random functions X_i. Here we use the eigenfunctions obtained from the FPCA to construct the expansion basis. The eigenfunctions are those of the integral operator associated with the integral kernel defined by the variance-covariance function of X, which is estimated for each $t, v \in [0, 1]$ as follows:

$$\hat{K}(t, v) = \frac{1}{n-1} \sum_{i=1}^{n} X_i(t) X_i(v). \tag{12.16}$$

A key step is the choice of the number p of functions used in the truncation strategy. To fulfill this task, we consider three criteria: the average squared error (ASE), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC), respectively.

The choice of p using AIC is consistent in the setting of FLM (see [31] for more details). Note that we use a preselected p based on the cumulative inertia. We focus on the selection of p from among those associated with cumulative inertia values lower than some threshold (here 95%).

As a measure of accuracy of the parameter function (see [33]), the usual integrated mean square error (IMSE),

$$\text{IMSE} = \int_0^1 \left(\beta(t) - \hat{\beta}(t)\right)^2 dt, \tag{12.17}$$

is considered to compare the three choice strategies for p, namely, ASE, AIC, and BIC.

12.4.1 Monte Carlo Simulations

The main objective of the Monte Carlo Simulation is to investigate the finite sample behavior of the QMLEs of $\hat{\beta}_n(\cdot)$, $\hat{\lambda}_n$, and $\hat{\sigma}_n^2$. We consider two spatial scenarios

(see [34]) in a data grid with $G = 60 \times 60$ locations, where we randomly allocate n spatial units.

- **Scenario 1**: The spatial weight matrix W_n is constructed by taking the k neighbors of each unit using kNN method (k nearest neighbors algorithm). Let us take $k = \{4, 8\}$.
- **Scenario 2**: We consider a number of districts r (block or group) with m members in each district, where the units of the same district have the same weight. As in [35], we can define the spatial weight matrix as block diagonal $W_n = I_r \otimes B_m$, where $\otimes$ is the Kronecker product, $B_m = (l_m l_m' - I_m)/(m-1)$, and l_m is an m vector of 1.

The simulations are performed based on the following data:

$$Y_i = \lambda_0 \sum_{j=1}^{n} w_{ij} Y_j + \int_{\mathcal{T}} X_i(t)\beta^*(t)dt + U_i \tag{12.18}$$

where $\mathbf{U}_n \sim N(0, \sigma_0^2 I_n)$ is a vector of i.i.d. normal components.

We generate the functional covariate as in [31] using the Fourier orthonormal basis $\{\varphi_j(t) = \sqrt{2}\sin(j\pi t),\ t \in [0,1], j = 1, 2, \dots\}$. Let us use the first 20 functions of this basis to generate the explanatory random function:

$$X(t) = \sum_{j=1}^{20} \varepsilon_j \varphi_j(t), \tag{12.19}$$

where $\varepsilon_j \sim \mathcal{N}(0, 1/j)$ for $j \geq 1$. We define the parameter function as $\beta^*(t) = \sum_{j=1}^{20} \beta_j^* \varphi_j(t)$, with $\beta_j^* = 0$ for $j > 3$, $\beta^* = (\beta_1^*, \beta_2^*, \beta_3^*)' = (1, 1/2, 1/3)'$. With this parameter function and $\sigma_0^2 = 1$, different samples are generated using different values of the autoregressive parameter $\lambda_0 = 0.2; 0.4; 0.6;$ and 0.8.

We apply the truncation strategy to reduce the infinite dimensionality of our model $Y_i = \lambda_0 \sum_{j=1}^{n} w_{ij} Y_j + \sum_{j=1}^{p_n} \beta_j^* \varepsilon_j^{(i)} + \sum_{j=p_n+1}^{\infty} \beta_j^* \varepsilon_j^{(i)} + U_i,\ i = 1, \dots, n,$ $n = 1, 2, \dots$ to a p_n-finite linear approximation and compute the quasi-likelihood estimator. The parameters λ_0, σ_0^2, and $\beta_1^*, \dots, \beta_{p_n}^*$ are estimated by solving the score equations defined in Section 12.3. Different sample sizes, $n = \{100, 200, 400\}$, are tested for the first scenario; for the second, we take $r = \{10, 20, 30\}$ and $m = \{5, 10, 15\}$, with sample size $n = m \times r$.

The studied models are replicated 200 times, and the results of Scenario 1 are presented in Tables 12.1 and 12.2, respectively, for $k = 4$ and $k = 8$. For Scenario 2, the results are reported in Tables 12.3–12.6. Each table represents a specific model. In each table, the rows λ, σ^2, IMSE, and principal components (PCs) give the averages over these replications (with the standard deviation in brackets) of the autoregressive parameter estimate $\hat{\lambda}_n$, the standard deviation parameter $\hat{\sigma}_n^2$, the associated IMSE defined in (12.17) and the number p of eigenfunctions (used in the truncation), respectively. For the different models, the strategies used to select

Table 12.1 Estimation of parameters with $n = \{100,200,400\}$, $k = 4$.

		$n = 100$			$n = 200$			$n = 400$		
		ASE	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**
$\lambda_0 = 0.2$	λ	0.1783	0.1786	0.1800	0.2045	0.2046	0.2043	0.1955	0.1955	0.1956
		(0.1150)	(0.1160)	(0.1144)	(0.0727)	(0.0727)	(0.0731)	(0.0493)	(0.0494)	(0.0495)
	σ^2	0.9669	0.9732	0.9878	0.9835	0.9858	0.9913	0.9829	0.9834	0.9870
		(0.1438)	(0.1465)	(0.1511)	(0.1036)	(0.1040)	(0.1055)	(0.0710)	(0.0711)	(0.0710)
	IMSE	**0.1584**	0.1996	0.2595	**0.0796**	0.1141	0.1489	**0.0337**	0.0478	0.0860
		(0.1499)	(0.1332)	(0.1339)	(0.0654)	(0.0709)	(0.0747)	(0.0325)	(0.0476)	(0.0564)
	PCs	2.920	2.170	1.715	2.965	2.445	2.115	2.990	2.785	2.415
		(0.2720)	(0.6349)	(0.3637)	(0.1842)	(0.5463)	(0.5226)	(0.0997)	(0.4119)	(0.5139)
$\lambda_0 = 0.4$	λ	0.3952	0.3969	0.3979	0.3992	0.3996	0.3997	0.3945	0.3947	0.3947
		(0.0987)	(0.1428)	(0.0997)	(0.0581)	(0.0984)	(0.0580)	(0.0449)	(0.0447)	(0.0448)
	σ^2	0.9573	0.9609	0.9786	0.9786	0.9798	0.9865	0.9885	0.9888	0.9929
		(0.1432)	(0.1428)	(0.1503)	(0.0983)	(0.0984)	(0.1002)	(0.0723)	(0.0725)	(0.0448)
	IMSE	**0.1778**	0.2063	0.2786	**0.0880**	0.1067	0.1536	**0.0399**	0.0507	0.0977
		(0.1680)	(0.1574)	(0.1528)	(0.0830)	(0.0794)	(0.0908)	(0.0365)	(0.0464)	(0.0629)
	PCs	2.850	2.285	1.720	2.865	2.520	2.125	2.950	2.790	2.360
		(0.3850)	(0.6753)	(0.6662)	(0.3426)	(0.5108)	(0.5926)	(0.2185)	(0.4083)	(0.5309)
$\lambda_0 = 0.6$	λ	0.5859	0.5877	0.5884	0.5975	0.5990	0.5988	0.5979	0.5984	0.5985
		(0.0725)	(0.0722)	(0.0731)	(0.0452)	(0.0458)	(0.0455)	(0.0365)	(0.0366)	(0.0368)
	σ^2	0.9623	0.9605	0.9773	0.9872	0.9835	0.9916	0.9981	0.9970	1.0009
		(0.1357)	(0.1335)	(0.1387)	(0.0965)	(0.0947)	(0.0964)	(0.0743)	(0.0741)	(0.0744)
	IMSE	**0.1568**	0.1770	0.2428	**0.1080**	0.1092	0.1642	0.0508	**0.0506**	0.0912
		(0.1248)	(0.1191)	(0.1272)	(0.0844)	(0.0747)	(0.0942)	(0.0497)	(0.0462)	(0.0527)
	PCs	2.680	2.275	1.710	2.680	2.525	2.070	2.845	2.800	2.410
		(0.6160)	(0.6175)	(0.6387)	(0.5560)	(0.5393)	(0.5889)	(0.3764)	(0.4010)	(0.5032)
$\lambda_0 = 0.8$	λ	0.7863	0.7889	0.7884	0.7929	0.7940	0.7938	0.7990	0.7997	0.7998
		(0.0468)	(0.0461)	(0.0470)	(0.0312)	(0.0312)	(0.0313)	(0.0192)	(0.0190)	(0.0191)
	σ^2	0.9814	0.9632	0.9788	0.9978	0.9875	0.9953	0.9986	0.9892	0.9927
		(0.1519)	(0.1482)	(0.1536)	(0.0971)	(0.0952)	(0.0966)	(0.0741)	(0.0689)	(0.0696)
	IMSE	0.2303	**0.1976**	0.2422	0.1326	**0.1085**	0.1624	0.0932	**0.0520**	0.0898
		(0.1469)	(0.1329)	(0.1281)	(0.1177)	(0.0809)	(0.0937)	(0.1007)	(0.0468)	(0.0520)
	PCs	2.295	2.330	1.845	2.465	2.470	2.035	2.535	2.765	2.390
		(0.8007)	(0.6428)	(0.6581)	(0.7151)	(0.539)	(0.5525)	(0.6488)	(0.4251)	(0.4991)

Bold formatting indicate the minimum value of the IMSE values.

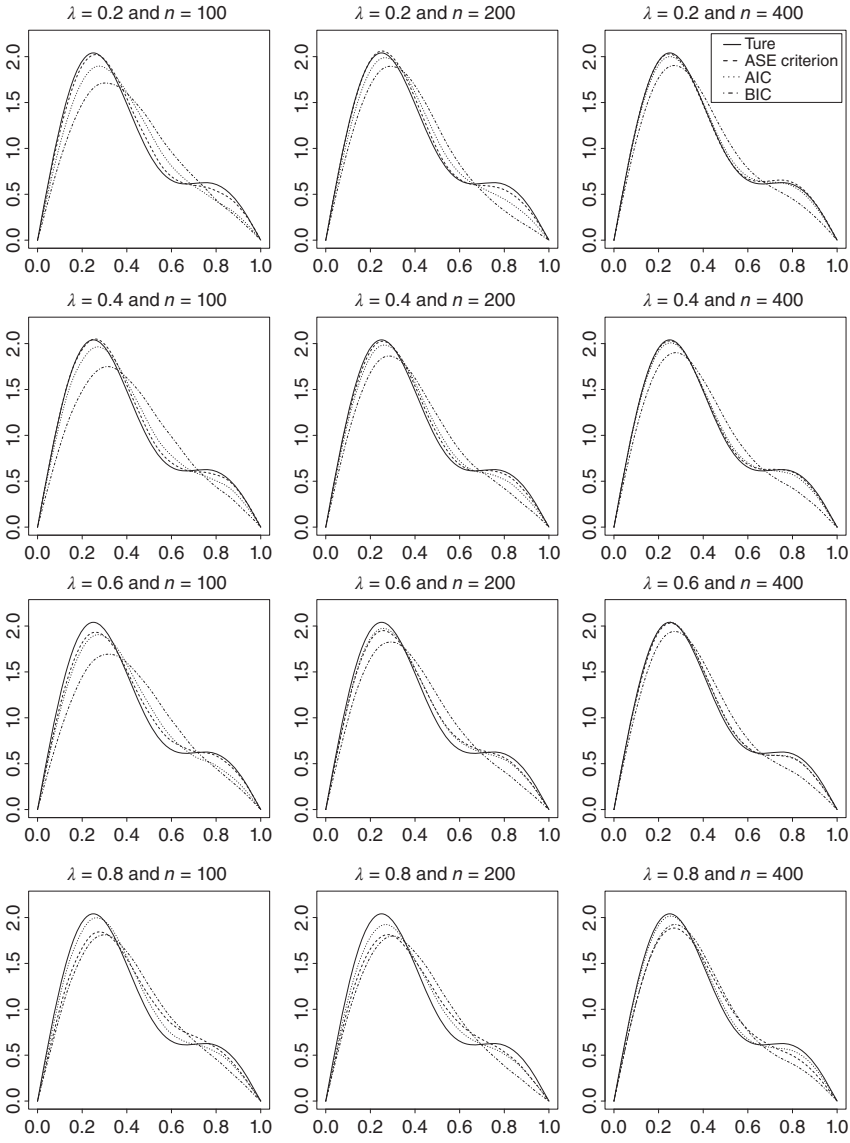


Figure 12.1 Estimated parameter function $\hat{\beta}(\cdot)$ with the different criteria and $k = 4$.

Table 12.2 Estimation of parameters with $n = \{100,200,400\}$, $k = 8$.

		$n = 100$			$n = 200$			$n = 400$		
		ASE	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**
$\lambda_0 = 0.2$	λ	0.1711	0.1709	0.1690	0.1876	0.1875	0.1886	0.1912	0.1912	0.1916
		(0.1604)	(0.1614)	(0.1439)	(0.1031)	(0.1037)	(0.1036)	(0.0800)	(0.0799)	(0.0801)
	σ^2	0.9656	0.9706	0.9892	0.9781	0.9797	0.9860	0.9833	0.9839	0.9871
		(0.1364)	(0.1385)	(0.1439)	(0.0995)	(0.1000)	(0.1010)	(0.0687)	(0.0688)	(0.0690)
	IMSE	**0.1612**	0.1920	0.2480	**0.0866**	0.1116	0.1517	**0.0394**	0.0548	0.0881
		(0.1731)	(0.1693)	(0.1560)	(0.0795)	(0.0840)	(0.0955)	(0.0409)	(0.0484)	(0.0476)
	PCs	2.925	2.275	1.705	2.950	2.540	2.190	2.980	2.725	2.395
		(0.2641)	(0.6256)	(0.1496)	(0.2185)	(0.5290)	(0.5964)	(0.1404)	(0.4476)	(0.4901)
$\lambda_0 = 0.4$	λ	0.3803	0.3809	0.3811	0.3859	0.3861	0.3859	0.3881	0.3880	0.3877
		(0.1416)	(0.1416)	(0.1413)	(0.0822)	(0.0822)	(0.0834)	(0.0705)	(0.0727)	(0.0710)
	σ^2	0.9593	0.9638	0.9782	0.9787	0.9800	0.9871	0.9945	0.0727	0.9985
		(0.1438)	(0.1456)	(0.1501)	(0.1019)	(0.1024)	(0.1048)	(0.0725)	(0.0518)	(0.0724)
	IMSE	**0.1541**	0.1821	0.2359	**0.0828**	0.1066	0.1490	**0.0426**	0.0518	0.0895
		(0.1111)	(0.1114)	(0.1274)	(0.0718)	(0.0801)	(0.0863)	(0.0389)	(0.0457)	(0.0556)
	PCs	2.855	2.180	1.730	2.925	2.555	2.165	2.940	2.800	2.445
		(0.3669)	(0.6632)	(0.6237)	(0.2641)	(0.5554)	(0.1240)	(0.2381)	(0.4010)	(0.5180)
$\lambda_0 = 0.6$	λ	0.5758	0.5791	0.5794	0.5924	0.5933	0.5935	0.5940	0.5947	0.5944
		(0.1061)	(0.1060)	(0.1072)	(0.0671)	(0.0672)	(0.0675)	(0.0496)	(0.0495)	(0.0494)
	σ^2	0.9719	0.9680	0.9844	0.9792	0.9790	0.9868	0.9932	0.9921	0.9950
		(0.1419)	(0.1398)	(0.1072)	(0.0982)	(0.0994)	(0.1020)	(0.0757)	(0.0749)	(0.0494)
	IMSE	0.2024	**0.2024**	0.2628	**0.0939**	0.1026	0.1540	**0.0477**	0.0485	0.9950
		(0.1581)	(0.1421)	(0.1414)	(0.0868)	(0.0739)	(0.0864)	(0.0476)	(0.0463)	(0.0755)
	PCs	2.600	2.290	1.760	2.780	2.530	2.110	2.855	2.795	2.465
		(0.6497)	(0.6542)	(0.6743)	(0.4612)	(0.5201)	(0.0864)	(0.3669)	(0.4047)	(0.5000)
$\lambda_0 = 0.8$	λ	0.7741	0.7777	0.7771	0.7890	0.7909	0.7905	0.7941	0.7950	0.7950
		(0.0630)	(0.0630)	(0.0633)	(0.0410)	(0.0411)	(0.0412)	(0.0321)	(0.0318)	(0.0321)
	σ^2	0.9852	0.9686	0.9840	1.0069	0.9984	1.0071	0.9957	0.9889	0.9925
		(0.1439)	(0.1374)	(0.1403)	(0.1037)	(0.1022)	(0.1044)	(0.0745)	(0.0720)	(0.0536)
	IMSE	0.2076	**0.1989**	0.2516	0.1199	**0.1027**	0.1609	0.0811	**0.0492**	0.0880
		(0.1378)	(0.1277)	(0.1403)	(0.1040)	(0.0695)	(0.0886)	(0.0970)	(0.0476)	(0.0536)
	PCs	2.245	2.200	1.720	2.545	2.505	2.035	2.615	2.775	2.405
		(0.7798)	(0.6725)	(0.6509)	(0.6858)	(0.5398)	(0.5616)	(0.6315)	(0.4186)	(0.5022)

Bold formatting indicate the minimum value of the IMSE values.

Table 12.3 Estimation of parameters associated with scenario 2 with $\lambda_0 = 0.2$.

		$m = 5$			$m = 10$			$m = 15$		
		ASE	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**
$r = 10$	λ	0.1457	0.1474	0.1483	0.1828	0.1842	0.1838	0.1734	0.1734	0.1746
		(0.1687)	(0.1700)	(0.1685)	(0.1466)	(0.1476)	(0.1468)	(0.1476)	(0.1479)	(0.1483)
	σ^2	0.9090	0.9209	0.9463	0.9583	0.9627	0.9759	0.9810	0.9836	0.9947
		(0.1897)	(0.1941)	(0.2047)	(0.1340)	(0.1353)	(0.1382)	(0.1076)	(0.1086)	(0.1108)
	IMSE	**0.3347**	0.3603	0.3778	**0.1655**	0.1925	0.2412	**0.1109**	0.1430	0.1973
		(0.2848)	(0.2541)	(0.2267)	(0.1515)	(0.1328)	(0.1251)	(0.1076)	(0.1103)	(0.1115)
	PCs	2.900	1.94	1.505	2.930	2.275	1.860	2.945	2.425	1.940
		(0.3170)	(0.7611)	(0.6497)	(0.2747)	(0.6010)	(0.6577)	(0.2286)	(0.5883)	(0.6232)
$r = 20$	λ	0.1794	0.1796	0.1788	0.1850	0.1851	0.1853	0.1917	0.1919	0.1914
		(0.0934)	(0.0938)	(0.0940)	(0.1079)	(0.1079)	(0.1073)	(0.1027)	(0.1023)	(0.1026)
	σ^2	0.9413	0.9450	0.9602	0.9748	0.9768	0.9841	0.9832	0.9840	0.9892
		(0.1429)	(0.1436)	(0.1498)	(0.1014)	(0.1018)	(0.1045)	(0.0809)	(0.1023)	(0.0823)
	IMSE	**0.1767**	0.2133	0.2686	**0.0725**	0.1032	0.1507	**0.0528**	0.0709	0.1164
		(0.1676)	(0.1620)	(0.1614)	(0.0666)	(0.0693)	(0.0874)	(0.0456)	(0.0561)	(0.0612)
	PCs	2.920	2.285	1.805	2.970	2.545	2.140	2.9850	2.690	2.280
		(0.2720)	(0.6900)	(0.7138)	(0.1710)	(0.5092)	(0.5585)	(0.1219)	(0.4637)	(0.5225)
$r = 30$	λ	0.1990	0.1985	0.1988	0.1942	0.1941	0.1943	0.1890	0.1890	0.1832
		(0.0853)	(0.0860)	(0.0869)	(0.0762)	(0.0816)	(0.0762)	(0.0867)	(0.0866)	(0.0867)
	σ^2	0.9668	0.9692	0.9797	0.9927	0.9938	0.9986	0.9900	0.9904	0.9930
		(0.1152)	(0.1156)	(0.0869)	(0.0816)	(0.0816)	(0.0829)	(0.0639)	(0.0638)	(0.0643)
	IMSE	**0.1112**	0.1446	0.1991	**0.0555**	0.0755	0.1143	**0.0330**	0.0452	0.0755
		(0.1047)	(0.1088)	(0.1130)	(0.0615)	(0.0651)	(0.0680)	(0.0298)	(0.0452)	(0.0643)
	PCs	2.920	2.455	1.990	2.980	2.6500	2.2750	2.9900	2.8100	2.5150
		(0.2720)	(0.5653)	(0.6340)	(0.1404)	(0.4782)	(0.5299)	(0.0997)	(0.3933)	(0.5010)

Bold formatting indicate the minimum value of the IMSE values.

p yield (on average) values close to the true parameter of $p = 3$, especially for ASE and AIC and large sample sizes (see the columns titled PCs in Tables 12.1–12.6). The parameter function estimates are given in Figures 12.2–12.3.

For all the models, the three methods used to select p and two spatial scenarios, the performance of the parameter function and the variance estimates varies with the sample size. A larger IMSE (the smallest is in bold) of order 0.2 is noted for sample size $n = 100$ and $k = 8$.

Table 12.4 Estimation of parameters associated with scenario 2 with $\lambda_0 = 0.4$.

		$m = 5$			$m = 10$			$m = 15$		
		ASE	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**
$r = 10$	λ	0.3590	0.3613	0.3619	0.3746	0.3756	0.3751	0.3739	0.3751	0.3748
		(0.1106)	(0.1134)	(0.1130)	(0.1184)	(0.1190)	(0.1186)	(0.1107)	(0.1110)	(0.1117)
	σ^2	0.9175	0.9271	0.9487	0.9642	0.9682	0.9845	0.9862	0.9890	0.9999
		(0.1891)	(0.1906)	(0.1943)	(0.1375)	(0.1399)	(0.1412)	(0.1245)	(0.1252)	(0.1276)
	IMSE	**0.3812**	0.4078	0.4122	**0.1702**	0.2024	0.2702	**0.1057**	0.1387	0.1976
		(0.3904)	(0.3665)	(0.3247)	(0.1452)	(0.1443)	(0.1459)	(0.0883)	(0.0856)	(0.1102)
	PCs	2.7300	1.9150	1.5150	2.8100	2.2200	1.7000	2.8950	2.3950	1.9300
		(0.5464)	(0.7816)	(0.3263)	(0.4414)	(0.6811)	(0.6650)	(0.3073)	(0.5750)	(0.6140)
$r = 20$	λ	0.3873	0.3883	0.3887	0.3733	0.3737	0.3736	0.3829	0.3830	0.3829
		(0.0749)	(0.0751)	(0.0749)	(0.0857)	(0.0861)	(0.0864)	(0.0777)	(0.0777)	(0.0775)
	σ^2	0.9587	0.9618	0.9769	0.9875	0.9890	0.9966	0.9914	0.9921	0.9967
		(0.1353)	(0.1341)	(0.1402)	(0.1043)	(0.1047)	(0.1070)	(0.0868)	(0.0870)	(0.0881)
	IMSE	**0.1700**	0.1980	0.2573	**0.0853**	0.1070	0.1563	**0.0570**	0.0734	0.1157
		(0.1368)	(0.1240)	(0.1271)	(0.0681)	(0.0696)	(0.0838)	(0.0455)	(0.0539)	(0.0699)
	PCs	2.780	2.275	1.795	2.905	2.530	2.115	2.90	2.670	2.300
		(0.4825)	(0.6414)	(0.1271)	(0.1277)	(0.5296)	(0.5861)	(0.3008)	(0.4714)	(0.5582)
$r = 30$	λ	0.3943	0.3952	0.3950	0.3867	0.3867	0.3871	0.3910	0.3911	0.3912
		(0.0670)	(0.0671)	(0.0676)	(0.0675)	(0.0675)	(0.0677)	(0.0647)	(0.0649)	(0.0654)
	σ^2	0.9706	0.9718	0.9832	0.9857	0.9863	0.9913	0.9870	0.9873	0.9906
		(0.1178)	(0.1176)	(0.1228)	(0.0840)	(0.0843)	(0.0854)	(0.0674)	(0.0676)	(0.0680)
	IMSE	**0.1150**	0.1343	0.1951	**0.0577**	0.0722	0.1122	**0.0374**	0.0461	0.0830
		(0.0903)	(0.0861)	(0.1100)	(0.0604)	(0.0687)	(0.0652)	(0.0343)	(0.0470)	(0.0512)
	PCs	2.810	2.395	0.915	2.895	2.690	2.290	2.960	2.830	2.465
		(0.4181)	(0.5750)	(0.6162)	(0.3073)	(0.4848)	(0.5169)	(0.1965)	(0.3897)	(0.5100)

Bold formatting indicate the minimum value of the IMSE values.

The methods using the ASE and AIC criteria outperform the other methods. The spatial structure, namely, the number of neighbors k (Scenario 1) and the number of observations m in each district (Scenario 2), has a slight impact on the performance of the spatial parameter estimator $\hat{\lambda}_n$. Better results are obtained for lower values, namely, $k = 4$ and $m = 5$, since the weights are more important in these cases. For a fixed value of k or m, the performance varies with sample size.

Table 12.5 Estimation of parameters associated with scenario 2 with $\lambda_0 = 0.6$.

		$m = 5$			$m = 10$			$m = 15$		
		ASE	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**
$r = 10$	λ	0.5867	0.5895	0.5903	0.5815	0.5843	0.5849	0.5736	0.5746	0.5747
		(0.0746)	(0.0752)	(0.0744)	(0.0838)	(0.0840)	(0.0835)	(0.0998)	(0.1002)	(0.1009)
	σ^2	0.9536	0.9524	0.9746	0.9617	0.9573	0.9718	0.9752	0.9736	0.9823
		(0.2158)	(0.2105)	(0.2150)	(0.1464)	(0.1463)	(0.1522)	(0.1121)	(0.1100)	(0.1115)
	IMSE	**0.3911**	0.4201	0.4261	**0.1919**	0.2053	0.2598	**0.1354**	0.1441	0.1922
		(0.3470)	(0.3498)	(0.3227)	(0.1480)	(0.1489)	(0.1418)	(0.1075)	(0.0988)	(0.1142)
	PCs	2.454	2.025	1.640	2.5750	2.2750	1.8150	2.6700	2.3700	1.9900
		(0.6558)	(0.7598)	(0.6948)	(0.6375)	(0.6256)	(0.6656)	(0.5501)	(0.5698)	(0.6179)
$r = 20$	λ	0.5875	0.5899	0.5899	0.5865	0.5884	0.5887	0.5851	0.5860	0.5860
		(0.0491)	(0.0493)	(0.0493)	(0.0571)	(0.0575)	(0.0574)	(0.0574)	(0.0580)	(0.0582)
	σ^2	0.9666	0.9580	0.9732	0.9838	0.9784	0.9866	0.9810	0.9785	0.9829
		(0.1403)	(0.1323)	(0.1385)	(0.1053)	(0.1005)	(0.1019)	(0.0791)	(0.0772)	(0.0582)
	IMSE	0.2148	**0.2138**	0.2629	0.1129	**0.1074**	0.1615	0.0723	**0.0685**	0.1062
		(0.1745)	(0.1755)	(0.1652)	(0.0932)	(0.0790)	(0.0893)	(0.0677)	(0.0517)	(0.0605)
	PCs	2.495	2.300	1.800	2.640	2.5500	2.0950	2.7450	2.6800	2.3300
		(0.6873)	(0.6650)	(0.6725)	(0.5934)	(0.5375)	(0.5724)	(0.4911)	(0.4676)	(0.5220)
$r = 30$	λ	0.5948	0.5964	0.5958	0.5879	0.5885	0.5883	0.5886	0.5888	0.5888
		(0.0425)	(0.0421)	(0.0428)	(0.0443)	(0.0445)	(0.0444)	(0.0479)	(0.0479)	(0.0481)
	σ^2	0.9846	0.9798	0.9899	0.9965	0.9953	1.0009	0.9964	0.9956	0.9994
		(0.1100)	(0.1077)	(0.1102)	(0.0803)	(0.0806)	(0.0822)	(0.0682)	(0.0683)	(0.0481)
	IMSE	**0.1293**	0.1404	0.1920	**0.0684**	0.0689	0.1184	0.0439	**0.0402**	0.0814
		(0.0988)	(0.0910)	(0.1035)	(0.0592)	(0.0555)	(0.0798)	(0.0538)	(0.0418)	(0.0527)
	PCs	2.630	2.420	1.995	2.8050	2.7150	2.2900	2.8900	2.8850	2.4800
		(0.5698)	(0.5790)	(0.5802)	(0.3972)	(0.4525)	(0.5723)	(0.3442)	(0.3198)	(0.5009)

Bold formatting indicate the minimum value of the IMSE values.

12.4.2 Real Data Application

The goal is to forecast ground-level ozone concentrations using observations from stations within the Southeastern United States over a span of 48 hours in the summer of 2005. The data are collected from monitoring stations across the United States and are available at https://www.epa.gov/outdoor-air-quality-data. We are given the ozone concentration for 106 stations (located at 106 different zip codes of several counties) for every hour from 12 a.m. 19 July to 11 p.m. 20 July 2015 (that is, 48 hours). We use linear interpolation to estimate the missing values.

Table 12.6 Estimation of parameters associated with scenario 2 with $\lambda_0 = 0.8$.

		$m = 5$			$m = 10$			$m = 15$		
		ASE	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**	**ASE**	**AIC**	**BIC**
$r = 10$	λ	0.7883	0.7905	0.7900	0.7921	0.7941	0.7941	0.7834	0.7857	0.7856
		(0.0474)	(0.0468)	(0.0474)	(0.0407)	(0.0404)	(0.0405)	(0.0432)	(0.0430)	(0.0430)
	σ^2	0.9461	0.9349	0.9596	0.9682	0.9549	0.9703	0.9917	0.9782	0.9883
		(0.2330)	(0.2326)	(0.2436)	(0.1353)	(0.1324)	(0.1379)	(0.1132)	(0.1073)	(0.1098)
	IMSE	**0.3333**	0.3607	0.3814	0.1946	**0.1890**	0.2367	0.1635	**0.1405**	0.1928
		(0.2556)	(0.2545)	(0.2152)	(0.1303)	(0.1239)	(0.1224)	(0.1248)	(0.1028)	(0.1132)
	PCs	2.265	1.950	1.515	2.340	2.275	1.785	2.415	2.420	1.975
		(0.7860)	(0.7749)	(0.6723)	(0.7464)	(0.6335)	(0.6088)	(0.7454)	(0.5703)	(0.6215)
$r = 20$	λ	0.7955	0.7968	0.7968	0.7943	0.7959	0.7960	0.7945	0.7956	0.7957
		(0.0297)	(0.0292)	(0.0296)	(0.0307)	(0.0302)	(0.0302)	(0.0285)	(0.0281)	(0.0280)
	σ^2	0.9782	0.9713	0.9871	0.9957	0.9821	0.9887	0.9951	0.9823	0.9872
		(0.1512)	(0.1527)	(0.1575)	(0.1096)	(0.1025)	(0.1055)	(0.0890)	(0.0835)	(0.0848)
	IMSE	0.1890	**0.1883**	0.2532	0.1340	**0.1006**	0.1449	0.1120	**0.0737**	0.1164
		(0.1541)	(0.1390)	(0.1445)	(0.1104)	(0.0645)	(0.0802)	(0.1114)	(0.0628)	(0.0676)
	PCs	2.470	2.250	1.735	2.430	2.570 0	2.200	2.515	2.715	2.300
		(0.7153)	(0.6706)	(0.6534)	(0.7265)	(0.5162)	(0.5931)	(0.7158)	(0.4525)	(0.5399)
$r = 30$	λ	0.7938	0.7947	0.7946	0.7948	0.7957	0.7957	0.7951	0.7959	0.7959
		(0.0240)	(0.0238)	(0.0240)	(0.0214)	(0.0211)	(0.0212)	(0.0223)	(0.0224)	(0.0224)
	σ^2	0.9946	0.9838	0.9949	1.0017	0.9905	0.9954	1.0027	0.9932	0.9965
		(0.1199)	(0.1149)	(0.1201)	(0.0873)	(0.0854)	(0.0866)	(0.0731)	(0.0700)	(0.0707)
	IMSE	0.1572	**0.1310**	0.1909	0.0962	**0.0638**	0.1074	0.0871	**0.0489**	0.0869
		(0.1366)	(0.1056)	(0.1207)	(0.0982)	(0.0532)	(0.0630)	(0.0923)	(0.0442)	(0.0481)
	PCs	2.4450	2.4400	1.9550	2.5500	2.7600	2.3450	2.5650	2.8100	2.4450
		(0.7414)	(0.5815)	(0.5956)	(0.6555)	(0.4397)	(0.5454)	(0.6307)	(0.3933)	(0.4982)

Bold formatting indicate the minimum value of the IMSE values.

Since the spatial region is a fixed subset of irregular shape, partinioned into a finite number of areal units (zip codes), it might be interesting to consider these data as functional lattice data instead of a functional geostatistical dataset. For instance, in health settings, some relevant outcome data are only available at the zip level due to personal privacy constraints. Thus, investigating the relationship between ozone exposure and some disease might be done in a zip or county level [36].

Thus in this application, we organize the original space-time series into a set of daily functional data to apply the proposed functional lattice methodology.

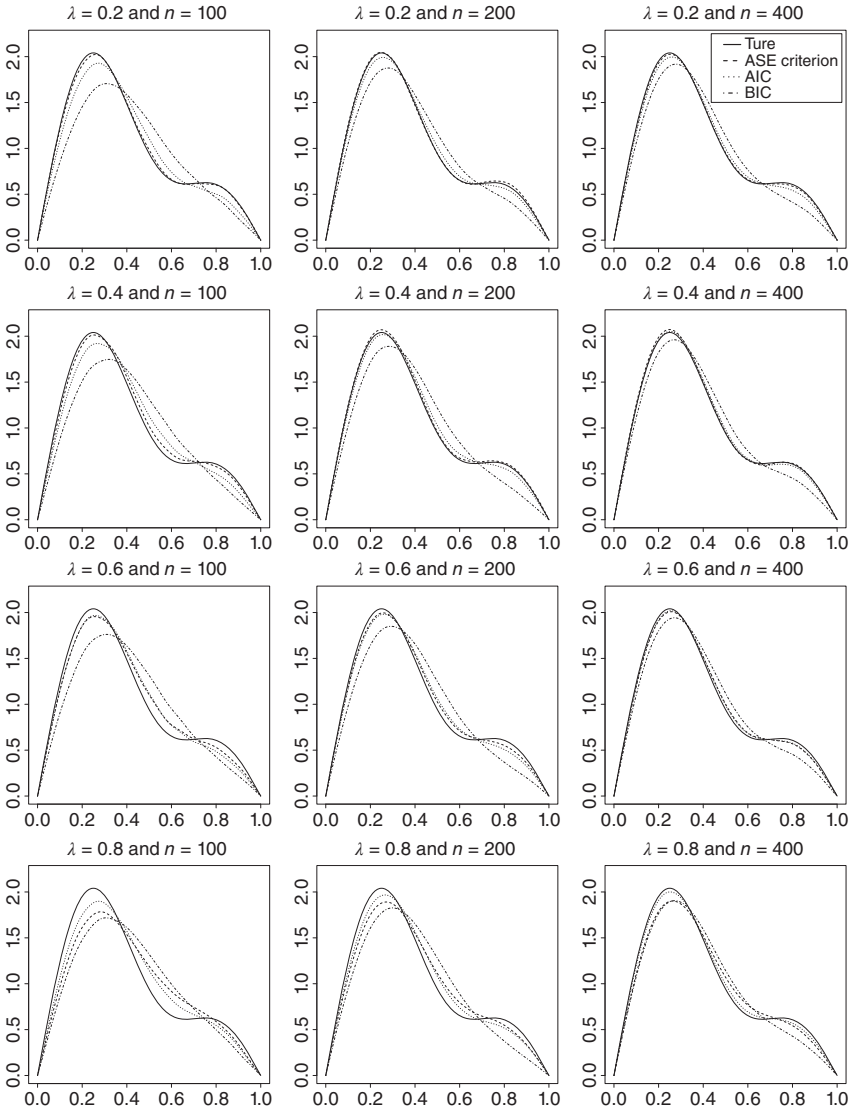


Figure 12.2 Estimated parameter function $\hat{\beta}(\cdot)$ with the different criteria and $k = 8$.

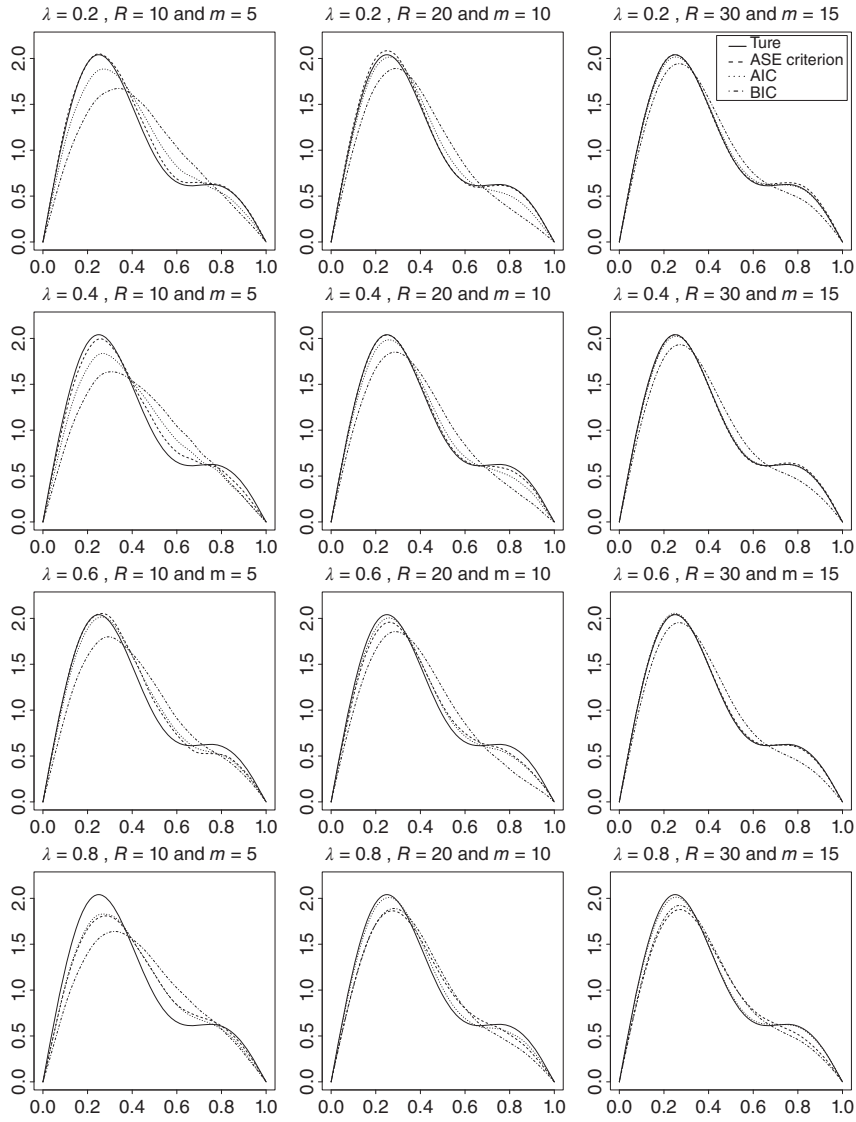


Figure 12.3 Estimated parameter function $\hat{\beta}(\cdot)$ with the different criteria in Scenario 2 for different values of r and m.

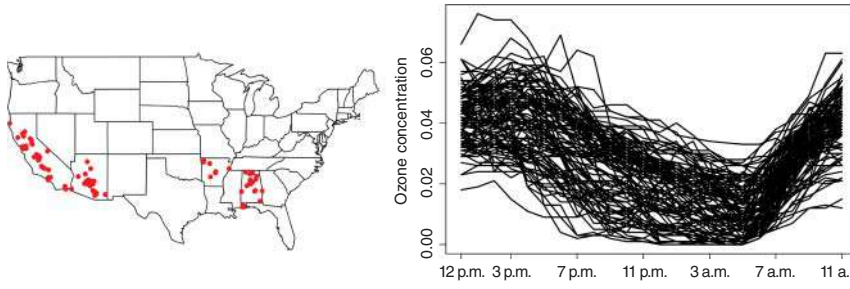


Figure 12.4 Locations and areas of the 106 stations (a) and corresponding ozone concentration curves from 12 p.m., 19 July to 11 a.m., 20 July (b).

Let us consider at each station a response variable Y as the ozone concentration at 12 p.m. on 20 July and a covariate function $\{X(t),\ t \in [0, 23]\}$ corresponding to the 24 records of ozone concentrations from 12 p.m. on 19 July to 11 a.m. on 20 July. Figure 12.4 presents the geographical positions of the 106 stations (points) and the curves of the ozone concentration from 12 p.m. 19 July to 11 a.m. 20 July.

To highlight the performance of the spatial functional spatial autoregressive model (FSARM) model, we compare with the usual FLM, that does not take into account any spatial structure in the estimation procedure.

The observations $(Y_i, \{X_i(t), t \in [0, 23]\}), i = 1, \dots, 106$, are then used to estimate, on the one hand, the parameter function and hypothetical intercept using the FLM methodology and, on the other hand, the parameter function and the autoregressive parameter using the FSARM methodology developed here. Even though the variance is estimated by the two methods, we do present it here but focus on the covariate and autoregressive parameters. We describe the spatial dependence between the stations using a 106×106 spatial weight matrix W_n. We follow the idea of [27] to define the elements of W_n by

$$w_{ij} = \begin{cases} \dfrac{1}{1 + d_{ij}} & \text{if } d_{ij} < \rho \\ 0 & \text{otherwise,} \end{cases}$$

where d_{ij} is the Euclidean distance between station i and station j, and ρ is some cut-off distance chosen such that each station has at least four neighbors. Other weight matrices have been tested, but we choose to present the results corresponding to this matrix.

Note that FPCA is used to smooth the curves before we reduce the dimension of the functional covariate using the eigenbasis, as explained above (see Figure 12.5). The AIC is used to select the number of eigenfunctions. For the two models, we

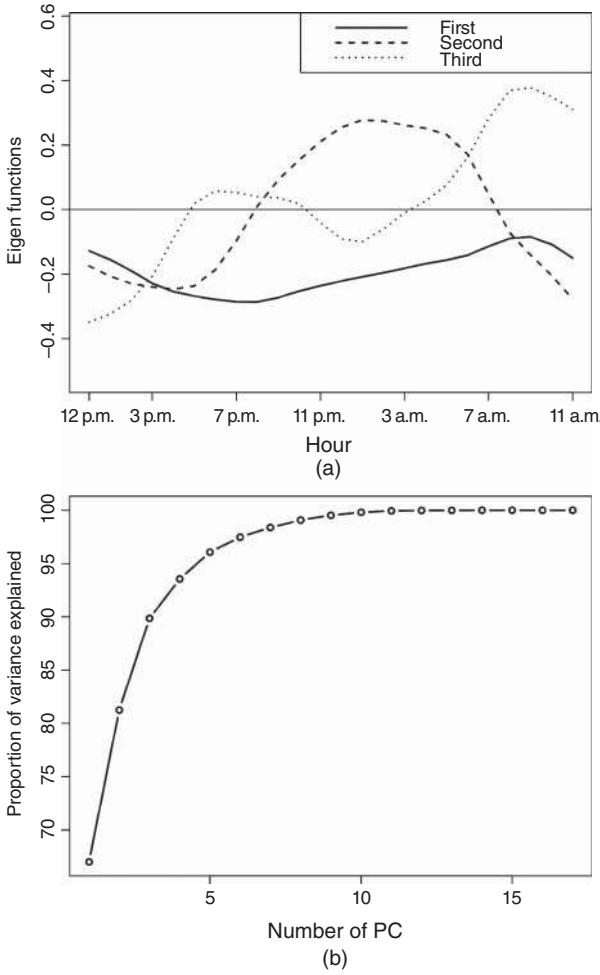


Figure 12.5 The three first eigenfunctions (a) and the proportion of explained variance (b).

have the same optimal number of eigenfunctions $p = 3$. Table 12.7 and Figure 12.6 give the estimation results of the FLM and FSARM. Note that the curves obtained by the two estimation methods are similar, with small differences around 12 p.m. and 7 p.m. The FLM gives an intercept estimate close to zero, while with FSARM, we have a spatial structure with an estimated autoregressive parameter close to 0.2.

Now, let us consider the following problem of prediction. At a given station s_0, we aim to predict the ozone concentration every hour, from 12 a.m. to 11 p.m., on 20 July 2015. For this aim, assume that at s_0, we observe only the

Table 12.7 Estimated parameters for FLM and functional spatial autoregressive linear model (FSRLM).

	PCs	Autoregressive parameter	Intercept
FSARLM	3	0.19	
FLM	3		0.006

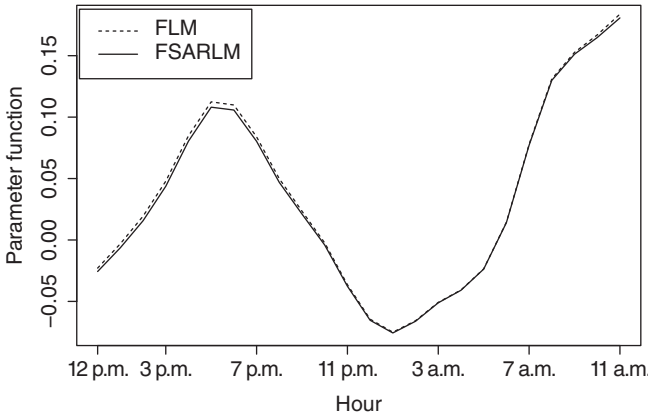


Figure 12.6 Estimated parameter functions.

24 records of ozone concentration from 12 a.m. to 11 p.m. on 19 July 2015, and we would like to predict the ozone concentration of the following day, that is, from 12 a.m. to 11 p.m. on 20 July 2015. To obtain these predictions, we proceed as follows:

1. For the prediction at 12 a.m. 20 July 2015, we estimate the parameters of FLM or FSARM, where the 105 observations (X_i, Y_i) are the following: $\{X_i(t), t \in \{0, \dots, 23\}\}$, the ozone concentrations from 12 a.m. to 11 p.m. on 19 July, and Y_i is the ozone concentration at 12 a.m., 20 July, at station i. The obtained estimated model is used to predict the ozone concentration at 12 a.m. 20 July at station s_0 (not contained in the sample), using the covariate $\{X_{s_0}(t), t \in \{0, \dots, 23\}\}$ composed of the ozone concentrations from 12 a.m. to 11 p.m. on 20 July. Let $\hat{Y}_{s_0}^{(1)}$ denote this prediction.
2. For the prediction at 2 a.m. 20 July 2015, let $X_i(t), t \in \{0, \dots, 23\}$ be the ozone concentrations from 1 a.m. 19 July to 12 p.m. 20 July and Y_i be the ozone concentration at 1 a.m. 20 July 2015, at station i. Use these observations to estimate the parameters of FLM or FSARM, and use them to predict the ozone

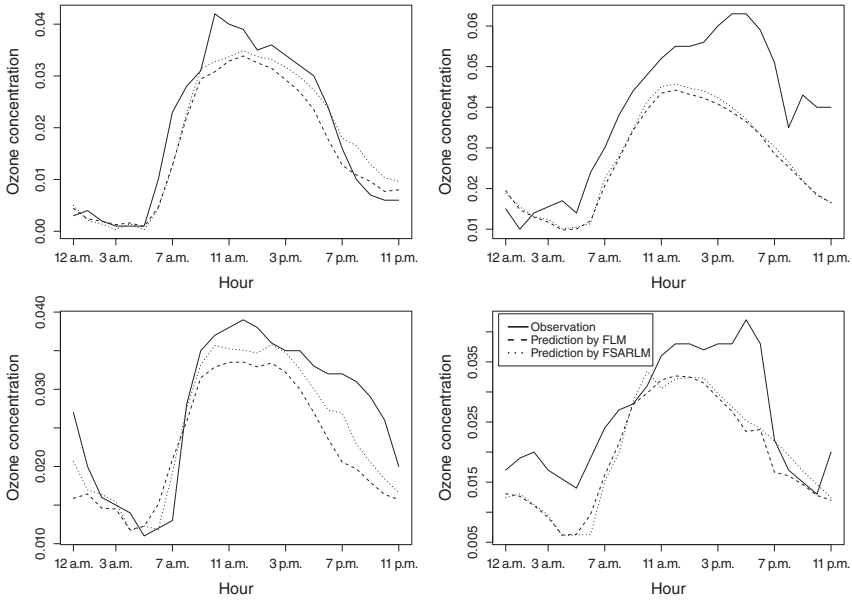


Figure 12.7 Ozone concentration (solid curves) at 4 stations selected randomly from the 106 stations and their predictions obtained using the FSAR model (dotted curves) and FLM (dashed curves).

concentration of station s_0 at 1 a.m. 20 July using $X_{s_0}(t), t \in \{0, \ldots, 23\}$, where the first 23 records are the real ozone concentrations from 1 a.m. to 11 p.m. 20 July and $X_{s_0}(23) = \hat{Y}_{s_0}^{(1)}$. Let $\hat{Y}_{s_0}^{(2)}$ denote the obtained prediction.

3. Repeat the above steps to obtain predictions from 2 a.m. to 11 p.m., 20 July 2015.

We randomly select 4 stations among the 106 and apply the prediction procedure. Figure 12.7 presents the prediction results; the true values are in solid curves, while the predictions are in dotted curves for the FSARM, and in dashed curves for the FLM (with no spatial structure). FSARM achieves some improvements, particularly around 12p.m., when the ozone concentration is higher.

12.5 Conclusion

This work proposes a functional spatial linear regression function for functional random field covariates. Our main theoretical contribution was to study the consistency and asymptotic normality of the estimator. One can see the proposed methodology and that of [20] as extensions of the real-valued SAR model to functional data. More precisely, it is apparent that the proposed estimation

approach based on a truncation technique is particularly well adapted to spatial regression estimation for functional data in the presence of spatial dependence. This good behavior is observed both from an asymptotic point of view and from a numerical study.

However, the present work focuses mainly on a single functional covariate and a real-valued response with endogenous interactive structures based on a weight matrix of distances. A number of interesting perspectives and applications can be investigated.

First, an adaptation of the proposed model to functional coefficient SAR model may be considered with interactions between locations based on nonparametric spatial interactive structures [14, 16], with elements of the weight matrix defined by an unknown function of distances (economical or physical distance) between locations. This might take into account more general interactive structures. Second, investigating mixed models allow the use of several covariates (functional and nonfunctional with eventually different domains). Notably, this could be done by using additive models to be fit efficiently by combining the model of [20] and that proposed in this work. Doing so permits models with endogenous/exogenous interactions and correlated effects which could be more general. Third, it can be of interest to investigate generalized functional linear spatial models (see, for instance, [21, 31]). Finally, it appears important in the future to investigate estimation of missing data (for instance a discrete or continuous part of a curve).

12.A Appendix

We start by showing the identifiability of the parameter λ_0 and the consistency of the estimator $\hat{\lambda}_n$ when the sequence h_n is bounded or not bounded. This is given in the following proposition:

Proposition 12.A.1 *Assume **Assumptions 12.1–12.3**.*

(i) If the sequence $\{h_n\}$ is bounded, λ_0 is identifiable and $\hat{\lambda}_n$ is consistent.
(ii) If the sequence $\{h_n\}$ is divergent, λ_0 is identifiable and $\hat{\lambda}_n$ is consistent.

Proof of Proposition 12.A.1

We give the main lines of the proof, and more details can be obtained from the authors upon request.

Proof of (i). Proving identification of λ_0 is equivalent to showing that the concentrated likelihood function $Q_n(\lambda)$ is maximum at λ_0. This can be done by checking the following uniqueness condition:

$$\text{for any } \epsilon > 0 \qquad \limsup_{n\to\infty} \max_{\lambda\in\overline{N}_\epsilon(\lambda_0)} \frac{1}{n}\left\{Q_n(\lambda) - Q_n(\lambda_0)\right\} < 0$$

where $\overline{N}_\epsilon(\lambda_0)$ is the complement of an open neighborhood of λ_0 in Λ with diameter ϵ. The rest of the proof relies on the convergence in probability of $\tilde{L}_n(\lambda)$ to $Q_n(\lambda)$ uniformly on λ in Λ, using Lemmas 12.A.1–12.A.3.

Proof of (ii):

The proof follows from the identification uniqueness condition and the uniform convergence (with the help of Lemma 12.A.3):

$$\frac{h_n}{n}\left\{\left(\tilde{L}_n(\lambda)-\tilde{L}_n(\lambda_0)\right)-\left(Q_n(\lambda)-Q_n(\lambda_0)\right)\right\}=o_p(1), \tag{12.A.1}$$

uniformly in $\lambda \in \Lambda$. □

Proof of Theorem 12.1

Identification and consistency of $\hat{\lambda}_n$ are given by Proposition 12.A.1. Let us now focus on the asymptotic normality of $\hat{\lambda}_n$.

Consider the first- and second-order derivatives of the concentrated log likelihood $\tilde{L}_n(\lambda)$:

$$\frac{\partial \tilde{L}_n(\lambda)}{\partial \lambda}=\frac{1}{\hat{\sigma}^2_{n,\lambda}}\mathbf{Y}_n'W_n'M_nS_n(\lambda)\mathbf{Y}_n-\operatorname{tr}\left(W_nS_n^{-1}(\lambda)\right),$$

and

$$\begin{aligned}\frac{\partial^2 \tilde{L}_n(\lambda)}{\partial \lambda^2}=&\frac{2}{n\hat{\sigma}^4_{n,\lambda}}\left[\mathbf{Y}_n'W_n'M_nS_n(\lambda)\mathbf{Y}_n\right]^2\\&-\frac{1}{\hat{\sigma}^2_{n,\lambda}}\mathbf{Y}_n'W_n'M_nW_n\mathbf{Y}_n-\operatorname{tr}\left(\left[W_nS_n^{-1}(\lambda)\right]^2\right).\end{aligned}$$

By Proposition 12.A.1 and Lemma 12.A.3, we have

$$\begin{aligned}&\frac{h_n}{n}\mathbf{Y}_n'W_n'M_nW_n\mathbf{Y}_n\\&\quad=\frac{h_n}{n}\mathbf{V}_n'G_n'M_nG_n\mathbf{V}_n+\frac{h_n}{n}\mathbf{U}_n'G_n'M_nG_n\mathbf{U}_n+o_p(1)\end{aligned} \tag{12.A.2}$$

and

$$\begin{aligned}&\frac{h_n}{n}\mathbf{Y}_n'W_n'M_nS_n(\lambda)\mathbf{Y}_n\\&\quad=\frac{h_n}{n}\mathbf{U}_n'G_n'M_n\mathbf{U}_n+(\lambda_0-\lambda)\frac{h_n}{n}\mathbf{V}_n'G_n'M_nG_n\mathbf{V}_n\\&\qquad+(\lambda_0-\lambda)\frac{h_n}{n}\mathbf{U}_n'G_n'M_nG_n\mathbf{U}_n+o_p(1)\\&\quad=O_p(1),\end{aligned}$$

by Proposition 12.A.1 and since under **Assumption 12.1**, Δ_n and $\operatorname{tr}(G_nB_n(\lambda))$ are of order $O_p(n/h_n)$, uniformly in λ.

From Proposition 12.A.1, we derived that $\hat{\sigma}^2_{n,\lambda} = \sigma^{*2}_{n,\lambda} + o_p(1)$. Thus, we have

$$\frac{h_n}{n}\frac{\partial^2 \tilde{L}_n(\lambda)}{\partial \lambda^2} = -\frac{1}{\sigma^{*2}_{n,\lambda}}\left[\frac{h_n}{n}\mathbf{V}_n' G_n' M_n G_n \mathbf{V}_n + \frac{h_n}{n}\mathbf{U}_n' G_n' M_n G_n \mathbf{U}_n\right] - \frac{h_n}{n}\,\mathrm{tr}\left(\left[W_n S_n^{-1}(\lambda)\right]^2\right) + o_p(1),$$

uniformly on Λ. For any $\tilde{\lambda}_n$ that converges in probability to λ_0, one can easily show that

$$\sigma^{*2}_{n,\tilde{\lambda}_n} - \sigma^{*2}_{n,\lambda_0} = o_p(1),$$

and as $\sigma^{*2}_{n,\lambda} \geq \sigma_0^2 > 0$ uniformly on Λ, we can conclude by the Taylor expansion

$$\begin{aligned}\frac{h_n}{n}\left[\frac{\partial^2 \tilde{L}_n(\tilde{\lambda}_n)}{\partial \lambda^2} - \frac{\partial^2 \tilde{L}_n(\lambda_0)}{\partial \lambda^2}\right] &= \frac{h_n}{n}\left[\mathrm{tr}\left(W_n S_n^{-1}(\tilde{\lambda}_n)\right)^2 - \mathrm{tr}\left(G_n^2\right)\right] + o_p(1)\\ &= -2(\tilde{\lambda}_n - \lambda_0)\frac{h_n}{n}\,\mathrm{tr}\left(G_n^3(\bar{\lambda}_n)\right) + o_p(1)\\ &= o_p(1),\end{aligned}$$

as under **Assumption 12.1**, $\mathrm{tr}\left(G_n^3(\lambda)\right)$ is of order $O(n/h_n)$ uniformly on Λ.

Finally, by Proposition 12.A.1 and the fact that $\sigma^{*2}_{n,\lambda_0} = \sigma_0^2$, we have

$$\frac{h_n}{n}\frac{\partial^2 \tilde{L}_n(\lambda_0)}{\partial \lambda^2} = -\frac{1}{\sigma_0^2}\frac{h_n}{n}\Delta_n - \frac{h_n}{n}\left[\mathrm{tr}(G_n' G_n) + \mathrm{tr}\left(G_n^2\right)\right] + o_p(1). \tag{12.A.3}$$

Let us now prove the asymptotic normality of $\sqrt{\frac{h_n}{n}}\frac{\partial \tilde{L}_n(\lambda_0)}{\partial \lambda}$.

Using the results of Lemma 12.A.3, we have

$$\sqrt{\frac{h_n}{n}}\mathbf{Y}_n' W_n' M_n S_n \mathbf{Y}_n = \sqrt{\frac{h_n}{n}}\left[\mathbf{V}_n' + \mathbf{U}_n'\right] G_n' M_n \mathbf{U}_n + o_p(1), \tag{12.A.4}$$

$$\text{and} \qquad \hat{\sigma}^2_{n,\lambda_0} = \frac{1}{n}\mathbf{Y}_n' S_n' M_n S_n \mathbf{Y}_n = \frac{1}{n}\mathbf{U}_n' M_n \mathbf{U}_n + o_p(1).$$

It follows that

$$\sqrt{\frac{h_n}{n}}\frac{\partial \tilde{L}_n(\lambda_0)}{\partial \lambda} = \frac{1}{\hat{\sigma}^2_{n,\lambda_0}}\sqrt{\frac{h_n}{n}}\left[\mathbf{V}_n' G_n' M_n \mathbf{U}_n + \mathbf{U}_n' C_n' M_n \mathbf{U}_n\right] + o_p(1),$$

where $C_n = G_n - \mathrm{tr}\left(\frac{G_n}{n}\right) I_n$. Again by Proposition 12.A.1, we have

$$\sqrt{\frac{h_n}{n}}\mathbf{U}_n' C_n' \xi_{p_n}\left(\xi_{p_n}' \xi_{p_n}\right)^{-1} \xi_{p_n}' \mathbf{U}_n = O_p\left(\frac{p_n}{\sqrt{n}}\right), \tag{12.A.5}$$

since under **Assumption 12.1**, the matrix C_n is uniformly bounded in both row and column sums, and $C_{ij} = O(1/h_n)$ uniformly in i and j.

Consider the following decomposition:

$$\begin{aligned}
&\xi_{p_n}' G_n' \xi_{p_n} \left(\xi_{p_n}' \xi_{p_n} \right)^{-1} \xi_{p_n}' \mathbf{U}_n \\
&= \left[\frac{\xi_{p_n}' G_n' \xi_{p_n}}{n} - \operatorname{tr}\left(\frac{G_n}{n} \right) \Gamma_{p_n} \right] \left[\frac{\xi_{p_n}' \xi_{p_n}}{n} \right]^{-1} \xi_{p_n}' \mathbf{U}_n \\
&\quad - \operatorname{tr}\left(\frac{G_n}{n} \right) \left[\frac{\xi_{p_n}' \xi_{p_n}}{n} - \Gamma_{p_n} \right] \left[\frac{\xi_{p_n}' \xi_{p_n}}{n} \right]^{-1} \xi_{p_n}' \mathbf{U}_n + \operatorname{tr}\left(\frac{G_n}{n} \right) \xi_{p_n}' \mathbf{U}_n \\
&= \operatorname{tr}\left(\frac{G_n}{n} \right) \xi_{p_n}' \mathbf{U}_n + O_p\left(\frac{p_n^2}{h_n} \right),
\end{aligned}$$

by Proposition 12.A.1 and Lemma 12.A.1. Thus,

$$\begin{aligned}
&\sqrt{\frac{h_n}{n}} \mathbf{V}_n' G_n' \xi_{p_n} \left(\xi_{p_n}' \xi_{p_n} \right)^{-1} \xi_{p_n}' \mathbf{U}_n \\
&= \frac{\sqrt{h_n}}{n} \operatorname{tr}\left(G_n \right) \frac{\mathbf{V}_n' \mathbf{U}_n}{\sqrt{n}} + O_p\left(\frac{p_n^2}{\sqrt{n h_n}} \right).
\end{aligned} \tag{12.A.6}$$

Consequently, (12.A.5) and (12.A.6) imply

$$\sqrt{\frac{h_n}{n}} \frac{\partial \tilde{L}_n(\lambda_0)}{\partial \lambda} = \frac{1}{\hat{\sigma}_{n,\lambda_0}^2} \sqrt{\frac{h_n}{n}} \left[\mathbf{V}_n' D_n' \mathbf{U}_n + \mathbf{U}_n' C_n' \mathbf{U}_n \right] + o_p(1),$$

with $D_n = G_n + \operatorname{tr}(\frac{G_n}{n}) I_n$.

Let $G_n^s = (G_n + G_n')/2$, $C_n^s = (C_n + C_n')/2$, and $D_n^s = (D_n + D_n')/2$. These matrices satisfy $C_{ij}^s = D_{ij}^s = G_{ij}^s$ for all $i \neq j$.

Now, because $\operatorname{tr}(C_n) = 0$, one can consider the decomposition:

$$\mathbf{V}_n' D_n' \mathbf{U}_n + \mathbf{U}_n' C_n' \mathbf{U}_n = \sum_{i=1}^{n} Z_{ni} \tag{12.A.7}$$

$$\text{with} \qquad Z_{ni} = D_{ii} U_i V_i + C_{ii} \left(U_i^2 - \sigma_0^2 \right) + 2 U_i \sum_{j=1}^{i-1} G_{ij}^s T_j,$$

where $T_i = V_i + U_i$, $i = 1, \ldots, n$. It is easy to show that

$$\begin{aligned}
\sum_{i=1}^{n} E\left(Z_{ni}^2 \right) &= \sigma_0^2 \left[E(V^2) + \sigma_0^2 \right] \operatorname{tr}\left(G_n \left(G_n' + G_n \right) \right) \\
&\quad + \left[3\sigma_0^2 E(V^2) + \sigma_0^4 - \mu_4 \right] \frac{1}{n} \operatorname{tr}^2(G_n) \\
&\quad + \left[\mu_4 - 3\sigma_0^4 - \sigma_0^2 E(V^2) \right] \sum_{i=1}^{n} G_{ii}^2.
\end{aligned}$$

Finally, let

$$s_Z^2 = \lim_{n \to \infty} \frac{h_n}{n} \sum_{i=1}^{n} E\left(Z_{ni}^2 \right) \quad \text{and} \quad \tilde{Z}_{ni} = \sqrt{\frac{h_n}{n}} \frac{Z_{ni}}{s_Z}.$$

Note that condition C.1 in Lemma 12.A.5 implies that $\{\tilde{Z}_{ni},\ i = 1, \dots, n;\ n = 1, 2, \dots\}$ form a triangular array of martingale differences sequences. According to (Theorem A.1, [37], p. 240) and under conditions C.2 and C.3 in Lemma 12.A.5, we have

$$\sqrt{\frac{h_n}{n}} \frac{\partial \tilde{L}_n(\lambda_0)}{\partial \lambda} = \frac{s_Z}{\hat{\sigma}^2_{n,\lambda_0}} \sum_{i=1}^{n} \tilde{Z}_{ni} + o_p(1) \to \mathcal{N}\left(0, \frac{s_Z^2}{\sigma_0^4}\right). \tag{12.A.8}$$

Finally, using (12.A.3) and (12.A.8), we can conclude by the Taylor expansion, that

$$\sqrt{\frac{n}{h_n}} \left(\hat{\lambda}_n - \lambda_0\right) \to \mathcal{N}\left(0, s_\lambda^2\right), \tag{12.A.9}$$

where

$$s_\lambda^2 = \lim_{n\to\infty} s_Z^2 \left\{ \frac{h_n}{n} \left[\Delta_n + \sigma_0^2 \mathrm{tr}\left(G_n \left(G_n' + G_n\right)\right)\right] \right\}^{-2}.$$

This concludes the proof of Theorem 12.A.1. □

Proof of Theorem 12.2

Let us consider the decomposition $S_n(\hat{\lambda}_n) = S_n + (\lambda_0 - \hat{\lambda}_n)W_n$ and note that

$$\begin{aligned}\hat{\sigma}^2_{n,\hat{\lambda}_n} &= \frac{1}{n}\mathbf{Y}_n' S_n' M_n S_n \mathbf{Y}_n + 2(\lambda_0 - \hat{\lambda}_n)\frac{1}{n}\mathbf{Y}_n' W_n' M_n S_n \mathbf{Y}_n \\ &\quad + (\lambda_0 - \hat{\lambda}_n)^2 \frac{1}{n}\mathbf{Y}_n' W_n' M_n W_n \mathbf{Y}_n.\end{aligned}$$

Lemma 12.A.3 and (12.A.5) imply that

$$\frac{1}{n}\mathbf{Y}_n' S_n' M_n S_n \mathbf{Y}_n = \frac{1}{n}\mathbf{U}_n' \mathbf{U}_n + o_p(1).$$

Thus,

$$\begin{aligned}\sqrt{n}\left(\hat{\sigma}^2_{n,\hat{\lambda}_n} - \sigma_0^2\right) &= \sqrt{\frac{n}{h_n}}(\lambda_0 - \hat{\lambda}_n)^2 \frac{\sqrt{h_n}}{n}\mathbf{Y}_n' W_n' M_n W_n \mathbf{Y}_n \\ &\quad - 2\sqrt{\frac{n}{h_n}}(\hat{\lambda}_n - \lambda_0)\frac{\sqrt{h_n}}{n}\mathbf{Y}_n' W_n' M_n S_n \mathbf{Y}_n \\ &\quad + \frac{1}{\sqrt{n}}\left(\mathbf{U}_n' \mathbf{U}_n - n\sigma_0^2\right).\end{aligned}$$

Note that Proposition 12.A.1, (12.A.2), (12.A.4), and (12.A.5) imply

$$\frac{\sqrt{h_n}}{n}\mathbf{Y}_n' W_n' M_n S_n \mathbf{Y}_n = \frac{\sqrt{h_n}}{n}\ \mathrm{tr}(G_n) + o_p(1) = O_p\left(\frac{1}{\sqrt{h_n}}\right), \tag{12.A.10}$$

$$\frac{\sqrt{h_n}}{n}\mathbf{Y}_n' W_n' M_n W_n \mathbf{Y}_n$$
$$= \frac{\sqrt{h_n}}{n}\Delta_n + \sigma_0^2 \frac{\sqrt{h_n}}{n}\,\mathrm{tr}\left(G_n G_n'\right) + o_p(1) = O_p\left(\frac{1}{\sqrt{h_n}}\right).$$

Consequently, the asymptotic normality of $\hat{\lambda}_n$ implies

$$\sqrt{\frac{n}{h_n}}(\lambda_0 - \hat{\lambda}_n)^2 \frac{\sqrt{h_n}}{n}\mathbf{Y}_n' W_n' M_n W_n \mathbf{Y}_n = o_p(1).$$

If $\lim_{n\to\infty} h_n = \infty$, (12.A.10) will be of order $o_p(1)$. Hence,

$$\sqrt{n}\left(\hat{\sigma}^2_{n,\hat{\lambda}_n} - \sigma_0^2\right) = \frac{1}{\sqrt{n}}\left(\mathbf{U}_n'\mathbf{U}_n - n\sigma_0^2\right) + o_p(1) \to \mathcal{N}\left(0, \mu_4 - \sigma_0^4\right).$$

Otherwise, we have

$$\sqrt{n}\left(\hat{\sigma}^2_{n,\hat{\lambda}_n} - \sigma_0^2\right) = \frac{1}{\sqrt{n}}\left(\mathbf{U}_n'\mathbf{U}_n - n\sigma_0^2\right)$$
$$-2\frac{\sqrt{h_n}}{n}\,\mathrm{tr}(G_n)\sqrt{\frac{n}{h_n}}(\hat{\lambda}_n - \lambda_0) + o_p(1). \qquad (12.A.11)$$

By the asymptotic normality proof of $\hat{\lambda}_n$ (see (12.A.3) and (12.A.7)), one can conclude

$$\sqrt{\frac{n}{h_n}}(\hat{\lambda}_n - \lambda_0) = -\delta_n\sqrt{\frac{h_n}{n}}\sum_{i=1}^{n} Z_{ni} + o_p(1),$$
$$\text{where} \qquad \delta_n = \frac{n}{h_n}\left[\Delta_n + \sigma_0^2 \mathrm{tr}\left(G_n\left(G_n' + G_n\right)\right)\right]^{-1}.$$

Therefore, on can rewrite (12.A.11) as

$$\sqrt{n}\left(\hat{\sigma}^2_{n,\hat{\lambda}_n} - \sigma_0^2\right) = 2\delta_n\frac{\sqrt{h_n}}{n}\,\mathrm{tr}(G_n)\sqrt{\frac{n}{h_n}}\sum_{i=1}^{n} Z^{\dagger}_{ni} + o_p(1), \qquad (12.A.12)$$
$$\text{where} \qquad Z^{\dagger}_{ni} = D_{ii}U_i V_i + \tilde{C}_{ii}\left(U_i^2 - \sigma_0^2\right) + 2U_i\sum_{j=1}^{i-1} G^s_{ij}T_j,$$

where $\tilde{C}_{ii} = C_{ii} + \frac{n}{2\delta_n\,\mathrm{tr}(G_n)}$, $\tilde{C}_{ii}$ is bounded uniformly in i, when h_n is bounded.

It is easy to show that

$$\sum_{i=1}^{n} E\left(Z^{\dagger 2}_{ni}\right) = \sum_{i=1}^{n} E\left(Z^2_{ni}\right) + n\left(\mu_4 - \sigma_0^4\right)\left[\frac{n}{2\delta_n\,\mathrm{tr}(G_n)}\right]^2.$$

Let

$$s^2_{Z^{\dagger}} = \lim_{n\to\infty}\frac{h_n}{n}\sum_{i=1}^{n} E\left(Z^{\dagger 2}_{ni}\right) \quad \text{and} \quad \tilde{Z}^{\dagger}_{ni} = \sqrt{\frac{h_n}{n}}\frac{Z^{\dagger}_{ni}}{s_{Z^{\dagger}}}.$$

Note that conditions C.1–C.3 in Lemma 12.A.5 hold when Z_{ni} and $\tilde{Z}_{ni}$ are replaced by $Z_{ni}^{\dagger}$ and $\tilde{Z}_{ni}^{\dagger}$, respectively. Therefore, (Theorem A.1, [37], p. 240) implies that

$$\sum_{i=1}^{n} \tilde{Z}_{ni}^{\dagger} \to \mathcal{N}(0,1). \tag{12.A.13}$$

Finally, by (12.A.12) and (12.A.13), we have

$$\sqrt{n}\left(\hat{\sigma}^2_{n,\hat{\lambda}_n} - \sigma_0^2\right) \to \mathcal{N}\left(0, s_\sigma^2\right)$$

$$\text{where} \quad s_\sigma^2 = \lim_{n\to\infty} h_n s^2_{Z^{\dagger}} \left[\frac{2\delta_n \ \mathrm{tr}(G_n)}{n}\right]^2 = \mu_4 - \sigma_0^4 + 4s^2_{\lambda} \lim_{n\to\infty} h_n \left[\frac{\mathrm{tr}(G_n)}{n}\right]^2.$$

This finishes the proof. □

Proof of Theorem 12.3

Recall that $S_n(\lambda)S_n^{-1} = I_n + (\lambda_0 - \lambda)G$, for all $\lambda \in \Lambda$, and

$$\hat{\beta}_{n,\hat{\lambda}_n} = \left(\xi'_{p_n}\xi_{p_n}\right)^{-1} \xi'_{p_n} S_n(\hat{\lambda}_n)\mathbf{Y}_n. \tag{12.A.14}$$

By Lemma 12.A.3, we have

$$\begin{aligned}
&\sqrt{n}\left(\hat{\beta}_{n,\hat{\lambda}_n} - \beta^*\right) \\
&\quad = \sqrt{n}(\lambda_0 - \hat{\lambda}_n)\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\left[\frac{\xi'_{p_n} G_n \xi_{p_n}}{n}\beta^* + \frac{\xi'_{p_n} G_n \mathbf{U}_n}{n}\right] \\
&\qquad + \left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\left[\frac{\xi'_{p_n}\mathbf{U}_n}{\sqrt{n}}\right] + o_p(1).
\end{aligned}$$

By Lemma 12.A.1, we have

$$\begin{aligned}
&\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\frac{\xi'_{p_n} G_n \xi_{p_n}}{n} \\
&\quad = \left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\left[\frac{\xi'_{p_n} G_n \xi_{p_n}}{n} - \mathrm{tr}\left(\frac{G_n}{n}\right)\Gamma_{p_n}\right] \\
&\qquad - \mathrm{tr}\left(\frac{G_n}{n}\right)\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\left[\frac{\xi'_{p_n}\xi_{p_n}}{n} - \Gamma_{p_n}\right] + \mathrm{tr}\left(\frac{G_n}{n}\right) I_{p_n} \\
&\quad = \mathrm{tr}\left(\frac{G_n}{n}\right) I_{p_n} + O_p\left(\frac{p_n^2}{h_n\sqrt{n}}\right).
\end{aligned}$$

The asymptotic normality result of $\hat{\lambda}_n$ and Proposition 12.A.1, imply that

$$\sqrt{n}(\lambda_0 - \hat{\lambda}_n)\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\frac{\xi'_{p_n} G_n \mathbf{U}_n}{n} = O_p\left(\frac{p_n}{\sqrt{nh_n}}\right).$$

Hence,

$$\sqrt{n}\left(\hat{\beta}_{n,\hat{\lambda}_n}-\beta^*\right)=\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\left[\frac{\xi'_{p_n}\mathbf{U}_n}{\sqrt{n}}\right]+\sqrt{n}(\lambda_0-\hat{\lambda}_n)\mathrm{tr}\left(\frac{G_n}{n}\right)\beta^*+o_p(1).$$

Therefore,

$$\begin{aligned}
&n\left(\hat{\beta}_{n,\hat{\lambda}_n}-\beta^*\right)'\Gamma_{p_n}\left(\hat{\beta}_{n,\hat{\lambda}_n}-\beta^*\right)\\
&=\left\{\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\left[\frac{\xi'_{p_n}\mathbf{U}_n}{\sqrt{n}}\right]\right\}'\Gamma_{p_n}\left\{\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\left[\frac{\xi'_{p_n}\mathbf{U}_n}{\sqrt{n}}\right]\right\}\\
&\quad+2\sqrt{n}(\lambda_0-\hat{\lambda}_n)\mathrm{tr}\left(\frac{G_n}{n}\right)\beta^{*\prime}\Gamma_{p_n}\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\left[\frac{\xi'_{p_n}\mathbf{U}_n}{\sqrt{n}}\right]\\
&\quad+n(\lambda_0-\hat{\lambda}_n)^2\mathrm{tr}^2\left(\frac{G_n}{n}\right)\beta^{*\prime}\Gamma_{p_n}\beta^*+o_p(1). \qquad (12.A.15)
\end{aligned}$$

Consider the last two terms in (12.A.15), we have by the asymptotic normality of $\hat{\lambda}_n$

$$n(\lambda_0-\hat{\lambda}_n)^2\mathrm{tr}^2\left(\frac{G_n}{n}\right)\beta^{*\prime}\Gamma_{p_n}\beta^*=O_p\left(\frac{1}{h_n}\right). \qquad (12.A.16)$$

In addition, by Proposition 12.A.1 and Lemma 12.A.1, we have

$$\sqrt{n}(\lambda_0-\hat{\lambda}_n)\mathrm{tr}\left(\frac{G_n}{n}\right)\beta^{*\prime}\Gamma_{p_n}\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\left[\frac{\xi'_{p_n}\mathbf{U}_n}{\sqrt{n}}\right]=O_p\left(\frac{1}{\sqrt{h_n}}\right). \qquad (12.A.17)$$

Let us now give the asymptotic distribution of the first term in (12.A.15). Let

$$\Psi_n=\Gamma_{p_n}^{\frac{1}{2}}\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\Gamma_{p_n}^{\frac{1}{2}},\quad \boldsymbol{\mathcal{X}}_n=\Gamma_{p_n}^{-\frac{1}{2}}\frac{\xi'_{p_n}\tilde{\mathbf{U}}_n}{\sqrt{n}},\quad \text{with}\quad \tilde{\mathbf{U}}_n=\sigma_0^{-1}\mathbf{U}_n,$$

and consider the following decomposition:

$$\begin{aligned}
&\left\{\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\left[\frac{\xi'_{p_n}\tilde{\mathbf{U}}_n}{\sqrt{n}}\right]\right\}'\Gamma_{p_n}\left\{\left(\frac{\xi'_{p_n}\xi_{p_n}}{n}\right)^{-1}\left[\frac{\xi'_{p_n}\tilde{\mathbf{U}}_n}{\sqrt{n}}\right]\right\}\\
&=\boldsymbol{\mathcal{X}}'_n\Psi_n^2\boldsymbol{\mathcal{X}}_n\\
&=\boldsymbol{\mathcal{X}}'_n\boldsymbol{\mathcal{X}}_n-2\boldsymbol{\mathcal{X}}'_n(I_{p_n}-\Psi_n)\boldsymbol{\mathcal{X}}_n\\
&\quad+\boldsymbol{\mathcal{X}}'_n(I_{p_n}-\Psi_n)^2\boldsymbol{\mathcal{X}}_n. \qquad (12.A.18)
\end{aligned}$$

We have, by **Assumptions 2**, **5**, **6**, and Proposition 7.1 of [31],

$$\frac{\boldsymbol{\mathcal{X}}'_n\boldsymbol{\mathcal{X}}_n-p_n}{\sqrt{2p_n}}\to\mathcal{N}(0,1).$$

Thus, we deduce by Proposition 12.A.1 and Lemma 12.A.4 that

$$\boldsymbol{\mathcal{X}}_n'(I_{p_n} - \Psi_n)\boldsymbol{\mathcal{X}}_n = o_p(\sqrt{p_n}) \quad \text{and} \quad \boldsymbol{\mathcal{X}}_n'(I_{p_n} - \Psi_n)^2\boldsymbol{\mathcal{X}}_n = o_p(\sqrt{p_n}).$$

Therefore,

$$\begin{aligned}&\frac{n\left(\hat{\beta}_{n,\hat{\lambda}_n} - \beta^*\right)'\Gamma_{p_n}\left(\hat{\beta}_{n,\hat{\lambda}_n} - \beta^*\right) - p_n}{\sqrt{2p_n}}\\ &\quad = \sigma_0^2\frac{\boldsymbol{\mathcal{X}}_n'\boldsymbol{\mathcal{X}}_n - p_n}{\sqrt{2p_n}} + O_p\left(\frac{1}{\sqrt{h_n p_n}}\right) \to \mathcal{N}(0, \sigma_0^4),\end{aligned}$$

by (12.A.15), (12.A.16), and (12.A.17). This yields (12.A.15) and completes the proof of Theorem 12.A.3. □

Lemma 12.A.1 *Assume that* $E\left(\varepsilon_i^4\right)$ *is finite, where* $\varepsilon_i = \int X(t)\varphi_i(t)dt$. *Under* ***Assumption 12.1****, we have*

$$\frac{\xi_{p_n}' G_n \xi_{p_n}}{n} - \operatorname{tr}\left(\frac{G_n}{n}\right)\Gamma_{p_n} = O_p\left(\frac{p_n + \sqrt{h_n}}{h_n\sqrt{n}}\right),$$

and

$$\left\|\frac{\xi_{p_n}' G_n \xi_{p_n}}{n}\right\|_2 = O_p\left(\frac{1}{h_n}\left[1 + \frac{p_n + \sqrt{h_n}}{\sqrt{n}}\right]\right).$$

Proof of Lemma 12.A.1

Note that $E\left(\varepsilon_r\varepsilon_s\right)^2 \le E\left(\varepsilon_r^2\right)E\left(\varepsilon_s^2\right)$, and $E\left(\varepsilon_s^2\right)$ is finite since $X(\cdot)$ is square integrable. Since $E\left(\varepsilon_s^4\right)$ is finite, $E\left(\varepsilon_r^2\varepsilon_s^2\right)$ is also finite.

Note that

$$\begin{aligned}E\left(\left\|\xi_{p_n}' G_n \xi_{p_n} - E\left(\xi_{p_n}' G_n \xi_{p_n}\right)\right\|_2^2\right) &= O\left(p_n^2\frac{n}{h_n^2} + \| G_n \|_2^2 + \left|\operatorname{tr}\left(G_n^2\right)\right|\right)\\ &= O\left(\frac{n}{h_n^2}\left(p_n^2 + h_n\right)\right),\end{aligned}$$

since $\| G_n \|_2^2$ and $\left|\operatorname{tr}\left(G_n^2\right)\right|$ are of order $O(n/h_n)$ by **Assumption 12.1**-ii. This concludes the proof. □

Lemma 12.A.2 *Assume that* $E\left(\varepsilon_i^4\right)$ *is finite, where* $\varepsilon_i = \int X(t)\varphi_i(t)dt$. *Under* ***Assumption 12.1****, we have*

$$\frac{\xi_{p_n}' G_n' \xi_{p_n}}{n}\left[\frac{\xi_{p_n}'\xi_{p_n}}{n}\right]^{-1}\frac{\xi_{p_n}' G_n \xi_{p_n}}{n} - \operatorname{tr}^2\left(\frac{G_n}{n}\right)\Gamma_{p_n} = O_p\left(\frac{p_n}{h_n^2\sqrt{n}}\left[1 + \frac{p_n^2}{\sqrt{n}}\right]\right).$$

Proof of Lemma 12.A.2

Note that

$$\begin{aligned}
&\frac{\xi'_{p_n} G'_n \xi_{p_n}}{n}\left[\frac{\xi'_{p_n}\xi_{p_n}}{n}\right]^{-1}\frac{\xi'_{p_n} G_n \xi_{p_n}}{n} - \mathrm{tr}^2\left(\frac{G_n}{n}\right)\Gamma_{p_n} \\
&= \left[\frac{\xi'_{p_n} G'_n \xi_{p_n}}{n} - \mathrm{tr}\left(\frac{G_n}{n}\right)\Gamma_{p_n}\right]\left[\frac{\xi'_{p_n}\xi_{p_n}}{n}\right]^{-1}\left[\frac{\xi'_{p_n} G_n \xi_{p_n}}{n} - \mathrm{tr}\left(\frac{G_n}{n}\right)\Gamma_{p_n}\right] \\
&\quad + 2\mathrm{tr}\left(\frac{G_n}{n}\right)\Gamma_{p_n}\left[\frac{\xi'_{p_n}\xi_{p_n}}{n}\right]^{-1}\left[\frac{\xi'_{p_n} G_n \xi_{p_n}}{n} - \mathrm{tr}\left(\frac{G_n}{n}\right)\Gamma_{p_n}\right] \\
&\quad + \mathrm{tr}^2\left(\frac{G_n}{n}\right)\Gamma_{p_n}\left[\frac{\xi'_{p_n}\xi_{p_n}}{n}\right]^{-1}\left[\Gamma_{p_n} - \frac{\xi'_{p_n}\xi_{p_n}}{n}\right] \\
&= O_p\left(\frac{p_n}{h_n^2\sqrt{n}}\left[1 + \frac{p_n^2}{\sqrt{n}}\right]\right),
\end{aligned}$$

by Lemma 12.A.1. This yields the proof. □

Lemma 12.A.3 *Under **Assumptions 12.1–12.2**, we have*

$$\sqrt{\frac{h_n}{n}}\mathbf{U}'_n G'_n M_n G_n \mathbf{R}_n = o_p(1), \tag{12.A.19}$$

$$\sqrt{\frac{h_n}{n}}\mathbf{R}'_n M_n G_n \xi_{p_n} = o_p(1), \tag{12.A.20}$$

$$\sqrt{\frac{h_n}{n}}\mathbf{R}'_n G'_n M_n G_n \mathbf{R}_n = o_p(1). \tag{12.A.21}$$

Proof of Lemma 12.A.3

Let

$$\pi_{n1} = \sum_{r=1}^{p_n} E\left(R^2\varepsilon_r^2\right) \quad \text{and} \quad \pi_{n2} = \sum_{r=1}^{p_n} E\left(R\varepsilon_r\right)^2.$$

Consider (F19), and note that by **Assumption 12.1**,

$$E\left(\left\|\mathbf{R}'_n G_n \xi_{p_n}\right\|_2^2\right) = O\left(\frac{n}{h_n^2}\left[h_n E(R^2) + \pi_{n1} + n\pi_{n2}\right]\right), \tag{12.A.22}$$

$$E\left(\left\|\mathbf{R}'_n \xi_{p_n}\right\|_2^2\right) = O\left(n\pi_{n1}\right), \quad \text{and} \quad E\left(\left[\mathbf{R}'_n \mathbf{U}_n\right]^2\right) = O\left(nE(R^2)\right). \tag{12.A.23}$$

Thus,

$$\mathbf{U}'_n G'_n M_n G_n \mathbf{R}_n = o_p\left(\sqrt{\frac{n}{h_n}}\right) + O_p\left(\frac{p_n}{h_n}\sqrt{h_n E(R^2) + \pi_{n1} + n\pi_{n2}}\right),$$

by (12.A.22) and (12.A.23).

Let us treat (12.A.20),

$$\mathbf{R}_n' G_n' M_n G_n \xi_{p_n} = O_p\left(\frac{\sqrt{n}}{h_n}\left[1+\frac{p_n}{h_n}\right]\sqrt{h_n E(R^2)+\pi_{n1}+n\pi_{n2}}\right).$$

Finally, considering (12.A.21), we have

$$\mathbf{R}_n' G_n' M_n G_n \mathbf{R}_n = O_p\left(\frac{p_n}{h_n^2}\left[h_n E(R^2)+\pi_{n1}+n\pi_{n2}\right]\right).$$

Therefore, the proof follows from **Assumption 12.2**. □

Lemma 12.A.4 *Under **Assumptions 12.2** and **12.5**, we have*

$$\| \Psi_n - I_{p_n} \|_2 = O_p(p_n^{-1}).$$

For the proof of this lemma, see (Lemma 7.2, [31], p. 28).

The following lemma gives conditions under which a martingale central limit theorem can be applicable to the triangular array of martingale difference sequences $\{Z_{ni},\ 1 \le i \le n,\ n \in \mathbb{N}\}$, for more of details see (Theorem A.1, [37], p. 240).

Lemma 12.A.5 *Under assumptions of Theorem 12.A.1, we have*

C.1. The random variables $\{Z_{ni},\ 1 \le i \le n,\ n \in \mathbb{N}\}$ form a triangular array of martingale difference sequence w.r.t. the filtrations
$(\mathcal{F}_{n,i}) = \sigma\left\{\varepsilon_r^{(j)}, U_j,\ 1 \le j \le i,\ 1 \le r \le p_n\right\}$ $(1 \le i \le n,\ n \in \mathbb{N})$.

C.2. Conditional normalization condition:

$$\sum_{i=1}^{n} E\left(\tilde{Z}_{ni}^2 \middle| \mathcal{F}_{n,i-1}\right) \to 1, \quad \text{in probability as } n \to \infty.$$

C.3. There exists a constant $\delta > 0$:

$$\sum_{i=1}^{n} E\left(\left|\tilde{Z}_{ni}\right|^{2+\delta}\right) \to 0, \quad n \to \infty.$$

(Lyapunov condition if $\delta = 2$).

Proof of Lemma 12.A.5

Proof of C.1 This is immediate because $E\left(Z_{ni}|\mathcal{F}_{n,i-1}\right) = 0$.

Proof of C.2

For each $i = 1, \ldots, n$, let

$$Q_{ni} = \sum_{j=1}^{i-1} G_{ij}^s T_j.$$

We have

$$E\left(Z_{ni}^2 \middle| \mathcal{F}_{n,i-1}\right) = \sigma_0^2 E(V^2) D_{ii}^2 + (\mu_4 - \sigma_0^4) C_{ii}^2 + 4\sigma_0^2 Q_{ni}^2,$$

hence,

$$E\left(\sum_{i=1}^{n}E\left(Z_{ni}^2\middle|\mathcal{F}_{n,i-1}\right)\right)=\sigma_0^2E(V^2)\sum_{i=1}^{n}D_{ii}^2+(\mu_4-\sigma_0^4)\sum_{i=1}^{n}C_{ii}^2 + 2\sigma_0^2E(T^2)\sum_{i=1}^{n}\sum_{j=1}^{i-1}G_{ij}^{s2}.$$

By definition of $\tilde{Z}_{ni}$,

$$E\left(\sum_{i=1}^{n}E\left(\tilde{Z}_{ni}^2\middle|\mathcal{F}_{n,i-1}\right)\right)=1+o(1).$$

Remark that

$$\mathrm{Var}\left(\sum_{i=1}^{n}E\left(Z_{ni}^2\middle|\mathcal{F}_{n,i-1}\right)\right)=16\sigma_0^4\mathrm{Var}\left(\sum_{i=1}^{n}Q_{ni}^2\right), \tag{12.A.24}$$

when U_i is normally distributed. Otherwise, result (12.A.27) remains valid.

Let us consider $\mathrm{Var}\left(\sum_{i=1}^{n}Q_{ni}^2\right)$. First, we have

$$\sum_{i=1}^{n}E\left(Q_{ni}^2\right)=E(T^2)\sum_{i=1}^{n}\sum_{j=1}^{i-1}G_{ij}^{s2}. \tag{12.A.25}$$

Let for all $1\le i\le j\le n$,

$$E\left(Q_{ni}^2Q_{nj}^2\right)=E\left(T^4\right)\sum_{k=1}^{i-1}G_{ik}^{s2}G_{jk}^{s2}+E(T^2)^2\sum_{k=1}^{i-1}\sum_{r=i}^{j-1}G_{ik}^{s2}G_{jr}^{s2} + E(T^2)^2\sum_{k\neq r=1}^{i-1}\left[G_{ik}^{s2}G_{jr}^{s2}+2G_{ik}^{s}G_{ir}^{s}G_{jk}^{s}G_{jr}^{s}\right].$$

We can rewrite (12.A.25) as follows:

$$\begin{aligned}&\left[2E(T^2)^2\right]^{-1}\left[E\left(\sum_{i=1}^{n}Q_{ni}^2\right)\right]^2\\&=\sum_{j=1}^{n}\sum_{i=1}^{j-1}\sum_{k=1}^{i-1}G_{ik}^{s2}G_{jk}^{s2}+\sum_{j=1}^{n}\sum_{i=1}^{j-1}\sum_{k\neq r=1}^{i-1}G_{ik}^{s2}G_{jr}^{s2}\\&\quad+\sum_{j=1}^{n}\sum_{i=1}^{j-1}\sum_{k=1}^{i-1}\sum_{r=i}^{j-1}G_{ik}^{s2}G_{jr}^{s2}.\end{aligned}$$

Therefore, we have

$$\mathrm{Var}\left(\sum_{i=1}^{n}Q_{ni}^2\right)=O\left[\frac{n}{h_n^2}\left(E\left(T^4\right)+h_nE(T^2)^2\right)\right]. \tag{12.A.26}$$

Then, by (12.A.24) and (12.A.26), we have

$$\mathrm{Var}\left(\sum_{i=1}^{n} E\left(\tilde{Z}_{ni}^2 \middle| \mathcal{F}_{n,i-1}\right)\right) = O\left(\frac{E\left(T^4\right), + h_n E(T^2)^2}{n}\right) = o(1) \qquad (12.A.27)$$

since $E(T^4) = O(E(V^4)) = O(p_n^2)$ and $E(T^2) = O(E(V^2)) = O(1)$. Hence, the result follows.

Proof of C.3

For any positive constants p and q such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\begin{aligned}
|Z_{ni}| &\leq |D_{ii}||V_i U_i| + |C_{ii}||U_i^2 - \sigma_0^2| + 2|U_i| \sum_{j=1}^{i-1} |G_{ij}^s||T_j| \\
&\leq |D_{ii}|^{\frac{1}{p}} |D_{ii}|^{\frac{1}{q}} |V_i U_i| + |C_{ii}|^{\frac{1}{p}} |C_{ii}|^{\frac{1}{q}} |U_i^2 - \sigma_0^2| \\
&\quad + \sum_{j=1}^{i-1} |G_{ij}^s|^{\frac{1}{p}} |G_{ij}^s|^{\frac{1}{q}} 2|T_j||U_i|.
\end{aligned}$$

Holder's inequality for inner products applied to the last term, implies that

$$|Z_{ni}|^q = O\left(|D_{ii}||V_i U_i|^q + |C_{ii}||U_i^2 - \sigma_0^2|^q + 2^q |U_i|^q \sum_{j=1}^{i-1} |G_{ij}^s||T_j|^q\right)$$

since under **Assumption 12.1**, D_{ii} and C_{ii} are of order $O(1/h_n)$ and G_n is uniformly bounded in row sums.

Let $q = 2 + \delta$, and note that

$$\sum_{i=1}^{n} E\left(\left|\tilde{Z}_{ni}\right|^{2+\delta}\right) = O\left(\frac{h_n^{\frac{\delta}{2}}}{n^{\frac{\delta}{2}}} \left[E\left(U^{4+2\delta}\right) + h_n E\left(|T|^{2+\delta}\right)\right]\right). \qquad (12.A.28)$$

Let $\delta = 2$, then (12.A.28) is of order $O\left(\frac{h_n^2 p_n^2}{n}\right)$, since $E(T^4) = O\left(p_n^2\right)$ and $E(U^8)$ is finite. This yields the proof as by assumption $h_n^4 = O(n)$ (when h_n is divergent) and $p_n^4 = o(n)$. □

References

1 Hastie, T. and Mallows, C. (1993). A discussion of a statistical view of some chemometrics regression tools. *Technometrics* 35 (1): 140–143.

2 Crambes, C., Kneip, A., and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics* 37 (1): 35–72. http://dx.doi.org/10.1214/07-AOS563.

3 Comte, F. and Johannes, J. (2012). Adaptive functional linear regression. *The Annals of Statistics* 40 (6): 2765–2797. http://dx.doi.org/10.1214/12-AOS1050.

4 Cai, T.T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association* 107 (499): 1201–1216. http://dx.doi.org/10.1080/01621459.2012.716337.

5 Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference* 147: 1–23. http://dx.doi.org/10.1016/j.jspi.2013.04.002.

6 Nerini, D., Monestiez, P., and Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis* 101 (2): 409–418. http://dx.doi.org/10.1016/j.jmva.2009.03.005.

7 Giraldo, R., Delicado, P., and Mateu, J. (2010). Continuous time-varying kriging for spatial prediction of functional data: an environmental application. *Journal of Agricultural, Biological, and Environmental Statistics* 15 (1): 66–82. http://dx.doi.org/10.1007/s13253-009-0012-z.

8 Giraldo, R., Delicado, P., and Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics* 18 (3): 411–426. http://dx.doi.org/10.1007/s10651-010-0143-y.

9 Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, *Springer Series in Statistics*. New York: Springer. http://dx.doi.org/10.1007/978-1-4614-3655-3.

10 Giraldo, R. (2014). Cokriging based on curves, prediction and estimation of the prediction variance. *InterStat* 2: 1–30.

11 Bohorquez, M., Giraldo, R., and Mateu, J. (2016). Optimal sampling for spatial prediction of functional data. *Statistical Methods & Applications* 25 (1): 39–54. http://dx.doi.org/10.1007/s10260-015-0340-9.

12 Bohorquez, M., Giraldo, R., and Mateu, J. (2017). Multivariate functional random fields: prediction and optimal sampling. *Stochastic Environmental Research and Risk Assessment* 31 (1): 53–70.

13 Cliff, A. and Ord, K. (1973). *Spatial Autocorrelation*. London: Pion Ltd.

14 Sun, Y. (2016). Functional-coefficient spatial autoregressive models with non-parametric spatial weights. *Journal of Econometrics* 195 (1): 134–153. http://www.sciencedirect.com/science/article/pii/S030440761630149X.

15 Koroglu, M. and Sun, Y. (2016). Functional-coefficient spatial Durbin models with non-parametric spatial weights: an application to economic growth. *Econometrics* 4 (1): 1–16. http://dx.doi.org/10.3390/econometrics4010006.

16 Sun, Y. and Malikov, E. (2018). Estimation and inference in functional-coefficient spatial autoregressive panel data models with fixed effects. *Journal of Econometrics* https://doi.org/10.1016/j.jeconom.2017.12.006.

17 Lee, L. and Yu, J. (2010). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics* 154 (2): 165–185. https://doi.org/10.1016/j.jeconom.2009.08.001.

18 Ruiz-Medina, M.D. (2011). Spatial autoregressive and moving average Hilbertian processes. *Journal of Multivariate Analysis* 102 (2): 292–305.

19 Ruiz-Medina, M.D. (2012). Spatial functional prediction from spatial autoregressive Hilbertian processes. *Environmetrics* 23 (1): 119–128. http://dx.doi.org/10.1002/env.1143.

20 Pineda-Ríos, W. and Giraldo, R. (2016). Functional SAR model. https://arxiv.org/pdf/1609.03680.pdf (accessed 28 January 2021).

21 Kelejian, H.H. and Prucha, I.R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics* 17 (1): 99–121.

22 Lee, Lf. (2007). GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics* 137 (2): 489–514. http://dx.doi.org/10.1016/j.jeconom.2005.10.004.

23 Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* 70 (349): 120–126.

24 Smirnov, O. and Anselin, L. (2001). Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. *Computational Statistics and Data Analysis* 35 (3): 301–319.

25 Lee, L.F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* 72 (6): 1899–1925. http://dx.doi.org/10.1111/j.1468-0262.2004.00558.x.

26 Yang, K. and Lee, L.-f. (2017). Identification and QML estimation of multivariate and simultaneous equations spatial autoregressive models. *Journal of Econometrics* 196 (1): 196–214. http://dx.doi.org/10.1016/j.jeconom.2016.04.019.

27 Pinkse, J. and Slade, M.E. (1998). Contracting in space: an application of spatial statistics to discrete-choice models. *Journal of Econometrics* 85 (1): 125–154.

28 Robinson, P.M. (2011). Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics* 165 (1): 5–19.

29 Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters* 45 (1): 11–22.

30 Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis* 92 (1): 24–41.

31 Müller, H.G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics* 33 (2): 774–805. http://dx.doi.org/10.1214/009053604000001156.

32 Cressie, N. (2015). *Statistics for Spatial Data*. Wiley.

33 Escabias, M., Aguilera, A.M., and Valderrama, M.J. (2007). Functional PLS logit regression model. *Computational Statistics and Data Analysis* 51: 4891–4902.

34 Su, L. (2004). Semiparametric GMM estimation of spatial autoregressive models. *Journal of Econometrics* 167 (6): 1899–1925.

35 Case, A.C. (1991). Spatial patterns in household demand. *Econometrica: Journal of the Econometric Society* 953–965.

36 Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

37 Kelejian, H.H. and Prucha, I.R. (2001). On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics* 104 (2): 219–257.

13

Spatial Prediction and Optimal Sampling for Multivariate Functional Random Fields

Martha Bohorquez[1], Ramón Giraldo[1], and Jorge Mateu[2]

[1] *Department of Statistics, National University of Colombia, Bogota, Colombia*
[2] *Department of Mathematics, University Jaume I of Castellon, Spain*

Spatial functional data analysis is an alternative approach to spatiotemporal modeling when the curves are time series varying spatially. Specifically, functional geostatistics allows carrying out optimal spatial prediction of the whole curve at unsampled sites. Along with the developments in functional geostatistics, the framework for optimal spatial sampling designs has been extended. This chapter is concerned with functional kriging, functional cokriging, and optimal sampling designs for spatial prediction of functional data. The chapter is structured as follows: In Section 13.1, we introduce the definition and some properties of multivariate functional random fields and review the theoretical framework of functional principal components (FPC). Several proposals for functional kriging are discussed in Section 13.2, and functional cokriging is presented in Section 13.3. In Section 13.4, we derive the design criteria to optimize the spatial prediction of curves using the predictors discussed in Sections 13.2 and 13.3. The methodological proposals are illustrated in Section 13.5, by an application to a dataset taken from the meteorological and air quality network of Mexico city. The chapter ends with some concluding remarks in Section 13.6.

13.1 Background

13.1.1 Multivariate Spatial Functional Random Fields

Let $D_{\boldsymbol{s}} \subset \mathbb{R}^d$ be the spatial index set, and let $\chi^1_{\boldsymbol{s}}(t), \ldots, \chi^P_{\boldsymbol{s}}(t),\ \boldsymbol{s} \in D_{\boldsymbol{s}}$, be P spatial functional square integrable random fields, such that $\chi^p_{\boldsymbol{s}}(t) \in L^2(\mathcal{B})\ p = 1, \ldots, P$. This chapter considers the case when $t \in \mathcal{B} \subset \mathbb{R}$, that is, the functional variable is a curve, see [1] for a good exposition. Note that $L^2(\mathcal{B})$ is a real separable Hilbert space $\mathcal{H}$.

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

A *multivariate spatial functional random field* is given by $\{\mathbf{\Xi}_{\boldsymbol{s}} : \boldsymbol{s} \in D_{\boldsymbol{s}} \subset \mathbb{R}^d\}$ where

$$\mathbf{\Xi}_{\boldsymbol{s}} = \left(\chi_{\boldsymbol{s}}^1(t), \ldots, \chi_{\boldsymbol{s}}^P(t)\right).$$

Now, let $\mathcal{H}^P = \mathcal{H} \oplus \cdots \oplus \mathcal{H}$ be the direct sum of the P real separable Hilbert spaces, then $\mathbf{\Xi}_{\boldsymbol{s}} \in \mathcal{H}^P$ [2, 3]. The sum, the scalar multiplication, and the inner product for the elements of $\mathcal{H}^P$ are defined by

$$\begin{aligned} \xi + \zeta &\equiv \left(\xi^1 + \zeta^1, \ldots, \xi^P + \zeta^P\right) \\ b\zeta &\equiv \left(b\zeta^1, \ldots, b\zeta^P\right) \\ [\xi, \zeta] &= \langle \xi^1, \zeta^1 \rangle + \cdots + \langle \xi^P, \zeta^P \rangle \end{aligned} \tag{13.1}$$

$\xi, \zeta \in \mathcal{H}^P$, $b \in \mathbb{R}$ and $\langle \xi^p, \zeta^p \rangle$ is the L^2-inner product.

A multivariate spatial functional dataset is an observation of $\mathbf{\Xi}_{\boldsymbol{s}}$ at a particular set of spatial sites, $S \subset D_{\boldsymbol{s}}$. If the P functional random fields can be measured at the same set of locations $S = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, we have

$$\left(\chi_{\boldsymbol{s}_i}^1(t), \ldots, \chi_{\boldsymbol{s}_i}^P(t)\right) \quad i = 1, \ldots, n$$

In other case, each spatial functional random field $\chi_{\boldsymbol{s}}^p(t)$ is observed at a different set S_p of n_p spatial locations $p = 1, \ldots, P$ as follows:

$$\left(\chi_{\boldsymbol{s}_1}^p(t), \ldots, \chi_{\boldsymbol{s}_{n_p}}^p(t)\right), \quad p = 1, \ldots, P$$

Usually, at least some locations are common for several variables.

13.1.2 Functional Principal Components

Assuming that the spatial functional random fields are random elements of $L^2(\mathcal{B})$ and that $E\left(\chi_{\boldsymbol{s}}(t)\right) = 0$, then the covariance operator C of $\chi_{\boldsymbol{s}}(t)$ is defined as

$$C(y) = E\left[\langle \chi_{\boldsymbol{s}}, y \rangle \chi_{\boldsymbol{s}}\right] \quad y \in L^2(\mathcal{B}) \tag{13.2}$$

Thus,

$$C(y)(t) = \int c(t, r) y(r) dr, \quad \text{where} \quad c(t, r) = E\left[\chi_{\boldsymbol{s}}(t) \chi_{\boldsymbol{s}}(r)\right]$$

with estimators given by

$$\hat{C}(y) = \frac{1}{n} \sum_{i=1}^{n} \left(\langle \chi_{\boldsymbol{s}_i}, y \rangle \chi_{\boldsymbol{s}_i}\right), \quad y \in L^2(\mathcal{B})$$

and

$$\hat{C}(y)(t) = \int \hat{c}(t, r) y(r) dr, \quad \text{where} \quad \hat{c}(t, r) = \frac{1}{n} \sum_{i=1}^{n} \chi_{\boldsymbol{s}_i}(t) \chi_{\boldsymbol{s}_i}(r)$$

A bounded continuous linear operator C on $\mathcal{H}$ is a covariance operator if and only if it is symmetric positive-definite and its eigenvalues η^k satisfy $\sum_{k=1}^{\infty} \eta^k < \infty$.

The FPC are defined as the eigenfunctions of the covariance operator (13.2), see [4]. The estimators of FPC are called the empirical functional principal components (EFPC).

13.1.3 The Spatial Random Field of Scores

Because the main interest is the reconstruction of the curve $\chi_{\boldsymbol{s}_i}$, a reasonable choice for a basis functions system is the EFPC formed by the eigenfunctions $\xi^k(t)$, $k = 1, \dots, K$ of the covariance operator C of $\chi_{\boldsymbol{s}}(t)$ with basis coefficients given by the associated principal component scores $f^k_{\boldsymbol{s}_i}$, defined as

$$f^k_{\boldsymbol{s}_i} = \left\langle \chi_{\boldsymbol{s}_i}, \xi^k \right\rangle, \quad k = 1, \dots, K, \quad i = 1, \dots, n \tag{13.3}$$

According to [4], the approximation of this basis is uniformly optimal, in the sense of minimizing $\hat{S}^2$ given by

$$\hat{S}^2 = \sum_{i=1}^{n} \left\| \chi_{\boldsymbol{s}_i}(t) - \sum_{k=1}^{K} f^k_{\boldsymbol{s}_i} \xi^k(t) \right\|^2. \tag{13.4}$$

It is possible to have a very good approximation using only a few EFPC. Denoting by η^k the corresponding eigenvalue, we choose K that ensures a minimum percentage of accumulated variability, previously established. The most frequently used value is 85%, but the user makes the appropriate decision. Using the Karhunen–Loève expansion [5], we assume the model takes the form:

$$\mathcal{Y}_{\boldsymbol{s}}(t) = \mu(t) + \chi_{\boldsymbol{s}}(t) = \mu(t) + \sum_{k=1}^{\infty} f^k_{\boldsymbol{s}} \xi^k(t), \quad \mathcal{Y}_{\boldsymbol{s}}(t) \in L^2 \tag{13.5}$$

where $E(\mathcal{Y}_{\boldsymbol{s}}(t)) = \mu(t)$. The mean function $\mu(t)$ is estimated by the sample mean function $\hat{\mu}(t) = \bar{\mathcal{Y}}_{\boldsymbol{s}}(t)$ with $\bar{\mathcal{Y}}_{\boldsymbol{s}}(t) = n^{-1} \sum_{i=1}^{n} \mathcal{Y}_{\boldsymbol{s}_i}(t)$. So, $E(\chi_{\boldsymbol{s}}(t)) = 0$ and from now on, we use the random variable $\chi_{\boldsymbol{s}}(t)$. In addition, for $k = 1, \dots, K$

$$E\left(f^k_{\boldsymbol{s}_i}\right) = E\left\langle \chi_{\boldsymbol{s}_i}, \xi^k \right\rangle = \left\langle 0, \xi^k \right\rangle = 0 \tag{13.6}$$

Note that for each k and $\boldsymbol{s} \in D_{\boldsymbol{s}}$, $f^k_{\boldsymbol{s}}$ is a scalar spatial random field, observed at locations $\boldsymbol{s}_1, \dots, \boldsymbol{s}_n$. So the corresponding data vector for each k is $\boldsymbol{f}^k_{\boldsymbol{s}} = (f^k_{\boldsymbol{s}_1}, \dots, f^k_{\boldsymbol{s}_n})$ and $\left(f^1_{\boldsymbol{s}}, \dots, f^K_{\boldsymbol{s}}\right)$ is a K-dimensional scalar spatial random field.

According to [3], $\boldsymbol{\Xi}_{\boldsymbol{s}} \in \mathcal{H}^P$ is a joint Gaussian $\mathcal{H}^P$-valued random field, if the real variable

$$[\boldsymbol{\Xi}_{\boldsymbol{s}}, \boldsymbol{\zeta}] = \left\langle \chi^1_{\boldsymbol{s}}, \zeta^1 \right\rangle + \cdots + \left\langle \chi^P_{\boldsymbol{s}}, \zeta^P \right\rangle \tag{13.7}$$

is Gaussian for all $\boldsymbol{\zeta} \in \mathcal{H}^P$.

Let $\xi^{p1}, \dots, \xi^{pK_p}$, $p = 1, \dots, P$ be the first K_p eigenfunctions of the covariance operator of $\chi_{\boldsymbol{s}}^{p}$ and, following the notation in (13.3), let the corresponding scores be

$$f_{\boldsymbol{s}}^{pKp} = \left\langle \chi_{\boldsymbol{s}}^{p}, \xi^{pK_p} \right\rangle. \tag{13.8}$$

For $b_{11}, \dots, b_{PK_p}$ arbitrary real numbers, let the vector $\boldsymbol{\zeta}$ be given by

$$\boldsymbol{\zeta} = \left(b_{11}\xi^{11} + \cdots + b_{1K_1}\xi^{1K_1}, \dots, b_{P1}\xi^{P1} + \cdots + b_{PK_p}\xi^{PK_P} \right)$$

We thus have that

$$[\boldsymbol{\Xi}_{\boldsymbol{s}}, \boldsymbol{\zeta}] = b_{11}f_{\boldsymbol{s}}^{11} + \cdots + b_{1K_1}f_{\boldsymbol{s}}^{1K_1} + \cdots + b_{P1}f_{\boldsymbol{s}}^{P1} + \cdots + b_{PK_P}f_{\boldsymbol{s}}^{PK_P} \tag{13.9}$$

is a real Gaussian variable, and therefore, the vector

$$\left(f_{\boldsymbol{s}}^{11}, \dots, f_{\boldsymbol{s}}^{1K_1}, \dots, f_{\boldsymbol{s}}^{P1}, \dots, f_{\boldsymbol{s}}^{PK_P} \right)$$

is a joint Gaussian multivariate random field in $\mathbb{R}^{K_1+\cdots+K_P}$.

13.2 Functional Kriging

In this section, we review some alternatives to carry out univariate spatial prediction of functional data. The ordinary kriging method uses the curves directly and models the trace-variogram, while the other proposals use the representation of the curves, in terms of basis functions and model, the spatial variability of the scalar random field is formed by the basis coefficients. Specifically, using the approximation of EFPC, the spatial variability is modeled through the random field formed by the score vectors.

13.2.1 Ordinary Functional Kriging (OFK)

The ordinary kriging method to predict functions is provided by Giraldo et al. [6], using nonparametric methods to build the curves. The predictor of the curve $\chi_{\boldsymbol{s}_0}(t)$ based on the set of functions $\chi_{\boldsymbol{s}_i}(t)$, $i = 1, \dots, n$, is given by

$$\tilde{\chi}_{\boldsymbol{s}_0}(t) = \sum_{i=1}^{n} \lambda_i \chi_{\boldsymbol{s}_i}(t) \quad t \in T, \lambda_1, \lambda_2, \dots, \lambda_n \in \mathbb{R} \tag{13.10}$$

The weights $\lambda_1, \lambda_2, \dots, \lambda_n$ in (13.10) are found as the solution of the minimization problem:

$$\min_{\lambda_1, \lambda_2, \dots, \lambda_n} \int_T \operatorname{Var}\left(\chi_{\boldsymbol{s}_0}(t) - \tilde{\chi}_{\boldsymbol{s}_0}(t) \right) dt \tag{13.11}$$

subject to the constraint $\sum_{i=1}^{n} \lambda_i = 1$ to ensure unbiasedness. The ordinary functional kriging, as in the scalar case, depends on the spatial dependence structure

which is modeled under the second-order stationarity assumption, through the trace-variogram function $\gamma(\chi_{\boldsymbol{s}_i}(t), \chi_{\boldsymbol{s}_{i'}}(t))$, defined as follows:

$$\gamma(\chi_{\boldsymbol{s}_i}(t), \chi_{\boldsymbol{s}_{i'}}(t)) = \frac{1}{2} \operatorname{Var}\left(\chi_{\boldsymbol{s}_i}(t) - \chi_{\boldsymbol{s}_{i'}}(t)\right) = \gamma(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|, t) \tag{13.12}$$

Once (13.12) has been integrated for every pair of curves, the variogram obtained, $\gamma(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|)$, is scalar and modeled with usual spatial variogram models which allow to include geometric anisotropy.

13.2.2 Functional Kriging Using Scalar Simple Kriging of the Scores (FK_{SK})

Spatial functional prediction under the assumption of a known mean function $\mu(t)$, and using a linear combination of the observed curves

$$\check{\chi}_{\boldsymbol{s}_0}(t) = \sum_{i=1}^{n} \lambda_i \chi_{\boldsymbol{s}_i}(t)$$

is solved for $\lambda_1, \lambda_2, \ldots, \lambda_n$ by the minimum least squares method [4],

$$E\left\|\chi_{\boldsymbol{s}_0}(t) - \check{\chi}_{\boldsymbol{s}_0}(t)\right\|^2 = E\left(\langle \chi_{\boldsymbol{s}_0}, \chi_{\boldsymbol{s}_0} \rangle\right) - 2\sum_{i=1}^{n} \lambda_i E\left(\langle \chi_{\boldsymbol{s}_i}, \chi_{\boldsymbol{s}_0} \rangle\right) + \sum_{i,i'=1}^{n} \lambda_i \lambda_{i'} E\left(\langle \chi_{\boldsymbol{s}_i}, \chi_{\boldsymbol{s}_{i'}} \rangle\right)$$

, where

$$E\left(\langle \chi_{\boldsymbol{s}_i}, \chi_{\boldsymbol{s}_{i'}} \rangle\right) = \sum_{k=1}^{\infty}\sum_{k'=1}^{\infty} E\left(f_{\boldsymbol{s}_i}^k f_{\boldsymbol{s}_{i'}}^{k'}\right)\left(\langle \xi^k(t), \xi^{k'}(t) \rangle\right) = \sum_{k=1}^{\infty} E\left(f_{\boldsymbol{s}_i}^k f_{\boldsymbol{s}_{i'}}^{k}\right) \tag{13.13}$$

Thus, due to the representation as a linear combination of the EFPC, see (13.4) and (13.5), the functional covariances between two locations $E(\langle \chi_{\boldsymbol{s}_i}, \chi_{\boldsymbol{s}_{i'}} \rangle)$ are completely determined by the sum of the spatial autocovariances of all score components $f_{\boldsymbol{s}}^k$ for the pair $(\boldsymbol{s}_i, \boldsymbol{s}_{i'})$, see (13.3). Note that this procedure does not need the cross-covariances between score vectors.

13.2.3 Functional Kriging Using Scalar Simple Cokriging of the Scores (FK_{CK})

Giraldo [7] and Nerini et al. [8] approximate each function in the dataset by K basis functions, and then perform ordinary cokriging at the unsampled site $\boldsymbol{s}_0$ with the scalar spatial process formed by the basis-coefficient vector. However, as the number of coefficients increases, so does the difficulty of the linear model of coregionalization (LMC), and this model can become intractable. Therefore, to make

this approach useful, the use of a basis functions system that ensures a reduced number of coefficients is required. The coefficients involved in the reconstruction of each function with the EFPC can be two or three in many practical cases, making the LMC a more feasible option. Thus, we present an alternative proposal using EFPC. According to Section 13.1.3 and Bohorquez et al. [9], we have $E\left(\chi_{\boldsymbol{s}}(t)\right) = 0$, $E\left(\boldsymbol{f}_{\boldsymbol{s}_i}^k\right) = 0, i = 1, \dots, n$ and $k = 1, \dots, K$ so we can use scalar simple cokriging [10] to predict the vector

$$\boldsymbol{f}_{\boldsymbol{s}_0} = \left(f_{\boldsymbol{s}_0}^1, \dots, f_{\boldsymbol{s}_0}^K\right)^T$$

at the unsampled location $\boldsymbol{s}_0$. Now, let $\boldsymbol{\xi}^T(t)$ be the vector containing the first K chosen eigenfunctions. The representation of the functions in terms of their functional principal components is given by

$$\chi_{\boldsymbol{s}_i}(t) = \boldsymbol{\xi}^T(t)\boldsymbol{f}_{\boldsymbol{s}_i}, \quad i = 1, \dots, n,$$

and our proposal to predict the curve $\chi_{\boldsymbol{s}_0}(t)$ is

$$\chi_{\boldsymbol{s}_0}^*(t) = \boldsymbol{\xi}^T(t)\boldsymbol{f}_{\boldsymbol{s}_0}^*, \quad i = 1, \dots, n$$

The simple cokriging predictor of the score vector at $\boldsymbol{s}_0$ is given by [10]

$$\boldsymbol{f}_{\boldsymbol{s}_0}^* = \sum_{i=1}^{n} \boldsymbol{f}_{\boldsymbol{s}_i}^T \Gamma_i,$$

where $\boldsymbol{f}_{\boldsymbol{s}_i} = \left(f_{\boldsymbol{s}_i}^1, \dots, f_{\boldsymbol{s}_i}^K\right)^T$ and Γ_i is a $K \times K$-matrix formed by the weights $\lambda_i^{kk'}$, representing the contribution of the k-th score at location $\boldsymbol{s}_i$ to the prediction of the k'-th score. Then the matrix $\boldsymbol{\Gamma} = \left(\Gamma_i\right)\ i = 1, \dots, n$ is the solution of the system

$$\begin{pmatrix} \boldsymbol{\Sigma}_{(\boldsymbol{s}_1,\boldsymbol{s}_1)} & \boldsymbol{\Sigma}_{(\boldsymbol{s}_1,\boldsymbol{s}_2)} & \cdots & \boldsymbol{\Sigma}_{(\boldsymbol{s}_1,\boldsymbol{s}_n)} \\ \boldsymbol{\Sigma}_{(\boldsymbol{s}_2,\boldsymbol{s}_1)} & \boldsymbol{\Sigma}_{(\boldsymbol{s}_2,\boldsymbol{s}_2)} & \cdots & \boldsymbol{\Sigma}_{(\boldsymbol{s}_2,\boldsymbol{s}_n)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{(\boldsymbol{s}_n,\boldsymbol{s}_1)} & \boldsymbol{\Sigma}_{(\boldsymbol{s}_n,\boldsymbol{s}_2)} & \cdots & \boldsymbol{\Sigma}_{(\boldsymbol{s}_n,\boldsymbol{s}_n)} \end{pmatrix} \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \\ \vdots \\ \Gamma_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{(\boldsymbol{s}_0,\boldsymbol{s}_1)} \\ \boldsymbol{\Sigma}_{(\boldsymbol{s}_0,\boldsymbol{s}_2)} \\ \vdots \\ \boldsymbol{\Sigma}_{(\boldsymbol{s}_0,\boldsymbol{s}_n)} \end{pmatrix}; \Gamma_i = \begin{pmatrix} \lambda_i^{11} & \lambda_i^{12} & \cdots & \lambda_i^{1K} \\ \lambda_i^{21} & \lambda_i^{22} & \cdots & \lambda_i^{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_i^{K1} & \lambda_i^{K2} & \cdots & \lambda_i^{KK} \end{pmatrix} \tag{13.14}$$

and $\boldsymbol{\Sigma}_{(\boldsymbol{s}_i,\boldsymbol{s}_{i'})} = \left(\mathrm{Cov}\left(f_{\boldsymbol{s}_i}^k, f_{\boldsymbol{s}_{i'}}^{k'}\right)\right),\ k, k' = 1, \dots, K$. Note that although we use an orthonormal basis, the cross-covariances between the respective score coefficients depend on the cross-covariance between the observed functions, so in general terms, there is no reason to assume independence between the score vectors. Indeed, note that

$$\begin{aligned} E\left(f_{\boldsymbol{s}_i}^k f_{\boldsymbol{s}_{i'}}^{k'}\right) &= E\left(\left\langle \chi_{\boldsymbol{s}_i}, \xi^k \right\rangle \left\langle \chi_{\boldsymbol{s}_{i'}}, \xi^{k'} \right\rangle\right) \\ &= E\left(\int \chi_{\boldsymbol{s}_i}(t)\xi^k(t)dt \int \chi_{\boldsymbol{s}_{i'}}(r)\xi^{k'}(r)dr\right) \end{aligned}$$

$$
\begin{aligned}
&= E\left(\iint \chi_{\boldsymbol{s}_i}(t)\xi^k(t)\chi_{\boldsymbol{s}_{i'}}(r)\xi^{k'}(r)dtdr\right)\\
&= \int \xi^{k'}(r)\left(\int E\left(\chi_{\boldsymbol{s}_i}(t)\chi_{\boldsymbol{s}_{i'}}(r)\right)\xi^k(t)dt\right)dr\\
&= \int \xi^{k'}(r)\left(\int c_{\boldsymbol{s}_i,\boldsymbol{s}_{i'}}(t,r)\xi^k(t)dt\right)dr\\
&= \int \xi^{k'}(r)C_{\boldsymbol{s}_i,\boldsymbol{s}_{i'}}(\xi^k)dr\\
&= \langle C_{\boldsymbol{s}_i,\boldsymbol{s}_{i'}}(\xi^k),\xi^{k'}\rangle
\end{aligned}
\tag{13.15}
$$

As a particular case, when $i = i'$, the covariance for different scores is 0 as in the multivariate case,

$$
\begin{aligned}
E\left(f^k_{\boldsymbol{s}_i}f^{k'}_{\boldsymbol{s}_i}\right) &= \int \xi^{k'}(r)\left(\int c_{\boldsymbol{s}_i,\boldsymbol{s}_i}(t,r)\xi^k(t)dt\right)dr\\
&= \int \xi^{k'}(r)C_{\boldsymbol{s}_i,\boldsymbol{s}_i}(\xi^k)dr\\
&= \int \xi^{k'}(r)\eta^k\xi^k(r)dr\\
&= \eta^k\left\langle \xi^{k'},\xi^k\right\rangle\\
&= \begin{cases}\eta^k, & \text{if } k = k'\\ 0 & \text{if } k \neq k'\end{cases}
\end{aligned}
\tag{13.16}
$$

Therefore, the spatial autocovariance function for each score vector k is given by

$$
E\left(f^k_{\boldsymbol{s}_i}f^k_{\boldsymbol{s}_{i'}}\right) = \begin{cases}\eta^k, & \text{if } i = i'\\ \eta^k\rho^k\left(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|;\Theta\right) & \text{if } i \neq i'\end{cases}
\tag{13.17}
$$

where $\rho^k(\cdot)$ is the correlation function of the spatial scalar field $\boldsymbol{f}^k_{\boldsymbol{s}}$. Consequently, (13.17) shows that the covariance structure is second-order stationary such that the variance of each score vector is the corresponding eigenvalue, and so the covariance model for each $\boldsymbol{f}^k_{\boldsymbol{s}}$ has finite and known variance (sill). The matrices in the main diagonal of (13.14) can be denoted in more general form as Σ_0, that is, $\Sigma_0 = \boldsymbol{\Sigma}_{(\boldsymbol{s}_i,\boldsymbol{s}_i)} = \left(\text{Cov}\left(f^k_{\boldsymbol{s}_i},f^{k'}_{\boldsymbol{s}_i}\right)\right),\ k,k' = 1,\ldots,K$. Therefore, its trace is constant and given by

$$
\text{Tr}\left(\Sigma_0\right) = \sum_{k=1}^{K}\eta^k
\tag{13.18}
$$

The variance of the prediction error can be obtained as

$$
\begin{aligned}
\text{Var}\left(\chi_{\boldsymbol{s}_0}(t) - \chi^*_{\boldsymbol{s}_0}(t)\right) &= \text{Var}\left(\boldsymbol{\xi}^T(t)\boldsymbol{f}_{\boldsymbol{s}_0} - \boldsymbol{\xi}^T(t)\boldsymbol{f}^*_{\boldsymbol{s}_0}\right)\\
&= \boldsymbol{\xi}^T(t)\text{Var}\left(\boldsymbol{f}_{\boldsymbol{s}_0} - \boldsymbol{f}^*_{\boldsymbol{s}_0}\right)\boldsymbol{\xi}(t)
\end{aligned}
$$

$$= \boldsymbol{\xi}^T(t)\left(\mathrm{Tr}\left(\Sigma_0\right) - \mathrm{Tr}\left(\sum_{i=1}^{n}\left(\boldsymbol{\Sigma}_{(\boldsymbol{s}_0,\boldsymbol{s}_i)}\Gamma_i\right)\right)\right)\boldsymbol{\xi}(t)$$

$$= \sigma^2_{\boldsymbol{f}_{\boldsymbol{s}_0}-\boldsymbol{f}^*_{\boldsymbol{s}_0}}\boldsymbol{\xi}^T(t)\boldsymbol{\xi}(t) \tag{13.19}$$

where

$$\sigma^2_{\boldsymbol{f}_{\boldsymbol{s}_0}-\boldsymbol{f}^*_{\boldsymbol{s}_0}} = \mathrm{Tr}\left(\Sigma_0\right) - \mathrm{Tr}\left(\sum_{i=1}^{n}\left(\boldsymbol{\Sigma}_{(\boldsymbol{s}_0,\boldsymbol{s}_i)}\Gamma_i\right)\right)$$

is the accumulated variance of the scalar cokriging predictor of the vector $\boldsymbol{f}_{\boldsymbol{s}_0}$, where $\mathrm{Tr}\left(\Sigma_0\right)$ is constant, see (13.18).

13.3 Functional Cokriging

In this section, we present the spatial prediction of a functional variable at unsampled sites, using spatial functional covariates. We show that through the representation of each function in terms of its EFPC, the functional cokriging method only depends on the autocovariance and cross-covariance of the associated score vectors, which are scalar random fields, see Section 13.1.3. The functional cokriging method is developed first for the case of two spatial functional random fields and then for the case of P spatial functional random fields.

13.3.1 Cokriging with Two Functional Random Fields

Let $\chi^1_{\boldsymbol{s}}(t)$ and $\chi^2_{\boldsymbol{s}}(t)$ be two spatial functional random fields such that $E\left(\chi^1_{\boldsymbol{s}}(t)\right) = 0$ and $E\left(\chi^2_{\boldsymbol{s}}(t)\right) = 0$. The cokriging predictor of $\chi^1_{\boldsymbol{s}_0}(t)$ for an unsampled site $\boldsymbol{s}_0$, using $\chi^2_{\boldsymbol{s}}(t)$ as a spatial covariate, is given by

$$\check{\chi}^1_{\boldsymbol{s}_0}(t) = \sum_{i=1}^{n_1}\lambda_i^{11}\chi^1_{\boldsymbol{s}_i}(t) + \sum_{j=1}^{n_2}\lambda_j^{12}\chi^2_{\boldsymbol{s}_j}(t)$$

where λ_i^{11} $i = 1, \ldots, n_1$ are the weights of the n_1 observations of $\chi^1_{\boldsymbol{s}}(t)$ and λ_j^{12} $j = 1, \ldots, n_2$ are the weights of the n_2 observations of $\chi^2_{\boldsymbol{s}}(t)$. Note that it is not required that both processes are measured at the same places. In addition, the unbiasedness of the predictor is ensured given that the mean is known,

$$E\left(\check{\chi}^1_{\boldsymbol{s}_0}(t) - \chi^1_{\boldsymbol{s}_0}(t)\right) = E\left(\sum_{i=1}^{n_1}\lambda_i^{11}\chi^1_{\boldsymbol{s}_i}(t) + \sum_{j=1}^{n_2}\lambda_j^{12}\chi^2_{\boldsymbol{s}_j}(t)\right) = 0$$

and $\lambda = \left(\lambda_i^{11}\right)$ $i = 1, \ldots, n_1$ and $\beta = \left(\lambda_j^{12}\right)$ $j = 1, \ldots, n_2$ are constants that minimize

$$Q = E\left\|\chi^1_{\boldsymbol{s}_0}(t) - \check{\chi}^1_{\boldsymbol{s}_0}(t)\right\|^2$$

Now, we carry out the minimization of Q to obtain the system of cokriging equations:

$$\begin{aligned}
Q &= E\left\|\chi^1_{\boldsymbol{s}_0}(t) - \check{\chi}^1_{\boldsymbol{s}_0}(t)\right\|^2 \\
&= E\langle \chi^1_{\boldsymbol{s}_0}, \chi^1_{\boldsymbol{s}_0}\rangle - 2E\langle \chi^1_{\boldsymbol{s}_0}, \check{\chi}^1_{\boldsymbol{s}_0}\rangle + E\langle \check{\chi}^1_{\boldsymbol{s}_0}, \check{\chi}^1_{\boldsymbol{s}_0}\rangle \\
&= E\langle \chi^1_{\boldsymbol{s}_0}, \chi^1_{\boldsymbol{s}_0}\rangle - 2\sum_{i=1}^{n_1}\lambda_i^{11}E\langle \chi^1_{\boldsymbol{s}_i}, \chi^1_{\boldsymbol{s}_0}\rangle - 2\sum_{j=1}^{n_2}\lambda_j^{12}E\langle \chi^2_{\boldsymbol{s}_j}, \chi^1_{\boldsymbol{s}_0}\rangle \\
&\quad + \sum_{i=1}^{n_1}\sum_{i'=1}^{n_1}\lambda_i^{11}\lambda_{i'}^{11}E\langle \chi^1_{\boldsymbol{s}_i}, \chi^1_{\boldsymbol{s}_{i'}}\rangle + 2\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\lambda_i^{11}\lambda_j^{12}E\langle \chi^2_{\boldsymbol{s}_j}, \chi^1_{\boldsymbol{s}_i}\rangle \\
&\quad + \sum_{j=1}^{n_2}\sum_{j'=1}^{n_2}\lambda_j^{12}\lambda_{j'}^{12}E\langle \chi^2_{\boldsymbol{s}_j}, \chi^2_{\boldsymbol{s}_{j'}}\rangle
\end{aligned} \tag{13.20}$$

Thus, for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$ the partial derivatives are given by

$$\frac{\partial Q}{d\lambda_i^{11}} = -2E\langle \chi^1_{\boldsymbol{s}_i}, \chi^1_{\boldsymbol{s}_0}\rangle + 2\sum_{i'=1}^{n_1}\lambda_{i'}^{11}E\langle \chi^1_{\boldsymbol{s}_i}, \chi^1_{\boldsymbol{s}_{i'}}\rangle + 2\sum_{j=1}^{n_2}\lambda_j^{12}E\langle \chi^2_{\boldsymbol{s}_j}, \chi^1_{\boldsymbol{s}_i}\rangle$$

and

$$\frac{\partial Q}{d\lambda_j^{12}} = -2E\langle \chi^2_{\boldsymbol{s}_j}, \chi^1_{\boldsymbol{s}_0}\rangle + 2\sum_{i=1}^{n_1}\lambda_i^{11}E\langle \chi^2_{\boldsymbol{s}_j}, \chi^1_{\boldsymbol{s}_i}\rangle + 2\sum_{j'=1}^{n_2}\lambda_{j'}^{12}E\langle \chi^2_{\boldsymbol{s}_j}, \chi^2_{\boldsymbol{s}_{j'}}\rangle$$

Therefore, the cokriging equations are

$$\sum_{i'=1}^{n_1}\lambda_{i'}^{11}E\langle \chi^1_{\boldsymbol{s}_i}, \chi^1_{\boldsymbol{s}_{i'}}\rangle = E\langle \chi^1_{\boldsymbol{s}_i}, \chi^1_{\boldsymbol{s}_0}\rangle - \sum_{j=1}^{n_2}\lambda_j^{12}E\langle \chi^2_{\boldsymbol{s}_j}, \chi^1_{\boldsymbol{s}_i}\rangle \tag{13.21}$$

and

$$\sum_{j'=1}^{n_2}\lambda_{j'}^{12}E\langle \chi^2_{\boldsymbol{s}_j}, \chi^2_{\boldsymbol{s}_{j'}}\rangle = E\langle \chi^2_{\boldsymbol{s}_j}, \chi^1_{\boldsymbol{s}_0}\rangle - \sum_{i=1}^{n_1}\lambda_i^{11}E\langle \chi^2_{\boldsymbol{s}_j}, \chi^1_{\boldsymbol{s}_i}\rangle \tag{13.22}$$

Replacing (13.21) and (13.22) in (13.20) we obtain

$$E\left\|\chi^1_{\boldsymbol{s}_0}(t) - \check{\chi}^1_{\boldsymbol{s}_0}(t)\right\|^2 = E\langle \chi^1_{\boldsymbol{s}_0}, \chi^1_{\boldsymbol{s}_0}\rangle - \sum_{i=1}^{n_1}\lambda_i^{11}E\langle \chi^1_{\boldsymbol{s}_i}, \chi^1_{\boldsymbol{s}_0}\rangle - \sum_{j=1}^{n_2}\lambda_j^{12}E\langle \chi^2_{\boldsymbol{s}_j}, \chi^1_{\boldsymbol{s}_0}\rangle \tag{13.23}$$

Let $f^{1k}_{\boldsymbol{s}}$ $k = 1, \ldots, K$ and $f^{2l}_{\boldsymbol{s}}$ $l = 1, \ldots, L$ be the scalar spatial random fields formed by the scores of the functions $\chi^1_{\boldsymbol{s}}(t)$ and $\chi^2_{\boldsymbol{s}}(t)$, respectively. Expressing each function in (13.23) in terms of their principal components, we obtain

$$E\langle \chi^1_{\boldsymbol{s}_0}, \chi^1_{\boldsymbol{s}_0}\rangle = \sum_{k=1}^{K}\sum_{k'=1}^{K}E\langle f^{1k}_{\boldsymbol{s}_0}f^{1k'}_{\boldsymbol{s}_0}\rangle\langle \xi^{1k}, \xi^{1k'}\rangle$$

$$\sum_{i=1}^{n_1}\lambda_i^{11}E\langle\chi_{\boldsymbol{s}_i}^1,\chi_{\boldsymbol{s}_0}^1\rangle=\sum_{i=1}^{n_1}\sum_{k=1}^{K}\sum_{k'=1}^{K}\lambda_i^{11}E\left(f_{\boldsymbol{s}_i}^{1k}f_{\boldsymbol{s}_0}^{1k'}\right)\langle\xi^{1k},\xi^{1k'}\rangle$$
$$\sum_{j=1}^{n_2}\lambda_j^{12}E\langle\chi_{\boldsymbol{s}_j}^2,\chi_{\boldsymbol{s}_0}^1\rangle=\sum_{j=1}^{n_2}\sum_{l=1}^{L}\sum_{k=1}^{K}\lambda_j^{12}E\left(f_{\boldsymbol{s}_i}^{2k}f_{\boldsymbol{s}_0}^{1k}\right)\langle\xi^{2k},\xi^{1k}\rangle$$

where $\xi^{1k},\ k=1,\dots,K$ are the eigenfunctions of the covariance operator of $\chi_{\boldsymbol{s}}^1(t)$ and $\xi^{2l},\ l=1,\dots,L$ are the eigenfunctions of the covariance operator of $\chi_{\boldsymbol{s}}^2(t)$. Due to the orthonormality of the EFPC, we have that

$$\langle\xi^{1k},\xi^{1k'}\rangle=\begin{cases}0 & \text{if} k\neq k'\\ 1 & \text{if} k=k'\end{cases}$$

Also note that

$$\sum_{k=1}^{K}E\left(f_{\boldsymbol{s}_0}^{1k}f_{\boldsymbol{s}_0}^{1k}\right)=\sum_{k=1}^{K}\eta^{1k}$$

Thus, $E\left\|\chi_{\boldsymbol{s}_0}^1(t)-\check{\chi}_{\boldsymbol{s}_0}^1(t)\right\|^2$ can be simplified as follows:

$$\begin{aligned}&\sum_{k=1}^{K}E\left(f_{\boldsymbol{s}_0}^{1k}f_{\boldsymbol{s}_0}^{1k}\right)-\sum_{i=1}^{n_1}\sum_{k=1}^{K}\lambda_i^{11}E\left(f_{\boldsymbol{s}_i}^{1k}f_{\boldsymbol{s}_0}^{1k}\right)-\sum_{j=1}^{n_2}\sum_{l=1}^{L}\sum_{k=1}^{K}\lambda_j^{12}c_{12}^{lk}E\left(f_{\boldsymbol{s}_i}^{2l}f_{\boldsymbol{s}_0}^{1k}\right)\\ &\quad=\sum_{k=1}^{K}\eta^{1k}-\sum_{i=1}^{n_1}\sum_{k=1}^{K}\lambda_i^{11}E\left(f_{\boldsymbol{s}_i}^{1k}f_{\boldsymbol{s}_0}^{1k}\right)-\sum_{j=1}^{n_2}\sum_{l=1}^{L}\sum_{k=1}^{K}\lambda_j^{12}c_{12}^{lk}E\left(f_{\boldsymbol{s}_i}^{2l}f_{\boldsymbol{s}_0}^{1k}\right)\end{aligned}\tag{13.24}$$

where $c_{12}^{lk}=\langle\xi^{2l},\xi^{1k}\rangle$. Hence, once the representation with the functional principal components of the functional variables involved is used, the variance and the equation system of functional cokriging depend only on the autocovariances and cross-covariances of the scores vectors, which are scalar processes.

13.3.2 Cokriging with *P* Functional Random Fields

The more general goal is the optimization of the spatial functional prediction of $\chi_{\boldsymbol{s}_0}^r(t)\ 1\le r\le P$ at the unsampled site $\boldsymbol{s}_0$ based on the P spatial functional variables,

$$\check{\chi}_{\boldsymbol{s}_0}^r(t)=\sum_{p=1}^{P}\sum_{i=1}^{n_p}\lambda_i^{rp}\chi_{\boldsymbol{s}_i}^p(t)$$

The interest is the minimization of the squared norm of the prediction error given by

$$Q=E\left\|\chi_{\boldsymbol{s}_0}^r(t)-\check{\chi}_{\boldsymbol{s}_0}^r(t)\right\|^2\tag{13.25}$$

For $m = 1, \ldots, P$, the derivatives and the cokriging equations take the form:

$$\frac{\partial Q}{d\lambda_i^{rm}} = -2E\langle \chi_{\boldsymbol{s}_j}^m, \chi_{\boldsymbol{s}_0}^r \rangle + 2\sum_{p=1}^{P}\sum_{i=1}^{n_p} \lambda_i^{rp} E\langle \chi_{\boldsymbol{s}_j}^m, \chi_{\boldsymbol{s}_i}^p \rangle, \quad j = 1, \ldots, n_m \tag{13.26}$$

and

$$E\langle \chi_{\boldsymbol{s}_j}^m, \chi_{\boldsymbol{s}_0}^r \rangle = \sum_{p=1}^{P}\sum_{i=1}^{n_p} \lambda_i^{rp} E\langle \chi_{\boldsymbol{s}_j}^m, \chi_{\boldsymbol{s}_i}^p \rangle, \quad j = 1, \ldots, n_m \tag{13.27}$$

respectively. Replacing (13.26) and (13.27) in the squared norm of the prediction error (13.25), we obtain

$$E\left\| \chi_{\boldsymbol{s}_0}^r(t) - \check{\chi}_{\boldsymbol{s}_0}^r(t) \right\|^2 = E\langle \chi_{\boldsymbol{s}_0}^r, \chi_{\boldsymbol{s}_0}^r \rangle - \sum_{p=1}^{P}\sum_{i=1}^{n_p} \lambda_i^{rp} E\langle \chi_{\boldsymbol{s}_i}^p, \chi_{\boldsymbol{s}_0}^r \rangle \tag{13.28}$$

Now, using the functional principal components representation, we show that (13.28) only depends on the autocovariances and cross-covariances between the score vectors chosen for each random field, that is,

$$E\left\| \chi_{\boldsymbol{s}_0}^r(t) - \check{\chi}_{\boldsymbol{s}_0}^r(t) \right\|^2 = \sum_{k=1}^{K} E\left(f_{\boldsymbol{s}_0}^{rk} f_{\boldsymbol{s}_0}^{rk} \right) - \sum_{i=1}^{n_p}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{p=1}^{P} \lambda_i^{rp} c_{rp}^{lk} E\left(f_{\boldsymbol{s}_i}^{pk} f_{\boldsymbol{s}_0}^{rl} \right) \tag{13.29}$$

where, as before, denoting by η^{rk}, $k = 1, \ldots, K$ the eigenvalues of the observation of $\chi_{\boldsymbol{s}}^r$, we have that

$$\sum_{k=1}^{K} E\left(f_{\boldsymbol{s}_0}^{rk} f_{\boldsymbol{s}_0}^{rk} \right) = \sum_{k=1}^{K} \eta^{rk}$$

and

$$c_{rp}^{kl} = \begin{cases} 1 & \text{If } p = r \text{ and } k = l \\ 0 & \text{If } p = r \text{ and } k \neq l \\ \langle \xi^{pk}, \xi^{rl} \rangle & \text{If } p \neq r \end{cases}$$

In most of the cases, it is sufficient with a few principal components, maybe one or two for each functional random field, due to the fact that the eigenvalue of the first principal component is too much larger than the rest of them $\eta^{1r} \gg \eta^{2r}$. This simplifies the use of the LMC. Note that all spatial processes of scores considered have constant mean, finite variance, and covariance structure depending only on the distance between locations, see (13.17), that is, all processes of score vectors are second-order stationary processes. Bohorquez et al. [11] illustrate and evaluate the performance of this proposal through a simulation study showing good results.

Finally, as a global measure for the quality of the optimal prediction of the functional random field $\chi_{\boldsymbol{s}_0}^r(t)$ at B unsampled sites, we can use

$$\sum_{b=1}^{B} \left(E\left\| \chi_{\boldsymbol{s}_0^b}^r(t) - \check{\chi}_{\boldsymbol{s}_0^b}^r(t) \right\|^2 \right). \tag{13.30}$$

13.4 Optimal Sampling Designs for Spatial Prediction of Functional Data

An optimal sampling design is the one that finds the best combination predictor-design or estimator-design, according to the optimization of a criterion previously established. Therefore, the optimal design criterion must be defined based on the aims of the study. Thus, we establish the methodology to determine the spatial sampling locations that allow to obtain univariate and multivariate optimal spatial functional prediction. For this purpose, we define design criteria for the predictors considered in Sections 13.2 and 13.3.

An optimal design S_n^* is defined by [12] as one that

$$S_n^* = \arg\max_{S_n \in \Xi_n} \Phi(\Theta, S_n) \tag{13.31}$$

where $\Phi(\Theta, S_n)$ is the design criterion and any scalar measure of information obtained with the design S_n that depends on the parameter vector Θ. The design criteria in the spatial sampling context depend on the uncertainty measure of the prediction, which, in turn depends on the spatial covariance structure [9, 11]. Ξ_n is the set of all n-observation designs. However, $D_{\boldsymbol{s}}$ is a continuous set, so there are infinite options for the new locations. Thus, in practice, the criterion is computed over a set $D'_{\boldsymbol{s}} \subset D_{\boldsymbol{s}}$ that contains a finite number of available possibilities previously determined. In addition, it does not make sense to take sites extremely close because the spatial correlation leads to a redundant information and therefore, to a waste of resources. Therefore, D'_s must be built according to some knowledge of the region conditions, the possibility of access, and maybe economic criteria. In other cases, the best option is the evaluation of the criterion over a fine regular grid.

For all spatial processes considered, we assume second-order stationarity, that is, constant mean and finite variance and covariance structure depending only on the distance between locations. We now set the procedure to select the optimal spatial configuration in the sense of minimum uncertainty of prediction. The goal is the optimal prediction of B curves in a set of interest locations $S_0 = \left\{\boldsymbol{s}_0^1, \dots, \boldsymbol{s}_0^B\right\}$. Let

$$S = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n\}$$

be the current set of sampling locations and

$$S_m = \{\boldsymbol{s}_{n+1}, \dots, \boldsymbol{s}_{n+m}\}$$

the set of new locations that must be determined. The enlarged network is then

$$S' = S \cup S_m = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n, \boldsymbol{s}_{n+1}, \dots, \boldsymbol{s}_{n+m}\}$$

Therefore, among all possible subsets of size m such that $S_m \subset D'_{\boldsymbol{s}}$, we must choose the one that minimizes the total variance of the prediction error.

13.4.1 Optimal Spatial Sampling for OFK

The uncertainty associated with the ordinary kriging prediction for an unsampled site $\boldsymbol{s}_0$, is called the trace-variance and is given by

$$\int_T \mathrm{Var}\left(\chi_{\boldsymbol{s}_0}(t) - \tilde{\chi}_{\boldsymbol{s}_0}(t)\right) = \sum_{i=1}^{n} \lambda_i \gamma(\|\boldsymbol{s}_i - \boldsymbol{s}_0\|) - \delta$$

where δ is a Lagrange multiplier. The trace-variance when m new locations $\{\boldsymbol{s}_{n+1}, \ldots, \boldsymbol{s}_{n+m}\}$ are added for the prediction at an unsampled site is $\sum_{i=1}^{n+m} \lambda_i \gamma(\|\boldsymbol{s}_i - \boldsymbol{s}_0\|) - \delta$. Now, the constraint to ensure unbiasedness turns into $\sum_{i=1}^{n+m} \lambda_i = 1$. According to the optimization in (13.11) and the trace-variogram model (13.12), the solution for the weights vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{n+m})$ and the Lagrange multiplier δ are given by

$$\lambda = \left(\gamma + \mathbf{1}\frac{1 - \mathbf{1}^T\Gamma^{-1}\gamma}{\mathbf{1}^T\Gamma^{-1}\mathbf{1}}\right)^T \Gamma^{-1} \quad \text{and} \quad \delta = -\frac{1 - \mathbf{1}^T\Gamma^{-1}\gamma}{\mathbf{1}^T\Gamma^{-1}\mathbf{1}}$$

respectively. Note that $\gamma^T = (\gamma(\|\boldsymbol{s}_1 - \boldsymbol{s}_0\|), \ldots, \gamma(\|\boldsymbol{s}_{n+m} - \boldsymbol{s}_0\|))$ and Γ is a $(n+m) \times (n+m)$ matrix whose (i, i')th element is $\gamma(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|)$. Therefore, the design criterion for the optimal prediction of B curves in a set $S_0 = \left\{\boldsymbol{s}_0^1, \ldots, \boldsymbol{s}_0^B\right\}$ of interest locations using the total trace-variance, is given by

$$\arg\min_{S_m \subset D'_s} \sum_{b=1}^{B} \left(\sum_{i=1}^{n+m} \lambda_i^b \gamma\left(\left\|\boldsymbol{s}_i - \boldsymbol{s}_0^b\right\|\right) - \delta_b \right) \tag{13.32}$$

and depends only on the distances. This predictor admits intrinsic stationarity for the trace-variogram. If the trace-variogram model has to be estimated from the data, its classical empirical estimator given by the method-of-moments as proposed in [6] takes the form

$$\hat{\gamma}(\|h\|) = \frac{1}{2|N(\|h\|)|} \sum_{|N(\|h\|)|} \int_T \left(\chi_{\boldsymbol{s}_i}(t) - \chi_{\boldsymbol{s}_{i'}}(t)\right)^2 dt$$

where for a fixed $\|h\|$, $N(\|h\|) = \{(\boldsymbol{s}_i, \boldsymbol{s}_{i'}) : \|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\| = \|h\|\}$, $i, i' = 1, \ldots, n+m$ with number of elements $|N(\|h\|)|$. Now, model parameters can be estimated as usual.

13.4.2 Optimal Spatial Sampling for FK_{SK}

The total squared norm of the prediction error for an unsampled site $\boldsymbol{s}_0$ based on the enlarged network $S \cup S_m$, applying (13.13), takes the form

$$\sum_{k=1}^{\infty} \eta^k - 2\varsigma\lambda + \lambda^T \Omega \lambda$$

where ς and Ω are the $(n+m)$-vector and the $(n+m)\times(n+m)$-matrix formed by the sum of the sequence of functional autocovariances between observations and the prediction site, and given by

$$\varsigma = \left(\sum_{k=1}^{\infty} E\left(f_{\boldsymbol{s}_i}^k f_{\boldsymbol{s}_0}^k\right)\right), \quad \Omega = \sum_{k=1}^{\infty}\Omega_k \text{ and } \Omega_k = E\left(f_{\boldsymbol{s}_i}^k f_{\boldsymbol{s}_{i'}}^k\right)$$

for $i = 1, \ldots, n+m$. The solution vector with simple kriging is $\lambda = \Omega^{-1}\varsigma$, hence,

$$\sum_{k=1}^{\infty}\eta^k - 2\varsigma\lambda + \lambda^T\Omega\lambda = \sum_{k=1}^{\infty}\eta^k - \varsigma\Omega^{-1}\varsigma$$

reducing the design criterion for the optimal prediction of B curves to

$$\arg\max_{S_m\subset D'_{\boldsymbol{s}}} \sum_{b=1}^{B}\varsigma_b\Omega^{-1}\varsigma_b \tag{13.33}$$

where $\varsigma_b = \left(\sum_{k=1}^{\infty} E\left(f_{\boldsymbol{s}_i}^k f_{\boldsymbol{s}_0^b}^k\right)\right)$, $b = 1, \ldots, B$. The covariance function that determines Ω_k and ς_b depends only on the distances between observations and prediction sites. The value of K that truncates the representation in terms of EFPC can be even more flexible, and we can include more terms until the cumulative variance reaches some prefixed threshold; this method does not use cross-covariances between score vectors and fitting the model for autocovariances is simpler.

13.4.3 Optimal Spatial Sampling for FK_{CK}

Given that $\xi(t)$ is known, the uncertainty of the prediction error for an unsampled site based on the enlarged network $S\cup S_m$ only depends on $\sigma^2_{\boldsymbol{f}_{\boldsymbol{s}_0}-\boldsymbol{f}^*_{\boldsymbol{s}_0}}$, see (13.19),

$$\sigma^2_{\boldsymbol{f}_{\boldsymbol{s}_0}-\boldsymbol{f}^*_{\boldsymbol{s}_0}} = \mathrm{Tr}\left(\Sigma_0\right) - \mathrm{Tr}\left(\sum_{i=1}^{n+m}\left(\boldsymbol{\Sigma}_{(\boldsymbol{s}_0,\boldsymbol{s}_i)}\Gamma_i\right)\right) = \sum_{k=1}^{K}\eta^k - \mathrm{Tr}\left(\sum_{i=1}^{n+m}\left(\boldsymbol{\Sigma}_{(\boldsymbol{s}_0,\boldsymbol{s}_i)}\Gamma_i\right)\right)$$

The design criterion for the optimal prediction of B curves in a set $S_0 = \left\{\boldsymbol{s}_0^1, \ldots, \boldsymbol{s}_0^B\right\}$ of interest locations using the total prediction error variance, goes through the calculation of

$$\arg\max_{S_m\subset D'_{\boldsymbol{s}}} \sum_{b=1}^{B}\left(\mathrm{Tr}\left(\sum_{i=1}^{n+m}\left(\boldsymbol{\Sigma}_{(\boldsymbol{s}_0^b,\boldsymbol{s}_i)}\Gamma_i\right)\right)\right) \tag{13.34}$$

Denoting by $\Delta_0^b = \left(\boldsymbol{\Sigma}_{(\boldsymbol{s}_0^b,\boldsymbol{s}_i)}\right)$ $i = 1, \ldots, n+m$, see (13.14), the cokriging solution is $\boldsymbol{\Gamma} = \sum^{-1}\Delta_0^b$. Therefore, the LMC and criterion (13.34) depend only on the distance between observations and prediction sites due to the assumption of second-order stationarity.

13.4.4 Optimal Spatial Sampling for Functional Cokriging

We need to design or redesign the p sets $S_p = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n_p}\}$ $p = 1, \ldots, P$, or at least those that can be changed, of observed spatial locations ensuring an optimal spatial functional prediction of $\chi^r_{\boldsymbol{s}}(t)$ in a set of interest locations $S_0 = \left\{\boldsymbol{s}_0^1, \ldots, \boldsymbol{s}_0^B\right\}$ based on the P spatially correlated functional random fields. We now set the procedure to select the optimal spatial configuration in the sense of the total minimum square norm of the prediction error for functional cokriging, see (13.30). Suppose first that m_p stations can be added for the observation of each random field $\chi^p_{\boldsymbol{s}}(t)$, $p = 1, \ldots, P$. Therefore, the enlarged network for each case is then

$$S'_p = S_p \cup S_{m_p} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n_p}, \boldsymbol{s}_{n_p+1}, \ldots, \boldsymbol{s}_{n_p+m_p}\}, \quad p = 1, \ldots, P$$

Let $\bigcup_{p=1}^{P} S_{m_p} = S_{m_1} \cup \cdots \cup S_{m_p} \subset D'_{\boldsymbol{s}}$ be the set of new locations that must be determined. Therefore, among all possible subsets $\bigcup_{p=1}^{P} S_{m_p}$, we must choose the one that minimizes the total minimum square norm of the prediction error for functional cokriging. Therefore, according to (13.29) and (13.30) the design criterion is given by

$$\arg\min_{\bigcup_{p=1}^{P} S_{m_p} \subset D'_{\boldsymbol{s}}} \sum_{b=1}^{B} E\left\| \chi^r_{\boldsymbol{s}_0^b}(t) - \check{\chi}^r_{\boldsymbol{s}_0^b}(t) \right\|^2 \tag{13.35}$$

where $\sum_{k=1}^{K} \eta^{rk}$ is constant and, therefore, the criterion (13.35) turns into

$$\arg\max_{\bigcup_{p=1}^{P} S_{m_p} \subset D'_{\boldsymbol{s}}} \sum_{b=1}^{B} \sum_{i=1}^{n_p+m_p} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{p=1}^{P} \lambda_i^{rp} c_{rp}^{lk} E\left(f_{\boldsymbol{s}_i}^{pk} f_{\boldsymbol{s}_0^b}^{rl}\right). \tag{13.36}$$

The criterion (13.36) establishes the general case, but, frequently, all random fields are measured at the same set of places $S = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$. Here, the optimization is over the possible sets $S_m \subset D'_{\boldsymbol{s}}$, and there is only one enlarged network $S \cup S_m \subset D'_{\boldsymbol{s}}$ for all random fields.

Once again, due to the second-order stationarity, the LMC and criterion (13.36) depend only on the distance between observations and prediction sites. The numbers K and L of the chosen principal components are usually small, 1 or 2; therefore, the iterative computation of the inverse of covariance matrix does not represent a high computational cost, if the spatial points are not too dense.

All design criteria shown in this section depend on parameter $\boldsymbol{\Theta}$. If this parameter is unknown and has to be estimated, the design is no longer optimum, it is only ***locally optimum*** because the optimization process is based also on $\boldsymbol{\Theta}$ and not only on the design S_n. Thus, (13.31) turns into $S_n^* = \arg\max_{S_n \in \Xi_n} \Phi(\hat{\boldsymbol{\Theta}}, S_n)$. For the covariance structures necessary for the predictors given in Sections 13.2 and 13.3, the classical empirical variogram and cross-variogram can be used, and the model parameters can be fitted by ordinary or weighted least squares to avoid

distributional assumptions, or by methods based on the likelihood function. We use the plug-in estimators in the LMC and trace-variogram models to carry on the optimization of the sampling criteria for each case. That is, in every place, where the terms $\eta^k \rho^k \left(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|; \boldsymbol{\Theta} \right)$ or $\gamma(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|; \boldsymbol{\Theta})$ appear we replace them with $\eta^k \rho^k(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|; \hat{\boldsymbol{\Theta}})$ and $\gamma(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|; \hat{\boldsymbol{\Theta}})$. Harville and Jeske [13] propose a correction of the kriging variance to incorporate the uncertainty due to the lack of knowledge of $\boldsymbol{\Theta}$. Zhu and Stein [14] find that this correction could be important only for weak spatial autocorrelation cases. Nevertheless, this correction is based on the Gaussian assumption, on the use of the maximum likelihood or Restricted Maximum Likelihood estimation and it depends on $\boldsymbol{\Theta}$. Therefore, the best option is to use the plug-in method as long as the spatial autocorrelation is moderate or strong, see [15].

13.5 Real Data Analysis

We analyze network data for air quality in México city during the dry season because in the rainy season all air pollutants diminish. The data correspond to consecutive hours from 01 January 2015, at 1:00 a.m. to 30 May 2015, at 12:00 a.m., at 23 environmental stations (see Figure 13.1a). The stations in the Mexico City's automatic air quality (RAMA) monitor hourly particulate matter up to 10 μm in size (PM10) and nitrogen dioxide (NO_2), among others. See [11] for details about the network and of the adverse effects of the PM10 and NO_2 on human health and human-made materials. The temperature (Temp) data are taken from the Mexico City's meteorological monitoring network (REDMET). RAMA and REDMET have 15 stations in common. The Secretariat of Environment of México currently operates the network of air quality monitoring in order to obtain, process, and disclose air quality to assess compliance with standards and the basis for the definition of policies pollution control. The data are obtained from the Automatic Monitoring System [16]. To convert the datasets to curves, we use *B-splines* basis functions of order 4 with equally spaced knots and a smoothing parameter 0.000 01. For the dataset of PM10, we use a set of 163 *B-splines* basis functions, for the dataset of NO_2, we use a set of 157 *B-splines* basis functions, and for the data set of Temp, we use 121 *B-splines* basis functions. Figure 13.1b–d shows the curves for the last week in the dataset. The first principal component explains 75%, 84.6%, and 85.7% of the variability for PM10, NO_2, and Temp, respectively, and the second principal component only accounts for 13.9%, 13.8%, and 13.1%. Thus, using 85% as a threshold, we include two score vectors for PM10, while only the first score vector is included for NO_2 and Temp. Our interest is the spatial functional prediction of PM10 using NO_2 and Temp as functional covariates. According to the notation in Section 13.3, PM10 is $\chi^1_{\boldsymbol{s}}(t)$, NO_2 is $\chi^2_{\boldsymbol{s}}(t)$, and

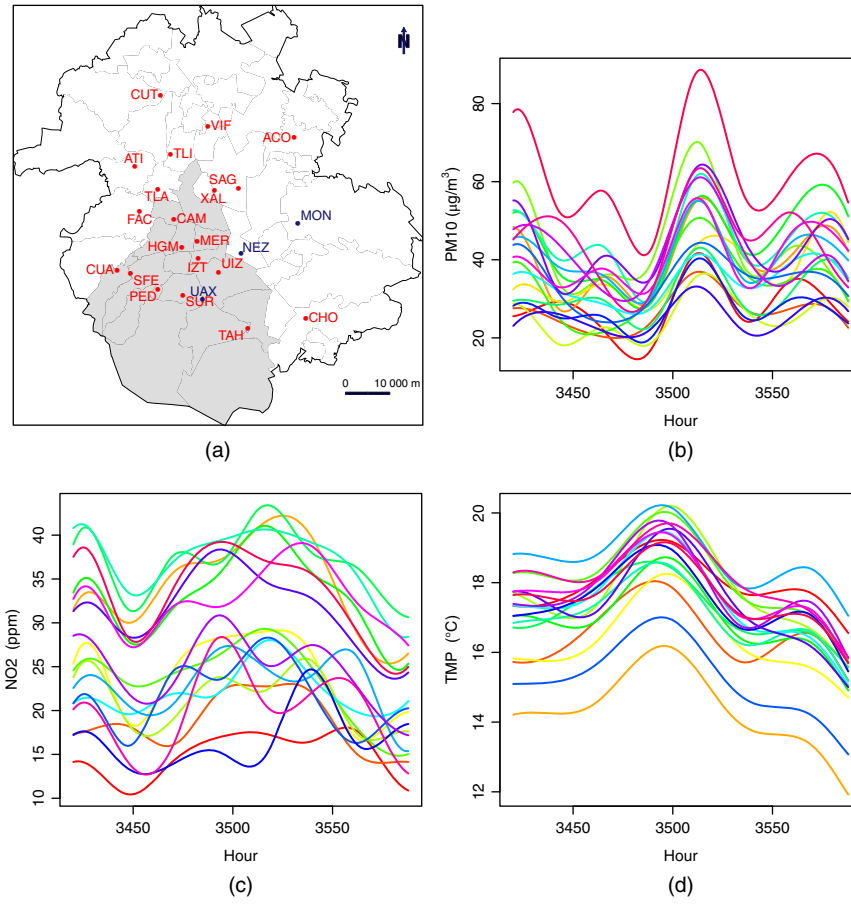


Figure 13.1 (a) México city. Air quality network RAMA (stations shown in light gray). The dark gray points are the stations that measure temperature but belong to REDMET. (b) PM10, 23–30 May 2015. (c) NO_2, 23–30 May 2015. (d) Temperature, 23–30 May 2015.

Temp is $\chi_{\boldsymbol{s}}^3(t)$. Figure 13.2 shows the empirical and theoretical variograms fitted according to the LMC. We use two nested Matérn structures linearly combined, with smoothing parameters 0.1 and 5, and ranges 3000 and 13 000. Thus, γ_f^{11} and γ_f^{12} are the variograms for the first two principal components of PM10, γ_f^{21} and γ_f^{31} are the variograms for the score vectors corresponding to the first principal component of NO_2 and Temp, respectively, and the rest of variograms in Figure 13.2 and in Table 13.1 are the cross-variograms between each pair of score vectors. From the empirical variograms, there is no reason to assume discontinuity at the origin, since there is no jump in $\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\| = 0$. Therefore, we kept the nugget

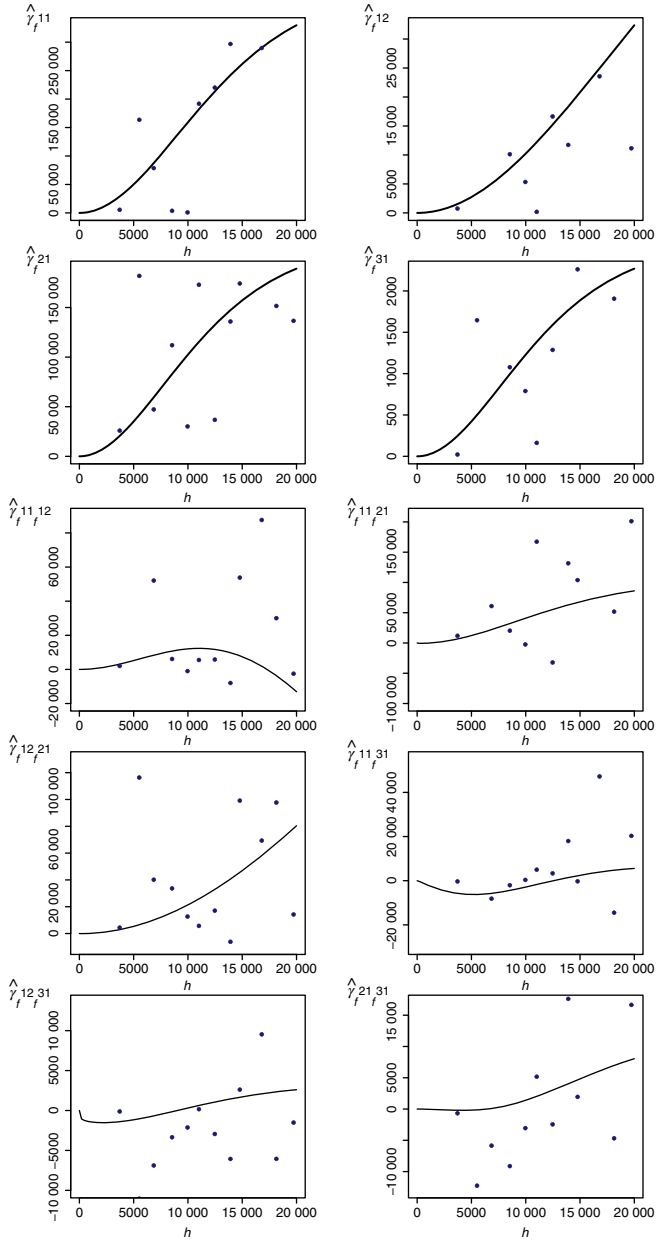


Figure 13.2 Empirical and theoretical variograms fitted according to the linear model of coregionalization.

Table 13.1 Nested variogram components of the linear model of coregionalization model using two nested Matérn structures.

$\hat{\gamma}_f^{pk}$	$\hat{\gamma}_{f^{11}}$	$\hat{\gamma}_{f^{21}}$	$\hat{\gamma}_{f^{31}}$	$\hat{\gamma}_{f^{11}f^{12}}$	$\hat{\gamma}_{f^{11}f^{21}}$	$\hat{\gamma}_{f^{11}f^{31}}$	$\hat{\gamma}_{f^{12}}$	$\hat{\gamma}_{f^{12}f^{21}}$	$\hat{\gamma}_{f^{12}f^{31}}$	$\hat{\gamma}_{f^{21}f^{31}}$
$\upsilon = 0.1$	216 778.0	24 415.3	56 381.7	−11 931	71 160.1	104 764.3	−11 931.2	31 932.8	−3508.696	−6856.4
$\upsilon = 5$	1 923 456.9	386 064.8	1 181 738.7	202 568	−680 077.6	−92 372.9	202 568.5	388 685.0	3251.8	164 083.5

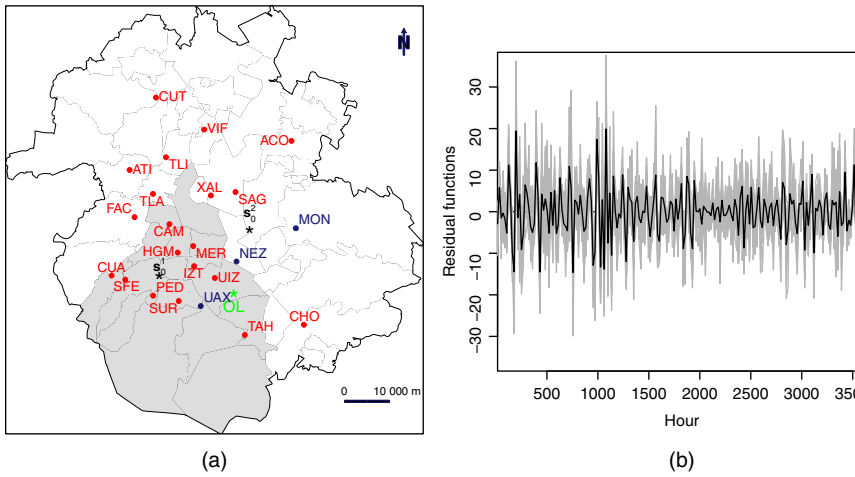


Figure 13.3 (a) Optimal location for one additional station. (b) Cross-validation residuals and residual mean.

parameter fixed and equal to zero. To illustrate the methodology for optimal sampling designs, we choose two interests locations for prediction, $\boldsymbol{s}_0^1$ and $\boldsymbol{s}_0^2$, see Figure 13.3a. As for the set $D'_{\boldsymbol{s}}$, we took the sampling grid of 375 spatial points separated by 2 km, 25 points west to east and 15 south to north, restricted to the area with stations. Figure 13.3a shows the locations of interest $\boldsymbol{s}_0^1$ and $\boldsymbol{s}_0^2$ and the optimal sampling location (OL) to add a new station keeping fixed the current network, and using (13.36). For the optimization procedure, we use simulated annealing [17] with state given by the spatial sampling design criterion applied at each iteration. The energy function is given by the criterion (13.36). In order to assess the quality of the spatial prediction based on the functional cokriging, we use the leave-one-out functional cross-validation method [18]. Although there are some large residuals at the beginning of the season due to the variation of pollutants in this period, the performance is good; the residual mean function varies close to zero, from −20 to 20 in most of the cases, see Figure 13.3b.

13.6 Discussion and Conclusions

We have presented univariate and multivariate spatial functional predictors based on the representation of functions in terms of its EFPC. Based on this approach, we have shown that the system of equations for functional kriging and functional cokriging depend only on the autocovariances and cross-covariances of the scalar random field of the associated score vectors. An additional advantage of our proposal is that it only uses the functional principal component representation of each random field and does not require multivariate functional principal component analysis. Thus, the advantages and drawbacks of the functional kriging and functional cokriging are the same of the scalar cokriging. The limitation of the dimension for the LMC is the most critical issue of the cokriging method. However, this difficulty is solved by the fact that the EFPC representation does not usually need too many eigenfunctions, even with one or two could be sufficient in most of the cases, making feasible to use this type of covariance model. The spatial optimal sampling design for spatial functional data is a natural extension of its counterpart with scalar variables. The criteria used here are useful when moving only one location or even lots of them. Once the covariance between curves has been modeled, the optimization process to find an optimal design has the same computational effort as in the scalar case. Its performance depends on the quality of the covariance parameter estimators, and on the optimization algorithm used. Some networks contain p mobile stations, $S = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_p, \boldsymbol{s}_{p+1}, \ldots, \boldsymbol{s}_n\}, p < n$, then the criterion is computed for all possible sets $S_p = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_p\} \subset D'_{\boldsymbol{s}}$. Finally, these criteria allow determining the performance of the whole set so that all sampling locations could be changed. Also, our statistical criteria can be used to decide the number of observation sites n. In this case, a maximum prediction variance is previously determined, and the optimization is carried out for $n = 1$. Then keeping this location fixed, the second location is optimized, and so on until finding that n that reaches the established threshold.

References

1 Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer-Verlag.
2 Reed, M. and Simon, B. (1980). *Methods of Modern Mathematical Physics I: Functional Analysis*. San Diego, CA: Academic Press, Inc.
3 Bongiorno, E.G., Salinelli, E., Goia, A., and Vieu, P. (2014). *Contributions in Infinite-Dimensional Statistics and Related Topics*. Societa Editrice Esculapio.
4 Horvath, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer.

5 Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*, vol. 149. Springer.

6 Giraldo, R., Delicado, P., and Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics* 18 (3): 411–426.

7 Giraldo, R. (2014). Cokriging based on curves, prediction and estimation of the prediction variance. *InterStat* 2: 1–30.

8 Nerini, D., Monestiez, P., and Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis* 101 (2): 409–418.

9 Bohorquez, M., Giraldo, R., and Mateu, J. (2015). Optimal sampling for spatial prediction of functional data. *Statistical Methods and Applications*. https://doi.org/10.1007/s10260-015-0340-9.

10 Myers, D. (1982). Matrix formulation of co-kriging. *Mathematical Geology* 14 (3): 249–257.

11 Bohorquez, M., Giraldo, R., and Mateu, J. (2017). Multivariate functional random fields: prediction and optimal sampling. *Stochastic Environmental Research and Risk Assessment* 31 (1): 53–70.

12 Müller, W. (2007). *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*. Springer-Verlag.

13 Harville, D. and Jeske, D. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association* 87 (419): 724–731.

14 Zhu, Z. and Stein, M. (2006). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics* 11 (1): 24–44.

15 Schabenberger, O. and Gotway, C. (2004). *Statistical Methods for Spatial Data Analysis*. CRC Press.

16 México City Air Quality Monitoring Network (Sistema de monitoreo atmosfé rico de la Ciudad de México) (2016). http://www.aire.df.gob.mx/default.php (accessed 30 July 2016).

17 Brooks, S. and Morgan, B. (1995). Optimization using simulated annealing. *Journal of the Royal Statistical Society: Series D (The Statistician)* 44 (2): 241–257.

18 Montero, J., Fernandez-Aviles, G., and Mateu, J. (2015). *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. Wiley.

Part III

Spatio–Temporal Functional Data

14

Spatio–temporal Functional Data Analysis

Gregory Bopp[1], John Ensley[1], Piotr Kokoszka[2], and Matthew Reimherr[1]

[1] *Department of Statistics, Pennsylvania State University, USA*
[2] *Department of Statistics, Colorado State University, USA*

14.1 Introduction

The objective of this chapter is to review some recent developments in inference for spatio–temporal functional data. We consider data of the form

$$X_n(\mathbf{s}_k; t_j), \quad 1 \le n \le N,\ 1 \le k \le K,\ 1 \le j \le J.$$

To focus attention, it is convenient to consider the following prototypical example. The data are available over N years; the index n refers to year. They are observed at K locations, so k is a location index. In each year, n, measurements or averages are available at times t_j, which are typically days or months, so $J = 365$ or $J = 12$. For example, $X_n(\mathbf{s}_k; t_j)$ can be the maximum daily temperature on day t_j of year n at location $\mathbf{s}_k$. Another example is average concentration of a pollutant in month t_j of year n at a monitoring station located at site $\mathbf{s}_k$. Similar data structures occur in brain studies, but n has a different interpretation; it is a subject index. So $X_n(\mathbf{s}_k; t_j)$ is a measurement on subject n, at a brain location $\mathbf{s}_k$ at time t_j. In abstract terms, such data can be viewed as a discretely observed sample of a spatio–temporal field $X(\mathbf{s}; t)$, with a continuous spatial index $\mathbf{s}$ and a continuous temporal index t. In climate and pollution studies, there is generally some dependence in the index n. In brain studies, independence in n can safely be assumed for unrelated subjects. For temperature data, N is generally about 100 as consistent climatological measurements start around the end of the nineteenth century. The number of locations, K, can be several hundred per large, developed country or continent. The meteorological stations however start operation at different times, some are closed or change locations, so these data have many gaps distributed unevenly over time and space. Generally, the number of gaps decreases with year n, but this is not the case for all types of data, see Section 14.5. In brain studies, N is generally small,

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

10–20 subjects, but K can be huge. The number of pixels can exceed several thousand. The number of temporal measurements can be large or small, depending on what is measured. In this chapter, we focus on methods developed for geophysical data; we think of relatively sparsely distributed locations $\mathbf{s}_k$, and relatively dense t_j. It is convenient to think that we observe functions $X_n(\mathbf{s}_k; t)$, with a continuous argument t, one function per year and per location. One could refer to such data as functional and spatio–temporal, with index n corresponding to time and $\mathbf{s}_k$ to space, with the units of observations being functions. Some research has recently been done for data of this form, and it is our objective in this chapter to survey and illustrate the methods, we are most familiar with.

Spatial (rather than spatio-temporal) functional data have been considered for at least two decades. Such data are a collection of curves at spatial locations $\{\mathbf{s}_k\}$. Each curve $X(\mathbf{s}_k; \cdot)$ can be the concatenation of the curves $X_n(\mathbf{s}_k; \cdot)$ considered above, but in most applications that stimulated the development of methodology for spatial functional data, $X(\mathbf{s}_k; \cdot)$ is an average of the N curves $X_n(\mathbf{s}_k; \cdot)$. For example, the $X(\mathbf{s}_k; \cdot)$ can be average annual temperature curves computed using data collected over several decades. The Canadian temperature data considered in [1] is a well-known example. Methodology for spatial functional data is presented in [2, 3], where many references are given.

The remainder of this chapter is organized as follows: We begin in Section 14.2 with a test of randomness whose null hypothesis is that the K dimensional vectors of functions $X_n(\mathbf{s}_k; \cdot)$ are independent and identically distributed across n. Such vectors are often called functional panels, and the test is intended to verify if these panels form a simple random sample. In particular, under the null hypothesis, there can be dependence in space, but no dependence in time n. The assumption of a simple random sample is strong, a weaker assumption is stationarity. A simple departure from stationarity is a change-point model. The data are assumed to have the form $X_n(\mathbf{s}, t) = \mu_n(\mathbf{s}; t) + \varepsilon_n(\mathbf{s}; t)$. The goal of a change-point test is to determine if the mean functions, μ_n, are the same across n or if they change at some unknown point. This problem is discussed in Section 14.3. The second-order structure of a random field $X(\cdot, \cdot)$ is described by the covariance function $\sigma(\mathbf{s}, \mathbf{s}'; t, t') = \mathrm{Cov}(X(\mathbf{s}, t), X(\mathbf{s}', t'))$. Theoretical and computational aspects of most procedures can be significantly simplified if one can assume that $\sigma(\mathbf{s}, \mathbf{s}'; t, t') = \mathcal{U}(\mathbf{s}, \mathbf{s}')\mathcal{V}(t, t')$, that is, that the spatio–temporal covariance function factors into the product of a purely spatial and purely temporal covariance functions. If the above decomposition holds, we say that the functional random field is *separable*. In particular, the spatial dependence structure is the same for any time t. Separability tests are explained in Section 14.4. Section 14.5 is devoted to the problem of testing for trends in in spatio–temporal functional data. We conclude with fairly recent research on extreme events defined in terms of the data $X_n(\mathbf{s}_k, t_j)$. This work, whose many aspects are still under development, is outlined in Section 14.6.

14.2 Randomness Test

In this section, we discuss summarize the work of [4], testing if the spatio–temporal observations, $X_n(\mathbf{s}, t)$, are independent and identically distributed across n. In settings where n represents year, it is maybe reasonable to assume independence, while if n represents days such assumption becomes more precarious. Regardless, it is useful to determine if this assumption holds so that further statistical analyses remain valid. Thus, the null hypothesis we aim to test is

$$H_0 \colon X_n(\mathbf{s}, t) \text{ are independent and identically distributed across } n.$$

While there are many potential alternative hypotheses, we focus on testing for serial dependence.

To test H_0, we outline a generalization of the Box–Ljung test from time series analysis, by considering lagged covariance operators in place of autocorrelations. In particular, we define

$$C_h(\mathbf{s}, t, \mathbf{s}', t') = \text{Cov}(X_n(\mathbf{s}, t), X_{n+h}(\mathbf{s}', t')).$$

To evaluate H_0, we test if C_h is zero for $h = 1, \dots, H$, where H is some predetermined number of lags. The approach used by Kokoszka et al. [4] is based on reducing the temporal dimension using functional principal component analysis (FPCA), and if needed, reducing the spatial domain using multivariate principal component analysis (PCA). The temporal dimension reduction is done at each spatial location separately, though one could also use a pooling approach as described in Section 14.4. Using the Karhunen–Loève (KL) expansion, we approximate

$$X_n(\mathbf{s}_k, t) \approx \mu(\mathbf{s}_k, t) + \sum_{i=1}^{p_k} \xi_{n;ki} \upsilon_{k,i}(t).$$

Here the eigenfunctions $\hat{\upsilon}_{k,i}$ and their corresponding eigenvalues, $\lambda_{k,i}$, are allowed to change with location, $\mathbf{s}_k$. Note that if one assumes separability as discussed in Section 14.4, then the population-level eigenfunctions are the same at every location and thus once can pool across space to estimate them, as opposed to estimating them individually at each spatial location. While our results are stated assuming the eigenfunctions are known, the asymptotic results do not change if they are replaced with consistent estimates.

We now work with the scores the $\{\xi_{n,k,i}\}$ to evaluate H_0, in particular, if H_0 is true, then the scores should also be iid across n. To test this assumption, we stack the scores into a single vector for each n:

$$\boldsymbol{\xi}_n = (\xi_{n;11}, \xi_{n;21}, \dots, \xi_{n;Kp})^\top.$$

We then estimate the lagged covariance terms

$$\hat{\mathbf{V}}_h = N^{-1} \sum_{n=1}^{N-h} \boldsymbol{\xi}_n \otimes \boldsymbol{\xi}_{n+h} \quad \text{and} \quad \hat{\mathbf{C}}_0 = N^{-1} \sum_{n=1}^{N} \boldsymbol{\xi}_n \boldsymbol{\xi}_n^\top,$$

where $\otimes$ here denotes a Kronecker product. If H_0 is true, then $\boldsymbol{\xi}_n$ is independent of $\boldsymbol{\xi}_{n+h}$ and one can show with some linear algebra that

$$\text{Cov}(\widehat{\mathbf{V}}_h) = \frac{(N-h)}{N^2}\mathbf{C}_0 \otimes \mathbf{C}_0,$$

where $\mathbf{C}_0 = E[\boldsymbol{\xi}_n \boldsymbol{\xi}_n^\top]$. Thus [4] proposed the following test statistic:

$$\widehat{Q}_n = N^{-1}\sum_{h=1}^{H} \widehat{\mathbf{V}}_h^\top (\widehat{\mathbf{C}}_0 \otimes \widehat{\mathbf{C}}_0)^{-1}\widehat{\mathbf{V}}_h,$$

note that $(\mathbf{C}_0 \otimes \mathbf{C}_0)^{-1} = \mathbf{C}_0^{-1} \otimes \mathbf{C}_0^{-1}$, which is more computationally convenient.

When H, p_k, and K are all relatively small then, under H_0, one can use the following asymptotic result to get a p-value:

$$\widehat{Q}_n \xrightarrow{d} \chi^2_{Hp^2},$$

where $p = \sum_k p_k$. However, if any of these quantities is large, one can instead use the result

$$\frac{\widehat{Q}_n - Hp^2}{2p\sqrt{H}} \xrightarrow{d} \mathcal{N}(0,1).$$

Note that this result is based on letting $p \to \infty$ with the sample size, see [4] for more details.

As an illustration, we will apply this test to a set of daily temperature observations from Russian meteorological stations. The full database contains daily records from 518 Russian weather stations numbered according to the names given by the World Meteorological Organization (WMO). It was put together by the Carbon Dioxide Information Analysis Center (CDIAC), who collaborated with National Oceanic and Atmospheric Administration (NOAA's) National Climatic Data Center (NCDC) and the All-Russian Research Institute for Hydrometeorological Information–World Data Center (RIHMI–WDC). The data was made available as a part of the Agreement on Protection of the Environment [5]. Observations exist from 1881 through 2010, although most stations do not have records for this entire period. It is freely available at https://cdiac.ess-dive.lbl.gov/ndps/russia_daily518.html.

To limit the amount of missing data, we selected a subset of 220 stations with observations from 1980 to 2009. None of these 220 stations have more than five missing daily observations in a given year, and most have no missing observations. A map of the locations of these stations is given in Figure 14.1. An example of five such curves form 2000 are given in Figure 14.2.

We take $X_n(\mathbf{s}_k, t)$ to correspond to the daily maximum temperature observed at location $\mathbf{s}_k$, in year n, on the tth day. The number of projections, p_k, are chosen separately for each series; each is chosen to explain at least 85% of the variance

Figure 14.1 Locations of the 220 Russian weather stations, with 14 stations around Moscow marked using solid circles.

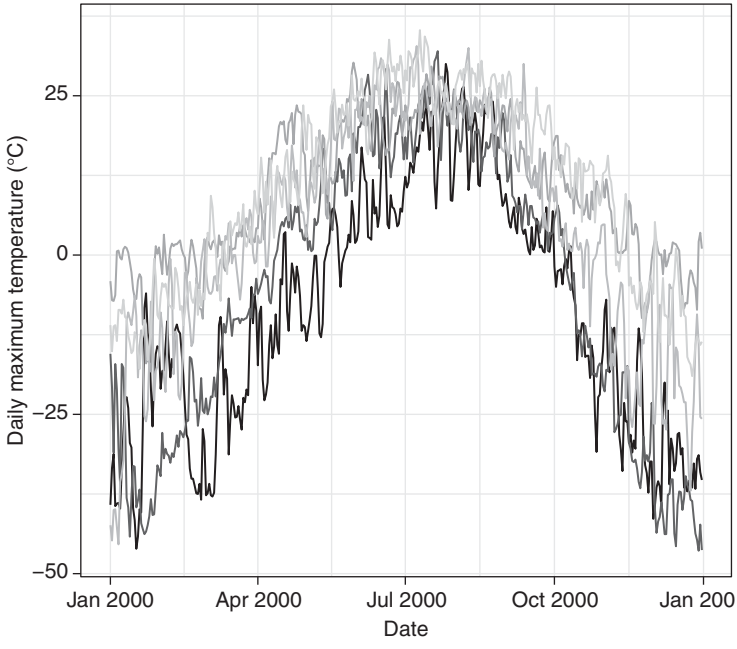


Figure 14.2 Daily temperature maxima for five weather stations during 2000.

of each curve. To minimize the potential for nonstationarity in the data, the functional objects are fit to the residuals of mean detrended data for a mean function which is linear in time (see Section 14.5). Randomness tests are performed using maximal lags H between 1 and 10. We carried out the test on two sets of stations, the first includes all 220, Table 14.1, while the second focuses on the 14 stations around Moscow, Table 14.2. Interestingly, there is strong evidence against the randomness assumption when looking across Russian, however, when focusing on

Table 14.1 Randomness test results applied to the Russian weather data.

Maximal lag	Test statistic	*p*-value
$H = 1$	87.6	$< 10^{-5}$
$H = 2$	120.2	$< 10^{-5}$
$H = 3$	146.2	$< 10^{-5}$
$H = 4$	167.0	$< 10^{-5}$
$H = 5$	185.2	$< 10^{-5}$
$H = 6$	201.1	$< 10^{-5}$
$H = 7$	215.5	$< 10^{-5}$
$H = 8$	228.2	$< 10^{-5}$
$H = 9$	240.7	$< 10^{-5}$
$H = 10$	258.2	$< 10^{-5}$

Table 14.2 Randomness test results applied to a subset of 14 Russian weather stations near Moscow.

Maximal lag	Test statistic	*p*-value
$H = 1$	0.441	0.330
$H = 2$	1.455	0.073
$H = 3$	1.564	0.059
$H = 4$	1.867	0.031
$H = 5$	1.811	0.035
$H = 6$	1.915	0.028
$H = 7$	1.855	0.032
$H = 8$	1.821	0.034
$H = 9$	1.732	0.042
$H = 10$	1.668	0.048

the region around Moscow the evidence becomes much weaker. This will be an ongoing theme in the chapter; the spatio–temporal dynamics are much simpler when focusing on smaller regions.

14.3 Change-Point Test

Change-point analysis has been an area of interest for statisticians for decades. Change-point methods allow researchers to determine if data collected over time is stationary, or if the underlying dynamics of the data has changed at some point. These tools can be used to justify stationarity assumptions so that other procedures that rely on stationarity remain valid. However, the changes themselves might also be of scientific import. Change-point detection among temperature extremes is a question of interest in climatology, as extreme temperatures can have devastating effects on ecosystems and crop yields [6]. In this chapter, we will describe a spatio–temporal change-point procedure and apply the discussed tools to the Russian data on daily maximum temperatures.

The approach described here was given by Aston et al. [7]. There is an accompanying R package, *scpt*, that can be downloaded from R-Forge and can be used to carry out the procedures described in this section. The goal here is to test if the mean function, $E[X_n(\mathbf{s}, t)] = \mu_n(\mathbf{s}, t)$ remains constant over n or if there is some sudden change. For example, in the Russian weather data, clearly temperature is going to vary with location, $\mathbf{s}$, and over the course of a year, t; however, it is of interest to know if the temperature is constant from year to year, n. The goal is therefore to test

$$H_0 : \mu_1(\mathbf{s}, t) = \ldots = \mu_N(\mathbf{s}, t) \quad \text{against}$$
$$H_A : \mu_1(\mathbf{s}, t) = \ldots = \mu_{n^*}(\mathbf{s}, t) \neq \mu_{n^*+1}(\mathbf{s}, t) = \ldots = \mu_N(\mathbf{s}, t).$$

We assume that the observations can be expressed as

$$X_n(\mathbf{s}, t) = \mu_n(\mathbf{s}, t) + \varepsilon_n(\mathbf{s}, t),$$

where ε_n is assumed to be iid across n with a finite covariance. We will also assume that the covariance of ε_n is separable, as describe in Section 14.4. Assuming separability is not strictly necessary, but makes it much easier to estimate the spatio–temporal covariance function, which will be important for the procedures described below. An additional feature of separability is that the eigenfunctions used for temporal FPCA, v_i, are the same across space, meaning that one can pool across space to estimate them, which typically results in much sharper estimates.

The test statistics described by Aston et al. [7] all revolved around the following CUSUM statistic:

$$\Gamma_r(\mathbf{s}_k, t) := \sum_{n=1}^{r} X_n(\mathbf{s}_k, t) - \frac{r}{N} \sum_{n=1}^{N} X_n(\mathbf{s}_k, t), \quad r = 1, \ldots, N.$$

Intuitively, if H_0 is true, then Γ_r should not be too large, regardless of r, whereas H_A is true, then Γ_r should be large when r is close to n^*. Three test statistics were presented in [7], each of which takes a slightly different approach to determining if Γ_r is large for some r:

$$\widehat{\Lambda}_1 = \frac{1}{N^2}\sum_{k=1}^{K}\hat{w}_k\sum_{i=1}^{p}\hat{\lambda}_i^{-1}\sum_{r=1}^{N}\left\langle\Gamma_r(\mathbf{s}_k),\hat{v}_i\right\rangle^2,$$

$$\widehat{\Lambda}_2 = \frac{1}{N^2}\sum_{k=1}^{K}\hat{w}_k\sum_{i=1}^{p}\sum_{r=1}^{N}\left\langle\Gamma_r(\mathbf{s}_k),\hat{v}_i\right\rangle^2,$$

$$\widehat{\Lambda}_2^{\infty} = \frac{1}{N^2}\sum_{k=1}^{K}\hat{w}_k\sum_{i=1}^{\infty}\sum_{r=1}^{N}\left\langle\Gamma_r(\mathbf{s}_k),\hat{v}_i\right\rangle^2 = \frac{1}{N^2}\sum_{k=1}^{K}w_k\sum_{r=1}^{N}\|\Gamma_r(\mathbf{s}_k)\|^2.$$

The terms $\{\hat{\lambda}_i,\hat{v}_i\}$ are the estimated (temporal) eigenvalues and eigenfunctions, which are estimated by pooling across space. How this pooling is done is up to the user as it will not affect the asymptotic distribution of the test statistics as long as the estimates are consistent. One could use the procedure described in Section 14.4; however, those estimates will be poor choices under H_A and may reduce power (since there is a change in the mean). An alternative as described in [7] is to work with the differences, with respect to n, of the series as this minimizes the effect of any change-points. Once the differences are taken, the procedures described in Section 14.4 can be used as well some alternative approaches described in [7]. If the ε_n are iid, then the covariance function of the difference is simply two times the covariance of the original, furthermore, under H_A, changes in the mean will have a relatively small impact as long as they are not too large.

The weights, $\hat{w}_k$, can be any data-driven choice as long as they converge in probability to some fixed values, w_k, as $N\to\infty$. Practically, the weights should be chosen to maximize the power of the test, though this is challenging as it will depend heavily on the alternative. One option that works well is to choose the weights so that the pooled estimate of $C(t,s)$, the temporal covariance, is unbiased and has the smallest possible variance. If one assumes that the data is Gaussian, then these weights are given by

$$\hat{\mathbf{w}} = (\mathbf{1}^\top\hat{\mathbf{\Sigma}}^{-1}\mathbf{1})^{-1}\hat{\mathbf{\Sigma}}\mathbf{1},$$

where $\hat{\mathbf{\Sigma}}$ is a matrix of the estimated spatial covariances. We refer the reader to [7] for more details.

The first procedure, $\hat{\Lambda}_1$, reduces the temporal dimension of the problem and at each spatial location applies a multivariate change-point test, which normalizes by the estimated eigenvalues, $\hat{\lambda}_i$. The second procedure is the same but removes the normalization by $\hat{\lambda}_i$. Often, one must be especially careful when normalizing by small eigenvalues as they can produce an unstable test statistic and harm the

specificity of the test. The third test is the same as the second, but without reducing the dimension in time, instead using the $L^2[0,1]$ norm.

Under appropriate assumptions, the asymptotic distribution of each of described test statistics is a weighted sum of sum of normed Brownian bridges:

$$\widehat{\Lambda}_1 \xrightarrow{D} \Lambda_1 := \sum_k^K w_k \sum_{i=1}^p \int B_{ik}(t)^2 \, dt,$$

$$\widehat{\Lambda}_2 \xrightarrow{D} \Lambda_2 := \sum_k^K w_k \sum_{i=1}^p \lambda_i \int B_{ik}(t)^2 \, dt,$$

$$\widehat{\Lambda}_2^\infty \xrightarrow{D} \Lambda_2^\infty := \sum_k^K w_k \sum_{i=1}^\infty \lambda_i \int B_{ik}(t)^2 \, dt,$$

where the $B_{ik}(t)$ satisfy $\mathrm{Cov}(B_{ik}(t), B_{i'k'}(s)) = \delta_{ii'}\Sigma_{kk'} \min\{t,s\}$, with $\delta_{ii'}$ being the Kronecker delta.

We now return to the Russian climate data. The emphasis of this analysis is on illustrating the change-point methods rather than conducting a thorough climatological analysis. Using the *scpt* package in R, we apply a Monte Carlo version of the described tests, using 1000 replicates. As we can see from Table 14.3, none of the tests suggest a detectable change-point. This conclusion still holds if we focus on the area around Moscow, Table 14.4. This suggests that there has been little change in the annual pattern of daily maximum temperatures in Russia over the last three decades. Of course, this is just one type of temperature pattern that one could test.

Table 14.3 *p*-Values for each change-point test applied to the Russian weather data.

Test statistic	$\hat{\Lambda}_1$	$\hat{\Lambda}_2$	$\hat{\Lambda}_2^\infty$
p-Value	0.663	0.515	0.353

None of the tests support the existence of a change-point between 1980 and 2009.

Table 14.4 *p*-Values for each change-point test applied to a subset of 14 Russian weather stations near Moscow.

Test statistic	$\hat{\Lambda}_1$	$\hat{\Lambda}_2$	$\hat{\Lambda}_2^\infty$
p-Value	0.118	0.153	0.118

None of the tests support the existence of a change-point between 1980 and 2009.

14.4 Separability Tests

Separability is a common assumption in spatio–temporal statistics. It is primarily used to simplify the dependence structure in the data, as it assumes that the spatio–temporal covariance function factors into a product of two functions, one depending only on space and the other depending only on time. In this way, one can model space and time separately, as well as pooling across one dimension to help estimate the parameters of the other. However, such an assumption must be verified to ensure that other inferential tools that rely on it are still valid. Separability tests for functional data have been explored in [8–10]. The tools discussed in [9] were specifically targeted at spatio–temporal functional data and will be the ones discussed here.

The goal of this section is to test the hypotheses

$$H_0 : \mathrm{Cov}(X_n(\mathbf{s}, t), X_n(\mathbf{s}', t')) = \mathcal{U}(\mathbf{s}, \mathbf{s}')\mathcal{V}(t, t'),$$

against the alternative that the covariance is not separable. To test this hypothesis, we first need to be able to estimate the functions $\mathcal{U}(\mathbf{s}, \mathbf{s}')$, and $\mathcal{V}(t, t')$. There are a few things to note before doing this though. First, clearly $\mathcal{U}(\mathbf{s}, \mathbf{s}')$ cannot be estimated where there is no data, thus, the most one could hope for is either an estimate of $\mathcal{U}(\mathbf{s}_k, \mathbf{s}_{k'})$ or with a dense enough sampling in space and/or spatial stationarity assumptions, one could also use smoothing techniques to estimate the entire $\mathcal{U}(\mathbf{s}, \mathbf{s}')$ function, as was done in [7]. The approach we take here is to estimate $\mathcal{U}(\mathbf{s}_k, \mathbf{s}_{k'})$ only. Second, the functions $\mathcal{U}$ and $\mathcal{V}$ are only identifiable up to a constant, and thus some constraint must be imposed to make the model identifiable. The approach we take here is to assume that $trace(\mathbf{U}) = K$, where the matrix $\mathbf{U}$ is a $K \times K$ matrix with entries $\mathcal{U}(\mathbf{s}_k, \mathbf{s}_{k'})$. Other constraints can be used and the test statistics we propose are independent of the chosen constraint.

Estimating $\mathcal{U}$ and $\mathcal{V}$ turns out to be challenging if one wants to do it optimally. The primary issue is that optimal estimation of $\mathcal{V}$ requires using $\mathcal{U}$ and vice versa. To address this, Constantinou et al. [9] explored a "flip-flop" estimation procedure where one iterates between estimating the two, while [8] and [10] chose instead to use a suboptimal estimate for the sake of simplicity, both in terms of computation and deriving mathematical properties.

If H_0 is true, then [9] proposed using a temporal FPCA to model the data using a lower dimension:

$$X_n(\mathbf{s}_k, t) \approx \mu(\mathbf{s}_k, t) + \sum_{i=1}^{p} \xi_{n;ik} v_i(t),$$

where v_i are the eigenfunctions of $\mathcal{V}$ and $\xi_{n,ik}$ is the ith score at location $\mathbf{s}_k$. To test H_0, it is then enough to test if the covariance of the scores is also separable. Namely, under H_0 we have that

$$\mathrm{Cov}(\xi_{n;ik}, \xi_{n;i'k'}) = V_{ii'} U_{jj'},$$

and, more specifically, $V_{ii'} = \lambda_i \delta_{ii'}$. In principle, any basis could be used for the dimension reduction; one need not use the FPCA basis, in which case $\mathbf{V}$ is no longer diagonal. Thus, Constantinou et al. [9] proposed first doing a dimension reduction in time and then testing of the resulting scores were separable. The initial dimension reduction was based off using the following pooled estimate of the temporal covariance function

$$\hat{\mathcal{V}}_0(t,s) = \frac{1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K}(X_n(\mathbf{s}_k,t) - \overline{X}(\mathbf{s}_k,t))(X_n(\mathbf{s}_k,t') - \overline{X}(\mathbf{s}_k,t')).$$

Next, use the eigenfunctions of $\hat{\mathcal{V}}_0$ to reduce the dimension of $X_n(\mathbf{s},t)$:

$$X_n(\mathbf{s},t) \approx \mu(\mathbf{s},t) + \sum_{i=1}^{p}\xi_{ni}(\mathbf{s}_k)\hat{v}_i(t).$$

Here $\xi_{ni}(\mathbf{s}_k) = \int (X_n(\mathbf{s}_k,t) - \overline{X}(\mathbf{s}_k,t))\hat{v}_i(t)\,dt$, are the scores. The covariance assuming separability is then estimated by iterating the equation

$$\hat{\mathbf{V}} = \frac{1}{NK}\sum_{n=1}^{N}\mathbf{\Xi}_n^\top \hat{\mathbf{U}}^{-1}\mathbf{\Xi}_n \quad \text{and} \quad \hat{\mathbf{U}} = \frac{1}{Np}\sum_{n=1}^{N}\mathbf{\Xi}_n \hat{\mathbf{V}}^{-1}\mathbf{\Xi}_n^\top,$$

where $\mathbf{\Xi}_n$ is the $K \times p$ matrix of scores for the n observation. The form of these estimates comes from the maximum likelihood estimators for multivariate version of this problem. Note that when H_0 is true, we have $\text{Cov}(\text{vec}(\mathbf{\Xi}_n)) \approx \mathbf{V} \otimes \mathbf{U}$. When the covariance is not separable, it is estimated using

$$\hat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}\text{vec}(\mathbf{\Xi}_n)^\top \text{vec}(\mathbf{\Xi}_n).$$

In [9] three different tests were then considered to see if $\hat{\mathbf{V}} \otimes \hat{\mathbf{V}}$ and $\hat{\mathbf{\Sigma}}$ are estimating the same thing (in which case H_0 holds) or if they are significantly different than each other. The first test uses the Frobenius norm (ℓ^2 norm for matrices) of the difference:

$$\hat{T}_F = N\|\hat{\mathbf{V}} \otimes \hat{\mathbf{U}} - \hat{\mathbf{\Sigma}}\|_F^2.$$

The second test tries to normalize by the covariance of the difference so that the test is pivotal, resulting in a Wald-type test:

$$\hat{T}_W = N\text{vec}\,(\hat{\mathbf{V}} \otimes \hat{\mathbf{U}} - \hat{\mathbf{\Sigma}})^\top \widehat{\mathbf{W}}^{+}\ \text{vec}(\hat{\mathbf{V}} \otimes \hat{\mathbf{U}} - \hat{\mathbf{\Sigma}}).$$

The term $\widehat{\mathbf{W}}$ is an estimate of the covariance of the difference. Since we are dealing with covariance matrices, they are symmetric and thus $\mathbf{W}$ does not have full rank; hence, we use a generalized inverse $\widehat{\mathbf{W}}^{+}$ which is equivalent to dropping the redundant entries. The last test is based on the likelihood ratio statistic (for the nonfunctional setting):

$$\hat{T}_L = N\left(p\log\det(\hat{\mathbf{U}}) + K\log\det(\hat{\mathbf{V}}) - \log\det(\hat{\mathbf{\Sigma}})\right).$$

Under H_0, both $\hat{T}_L$ and $\hat{T}_W$ are asymptotically chi-squared, while $\hat{T}_F$ is asymptotically a weighted sum of chi-squares, with weights given by the eigenvalues of $\mathbf{W}$. While estimating $\mathbf{W}$ is possible, it is quite involved, and we thus refer the interested reader to [9]. If the data is normally distributed, then $\hat{T}_L$ actually doesn't depend on the underlying parameters. Exploiting this, Mitchell et al. [11] provided a Monte-Carlo algorithm for calculating p-values of $\hat{T}_L$ for multivariate data, which can be applied here. Practically, $\hat{T}_F$ performs fairly well even when using the asymptotic distribution, while $\hat{T}_L$ and $\hat{T}_W$ often have poor specificity when using their asymptotic distributions. Thus using $\hat{T}_F$ or $\hat{T}_L$ in conjunction with Monte-Carlo often works best. Lastly, while we only described dimension reduction in time, for very large spatial data sets a reduction in space may be required as well, in which case one would reduce in both time and space, and then test the separability of the resulting scores as discussed.

Returning to the Russia data, we consider each test under varying reduced spatial and temporal dimensions. Temporal dimensions of $J = 3, 4, 5, 6$ and spatial dimensions of $K = 3, 4, 5$ are considered. The results of these tests are given in Table 14.5 for all 220 stations, while in Table 14.6, we provide the results for the 14 stations around Moscow. For $J > 3$ and except in the case of $K = 3, J = 4$, the tests indicate that the daily maximum temperature fields possess a nonseparable

Table 14.5 p-Values for norm-based separability test (T_F) applied to the Russian weather data with reduced spatial dimension K and temporal dimension J.

$K \backslash J$	3	4	5	6
3	0.26	0.053	0.025	0.017
4	0.08	0.031	0.014	0.011
5	0.072	0.019	0.011	0.0083

For $J > 3$ and except in the case of $K = 3, J = 4$, the tests give evidence of a nonseparable covariance structure.

Table 14.6 p-values for norm-based separability test (T_F) applied to a subset of 14 Russian weather stations near Moscow with reduced spatial dimension K and temporal dimension J.

K \ J	3	4	5	6
3	0.159	0.14	0.134	0.115
4	0.079	0.093	0.056	0.22
5	0.132	0.135	0.102	0.095

No tests give evidence against a separable covariance structure at an $\alpha = 0.05$ level.

covariance structure at an $\alpha = 0.05$ level when looking across all of Russia. However, the p-values are not overwhelmingly small, and we see that when focusing on the smaller region around Moscow, separability does indeed seem to hold.

14.5 Trend Tests

In this section, we discuss two trend tests for spatio–temporal functional data. They pertain to different statistical models motivated by two distinct applied problems: testing for the increase in the intensity of tropical storms and testing for the presence of a cooling trend in the ionosphere.

The formulation of the first testing problem does not involve locations $\mathbf{s}_k$. It is motivated by the question of whether the strength of tropical storms increases from year to year. The raw data do not have the form of curves; these are appropriately processed wind speeds of individual storm events. Examples of such data are shown in Figure 14.3. From these data, expectile curves can be constructed; examples are shown in Figure 14.4. The definition of the expectile curves would take too much space to be presented here; we refer to [12], who also provide a number of references. The idea is that for each year, a family of curves indexed by $\tau \in (0, 1)$ can be constructed. The index τ refers to the a quantity similar to a quantile level. The curve with $\tau = 0.5$ describes a sort of median of wind speeds of storms in a given year. The curve with $\tau = 0.9$, shows the pattern of the strongest storms.

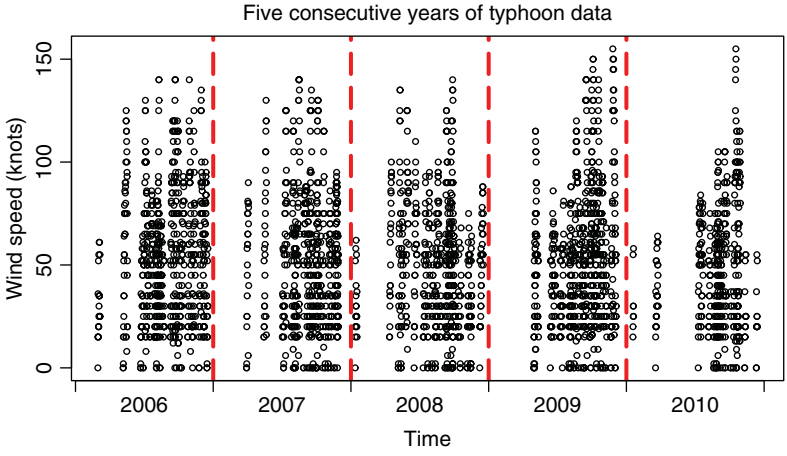


Figure 14.3 Five consecutive years (2006–2010) of typhoon data. The dots represent the wind speed measurements. Dashed vertical lines separate the years.

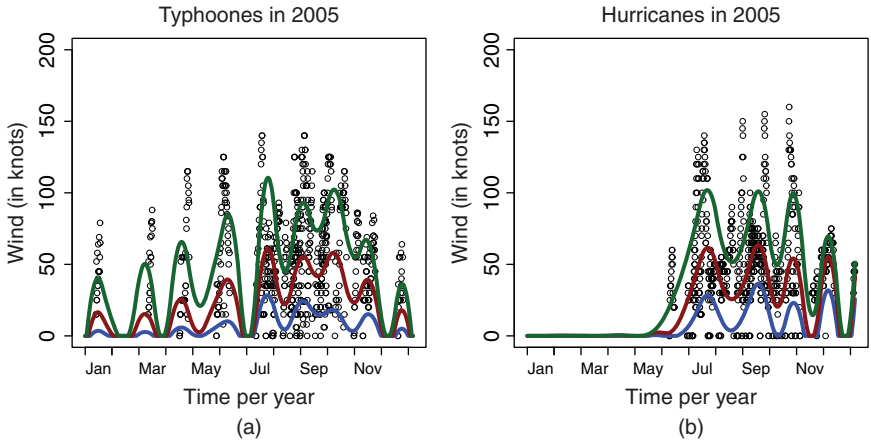


Figure 14.4 Typhoons (a) and hurricanes (b) data in 2005 with expectile curves for $\tau = 0.1, 0.5$, and 0.9. The dots represent the wind speed measurements. Generally, a vertical streak of dots represents one tropical storm event. The lines are the estimated expectile curves.

The statistical analysis is performed for each expectile level τ separately. For example, if we take $\tau = 0.9$, our objective is to determine if there is a trend in the strongest storms. For this purpose, we assume the following model for the expectile curves (dependence on τ is suppressed):

$$X_n(t) = \alpha(t) + \beta(t)n + \varepsilon_n(t). \tag{14.1}$$

We consider the testing problem:

$$H_0 : \beta = 0, \text{ vs. } H_A : \beta \neq 0.$$

In the above problem, the parameter β is a square integrable function. Denoting by $\hat{\beta}$ its suitably defined estimate, one can show that the statistic

$$\widehat{\Lambda}_N = \frac{N^3}{12} \int_0^1 \hat{\beta}(t)^2 dt$$

converges to a limit distribution under H_0 and exceeds the critical values of this distribution with probability approaching 1 under H_A. The limits are taken as the number of years, N, increases. The limit distribution does not have a closed form, but can be readily simulated.

The application of the test to typhoon and hurricane data shows that there is no significant trend in the strength of typhoons. There is a significant upward trend in the strength of hurricanes, but only for large values of τ, approximately for $\tau \geq 0.6$. This means that the wind speeds of typical or weak hurricanes are not increasing, but those of the strongest hurricanes are. The significance of

the function β is concluded from the p-values, the conclusion of an upward trend from the fact that $\hat{\beta}(t)$ is positive for almost all values of inter-year time t. The paper of [12] contains a more detailed discussion, comparison with related atmospheric science research, and several references to related trend tests.

The second test is motivated by an interesting and extensively studied problem of space physics. The account presented here is based on [13] and [14]. We first describe the space physics problem, and then explain the idea of the test.

Increased concentration of greenhouse gases in the upper atmosphere is associated with global warming in the lower troposphere (the atmosphere roughly below 10 km). Roble and Dickinson [15] suggested that the increasing amounts of these radiatively active gases, mostly CO_2 and CH_4, would lead to a global cooling in the ionosphere (atmosphere roughly 300 km above the Earth's surface). Rishbeth [16] pointed out that this would result in a thermal contraction of the ionosphere. The height of the ionosphere can be approximately computed using data from a radar-type instrument called the ionosonde. Relevant measurements have been made for many decades by globally distributed ionosondes. In principle, these observations could be used to quantitatively test the hypothesis of Roble and Dickinson. The difficulty in testing the contraction hypothesis comes from several sources. The height of the ionosphere depends on magnetic coordinates, the season, long-term changes in the strength and direction of the internal magnetic field, and, most importantly, on the solar cycle; more solar radiation leads to greater ionization. This is illustrated in Figure 14.5. Another difficulty stems from the fact that ionosonde records are not complete. Most observation stations do not operate continuously for many decades. They start and end

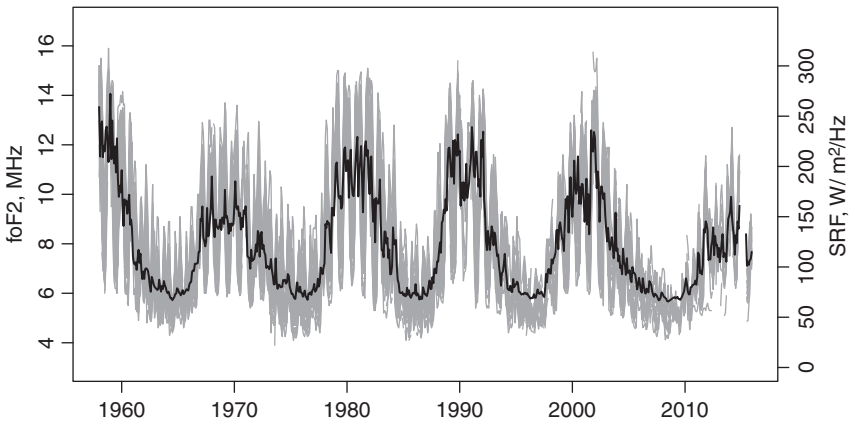


Figure 14.5 Gray lines represent ionosonde measurements obtained at observatories located in mid-latitude northern hemisphere, with the scale on the left-hand side. The black line represents the observed solar radio flux with the scale on the right-hand side.

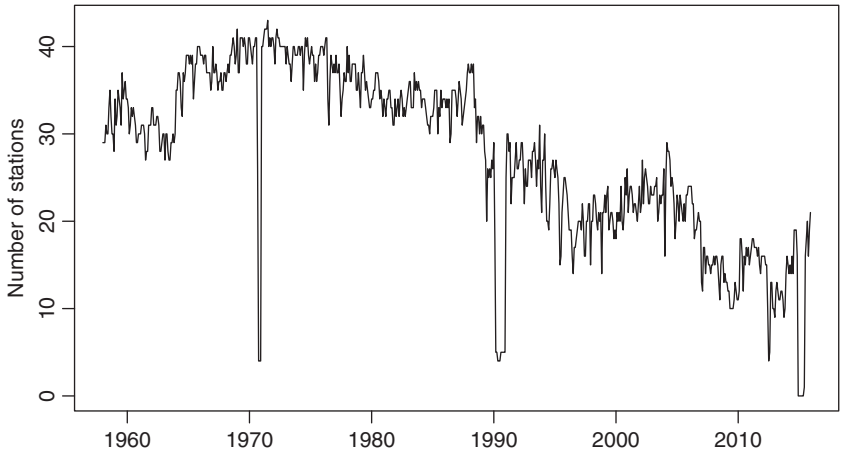


Figure 14.6 Number of available stations in the mid-latitude northern hemisphere.

operation at different times, some of them are out of service for many years, or even decades. In the mid-latitude northern hemisphere, there are 81 ionosonde stations, but at any given time, data from no more than 40 are available, as shown in Figure 14.6. This means estimation methods designed for complete records, developed by Gromenko et al. [17], cannot be used, as they require computation of integrals of products of functions. If the functions have disjoint supports, due to gaps, the integrands will be mostly zero, and the values of the integrals will introduce large biases to the estimators. More complex methods for incomplete records are needed.

Let $Y(\mathbf{s}_k; \tau_i)$ be the original record at location $\mathbf{s}_k$, measured from 1958 to 2015, possibly with long gaps. The set of all locations is $\{\mathbf{s}_k, 1 \le k \le K\}$, and the set of time points at which measurements may be available is $\{\tau_i, 1 \le i \le T\}$; in [14] these are months from January 1958 to December 2015. The following spatio–temporal model is postulated the entire time span

$$Y(\mathbf{s}_k; \tau_i) = \mu(\mathbf{s}_k; \tau_i) + \varepsilon(\mathbf{s}_k; \tau_i) + \delta_{ki}, \tag{14.2}$$

where δ_{ik} represents a random noise, which can be associated with measurement error. The second term, $\varepsilon(\mathbf{s}; \tau)$, describes the spatio–temporal variability away from the mean function, $\mu(\mathbf{s}, \tau)$. Stochastic modeling of this term is needed to develop inferential procedures. A simple form of the mean function relevant to the space physics problem is

$$\mu(\mathbf{s}; \tau) = \beta_1 + \beta_2 \tau + \beta_3 \, \mathrm{SRF}(\tau) + \beta_4 \, M(\mathbf{s}; \tau),$$

where $\mathrm{SRF}(\tau)$ is the solar radio flux, cf. Figure 14.5, and $M(\mathbf{s}; \tau)$ is a suitable function computed from the coordinates of the internal magnetic field. The interest

lies in the estimation of the mean function and testing if it contains a linear trend, i.e. testing $H_0 : \beta_2 = 0$. The function $\mu(\cdot, \cdot)$ is treated as known, except for the scalar parameters $\beta_1, \beta_2, \beta_3$, and β_3. The details of the estimation and testing procedures are too complex to describe here, but an important aspect is that the estimation proceeds by splitting records from 1958 to 2015 into years. This means that at the estimation stage, one works with the spatio–temporal functional data of the form $X_n(\mathbf{s}_k; t_j)$, which was discussed in the introduction. The conclusion is that β_2 is significantly negative, confirming the hypothesis of global ionospheric contraction. The software to perform the test is available and can be used to test for the presence of global trends in other data of this type, for example in near-surface temperatures. Records of this type also contain large gaps.

14.6 Spatio–Temporal Extremes

In this section, we summarize the work of [18] which deals with the computation of probabilities of heat waves. As before, the raw data are spatially indexed time series of daily temperature measurements. As argued above, due to the natural annual climate cycle, for each site, we partition the data into year and view the resulting 365-dimensional vectors as samples from a functional time series:

$$X_n(\mathbf{s}_k; \cdot) = \left\{ X_n(\mathbf{s}_k; t_i),\ i \in \{1, 2, \dots, 365\} \right\}. \tag{14.3}$$

Here, $t \mapsto X_n(\mathbf{s}_k; t)$ is the temperature curve at site $\mathbf{s}_k$ for year n, viewed as a function of time t in days. In contrast to the setting of previous sections, the data used by French et al. [18] are not historical records, but data generated by a computer climate model. These artificial data are of much higher quality than historical records; there are no gaps, and the daily records are available at 16 100 locations forming a grid covering much of North America. It is, at this point, not clear how to extend the methodology of [18] to historical records. The advantage of using computer model data is that they are predicted future temperatures ([18] work with the period 2041–2070), which are more relevant to the prediction of future heat waves. On the other hand, these data do depend on a model, and the poor-quality, geostatistical historical records are the real data.

Many functionals were proposed in [18] that can quantify a heat wave, but here we focus on one specific approach that explains the general idea. A heat wave is characterized by its spatial and temporal extents and by its intensity. The intensity is typically quantified by a threshold. Public health concerns call for a fixed threshold, like 105 °F. However, in climate studies of large spatial regions, with many climatic zones, such a fixed threshold is not appropriate. Also the variability of temperatures depends greatly on the geographical location, with coastal locations

exhibiting much smaller variability than locations far away from large bodies of water. It is therefore reasonable to work with standardized temperatures

$$Z_n(\mathbf{s}_k, t_i) = \frac{X_n(\mathbf{s}_k, t_i) - \overline{X}(\mathbf{s}_k, t_i)}{\mathrm{SD}(\mathbf{s}_k, t_i)}, \tag{14.4}$$

where

$$\overline{X}(\mathbf{s}_k, t_i) = \frac{1}{N}\sum_{n=1}^{N} X_n(\mathbf{s}_k, t_i) \text{ and} \tag{14.5}$$

$$\mathrm{SD}^2(\mathbf{s}_k, t_i) = \frac{1}{N-1}\sum_{n=1}^{N} (X_n(\mathbf{s}_k, t_i) - \overline{X}(\mathbf{s}_k, t_i))^2.$$

If the $Z_n(\mathbf{s}_k, t_i)$ exceed a fixed threshold z, e.g. $z = 2$, for a number of neighboring locations and over a period of time, then we have observed a heat wave (the $Z_n(\mathbf{s}_k, t_i)$ are practically normal). The severity of a heat wave increases with the size of the region, the duration, and the threshold z that is exceeded. Suppose the $X_n(\mathbf{s}_k, t_i)$ are maximum daily temperatures, and set

$$Z_n^\star(\mathbf{s}_k, t_j) = \frac{1}{\ell} \sum_{t_j - \ell < t_i \le t_j} Z_n(\mathbf{s}_k, t_i).$$

This is the average maximum temperature over the ℓ days preceding day t_j. Next, define

$$Z_n^\star(t_j) = \min_{1 \le k \le K} Z_n^\star(\mathbf{s}_k, t_j).$$

If $Z_n^\star(t_j) > z$, then the average maximum temperature over ℓ days over K (neighboring) locations exceeds, z; this corresponds to a heat wave defined by this specific functional. We are interested in the probability of a heat wave in any given year. We assume that this probability does not depend on year n. We thus want to compute, for some relevant $z > 0$,

$$p(z) = P(\exists\, j : Z_n(t_j) > z) = P\left(\max_{1 \le j \le J} Z_n^\star(t_j) > z\right) = P(M_J > z),$$

where $J = 365$, and

$$M_J \overset{d}{=} M_{J,n} := \max_{1 \le j \le J} Z_n^\star(t_j).$$

The concatenated sequence $Z^\star(t_j)$ is stationary and weakly dependent, so (see, e.g. [19], Chapter 10), there are sequences a_J and b_J such that

$$\lim_{J \to \infty} P\left(\frac{M_J - b_J}{a_J} \le z\right) \to H(z),$$

where H is a univariate Generalized Extreme Value distribution function. The function H depends on three parameters, which can be estimated, together with the constants a_J and b_J, using now standard R implementations.

Figure 14.7 A map of the neighborhood structures for different locations using 50, 150, and 450 nearest neighbors. Each × marks a neighborhood centroid and the sequences of gray shading mark the extents of the increasing neighborhood sizes.

Figure 14.7 shows examples of regions corresponding to 50, 150, and 450 neighboring locations. Figure 14.8 shows a map of the probability of a heat wave for $d = 50$ for three durations ℓ, with (a) corresponding to $\ell = 3$, (b) to $\ell = 10$, and (c) to $\ell = 30$. When $\ell = 3$, there is a surprisingly high probability of localized heat waves over the Labrador Peninsula. Such short heat spells may occur with probability approaching 50%, that is on average every second year. While our extreme value theory (EVT) approximation may break down for such high probabilities, it is, nevertheless, obvious that part of Canada will see heat spells much more frequently than in the past. Generally, we see that the area around the Hudson Bay will experience an increased frequency of hot spells lasting a few days. There is a noticeable drop in the probability of such a heat wave around the Rocky Mountain Range. The probability is also very low along the Eastern seaboard of the United States. Increasing the duration to $\ell = 10$ days, dramatically reduces the probability of a heat wave of the corresponding magnitude. The reader will note the different probability scale. Many parts of Canada once again show an increased probability of a heat wave of this magnitude, as well as parts of Iowa and Illinois, certain regions in Texas, and, most visibly, the Pacific Ocean off the Southern California coast. Increasing the duration to approximately one month ($\ell = 30$) causes the probability of a heat wave to drop even further; generally, throughout North America, heat waves of this magnitude will occur with probability of less than 1%, i.e. once per one hundred years, on average. Over the Canadian Plains and the Canadian Rockies, this probability increases only slightly to about 1.5%. There are two patches, in Arizona and Southern Texas, with probabilities elevated to 2–3%.

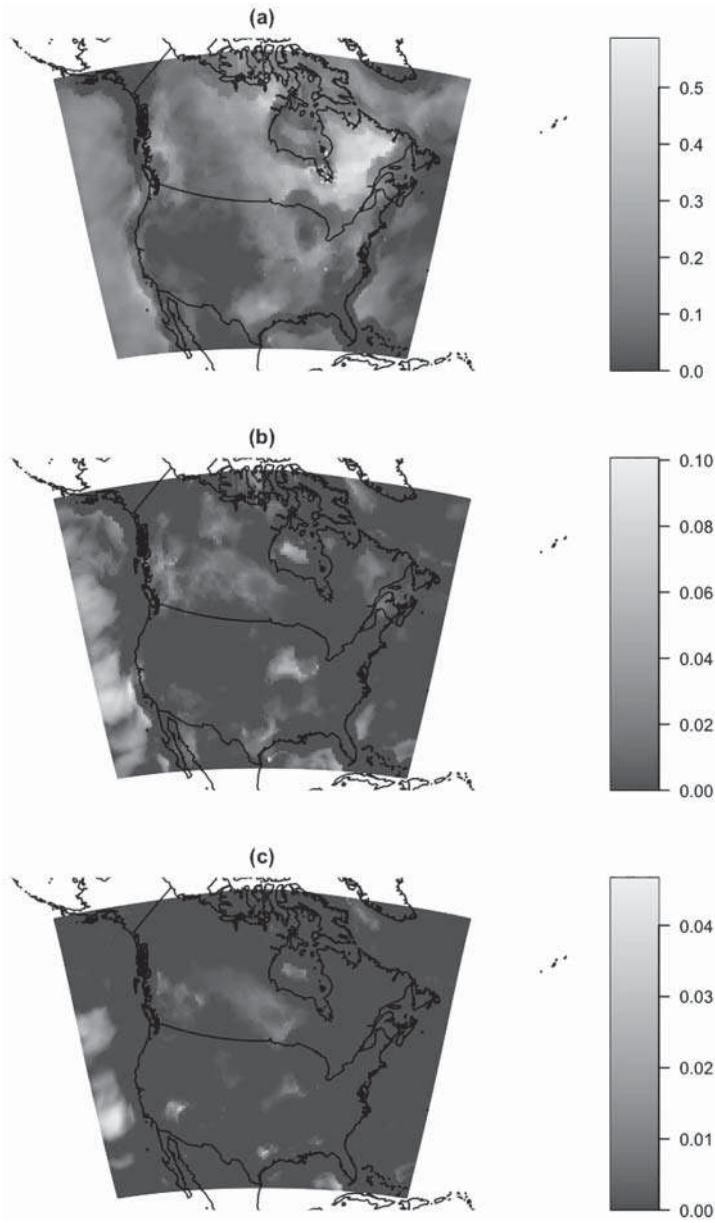


Figure 14.8 Probability of a heat wave with amplitude more than two standard deviations above the mean for spatial extent $d = 50$ and durations of (a) $\ell = 3$, (b) $\ell = 10$, and (c) $\ell = 30$.

References

1 Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*. Springer.
2 Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. CRC Press.
3 Kokoszka, P. and Reimherr, M. (2018). Some recent developments in inference for geostatistical functional data. *Revista Colombiana de Estadística* 42: 101–122, 2019.
4 Kokoszka, P., Reimherr, M., and Wölfing, N. (2016). A randomness test for functional panels. *Journal of Multivariate Analysis* 151: 37–53.
5 Tatusko, R. (1990). Cooperation in climate research: an evaluation of the activities conducted under the US-USSR Agreement for Environmental Protection since 1974. Washington, DC: National Oceanic and Atmospheric Administration.
6 Liu, X., Yin, Z.Y., Shao, X., and Qin, N. (2006). Temporal trends and variability of daily maximum and minimum, extreme temperature events, and growing season length over the Eastern and Central Tibetan Plateau during 1961–2003. *Journal of Geophysical Research: Atmospheres* 111 (D19). 1–19.
7 Gromenko, O. and Kokoszka, P. and Reimherr, M. (2017). Detection of change in the spatiotemporal mean function. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (1), 29–50, Wiley Online Library
8 Aston, J.D.A., Pigoli, D., and Tavakoli, S. (2017). Tests for separability in nonparametric covariance operators of random surfaces. *Annals of Statistics* 45 (4): 1431–1461.
9 Constantinou, P., Kokoszka, P., and Reimherr, M. (2017). Testing separability of space–time functional processes. *Biometrika* 104: 425–427.
10 Constantinou, P., Kokoszka, P., and Reimherr, M. (2018). Testing Separability of Functional Time Series. *Journal of Time Series Analysis* 39 (5), 731–747, Wiley Online Library.
11 Mitchell, M.W., Genton, M.G., and Gumpertz, M.L. (2006). A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis* 97: 1025–1043.
12 Burdejova, P., Härdle, W., Kokoszka, P., and Xiong, Q. (2017). Change point and trend analyses of annual expectile curves of tropical storms. *Econometrics and Statistics* 1: 101–117.
13 Gromenko, O. and Kokoszka, P. (2013). Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination. *Computational Statistics and Data Analysis* 59: 82–94.
14 Gromenko, O., Kokoszka, P., and Sojka, J. (2017). Evaluation of the global cooling trend in the ionosphere using functional regression models with incomplete curves. *The Annals of Applied Statistics* 11: 898–918.

15 Roble, R.G. and Dickinson, R.E. (1989). How will changes in carbon dioxide and methane modify the mean structure of the mesosphere and thermosphere? *Geophysical Research Letters* 16: 1441–1444.

16 Rishbeth, H. (1990). A greenhouse effect in the ionosphere? *Planetary and Space Science* 38: 945–948.

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

18 French, J., Kokoszka, P., Stoev, S., and Hall, L. (2019). Quantifying the risk of heat waves using extreme value theory and spatio-temporal functional data. *Computational Statistics & Data Analysis* 131, 176–193, Elsevier.

19 Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2006). *Statistics of Extremes: Theory and Applications*. Wiley.

15

A Comparison of Spatiotemporal and Functional Kriging Approaches

Johan Strandberg[1], Sara Sjöstedt de Luna[2], and Jorge Mateu[3]

[1] *Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden*
[2] *Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden*
[3] *Department of Mathematics, University Jaume I of Castellon, Spain*

15.1 Introduction

In many fields, such as environmental, forestry, climatology, meteorology, and medical sciences, the spatial variation of objects in the form of curves is of interest to study. It could be, e.g. ocean temperature, salinity, or other variables measured over time (or at different depths) at a set of spatial locations. With today's modern technology and huge storage capability, it is, in principal, possible to observe entire curves by recording them over a dense raster of time (depth) points. In particular, it may be of interest to predict a curve at a new spatial location given that such curves have been observed at n other locations, utilizing the information inherent in the spatial dependence between curves.

Kriging predictors have a long history of being used to predict objects at new locations based on information observed at a set of other locations. There is a rich geostatistical literature about kriging prediction when the objects are real- or vector-valued, see, e.g. Chiles and Delfiner [1], Cressie [2], Cressie and Wikle [3], and references therein. Functional kriging predictors, used when the objects are functions with infinite dimension, have been proposed in the last decades, see, e.g. pioneering work by Goulard and Voltz [4], and further the work of Giraldo et al. [5, 6] and Nerini et al. [7]. They assume that the expected value of the curves is the same irrespective of the spatial location, so-called "ordinary functional kriging." More recently, functional kriging methods where the expected value of the curves may also depend on location are investigated, e.g. by Caballero et al. [8], Menafoglio et al. [9], Ignaccolo et al. [10], and Reyes et al. [11].

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

A kriging predictor is typically defined to be the best linear unbiased predictor (BLUP) minimizing the mean squared prediction error (MSPE). A kriging predictor is thus a weighted sum of the objects observed at the n spatial locations. The optimal kriging weights to be determined turn out to be functions of the (spatial) dependence structure of the objects, which in practice is not known and needs to be estimated. Typically, estimators of the dependence structure rely on stationarity assumptions, unless parametric and distributional assumptions are made.

Here, we will compare two kriging approaches to predict spatial functional random processes. A functional random process, $\{\chi_s : s \in D \subset \mathbf{R}^d\}$, is a process with stochastic functional objects (curves) $\chi_s = \chi_s(t), t \in T \subset \mathbf{R}$ over the "time" domain T at each spatial location $s \in D$. Such a process can also be viewed as a spatiotemporal (Sp.T.) stochastic process $\{Z(s,t) = \chi_s(t), (s,t) \in D \times T\}$. Given that the process has been observed at n different locations, a curve at a new location s_0, χ_{s_0}, can be predicted by a functional kriging approach, i.e. as a linear combination of the n observed curves. As an alternative, a Sp.T. kriging approach could be used. The curve $\chi_{s_0}(t), t \in T$ would then be predicted at a dense grid of values over T, based on a linear combination of a time-grid of values over the observed curves. As pointed out by Delicado et al. [12], the question of which approach, functional kriging or Sp.T. kriging, should be used to analyze a particular data set is an important one (with no complete answer). Here, we compare the two approaches with respect to prediction performance and computation time. The presented material in this section comes to a large extent from the article written by Strandberg et al. [13], with some modifications.

In Section 15.2, notation and definitions are given. Section 15.3 presents the two kriging approaches for Sp.T. processes. In Section 15.3.1, we describe functional kriging methods in more detail, including how to estimate the dependence structure. We also discuss how the functional kriging methods relate to each other and under which circumstances they may coincide. The Sp.T. kriging approach is described in Section 15.3.2, with a discussion on how to estimate the Sp.T.-dependence structure. In Section 15.4, the two kriging approaches are evaluated by a simulation study. Sp.T. stationary (isotropic) processes with separable and nonseparable covariance functions are considered, as well as some nonstationary Sp.T. (but stationary functional) random processes with constant mean. We also apply both kriging approaches to Canadian temperature data in Section 15.5. A discussion and concluding remarks are found in Section 15.6.

15.2 Preliminaries

A *spatial functional random process* $\{\chi_s : s \in D \subset \mathbf{R}^d\}$ [5, 12], is a process where for each given $s \in D$, the observed random element is a functional random

variable, χ_s, taking values in an infinite dimensional space (or function space). We will consider the case where χ_s, for every fixed s, is a real-valued function, $\chi_s(t)$, $t \in T \subset \mathbf{R}$, from the compact set T to $\mathbf{R}$ and with $s \in D \subset \mathbf{R}^2$. It is usually assumed that the realizations of the curves (functions) $\chi_s(t), t \in T, s \in D$ belong to a separable Hilbert space $\mathbf{H}$ of square integrable functions defined on T. Let the mean function be denoted by $m_s(t) = E[\chi_s(t)]$, the covariance function (covariogram) by $C(s, r, v, t) = \mathrm{Cov}[\chi_s(r), \chi_v(t)]$, and the semivariogram by $\gamma(s, r, v, t) = V[\chi_s(r) - \chi_v(t)]/2$.

The spatial functional random process is said to be *second-order stationary* if for each $t \in T$ the corresponding spatial random process $\{\chi_s(t), s \in D\}$ is second-order stationary, i.e. if

(i) $E[\chi_s(t)] = m(t)$ and $V[\chi_s(t)] = \sigma^2(t)\ \forall s \in D$ and $\forall t \in T$,
(ii) $\mathrm{Cov}[\chi_s(r), \chi_v(t)] = C(s - r, v, t)\ \forall s, v \in D$ and $\forall r, t \in T$.

For (second-order) stationary functional processes, the covariance structure can equivalently be described by the variogram

$$V[\chi_s(r) - \chi_v(t)] = 2\gamma(s - v, r, t),$$

via the relation

$$2C(s - v, r, t) = \sigma^2(r) + \sigma^2(t) - 2\gamma(s - v, r, t). \tag{15.1}$$

Here, we will mainly focus on spatial functional random processes that are *second-order isotropically stationary*, i.e.

$$\begin{array}{ll} \text{(i)} & E[\chi_s(t)] = m(t) \text{ and } V[\chi_s(t)] = \sigma^2(t)\ \forall s \in D \text{ and } \forall t \in T, \\ \text{(ii)} & \mathrm{Cov}[\chi_s(r), \chi_v(t)] = C(\|s - v\|, r, t)\ \forall s, v \in D \text{ and } \forall r, t \in T, \end{array} \tag{15.2}$$

where $\|\cdot\|$ denotes the Euclidean distance. For any given $t \in T$, $\gamma_t(h) := V[\chi_s(t) - \chi_v(t)]/2$, $h = \|s - v\|$, is the semivariogram of the spatial random process $\{\chi_s(t) : s \in D\}$, satisfying $\gamma_t(h) = \sigma^2(t) - C_t(h)$, where $C_t(h) = C(h, t, t)$ is the corresponding covariogram. In order to ensure that $V\left[\sum_{i=1}^n l_i \chi_{s_i}(t)\right] \geq 0$ for any set of constants $l_1, \ldots, l_n \in R, n = 1, 2, \ldots$, the variogram (as a function of h) needs to be a conditional negative definite function and the covariogram needs to be a positive definite function, see, e.g. [2].

As previously mentioned, a spatial functional random process can also be viewed as a Sp.T. process $Z(s, t) = \chi_s(t)$, where $Z(s, t)$ takes values in $\mathbf{R}$ and is mapped from $(s, t) \in D \times T$, cf. [3]. A Sp.T. process is said to be *second-order stationary and spatially isotropic* if

$$\begin{array}{ll} \text{(i)} & E[Z(s, t)] = m \text{ and } V[Z(s, t)] = \sigma_Z^2\ \forall s \in D \text{ and } \forall t \in T, \\ \text{(ii)} & \mathrm{Cov}[Z(s, r), Z(v, t)] = C_Z(\|s - v\|, \mid r - t \mid)\ \forall s, v \in D \text{ and } \forall r, t \in T. \end{array} \tag{15.3}$$

Note that the concept of stationarity differs between functional and Sp.T. random processes: A stationary Sp.T. process implies that the corresponding

functional random process also is stationary, while the opposite may not be true. Hence, the class of stationary Sp.T. processes is a subset of the class of stationary functional random processes.

15.3 Kriging

In this section, we describe the functional and Sp.T. kriging approaches, in Sections 15.3.1 and 15.3.2, respectively. A way to evaluate prediction performance using functional cross-validation (FCV) is given in Section 15.3.3.

15.3.1 Functional Kriging

For the presentation below, unless otherwise stated, we will assume that the spatial functional random process is second-order stationary and isotropic. Within the functional kriging framework, it is of interest to predict the complete random function $\chi_{s_0}(t), t \in T$, at a new location s_0, given that a sample of random functions have been observed at n different locations, $s_1, \ldots, s_n$. We will consider unbiased functional predictors, $\hat{\chi}_{s_0}(t),\ t \in T$, that minimize the mean integrated squared error (MISE).

$$\text{MISE}(s_0) = E\left[\int_T (\hat{\chi}_{s_0}(t) - \chi_{s_0}(t))^2 dt\right]. \tag{15.4}$$

By Fubini's theorem (assuming that the realizations of the random functions are square integrable), we may change the order of expectation in MISE(s_0), and further, due to the unbiasedness of the kriging predictor, i.e. $E[\hat{\chi}_{s_0}(t) - \chi_{s_0}(t)] = 0$ for all $t \in T$, we have that (15.4) equals:

$$\text{MISE}(s_0) = \int_T E\left[(\hat{\chi}_{s_0}(t) - \chi_{s_0}(t))^2\right] dt = \int_T V[\hat{\chi}_{s_0}(t) - \chi_{s_0}(t)] dt. \tag{15.5}$$

15.3.1.1 Ordinary Kriging for Functional Data

One of the first functional kriging predictors was proposed by Goulard and Voltz [4], the so-called *curve kriging predictor* being of the form:

$$\hat{\chi}_{s_0}(t) = \sum_{i=1}^{n} \lambda_i \chi_{s_i}(t), \quad t \in T. \tag{15.6}$$

Giraldo et al. [6, 14] further discussed this predictor and gave it the name *ordinary kriging for functional data* (OKFD). The optimal weights $\lambda_1, \ldots, \lambda_n \in \mathbf{R}$ that minimizes MISE(s_0) given that the predictor is unbiased, are called the kriging

weights. Under assumption (15.2) unbiasedness of (15.6) implies that $\sum_{i=1}^{n} \lambda_i = 1$. From this fact, combined with (15.5) and (15.1), it follows that

$$\begin{aligned}\mathrm{MISE}(s_0) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \int_T \mathrm{Cov}[(\chi_{s_i}(t), \chi_{s_j}(t)]dt + \int_T \sigma^2(t)dt \\ &\quad - 2\sum_{i=1}^{n}\lambda_i \int_T \mathrm{Cov}[\chi_{s_i}(t), \chi_{s_0}(t)]dt \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\|s_i - s_j\|) + 2\sum_{i=1}^{n}\lambda_i\gamma(\|s_i - s_0\|),\end{aligned}$$

where

$$\gamma(h) = \frac{1}{2}E\left[\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt\right] = \int_T \gamma_t(h)dt, \quad \forall s_i, s_j \in D, \tag{15.7}$$

and $h = \|s_i - s_j\|$, is called the *(isotropic) trace-semivariogram*. The second equality in (15.7) holds by Fubini's theorem under the assumption that the realizations of the random functions are square integrable. Since $\mathrm{MISE}(s_0)$ only depends on the trace-semivariogram, so will the optimal λ_i's, for a more detailed derivation see [6]. Note that the trace-semivariogram often has the same property as the classical semivariogram, being a conditional negative definite function [9].

In practice, the trace-variogram is unknown and thus needs to be estimated from the data. To estimate the trace-variogram under assumption (15.2), first a (consistent) method of moments estimator of (15.7) is formed as

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{i,j\in N(h)} \int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt, \tag{15.8}$$

where $N(h) = \{(s_i, s_j) : \|s_i - s_j\| = h\}$, and $|N(h)|$ is the number of distinct elements in $N(h)$. For irregularly spaced observations, it is rare to have several pairs of observations separated at exactly distance h and then $N(h)$ is modified to $\{(s_i, s_j) : \|s_i - s_j\| \in (h - \epsilon, h + \epsilon)\}$, with $\epsilon > 0$ being some small positive value, in order to obtain a more stable estimate. To obtain a valid (variogram) estimate for any h, a parametric variogram model $\gamma(h \mid \theta)$, e.g. the spherical, exponential, or stable model, is fitted to a set of estimated values $\{\hat{\gamma}(h_l), h_l\}$, $l = 1, \ldots, L$, by a least squares method, cf. [2]. Here we use the ordinary least squares (OLS) method to estimate θ.

Typically, the random functions $\chi_{s_i}(t)$ are observed only at a finite number of time points $t_{i1}, \ldots, t_{im_i}$, $i = 1, \ldots, n$. Goulard and Voltz [4] suggested to fit a parametric model $\chi_{s_i}(\cdot, \alpha_i)$ to the observed values and replace $\chi_{s_i}(t)$ by $\chi_{s_i}(t, \hat{\alpha}_i)$ in (15.6) and (15.8). Giraldo et al. [6] instead suggested a nonparametric approach, where

the observed random functions are represented (approximated) by linear combinations of p known basis functions, $\mathbf{B}(t) = (B_1(t), \dots, B_p(t))^\intercal$, as

$$\tilde{\chi}_{s_i}(t) = \sum_{k=1}^{p} a_{ik} B_k(t) = \mathbf{a}_i^\intercal \mathbf{B}(t). \tag{15.9}$$

The basis functions could, e.g. be B-splines, Fourier basis, or wavelets. The coefficients ($\mathbf{a}_i$'s) are typically determined by the least squares method. Giraldo et al. [6] suggested to choose the number of basis functions p by cross-validation. In the final ordinary kriging predictor (15.6), the estimated trace-variogram values are plugged into the kriging weights (λ_i's), and the $\tilde{\chi}_{s_i}(t)$'s replacing the $\chi_{s_i}(t)$'s.

15.3.1.2 Pointwise Functional Kriging

To allow more flexibility than the OKFD predictor (15.6), Giraldo et al. [5, 15] suggested the *pointwise functional kriging predictor* (PWFK) which allows the λ_i's to depend on t and is defined as

$$\hat{\chi}_{s_0}(t) = \sum_{i=1}^{n} \lambda_i(t) \chi_{s_i}(t), \quad t \in T.$$

The BLUP minimizing the mean squared integrated prediction error is found by choosing the $\lambda_i(t)$-functions such that (15.4) is minimized subject to the unbiasedness constraint of the predictor, $\sum_{i=1}^{n} \lambda_i(t) = 1$, for all $t \in T$. In order to solve the optimization problem, the $\lambda_i(t)$-functions are represented by a linear combination of K known basis functions:

$$\lambda_i(t) = \sum_{k=1}^{K} b_{ik} B_{\lambda k}(t) = \mathbf{b}_i^\intercal \mathbf{B}_\lambda(t), \quad i = 1, \dots, n, \tag{15.10}$$

where the $\mathbf{b}_i$'s are to be determined. Moreover, the $\chi_{s_i}(t)$'s are represented as in (15.9), implying that $E[\chi_{s_i}(t)] = E[\mathbf{a}_i]^\intercal \mathbf{B}(t)$ and $\mathrm{Cov}[\chi_{s_i}(t), \chi_{s_j}(u)] = \mathbf{B}(t)^\intercal \mathrm{Cov}[\mathbf{a}_i, \mathbf{a}_j] \mathbf{B}(u)$. The optimization problem then reduces the infinite dimensional problem to a multivariate geostatistics problem. Given that the weights satisfy (15.10), the unbiasedness condition implies that

$$\sum_{i=1}^{n} \lambda_i(t) = \sum_{i=1}^{n} \mathbf{b}_i^\intercal \mathbf{B}_\lambda(t) = \mathbf{c}^\intercal \mathbf{B}_\lambda(t) = 1, \quad \text{for all } t \in T, \tag{15.11}$$

where $\mathbf{c} = \sum_{i=1}^{n} \mathbf{b}_i$. Hence, only basis functions $\mathbf{B}_\lambda(t)$ that satisfy (15.11) for some constant vector $\mathbf{c}$ give admissible solutions to the kriging optimization problem. When $\mathbf{B}_\lambda(t)$ are B-splines, (15.11) is fulfilled when $\mathbf{c} = \mathbf{1}$, and for Fourier basis functions when $\mathbf{c} = (1, 0, \dots, 0)^\intercal$. In fact, any set of basis functions where one (the first say) basis function is a constant, $B_{\lambda 1}(t) = k$, satisfies (15.11) for $\mathbf{c} = (1/k, 0, \dots, 0)^\intercal$. The full derivation of the equation system to be solved in order to find the $\mathbf{b}_i$'s is given by Giraldo et al. [5] when $\mathbf{B}_\lambda(t) = \mathbf{B}(t)$, and for general $\mathbf{B}_\lambda(t)$ in [13].

It turns out that the $\mathbf{b}_i$'s are functions of the covariances between the various $\mathbf{a}_i$'s, which in practice are not known and thus need to be estimated, e.g. by the least squares method. If $\mathbf{a}_i = \mathbf{a}(s_i)$, and $\mathbf{a}(s) = [a_1(s), \dots, a_p(s)]^\intercal$ is a p-variable second-order isotropically stationary spatial random field for all $s \in D$, with $E[\mathbf{a}(s)] = \mathbf{m}_a$ and $\text{Cov}[\mathbf{a}(s_i), \mathbf{a}(s_j)] = \Sigma(\|s_i - s_j\|) = \{c_{kl}(h_{ij})\} \in \mathbf{R}^{p\times p}$, where $c_{kl}(h_{ij}) = \text{Cov}[a_k(s_i), a_l(s_j)]$, $h_{ij} = \|s_i - s_j\|$, it follows that $\{\chi_s(t) = \mathbf{a}(s)^\intercal \mathbf{B}(t), s \in D, t \in T\}$ satisfies (15.2). Under this assumption, [5] suggest estimating the covariograms and cross-covariograms (the $c_{kl}(\cdot)$'s) via a linear model of coregionalization [16]. This means that $\mathbf{a}(s)$ can be expressed as $\mathbf{a}(s) = \mathbf{P}\mathbf{r}(s)$, where $\mathbf{P} \in \mathbf{R}^{p\times q}$ and $\mathbf{r}(s) = (r_1(s), \dots, r_q(s))^\intercal$ are q latent univariate (second-order isotropically stationary) random fields, typically assumed to be independent. Given available data, $\mathbf{a}_i = \mathbf{a}(s_i), i = 1, \dots n$, the $c_{kl}(\cdot)$'s (and $\mathbf{P}$) can be estimated using the R-package gstat [17]. In order to perform the estimation, the value of q and the variogram models of the $r_i(s)$'s need to be specified.

The PWFK may have the potential to give better prediction performance than OKFD since it allows more flexible kriging weights. In which situations this could be true is still not completely known. Strandberg et al. [13] have confirmed situations in which PWFK and OKFD coincide: Suppose that $E[\chi_s(t)] = m(t)$ and $Cov[\chi_s(t), \chi_v(t)] = w(t)C(s, v)$ $\forall s, v \in D$ and $\forall t \in T$, where $w(t)$ is a real-valued deterministic function. These type of processes include second-order stationary spatial functional random processes and also some non-stationary variants. Then, under some mild conditions on $\mathbf{B}_\lambda(t)$, that ensures a unique solution, the optimal kriging weights of PWFK satisfy $\lambda_i(t) = \lambda_i$, for all $i = 1, \dots, n$, and thus coincide with those of OKFD.

Strandberg et al. [13] experienced that the computational time of PWFK (using R-code kindly provided by Giraldo et al. [5]) was much larger than that of OKFD (using the R-package geofd, Giraldo et al. [18]). They also found bugs in the PWFK R-code, which after correction gave constant weights for PWFK when $\mathbf{B}_\lambda(t) = \mathbf{B}(t)$ were B-splines and Fourier basis functions.

15.3.1.3 Functional Kriging Total Model

The OKFD and PWFK methods predicts the spatial functional process at a new location s_0 and time point t by a linear combination of the n observed functions ($\chi_{s_i}(t)$'s) at the same time point t, but does not utilize the information from other time points of the observed functions. A third functional kriging method, proposed by Giraldo and [19, 20], and independently by Nerini et al. [7], addresses this and allows the usage of all time points of the observed functions. The method is called the *functional kriging total model* (FKTM), and the predictor defined as

$$\hat{\chi}_{s_0}(t) = \sum_{i=1}^{n} \int_T \lambda_i(t, v)\chi_{s_i}(v)dv, \quad t \in T. \tag{15.12}$$

This modeling approach is coherent with the functional linear model for functional responses (total model), see e.g. [21]. Assuming that the random functions $\chi_{s_i}(t)$'s satisfy (15.9) and that the kriging weights satisfy

$$\lambda_i(t,v) = \sum_{k=1}^{p}\sum_{l=1}^{p} c_{ik}^{l} B_k(t) B_l(v) = \mathbf{B}(t)^{\mathsf{T}} \mathbf{C}_i \mathbf{B}(v), \quad i = 1, \dots, n,$$

Giraldo [20] proposed a way to determine the $\lambda_i(t,v)$'s (i.e. the $\mathbf{C}_i$'s) such that the predictor (15.12) is unbiased and minimizes (15.4). Also here, the $\mathbf{C}_i$'s turn out to be functions of the covariances between the various $\mathbf{a}_i$'s, which in practice are not known and can be estimated as proposed in Section 15.3.1.2. For more detailed derivations, see [20].

The FKTM method is, just like PWFK, computationally heavy in comparison with OKFD [19]. Moreover, Menafoglio and Petris [22] showed that if the realizations of $\chi_s(t)$ belong to the Hilbert space of square integrable functions on T, and the functional second-order stationary random process is Gaussian, then the kriging weights of FKTM and OKFD agree almost surely for any orthonormal base $\mathbf{B}(t)$.

15.3.2 Spatiotemporal Kriging

Alternatively, the spatial functional process can be viewed as a Sp.T. process, $Z(s,t) = \chi_s(t)$, taking values in $(s,t) \in D \times T$, and hence, be predicted by Sp.T. kriging methods. The Sp.T. kriging predictor at location s_0 and time point $t \in T$, given the observations $Z(s_i, t_{ij}), j = 1, \dots, m_i, i = 1, \dots, n$, is of the form

$$\hat{Z}(s_0,t) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \lambda_{ij}^{t} Z(s_i, t_{ij}), \tag{15.13}$$

being the BLUP minimizing the MSPE

$$\mathrm{MSPE}(s_0,t) = E[(\hat{Z}(s_0,t) - Z(s_0,t))^2]. \tag{15.14}$$

Note that for each s_0, the Sp.T. kriging weights (λ_{ij}^{t}'s) are allowed to change for each $t \in T$. When the mean value of the process is constant, the unbiasedness condition implies that $\sum_{i=1}^{n}\sum_{j=1}^{m_i} \lambda_{ij}^{t} = 1$. Moreover, if the constant mean value of the process is unknown, the kriging weights depend on the Sp.T. covariance structure solely, and (15.13) is referred to as the so-called *Sp.T. ordinary kriging predictor*. In practice, the dependence structure is unknown too and needs to be estimated from the data and then plugged into the kriging weights (λ_{ij}^{t}'s). Under the assumption that the Sp.T. process is second-order stationary and spatially isotropic, satisfying (15.3), the dependence structure, given by the (spatially isotropic) Sp.T. variogram,

$$E[(Z(s,r) - Z(v,t))^2] = 2\gamma_Z(\|s-v\|, |r-t|), \quad s, v \in D \text{ and } r, t \in T,$$

where $h = \|s - v\|$ and $u = |r - t|$ is typically estimated via the following steps: First, an empirical (spatially isotropic) Sp.T. semivariogram is computed from lag classes as

$$\hat{\gamma}_Z(h,u) = \frac{1}{2|N(h,u)|} \sum_{(i,j,k,l)\in N(h,u)} (Z(s_i,t_{ik}) - Z(s_j,t_{jl}))^2,$$

where $N(h,u) = \{(s_i,t_{ik}),(s_j,t_{jl}) : \|s_i - s_j\| \in (h-\epsilon, h+\epsilon)$, and $|t_{ik} - t_{jl}| \in (u-\delta, u+\delta)\}$, for some $\epsilon, \delta > 0$, and $|N(h,u)|$ is the number of distinct elements in $N(h,u)$. A parametric semivariogram model, $\gamma(h,u|\theta)$, is then fitted to a set of $\{\hat{\gamma}(h_l,u_l),(h_l,u_l)\}, l = 1,\dots,L$ by a least squares method. Three types of stationary Sp.T. semivariogram (covariogram) models that are commonly used to estimate the Sp.T.-dependence structure is described below: the separable, product-sum, and metric models. In Gräler et al. [23], it is illustrated how to perform Sp.T. ordinary kriging prediction with these three models in the R-package gstat.

The separable model assumes that the Sp.T. covariance function can be modeled by the product of the spatial and the temporal covariance functions:

$$C_Z(h,u) = C_s(h)C_t(u). \tag{15.15}$$

The corresponding semivariogram is given by

$$\gamma_Z(h,u) = \sigma_Z^2(\overline{\gamma}_s(h) + \overline{\gamma}_t(u) - \overline{\gamma}_s(h)\overline{\gamma}_t(u)),$$

where $\overline{\gamma}_s(h)$ and $\overline{\gamma}_t(u)$ are standardized spatial and temporal semivariograms with separate nugget effects and (joint) sill of 1. The parameter σ_Z^2 is the overall sill, i.e. the variance of the process $Z(s,t)$. This model has the computational advantage of being able to express the covariance matrix as the Kronecker product between two covariance matrices (space and time) which simplifies and speeds up the computation of its determinant and inverse.

The product-sum model is an extension of the separable model and relies on the assumption that the covariance function can be written as follows:

$$C_Z(h,u) = kC_s(h)C_t(u) + C_s(h) + C_t(u),$$

where $k > 0$. The corresponding semivariogram is given by

$$\gamma_Z(h,u) = (k\sigma_t^2 + 1)\gamma_s(h) + (k\sigma_s^2 + 1)\gamma_t(u) - k\gamma_s(h)\gamma_t(u),$$

where σ_t^2 and σ_s^2 are the temporal and spatial sills, respectively. Moreover, the value of k needs to satisfy $\sigma_Z^2 = \sigma_t^2 + \sigma_s^2 + k\sigma_t^2\sigma_s^2$.

The metric model is another way of modeling the covariance function. Here the covariance function is a function of the (weighted) Euclidean distance between two observations. To treat the spatial and temporal distances equally, the spatial and temporal dimensions are matched by an anisotropy parameter κ. The metric Sp.T. covariance model is given by

$$C(h,u) = C_{joint}(\sqrt{h^2 + (\kappa u)^2}).$$

The corresponding semivariogram becomes

$$\gamma(h,u) = \gamma_{joint}(\sqrt{h^2 + (\kappa u)^2}) = \sigma_Z^2 - C_{joint}(\sqrt{h^2 + (\kappa u)^2}).$$

Note that when $\kappa = 1$, this covariance model corresponds to an isotropic second-order stationary random process in $\mathbf{R}^3$.

More generally, in Sp.T. kriging modeling, the process is often described as

$$Z(s,t) = \mu(s,t) + \epsilon(s,t),$$

where $\mu(s,t) = m_s(t)$ is a deterministic trend and $\epsilon(s,t)$ is a mean zero Sp.T. random field, usually assumed stationary. The trend is typically modeled by

$$\mu(s,t) = \boldsymbol{\beta}^\intercal \mathbf{x}(s,t),$$

where $\mathbf{x}(s,t) \in \mathbf{R}^M$ is a set of M known covariates, often chosen to be polynomials of s and t, and $\boldsymbol{\beta} \in \mathbf{R}^M$ is an unknown parameter to be determined.

When the Sp.T. process has a deterministic (unknown) nonconstant trend, then the BLUP (15.13) that minimizes (15.14) is called the *Sp.T. universal kriging predictor*, and the kriging weights are functions of both the dependence structure and the covariates evaluated at the observed and predicted locations, see e.g. [3] Chapter 4.1.2, page 148. In order to estimate $\boldsymbol{\beta}$ and the Sp.T. variogram parameter θ, without relying on distributional assumptions, an iterative weighted least squares method may be used. First, $\boldsymbol{\beta}$ is estimated by the OLS method, minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Z(s_i,t_{ij}) - \boldsymbol{\beta}^\intercal \mathbf{x}(s_i,t_{ij}))^2.$$

Based on the resulting regression residuals, the Sp.T. semivariogram is then estimated by fitting a parametric Sp.T. semivariogram model to the corresponding empirical Sp.T. semivariogram by a least squares method. The parameter $\boldsymbol{\beta}$ is then re-estimated using a weighted least squares method, taking into account the estimated dependence structure of the residuals [2]. The dependence structure (variogram) is again estimated based on the updated residuals and the whole procedure iterated until convergence.

Note that if there is a deterministic time trend, but no spatial, such that $\mu(s,t) = m_s(t) = m(t)$, the functional kriging methods do not need to specify and estimate the trend, whereas the Sp.T. kriging methods need to.

15.3.3 Evaluation of Kriging Methods

A common way of evaluating the prediction performance of prediction methods for functional data is *FCV*, as suggested by Giraldo et al. [5, 6]. In FCV, the data from each observed spatial location is removed, one at a time, and then it is

predicted at all observed time points by the prediction method using the observed functional data at the remaining locations. The MSPE is computed as

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Z(s_i,t_{ij}) - \hat{Z}^{-i}(s_i,t_{ij}))^2/m_i, \tag{15.16}$$

where $\hat{Z}^{-i}(s_i,t_{ij})$ denotes the predicted value at location (s_i,t_{ij}) based on the functional data with the observations $Z(s_i,t_{ij}), j = 1,\dots,m_i$ excluded.

15.4 A Simulation Study

Here we present a simulation study that aims to shed light over the relative merits of Sp.T. and functional kriging, with particular focus on Gaussian second-order stationary functional processes in $\mathbf{R}^2$. Since the functional kriging methods OKFD, PWFK, and FKTM often coincide for such processes (see Sections 15.3.1.2 and 15.3.1.3), we restrict our comparisons to Sp.T. kriging vs. OKFD. We simulate data from three main types of Gaussian processes. The first two scenarios correspond to stationary isotropic Sp.T. processes with separable and nonseparable covariance functions, respectively. The third scenario corresponds to second-order stationary functional (but nonstationary Sp.T.) processes with constant mean. For all three scenarios, several different cases are simulated, with varying strengths of spatial and temporal dependence. All the considered cases in the study are presented in Table 15.1, where the different parameters control the Sp.T. correlation structure in the three different main scenarios. Examples of simulated data for six cases are illustrated in Figure 15.1.

For each case in Table 15.1, data at $m = 50$ time points, equally distributed on $[0,1]$, were generated at each of $n = 6\times 6 = 36$ spatial locations, located on a regular grid in $[0,1]\times[0,1]$. For each case, $N = 100$ independent realizations were generated. An extended version of this simulation study is presented by Strandberg et al. [13], where two other sample sizes also were considered; *small* referring to $n = 6\times 6$ spatial locations and $m = 12$ time points, and *large* referring to $n = 15\times 15$ spatial locations and $m = 50$ time points. The sample size presented here is referred to as *medium*. Moreover, the presence of a deterministic time trend, $m(t) = 9 + 3\sin(2\pi t)$, for cases 1–18 was also investigated by Strandberg et al. [13]. Below we present each of the three main scenarios in more detail, together with the simulated results.

15.4.1 Separable

The first nine cases in Table 15.1 were simulated using the R-package RandomFields [24] and are Gaussian stationary Sp.T. processes with separable covariance

Table 15.1 The 24 different types (cases) of simulated Gaussian processes and their parameters: isotropic second-order stationary Sp.T. processes with separable (cases 1–9) and nonseparable (cases 10–18) covariance functions, and second-order stationary functional (but non-stationary Sp.T.) processes (cases 19–24) with constant means.

Generated data				Generated data				Generated data			
Case	**Type**	α	β	**Case**	**Type**	α	β	**Case**	**Type**	**#bases (p)**	α
1	Separable	0.1	0.1	10	Non-separable	0.1	0.1	19	Non-stationary	7	0.1
2			1	11			1	20			0.5
3			10	12			10	21			2
4		0.5	0.1	13		0.5	0.1	22		15	0.1
5			1	14			1	23			0.5
6			10	15			10	24			2
7		2	0.1	16		2	0.1				
8			1	17			1				
9			10	18			10				

The larger the value of α and β the weaker the spatial and temporal correlation, respectively.

functions (15.15). The spatial covariance function $C_s(h)$ in Eq. (15.15) was set to the exponential covariance function with nugget effect, taking the form:

$$C_s(h) = (1 - \nu)\exp(-\alpha h) + \nu I\{h = 0\},$$

where $\nu \in (0, 1]$ is the nugget effect and α controls the strength of the spatial correlation structure. The parameter ν was set to be 0.04, while the following values of α were considered; 0.1, 0.5, and 2, corresponding to effective ranges 30, 6, and 1.5 (very strong, medium and weak spatial correlation), respectively. The temporal covariance function $C_t(u)$ in (15.15) was given by the stable covariance function:

$$C_t(u) = \exp(-(\beta u)^\gamma), \tag{15.17}$$

where β controls the strength of the temporal correlation structure and γ is a parameter that should be in the interval $(0, 2]$ in order to provide a valid covariance function. Here, γ was fixed to be 0.5, while the following values of β were considered: 0.1, 1, and 10, corresponding to the effective ranges 90, 9, and 0.9 (very strong, medium, and weak temporal correlation), respectively.

Estimation of the OKFD model was performed using the R-package geofd [18]. Given the generated data $Z(s_i, t_j), i = 1, \ldots, 36, j = 1, \ldots, 50$, the OKFD model was estimated using two types of basis functions for $p = 5, 15, 25, 35, 45, 47$, and 49, in order to study the effect of changing the number of basis functions in (15.9). We used Fourier (the first p) and cubic B-spline basis functions. The (p) cubic B-splines

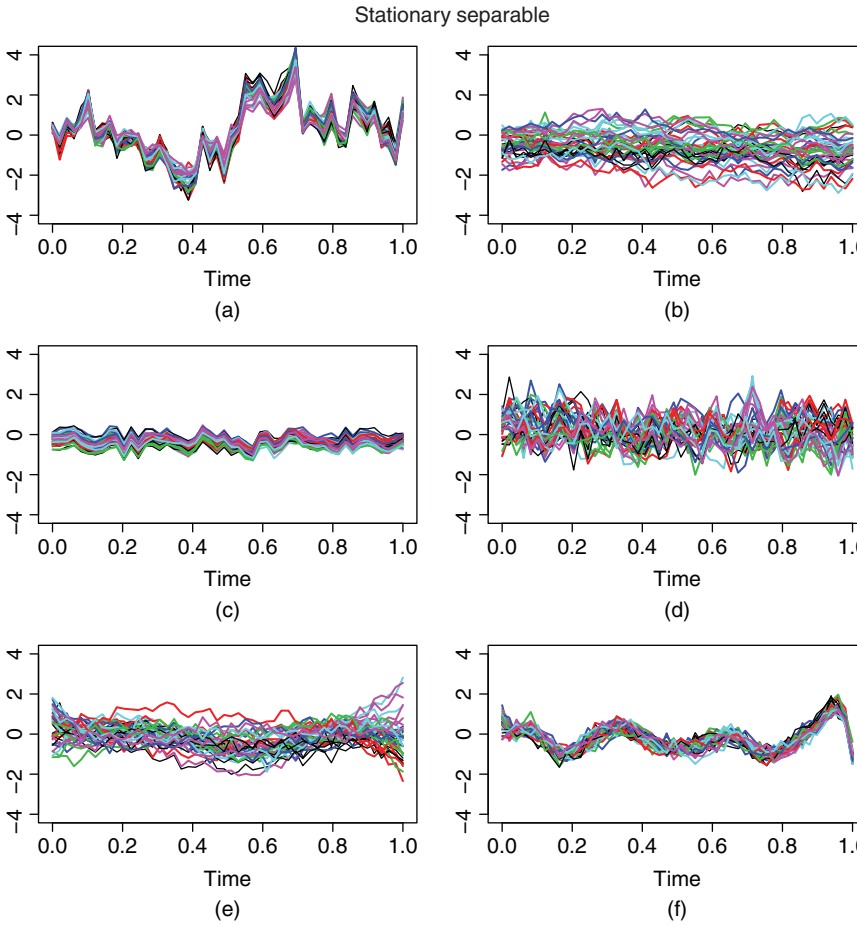


Figure 15.1 Examples of simulated data for: (a) case 3 ($\alpha = 0.1$, $\beta = 10$), (b) case 7 ($\alpha = 2$, $\beta = 0.1$), (c) case 10 ($\alpha = 0.1$, $\beta = 0.1$), (d) case 18 ($\alpha = 2$, $\beta = 10$), (e) case 21 ($\alpha = 2$, $p = 7$), and (f) case 22 ($\alpha = 0.1$, $p = 15$). The larger the value of α and β, the weaker the spatial and temporal correlation, respectively.

were constructed based on $p - 4$ equally distributed interior knots on the interval $[0, 1]$. For each p and type of basis function, the exponential, spherical, and stable semivariogram models were fitted to the empirical trace-semivariogram. For each case (1–9), a total of $2 \times 7 \times 3 = 42$ different estimated OKFD models were thus fitted to the generated data, predictions made and evaluated by FCV in terms of the MSPE (15.16), and the minimum MSPE over the 42 models registered. The *overall MSPE* for each case was computed as the average minimum MSPEs over the 100 replicates.

Estimation of the Sp.T. kriging models was performed using the R-packages gstat [17] and space-time [25]. Separable semivariogram models (15.15) were fitted to the empirical Sp.T. semivariogram, being all pairwise combinations of the exponential, spherical, and stable variograms for the spatial (isotropic) and temporal variogram models. Hence, a total of $3 \times 3 = 9$ Sp.T. kriging models were fitted to the data, predictions evaluated by FCV, minimum MSPE registered over the nine models, and overall MSPE computed for each case (1–9).

The overall MSPEs for the OKFD and the Sp.T. separable model for cases 1–9 with medium sample size are presented in Table 15.2. The numbers highlighted

Table 15.2 Prediction performance in terms of MSPEs for the simulated cases 1–18.

Generated data				Overall MSPE		Comparison	
Case	**Type**	α	β	**OKFD**	**Sp.T. separable**	**#Times**	***P*-value**
1	Separable	0.1	0.1	**0.061** (0.2)	0.062 (26.7)	27	0.552
2			1	0.068 (0.2)	**0.067** (26.1)	23	0.059
3			10	0.069 (0.2)	**0.066** (24.7)	13	<0.001
4		0.5	0.1	**0.134** (0.2)	0.143 (23.6)	56	<0.001
5			1	**0.131** (0.2)	0.135 (29.8)	42	0.011
6			10	0.139 (0.2)	**0.137** (27.6)	24	0.044
7		2	0.1	**0.334** (0.2)	0.357 (29.0)	64	<0.001
8			1	**0.368** (0.2)	0.400 (29.3)	65	<0.001
9			10	**0.372** (0.2)	0.386 (28.7)	54	0.001
10	Non-Separable	0.1	0.1	**0.066** (0.2)	0.067 (26.2)	38	0.050
11			1	0.066 (0.2)	**0.065** (25.7)	25	0.082
12			10	0.065 (0.2)	**0.064** (24.4)	27	0.075
13		0.5	0.1	**0.128** (0.2)	0.139 (25.8)	49	0.001
14			1	**0.134** (0.2)	0.140 (24.2)	55	<0.001
15			10	**0.137** (0.2)	0.140 (27.4)	41	0.006
16		2	0.1	**0.366** (0.2)	0.398 (26.8)	67	<0.001
17			1	**0.354** (0.2)	0.390 (24.8)	63	<0.001
18			10	**0.373** (0.2)	0.391 (27.4)	52	0.003

The smallest overall MSPE for each case is highlighted in bold. The numbers in parentheses represent the average computational time in seconds over the corresponding estimated models and replications. The column #Times represents the number of times, out of the 100 realizations, that OKFD had lower (minimum) MSPE than the Sp.T. separable model. The last column shows *P*-values from two-sided paired *t*-tests comparing the overall MSPEs between the OKFD and the Sp.T. separable models.

in bold correspond to the smallest overall MSPE (per case), while the numbers in parentheses report the average computation time (for estimation and FCV) in seconds over all estimated models and replications (when run on a 3.5 GHz Intel Core i7 processor with 32 GB RAM memory). The second to last column presents the number of times, out of the 100 realizations, that OKFD had lower (minimum) MSPE than the Sp.T. separable model. The last column in Table 15.2 reports p-values from paired two-sided t-tests comparing the overall MSPEs between the OKFD and the Sp.T. separable models, and thus reflects for which cases significant differences occur.

It is interesting to note that for cases 1–9, the overall MSPE was often (significantly) lower for OKFD (cf. Table 15.2), even though the simulated data were generated from Sp.T. models with separable covariance functions. A closer look at the overall MSPEs reveals that the weaker the spatial correlation and the stronger the temporal correlation, the better the OKFD performs and the worse the Sp.T. separable model performs. As an example, for case 3, corresponding to strong spatial and weak temporal correlation, the Sp.T. separable model has a significantly lower overall MSPE compared to the OKFD model ($p < 0.001$), while the result is reversed ($p < 0.001$) for case 7, which corresponds to data generated with weak spatial and strong temporal correlation. Moreover, a comparison of the computation time shows that prediction by and estimation of an OKFD model is substantially (over 100 times) faster than the Sp.T. separable model for cases 1–9 (Table 15.2).

Figure 15.2 presents how the type and number of basis functions used in the OKFD model affect the prediction performance (minimum MSPE over the three trace-semivariogram models, averaged over the 100 realizations) for cases 3 and 7. The number of basis functions turns out to be an important factor for prediction

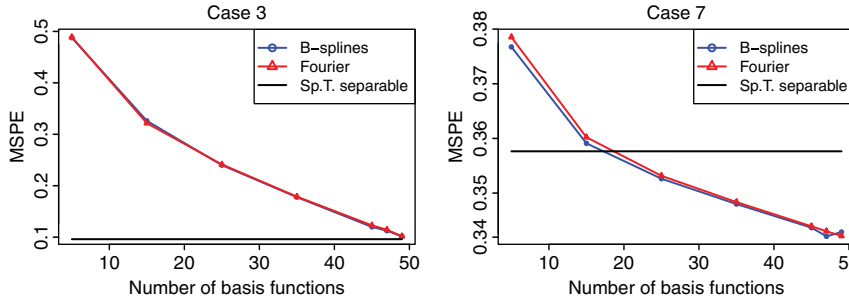


Figure 15.2 Prediction performance (minimum MSPE over the three trace-semivariogram models, averaged over the 100 realizations) for cases 3 and 7 when the estimated OKFD model is based on different numbers (p) of basis functions, being both Fourier and cubic B-spline bases. The solid black lines represent the corresponding overall MSPE of the Sp.T. separable model.

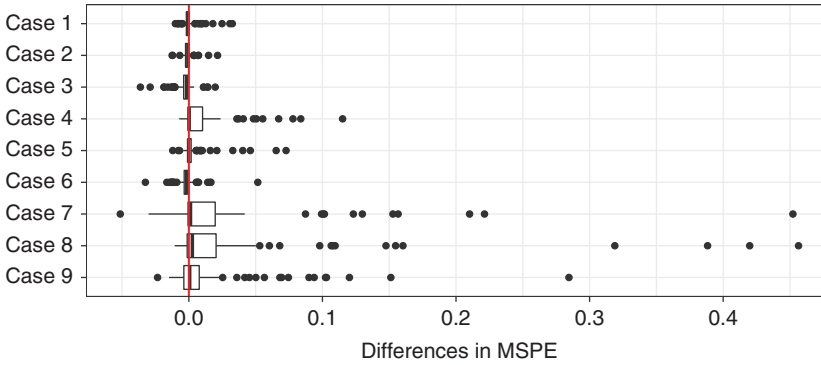


Figure 15.3 Box plots for cases 1–9 of the differences in (minimum) MSPE between the two kriging approaches (MSPE(Sp.T)−MSPE(OKFD)) for the 100 replicates, for medium sample size.

performance; in general using more basis functions results in a smaller prediction error. On the other hand, the type of basis functions, Fourier or cubic B-splines, is of less importance. These findings are consistent with all cases (1–9) and for small, medium, and large sample sizes [13].

To see how prediction performance may vary between replicates, Figure 15.3 presents box plots of the differences in (minimum) MSPE between the two kriging approaches (MSPE(Sp.T)–MSPE(OKFD)) for the 100 replicates. Here it becomes clear that OKFD produces more robust predictions. The Sp.T. kriging method (with estimated separable covariance function) produced much higher MSPEs than OKFD (casewise) for many realizations, especially for cases 1, 4, 5, 7, 8, and 9.

The study made by Strandberg et al. [13] further indicates that, in general, the larger the sample size, the more likely it is that the estimated Sp.T. (separable) models perform better than OKFD. They also show that the presence and estimation of a deterministic (sinusoidal) time trend did not have a large effect on the prediction performance, and more or less gave the same conclusions with respect to the relative performance of the two kriging approaches.

15.4.2 Non-separable

Cases 10–18 in Table 15.1 were simulated using the R-package RandomFields, and correspond to Gaussian stationary Sp.T. processes with nonseparable covariance functions of the form

$$\mathrm{Cov}_{\mathrm{NSEP}}(h,u) = (1-\nu)(2-C_t(u))^{-\delta/2}\exp\left(-\frac{\alpha h}{\sqrt{2-C_t(u)}}\right) + \nu I\{h=0\}.$$

Here δ is a parameter that must be greater than or equal to the spatial dimension of the field. The parameter δ was fixed to be 2, $\nu = 0.04$, and $\alpha = 0.1, 0.5$, and 2. The covariance function $C_t(u)$ was set to be the stable covariance function (15.17) with $\gamma = 0.5$ and $\beta = 0.1, 1$, and 10.

The same (42) OKFD models and (9) Sp.T. kriging models (with separable variograms) as estimated in Section 15.4.1 were fitted to the simulated data sets of cases 10–18 using the R-packages geofd, gstat, and space-time, each with 100 realizations. In addition to the Sp.T. separable kriging models, Strandberg et al. [13] also fitted Sp.T. kriging models with product sum and metric variogram models.

Prediction performance was evaluated in the same way as described in Section 15.4.1 and is summarized in Table 15.2 for medium sample size. In general, when comparing the two kriging approaches for the nonseparable cases 10–18, similar conclusions as for the separable cases 1–9 were drawn; the weaker the spatial correlation and the stronger the temporal correlation, the better the OKFD performs and the worse the Sp.T. separable model performs; OKFD works better for smaller sample sizes, whereas fitted Sp.T. separable kriging models performs better for large sample sizes, cf. [13]; more basis functions in OKFD generally improve prediction performance; computation times are much shorter for OKFD; the presence of a deterministic time trend did not change the conclusions. For more details see [13].

A more detailed comparison of the overall MSPEs in Table 15.2 reveals that prediction performance of OKFD in general improves in comparison to the Sp.T. separable kriging models for the simulated data sets with nonseparable covariance functions (cases 10–18) compared to those simulated from separable covariance functions (cases 1–9). This result is to be expected, since none of the fitted (Sp.T.) kriging models coincide with the models that generated the data for cases 10–18. This tendency holds also for small and large sample sizes [13].

15.4.3 Nonstationary

Generation of simulated data sets of second-order isotropic stationary functional, but nonstationary Sp.T. Gaussian processes with constant mean (cases 19–24 in Table 15.1) were based on the model:

$$\chi_{s_i}(t) = \mathbf{a}_i^\intercal \mathbf{B}(t) + \epsilon_{s_i}(t), \quad i = 1, \dots, n. \tag{15.18}$$

Here $\mathbf{B}(t) = (B_1(t), \dots, B_p(t))^\intercal$ are $p = 7$, and 15 cubic B-spline basis functions, defined on equally space knots on the interval $[0, 1]$. Moreover, $\mathbf{a}_i = (a_1(s_i), \dots, a_p(s_i))^\intercal$, where $a_k(s), k = 1, \dots, p$, were chosen to be p independent identically distributed second-order stationary isotropic mean zero Gaussian processes in $\mathbf{R}^2$ with exponential covariance function $C(h) = \exp(-\alpha h)$ with α set to 0.1, 0.5, and 2. Hence, the vectors $(a_k(s_1), \dots, a_k(s_n))^\intercal, k = 1, \dots, p$, are

p independent realizations of a multivariate Gaussian random variable $N_n(\mathbf{0}, \mathbf{\Sigma})$, where the $n \times n$ covariance matrix equals $\mathbf{\Sigma} = \{\exp(-\alpha \|s_i - s_j\|)\}$. Finally, the $\epsilon_{s_i}(t)$'s correspond to white noise measurement errors, assumed to be independent identically normally distributed random variables with mean 0 and variance 0.04 for all i and t, i.e. $\epsilon_{s_i}(t) \sim N(0, 0.04)$. For each of the $2 \times 3 = 6$ cases (19–24), 100 independent realizations were generated using the R-package fda [26].

We fitted the same OKFD models as those fitted in Section 15.4.1 to the data sets using the R-package geofd. However, this time we extended the choices of number of basis functions to 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 35, 45, 47, and 49, using both Fourier and cubic B-splines basis functions, yielding a total of $2 \times 15 \times 3 = 90$ different estimated OKFD models. For each case (19–24) and realization, predictions were made and evaluated by FCV for all 90 models, and the minimum MSPE over the 90 models registered. The overall MSPE for each case was finally computed as the average minimum MSPE over the 100 replicates. Furthermore, the nine Sp.T. ordinary kriging models with separable covariance functions, fitted to the data in Section 15.4.1, were also estimated for these data sets. Additionally, Sp.T. universal kriging models were fitted, with a deterministic time trend specified by a linear combination of the same basis functions that were used to generate the data set. Hence, a total of $2 \times 9 = 18$ Sp.T. kriging models were fitted to the data, predictions evaluated by FCV, minimum MSPE registered over the 18 models, and overall MSPE computed for each case (19–24).

Note that these simulated data sets have time-varying variances and covariances, which the Sp.T. kriging approach is not designed to capture, whereas the OKFD model can handle such situations. For cases 19–24, we would therefore expect OKFD to perform better than the Sp.T. kriging approach. This is indeed the case, as can be seen in Table 15.3, which summarizes the prediction performance of the two kriging approaches for medium sample size. In fact, OKFD has significantly lower overall MSPE for all cases 19–24. This is generally also true for small and large sample sizes [13]. Moreover, we again note that the computation time for OKFD is much shorter than for the Sp.T. separable model.

Figure 15.4 illustrates how the type and number of basis functions used in the fitted OKFD models affect the prediction performance (minimum MSPE over the three trace-semivariogram models, averaged over the 100 realizations) for cases 21 and 22. Case 21 corresponds to simulated data generated by 7 B-splines with weak spatial dependence, whereas case 22 corresponds to simulated data generated by 15 B-splines with strong spatial dependence. In contrast to the simulated stationary Sp.T. models (cases 1–18) where prediction performance typically increases with the number of basis functions used in the fitted OKFD models, here we observe this phenomena only when Fourier basis are used in the fitted OKFD models. For B-spline bases, the best prediction performance is (naturally) achieved using the same number of B-splines in the OKFD fitted models as used to generate

Table 15.3 Prediction performance in terms of MSPEs for the simulated cases 19–24.

Generated data				Overall MSPE		Comparison	
Case	**Type**	**#bases (p)**	α	**OKFD**	**Sp.T. separable**	**#Times**	**P-value**
19	Non-stationary	7	0.1	**0.050** (0.2)	0.055 (24.8)	100	<0.001
20			0.5	**0.083** (0.2)	0.092 (24.4)	100	<0.001
21			2	**0.202** (0.2)	0.212 (28.7)	98	<0.001
22		15	0.1	**0.052** (0.2)	0.056 (26.1)	100	<0.001
23			0.5	**0.087** (0.2)	0.094 (28.0)	100	<0.001
24			2	**0.209** (0.2)	0.218 (28.2)	100	<0.001

The smallest overall MSPE for each case is highlighted in bold. The numbers in parentheses represent the average computational time in seconds over the corresponding estimated models and replications. The column #Times represents the number of times, out of the 100 realizations, that OKFD had lower (minimum) MSPE than the Sp.T. separable model. The last column shows P-values from two-sided paired t-tests comparing the overall MSPEs between the OKFD and the Sp.T. separable models.

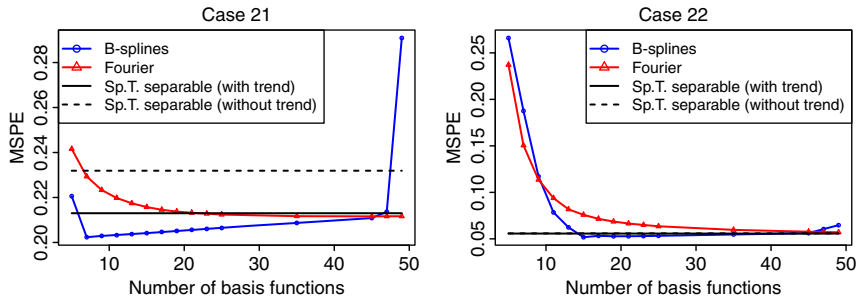


Figure 15.4 Prediction performance (minimum MSPE over the three trace-semivariogram models, averaged over the 100 realizations) for cases 21 and 22 when the estimated OKFD model is based on different numbers (p) of basis functions, being both Fourier and cubic B-spline bases. The solid and dashed black lines represent the corresponding overall MSPE of the Sp.T. separable model with and without an estimated deterministic time trend, respectively.

the simulated data set (7 for case 21 and 15 for case 22). In fact, using too many B-splines may give substantially poorer predictions, especially when the spatial dependence is weak, as for case 21, cf. Figure 15.4. It can also be noted that the best OKFD model using B-splines has significantly smaller MSPE than the best OKFD model using Fourier basis. If the simulated data sets would have been generated by a set of Fourier basis instead, we would most likely see the opposite behavior, i.e. that the same number of Fourier basis in the fitted OKFD model as in the data

generation model would probably give the best prediction performance, and do better than the OKFD models using B-splines.

For the Sp.T. separable kriging models, it turned out that it was advantageous to use universal kriging (estimating a deterministic time trend), especially for the cases with weak spatial dependence, see Figure 15.4, whereas the prediction performance was about the same for cases with strong spatial dependence. This was true also for small and large sample sizes [13]. It was also noted by Strandberg et al. [13] that Sp.T. kriging models with fitted metric variograms sometimes had better prediction performance than the Sp.T. separable kriging models, but still worse than the best OKFD models.

15.5 Application: Spatial Prediction of Temperature Curves in the Maritime Provinces of Canada

Here, we analyze meteorological data with the same structure as the Canadian weather data set introduced in [21] but observed over a smaller region. Our data set, available in the R package geofd [18], consists of temperature measurements recorded at $n = 35$ weather stations (spatial locations) at Canada's Atlantic coast in the Maritime Provinces (Figure 15.5a). An additional weather station, Moncton, which is not used to estimate the kriging models, is also considered to illustrate out-of-sample prediction performance. For each station, the data contain information of the daily mean temperature averaged over the period 1960–1994 (29 February combined with 28 February). The resulting functional data (365 measurements for each of the 35 stations) are displayed in Figure 15.5b, connected by light gray lines. These data have previously been analyzed by e.g. Giraldo [19], Giraldo et al. [5], and Menafoglio et al. [9] to illustrate and compare the ordinary functional kriging approaches OKFD, PWFK, FKTM, and a universal kriging for functional data (UKFD). Here we use the data to compare the prediction performance of Sp.T. kriging models with the OKFD model.

First the data was predicted by the OKFD model using the R-package geofd. In coherence with the abovementioned analyses, we represented the functional data at each weather station by a linear combination of Fourier basis functions using (15.9). The OKFD model was estimated using 51, 101, 151, 201, 251, 301, and 351 Fourier bases, in order to study how the number of basis functions affected the prediction performance. Moreover, three semivariogram models (exponential, spherical, and stable) were fitted to the empirical trace-semivariogram by the OLS method. Thus, in total, we estimated $7 \times 3 = 21$ OKFD models. Predictions were then made and evaluated by FCV in terms of their MSPEs (15.16).

The stable trace-semivariogram (Figure 15.6a) resulted in the best prediction performance for all considered numbers of Fourier bases. Figure 15.6b presents

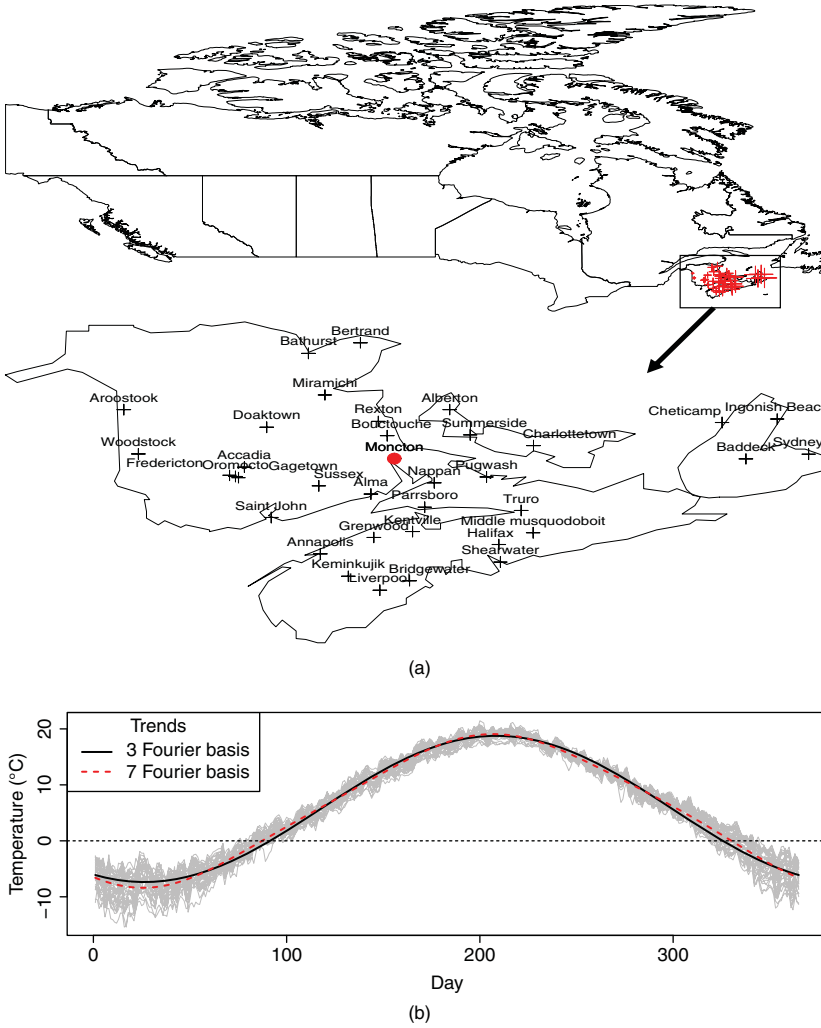


Figure 15.5 The locations of the 36 weather stations in the Canadian Maritime provinces (a) where the average (over 30 years) daily temperature curves (b) were registered. The bottom panel also presents the estimated common time trend specified as linear combinations of the first 3 and 7 Fourier basis functions, respectively.

how the number of Fourier basis functions used in the fitted OKFD models affects the prediction performance (minimum MSPE over the three trace-semivariogram models). The figure clearly reveals that the prediction performance increases with the number of bases. Thus, the best performance was attained with 351 Fourier bases and its MSPE was 0.5738. Furthermore, it was also noted that the

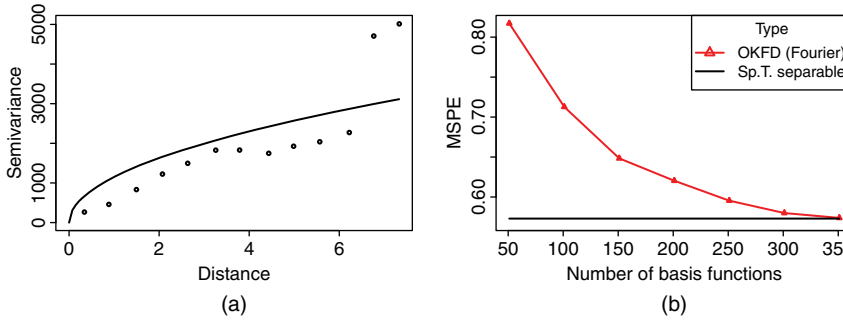


Figure 15.6 (a) The empirical trace-semivariogram and the best fitted stable model for the Canadian temperature curves, represented by 351 Fourier basis functions. (b) Minimum MSPE over the three trace-semivariogram models for OKFD, based on different number of Fourier basis functions. The solid black line represents the MSPE of the best Sp.T. model.

computation time for running OKFD increased slightly with the number of bases. The average computation time for an estimated OKFD model based on 51 and 351 Fourier bases was less than one and three seconds, respectively.

Next, the data was predicted using Sp.T. kriging. Since the data show a clear time trend, universal Sp.T. kriging was first applied. The deterministic time trend was modeled by a linear combination of the 3 (and 7) first Fourier bases,

$$m(t) = \beta_0 + \sum_{k=1}^{p} \left(\beta_{k1} \cos\left(\frac{2k\pi(t-1)}{364} \right) + \beta_{k2} \sin\left(\frac{2k\pi(t-1)}{364} \right) \right),$$

where $t \in T = [1,365]$, for $p = 1$ (and 3), and estimated by the OLS method. The dependence structure of the residuals obtained from the fitted trend was then estimated by fitting Sp.T. second-order stationary and isotropic semivariogram models to the empirical Sp.T. semivariogram of the residuals (see Figure 15.7a). The Sp.T. semivariogram models (separable, product-sum, and metric) described in Section 15.3.2 were estimated, letting their corresponding spatial, temporal, and joint semivariogram models be altered between the exponential, spherical, and stable semivariogram models. This resulted in nine separable, nine product-sum, and three metric Sp.T. semivariogram models. As a comparison, we also predicted the original data by Sp.T. ordinary kriging which assumes a constant deterministic trend. The Sp.T. semivariogram models fitted to the empirical Sp.T. semivariogram based on the original data were the same as those used for the Sp.T. universal kriging models. Thus, in total, we investigated $(9 + 9 + 3) \times 3 = 63$ Sp.T. kriging models. All models were fitted to the data and predictions evaluated by FCV.

Figure 15.7a illustrates the dependence structure in terms of the empirical Sp.T. semivariogram, computed based on the residuals after estimating a deterministic

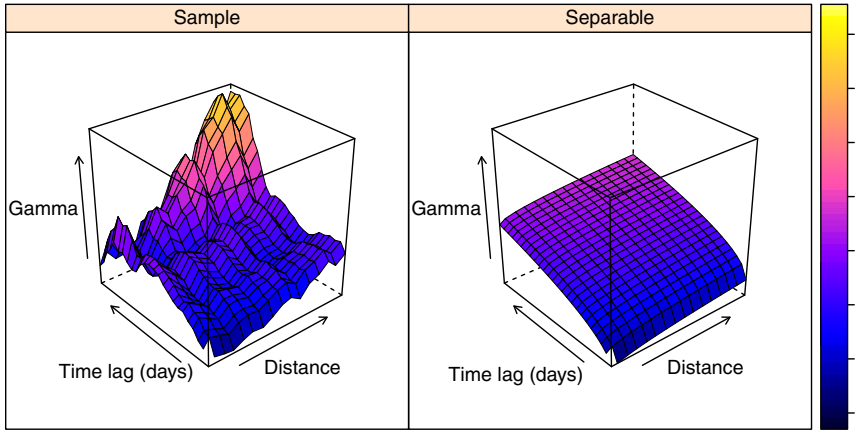


Figure 15.7 The empirical Sp.T. semivariogram (a) and the best-fitted Sp.T. separable model (with the stable variogram models used for both the spatial and temporal variograms) based on the residuals obtained from a deterministic time trend with three Fourier bases (b), for the Canadian weather data.

time trend with three Fourier basis functions. The corresponding fitted Sp.T. separable semivariogram model, yielding the best prediction performance, is given in Figure 15.7b. The increasing values of the Sp.T. semivariograms in both the spatial and temporal dimensions indicate that there is dependence structure left in the residuals.

The best Sp.T. models, in terms of smallest MSPE, within each of the three groups of dependence structure (separable, product-sum, and metric), and for each type of trend are presented Table 15.4. The numbers in parentheses report the corresponding average computation time in seconds over the estimated

Table 15.4 Prediction performance of different Sp.T. kriging models for the Canadian weather data.

	MSPE		
Trend	**Sp.T. separable**	**Sp.T. product-sum**	**Sp.T. metric**
No trend	$0.5730\ (1.8 \cdot 10^2)$	$0.5861\ (1.3 \cdot 10^4)$	$0.5730\ (1.3 \cdot 10^4)$
3 Fourier basis	$0.5730\ (1.8 \cdot 10^2)$	$0.5731\ (1.3 \cdot 10^4)$	$1.1126\ (1.4 \cdot 10^4)$
7 Fourier basis	$0.5734\ (1.6 \cdot 10^2)$	$0.5731\ (1.3 \cdot 10^4)$	$1.0670\ (1.4 \cdot 10^4)$

For each type of trend and Sp.T. variogram model, the (minimum) MSPE is reported. The numbers in parentheses represent the average computational time in seconds over the corresponding estimated models.

models. Among the Sp.T. models, many of them have about the same prediction performance in terms of (minimum) MSPE, the exceptions being the Sp.T. metric models with estimated trend, cf. Table 15.4, which worked less well. Note that the best Sp.T. models have approximately the same magnitude of MSPE as the best OKFD model (MSPE being 0.5738). In terms of computation time, an OKFD model (taking one to three seconds to compute) was 100–10 000 times faster to compute compared to a Sp.T. kriging model (Table 15.4).

Figure 15.8A presents the FCV residuals obtained by the best OKFD (a) and Sp.T. (b) models, per location. The dark gray corresponds to the residuals at one of the locations, Bertrand, the black solid line to the pointwise residual means and the dashed line to the pointwise standard errors. The spatial distribution of the corresponding MSPEs (averaged over time) at each location, are given in Figure 15.8B,

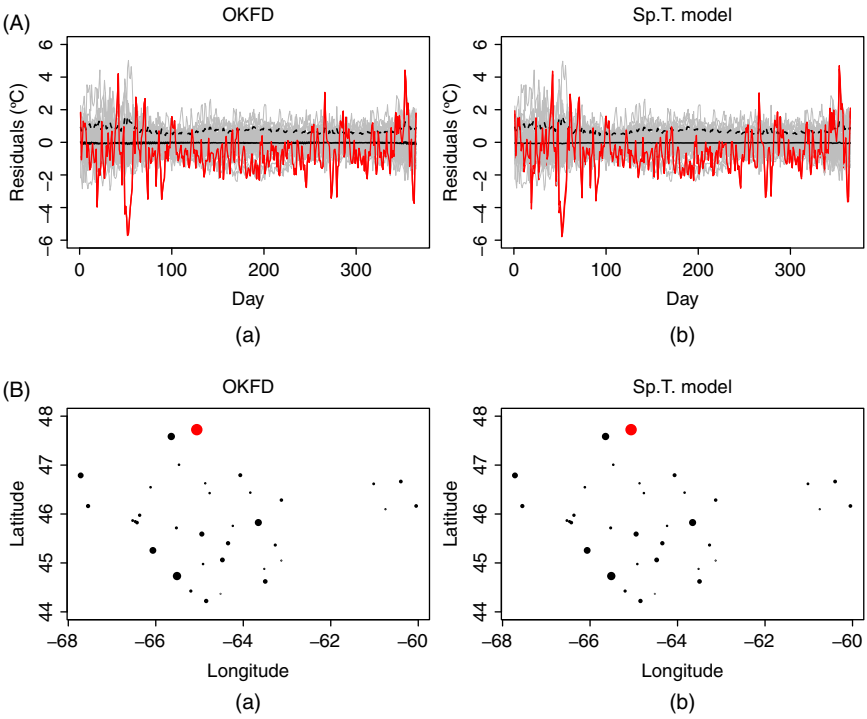


Figure 15.8 (A) Functional cross-validation residuals (gray lines) resulting from the best OKFD (a) and Sp.T. (b) model together with residual means (solid black lines) and standard deviations (dashed black lines) of the respective methods. (B) The MSPE (averaged over time) over the different sites for the best OKFD (a) and Sp.T. (b) model. The size of the points is proportional to the MSPE. The location with the highest MSPE (Bertrand) and its corresponding residual curve is highlighted in dark gray.

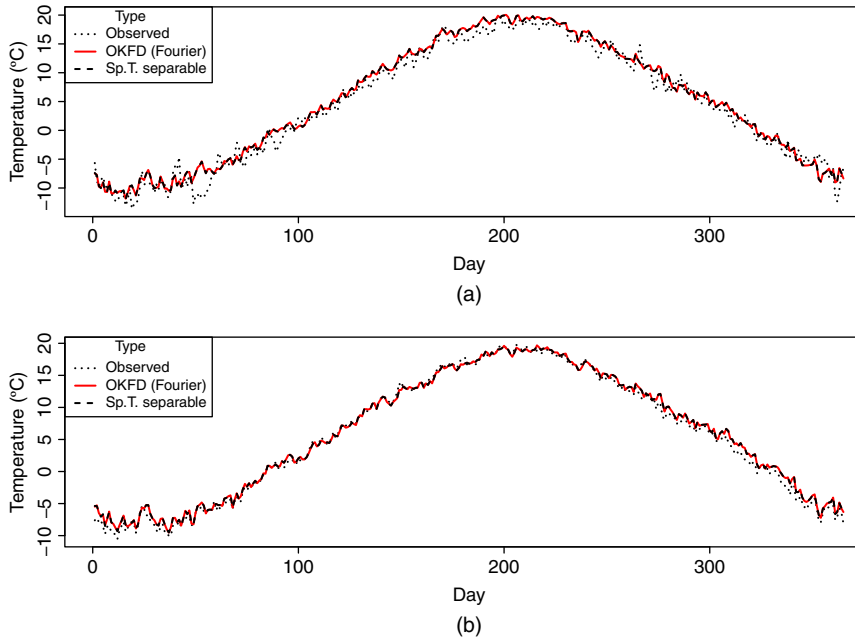


Figure 15.9 Predicted temperatures at locations Bertrand (a) and Moncton (b) obtained by the best OKFD model (solid gray line) and the best Sp.T. model (dashed black line) together with the observed (dotted) values.

where the size of each circle shows the magnitude of the MSPE. Bertrand has the largest prediction error as indicated by the dark gray circle in Figure 15.8B. Indeed, it can be noted that there are very small differences between the best OKFD and Sp.T. models in terms of prediction performance. This is emphasized in Figure 15.9, where the observed daily temperatures at locations Bertrand and Moncton are presented together with the corresponding predicted values using the best OKFD and Sp.T. kriging models.

Giraldo [19] analyzed this data set for the ordinary functional kriging methods OKFD, PWFK, and FKTM. The number and type of basis functions used in (15.9) to represent the $\chi_{s_i}(t)$'s were chosen to be the first 65 Fourier basis functions, determined by FCV. Giraldo [19] concluded that the three methods have similar FCV prediction performance. Menafoglio et al. [9] investigated the effect of using UKFD instead of OKFD for the Maritime data set, also representing the functional data by 65 Fourier basis functions. Menafoglio et al. [9] concluded that UKFD is better performing in terms of FCV prediction performance compared to OKFD. The FCV performance was there computed with respect to the fitted data, thus differing from previous work, including ours, where raw data was used.

15.6 Concluding Remarks

In this section, we have presented and compared functional and Sp.T. kriging approaches with respect to prediction performance and computation time, mainly by a simulation study but also on a real data set. The comparisons were restricted to Sp.T. kriging vs. the functional kriging method OKFD, since the more flexible functional kriging approaches PWFK and FKTM coincide with OKFD in several situations (Sections 15.3.1.2 and 15.3.1.3).

First, we noted that the prediction performance of OKFD (in terms of FCV) normally was improved when the number of basis functions used to represent the functional data increased. Second, it turned out that OKFD typically performed similarly or better than the Sp.T. kriging models for small and medium sample sizes. This is likely due to that it is more complicated to find good estimates of the Sp.T. variogram compared to the trace-variograms used in OKFD, since it has one dimension less. The large number of choices of Sp.T. variogram models and parameters to estimate makes the Sp.T. estimation process more vulnerable, especially for small data sets. For larger sample sizes, the Sp.T. kriging starts to perform better for the stationary Sp.T. processes, whereas OKFD continues to work best for the nonstationary Sp.T. (but stationary functional) processes. Furthermore, it was noted that OKFD performed better relative to Sp.T. kriging, the stronger the temporal- and the weaker the spatial dependence considered.

For all considered cases, OKFD was computationally much faster than the Sp.T. kriging models. This is mainly related to the large matrices that need to be inverted in order to perform Sp.T. kriging prediction at each location. One way of reducing the computation time for the Sp.T. kriging models is to use only the local neighborhood (e.g. the k closest neighboring locations) when prediction is made. This can often be done without much loss in prediction performance.

To conclude, the prediction performance of the two kriging approaches (functional and Sp.T.) is in general rather equal, with a tendency for functional kriging to work better for small sample sizes and Sp.T. kriging to work better for large sample sizes, when data are generated from stationary Sp.T. processes. However, from a computational perspective, OKFD is substantially faster than Sp.T. kriging. OKFD also has the possibility to give good predictions for a class of nonstationary Sp.T. processes where Sp.T. kriging may have problems. On the other hand, the functional kriging methods are designed to work on a common time domain, whereas this is not an issue for Sp.T. kriging.

References

1 Chiles, J.P. and Delfiner, P. (2009). *Geostatistics: Modeling Spatial Uncertainty*, vol. 497. Wiley.

2 Cressie, N. (2015). *Statistics for Spatial Data*. Wiley.

3 Cressie, N. and Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data*. Wiley.

4 Goulard, M. and Voltz, M. (1993). Geostatistical interpolation of curves: a case study in soil science. In: (ed. A.O. Soares)*Geostatistics Tróia'92*, 805–816. Springer.

5 Giraldo, R., Delicado, P., and Mateu, J. (2010). Continuous time-varying kriging for spatial prediction of functional data: an environmental application. *Journal of Agricultural, Biological, and Environmental Statistics* 15 (1): 66–82.

6 Giraldo, R., Delicado, P., and Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics* 18 (3): 411–426.

7 Nerini, D., Monestiez, P., and Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis* 101 (2): 409–418.

8 Caballero, W., Giraldo, R., and Mateu, J. (2013). A universal kriging approach for spatial functional data. *Stochastic Environmental Research and Risk Assessment* 27 (7): 1553–1563.

9 Menafoglio, A., Secchi, P., and Dalla Rosa, M. (2013). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics* 7: 2209–2240.

10 Ignaccolo, R., Mateu, J., and Giraldo, R. (2014). Kriging with external drift for functional data for air quality monitoring. *Stochastic Environmental Research and Risk Assessment* 28 (5): 1171–1186.

11 Reyes, A., Giraldo, R., and Mateu, J. (2015). Residual kriging for functional spatial prediction of salinity curves. *Communications in Statistics - Theory and Methods* 44 (4): 798–809.

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

13 Strandberg, J., Sjöstedt-de Luna, S., and Mateu, J. (2019). Prediction of spatial functional random processes: comparing functional and spatio-temporal kriging approaches. *Stochastic Environmental Research and Risk Assessment* 33 (10): 1699–1719.

14 Giraldo, R., Delicado, P., and Mateu, J. (2007). Geostatistics for Functional Data: An Ordinary Kriging Approach. *Tech. Rep*. Universitat Politècnica da Catalunya. http://hdl.handle.net/2117/1099 (accessed 01 February 2021).

15 Giraldo, R., Delicado, P., and Mateu, J. (2008). Continuous Time-Varying Kriging for Spatial Prediction of Functional Data: An Environmental Application. *Tech. Rep*. Universitat Politècnica da Catalunya. http://hdl.handle.net/2117/2167 (accessed 01 February 2021).

16 Goulard, M. and Voltz, M. (1992). Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Mathematical Geology* 24 (3): 269–286.

17 Pebesma, E.J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences* 30 (7): 683–691.

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

19 Giraldo, R. (2009). Geostatistical analysis of functional data. PhD thesis. Barcellona: Universitat Politècnica da Catalunya.

20 Giraldo, R. (2014). Cokriging based on curves, prediction and estimation of the prediction variance. *InterStat* 2: 1–30.

21 Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*, 2e. Springer, New York.

22 Menafoglio, A. and Petris, G. (2016). Kriging for hilbert-space valued random fields: the operatorial point of view. *Journal of Multivariate Analysis* 146: 84–94.

23 Gräler, B., Pebesma, E., and Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *R Journal* 8 (1): 204–218.

24 Schlather, M., Malinowski, A., Menck, P.J. et al. (2015). Analysis, simulation and prediction of multivariate random fields with package randomfields. *Journal of Statistical Software* 63 (8): 1–25.

25 Pebesma, E. (2012). spacetime: Spatio-temporal data in R. *Journal of Statistical Software* 51 (7): 1–30.

26 Ramsay, J.O., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer Science & Business Media.

16

From Spatiotemporal Smoothing to Functional Spatial Regression: a Penalized Approach

Maria Durban[1], Dae-Jin Lee[2], María del Carmen Aguilera Morillo[3], and Ana M. Aguilera[4]

[1] *Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain*
[2] *BCAM—Basque Center for Applied Mathematics, Bilbao, Basque Country, Spain*
[3] *Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València, Spain*
[4] *Department of Statistics and Operational Research, University of Granada, Granada, Spain*

16.1 Introduction

In the context of spatial data, the ideas of smoothing has been present for many years. For example, Kammann and Wand [1] combined kriging with low- rank smoothing for continuous covariates. Most recently, Cressie and Johannesson [2] built up on this idea and developed a spatial random effects model and fixed rank kriging, based on predicting the coefficients of a set of basis functions. In a similar context, Lee and Durbán [3] proposed the use of two-dimensional *B*-spline basis with random coefficients whose covariance structure was given by a penalty matrix controlled by separate smoothing parameters for each dimension.

In the last 20 years, several approaches have been developed in the context of spatiotemporal data. Some of them are based on the idea of Kalman filters [4, 5]. Others have a Bayesian perspective, in particular, hierarchical models based on Markov Random fields have become very popular [6, 7]. Here, we adopt the method introduced by Lee and Durbán [8] based on penalized smooth mixed models. These models use *B*-spline basis and discrete penalties [9] in a multidimensional setting [9, 10], but the methodology can be applied for any basis and quadratic penalty. Penalized splines do not rely on stochastic processes, and so, there is no need to model the covariance function. Although the number and position of the knots for the basis functions have to be chosen, the use of penalties relaxes the importance of the knots placement [11]. The mixed model representation of P-splines [12] solves the problems of the selection of the

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.

smoothing parameter (since they become a ratio of variance parameters), and the nonidentifiability common in additive models (see [13] for further details).

Smoothing is also a key tool in the context of FDA where sample functions are usually observed with error and need to be presmoothed. An interesting review of different ways of including smoothing in FDA methodologies can be seen in [14]. In this context P-splines were used for smoothing the sample curves and estimating different FDA models such as principal component analysis (PCA), functional logit regression, and functional partial least squares (PLSs), among others [15–18].

Alternative approaches for predicting spatiotemporal data are based on using different FDA methodologies for modeling a set of continuous-time curves with spatial dependence. On the one hand, classical geostatistical tools such as kriging were extended for this purpose in [19–23]. On the other hand, functional regression models with a functional response have been recently applied in [24]. The spatial information is introduced in terms of scalar covariates and considering a three dimensional P-spline penalty that combines the two-dimensional P-spline discrete penalty used for spatial regression [3, 25] with the continuous penalty (based on the second-order squared derivatives of the parameter functions) used for functional regression [14].

The rest of the chapter is organized as follows: Section 16.2 introduces a penalized approach for smoothing spatial data and the reparametrization of this approach into a mixed model is presented in Section 16.3. Section 16.4 gives a general framework for smoothing spatiotemporal data using an (analysis-of-variance) ANOVA-type decomposition. The benefits of this approach are shown in a small simulation study. P-spline functional spatial regression is introduced in Section 16.5. Finally, in Section 16.6, we analyze an air pollution dataset in Spain using both the functional spatial regression and spatiotemporal smoothing approach.

16.2 Smoothing Spatial Data via Penalized Regression

Suppose, for simplicity, that we observe a response variable, y_i, at a finite set of spatial locations $s_i = (u_i, v_i)$, $i = 1, \dots, n$, and y_i is normally distributed. There are many different approaches to smoothing and predicting spatial data: geostatistical models [26], Bayesian hierarchical models [27], or penalized regression [28], among others. Although they approach the smoothing problem from different perspectives, they are intimately related, for example, penalized splines can be interpreted as Bayes estimates with a suitable Gaussian process prior [29, 30], and spline fitting is well known to be a special case of kriging [31]. In this section, we focus on the use of penalized regression splines (this will help us to see immediately the links with spatial functional regression).

Penalized regression splines or P-splines [9, 32] are based on the use of a rich basis for regression and a penalty on the coefficients to control the smoothness of the fit. There are many possibilities for the choice of basis (*B*-splines, thin plate regression splines, etc.) and penalties (differenced or derivative based penalties), we will illustrate (without loss of generality) the methodology using *B*-spline basis and second-order difference penalties. The model proposed to capture the spatial dependence is

$$y_i = f(u_i, v_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \tag{16.1}$$

where the smooth function f can be expressed in terms of a number of basis functions. Some authors suggest the use of radial basis functions or thin plate splines, or a more computationally efficient version of thin-plate regression splines proposed by Wood [33]. These bases have the limitation of being isotropic smoothers and the selection of knots to construct the basis is not trivial. We follow the approach of [34], and use tensor product of *B*-spline basis with equally spaced knots. Although the e domain of the tensor product smooth is a rectangle or cuboid, it is often the case that the covariates only occupy part of the domain; in that case, a simple solution is to drop the basis functions that are to be evaluated at zero, and the corresponding components of the penalty. In the case of scattered data, the basis is constructed from the tensor product of marginal *B*-spline basis defined in [35] so that

$$f(u, v) = \sum_{k=1}^{q} \sum_{l=1}^{r} a_{kl} B_k^U(u) B_l^V(v) \tag{16.2}$$

where $\left\{B_k^U(u) : k = 1, \ldots, q\right\}$ and $\left\{B_l^V(v) : l = 1, \ldots, r\right\}$ are the marginal *B*-spline basis for each spatial coordinate. Let us denote by $\mathbf{B}^U$ the $n \times q$ matrix of values of the *B*-spline basis along u evaluated at the sample spatial locations u_i and by $\mathbf{B}^V$ the $n \times r$ matrix of values of the *B*-spline basis along v evaluated at the sample spatial coordinates v_i.

If unpenalized regression was used, then, the coefficients a_{kl} could be chosen by minimizing the least squared problem:

$$S = \sum_{i}^{n} (y_i - f(u_i, v_i))^2 = \sum_{i} \left(y_i - \sum_{k=1}^{q} \sum_{l=1}^{r} a_{kl} B_k^U(u_i) B_l^V(v_i) \right)^2. \tag{16.3}$$

In this case, the smoothness of the spatial surface is controlled by the number of *B*-spline basis in each dimension. As an alternative approach, it is possible to introduce a penalty that constraints coefficients that are next to each other to be similar. By construction, the domain of the tensor product smooth is a rectangle, and the coefficients a_{kl} are arranged in a matrix $\mathbf{A}$ of size $q \times r$, and so, we penalized the coefficients along the rows and columns of that matrix, i.e.

$$\text{PEN}(\mathbf{A}) = \lambda_u \sum_{l=1}^{r} \|\mathbf{P_u a_{.l}}\|^2 + \lambda_v \sum_{k=1}^{q} \|\mathbf{P_v a_{k.}}\|^2, \tag{16.4}$$

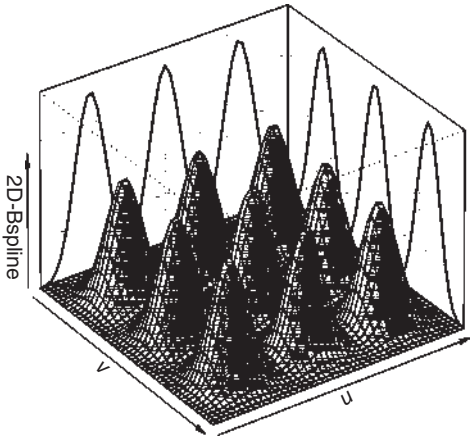


Figure 16.1 Portion of the *B*-spline basis (tensor product of nine cubic splines) in the case of data in a regular grid.

where $\mathbf{P_u}$ imposes a penalty on the columns of $\mathbf{A}$ ($\mathbf{a_1}$ corresponds to column l) and $\mathbf{P_v}$ imposes a penalty on each row of $\mathbf{A}$ ($\mathbf{a_k}$). One important feature of (16.4) is the fact that λ_u and λ_v can be different. This allows different amounts of smoothing along the two dimensions.

Using expression (16.2, model (16.1) can be expressed in matrix form as follows:

$$\mathbf{y} = f(\mathbf{u}, \mathbf{v}) + \boldsymbol{\epsilon} = \mathbf{Ba} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2),$$

where the "row-Tensor" product of two matrices denoted by symbol □ is defined as follows:

$$\mathbf{B} = \mathbf{B}^U \square\, \mathbf{B}^V = (\mathbf{B}^V \otimes \mathbf{1}_q') \odot (\mathbf{1}_r' \otimes \mathbf{B}^U)\ . \tag{16.5}$$

The basis $\mathbf{B}$ is of dimension $n \times qr$, the operator $\odot$ is the *Hadamard or "element-wise" matrix product*, and $\mathbf{1}_q$ and $\mathbf{1}_r$ are column vectors of ones of length q and r, respectively. In the case of data on a regular grid, the basis is calculated as the Kronecker product of the marginal basis, $\mathbf{B} = \mathbf{B}^U \otimes \mathbf{B}^V$. Figure 16.1 plots a portion of the basis functions in this case.

Then, the penalized least squares problem is written as follows:

$$\mathbf{S}(\mathbf{a}; \mathbf{y}, \lambda_u, \lambda_v) = (\mathbf{y} - \mathbf{Ba})'(\mathbf{y} - \mathbf{Ba}) + \mathbf{a}'\mathbf{Pa}, \tag{16.6}$$

where $\mathbf{P}$ is expressed as follows:

$$\mathbf{P} = \lambda_u \mathbf{I}_r \otimes \mathbf{P}_u + \lambda_v \mathbf{P}_v \otimes \mathbf{I}_q. \tag{16.7}$$

In particular, when $\mathbf{P}_u$ and $\mathbf{P}_v$ are based on second-order differences, i.e. $\mathbf{P}_u = \Delta'\Delta$ (and Δ a matrix that forms differences of order 2), the structure imposed by this penalty is such that each coefficient a_{kl} depends on the eight next neighboring coefficients along the coordinate axes (in the Bayesian approach, this would be the covariance structured of the coefficients). Of course, the dependence can be easily modified if differences are necessary along other directions.

Conditional on the values of λ_u and λ_v,

$$\hat{\mathbf{y}} = \hat{f}(\mathbf{u}, \mathbf{v}) = \mathbf{B}(\mathbf{B}'\mathbf{B} + \mathbf{P})^{-1}\mathbf{B}'\mathbf{y} = \mathbf{H}\mathbf{y}.$$

The matrix $\mathbf{H}$ is called the hat matrix, and it is very useful tool. It shows that the smoother is linear, and its trace gives a measure of the effective dimension of the model [36].

Optimization of smoothing parameters can be done using leave-one-out cross-validation, information criteria, etc. We choose an approach that takes advantage of the connections between penalized splines and mixed models. Details are given in Section 16.3.

16.3 Penalized Smooth Mixed Models

The connection between nonparametric regression and mixed models was established many years ago [37, 38], but it became popular much later [12, 25]. This approach has many advantages: (i) the smoothing parameter is estimated via maximum likelihood and (ii) it can deal easily with the identifiability problems that appear in models with more than one smooth term.

Smoothing penalties can also be viewed as resulting from improper Gaussian prior distributions on the spline coefficients, i.e. model (16.1) can be expressed as follows:

$$\mathbf{y} = \mathbf{B}\mathbf{a} + \boldsymbol{\epsilon}, \quad \mathbf{a} \sim N(\mathbf{0}, \mathbf{P}^{-}), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2), \tag{16.8}$$

since $\mathbf{P}^{-}$ is the Moore–Penrose pseudo-inverse of $\mathbf{P}$ (since penalties based on differences or derivatives are semidefinite positives, within the number of zero eigenvalues equal to the order of the difference/derivative). To avoid the improper distribution we propose a reparametrization of the model in which we separate the penalized and unpenalized coefficients yielding a mixed model. Our aim will be to reformulate model (16.1) (and therefore model (16.8)) as follows:

$$f(\mathbf{u}, \mathbf{v}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}, \text{ with } \boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G}),$$

where the basis and coefficients are reparameterized as follows:

$$\mathbf{B} \to [\mathbf{X} : \mathbf{Z}] \quad \text{and} \quad \mathbf{a} \to (\boldsymbol{\beta}, \boldsymbol{\alpha}).$$

The transformation is based on the singular value decomposition (SVD) of the penalty matrix given in (16.7) (which is a function of the SVD of the marginal penalties $\mathbf{P}_u$ and $\mathbf{P}_v$). Using a similar approach to [13], we find that the transformation $\mathbf{T}$ needed to reparameterized the model bases is

$$\mathbf{T} = [\mathbf{U}_{un}u \otimes \mathbf{U}_{vn} : \mathbf{U}_{us} \otimes \mathbf{U}_{vn} : \mathbf{U}_{un} \otimes \mathbf{U}_{vs} : \mathbf{U}_{us} \otimes \mathbf{U}_{vs}], \tag{16.9}$$

where $\mathbf{U}_{un}$ and $\mathbf{U}_{us}$ are the eigenvectors corresponding to the zero and nonzero eigenvalues of $\mathbf{P}_u$ (and similarly for $\mathbf{P}_v$). Then, the matrices of fixed and random effects are $\mathbf{BT} = [\mathbf{X} : \mathbf{Z}]$,

$$\mathbf{X} = \left(\mathbf{X}^U \square \mathbf{X}^V\right), \tag{16.10}$$

$$\mathbf{Z} = \left(\mathbf{Z}^U \square \mathbf{X}^V : \mathbf{X}^U \square \mathbf{Z}^V : \mathbf{Z}^U \square \mathbf{Z}^V\right), \tag{16.11}$$

where $\mathbf{Z}^U = \mathbf{B}^U \mathbf{U}_{su}$. Then, columns of $\mathbf{X}$ span the polynomial null space of $\mathbf{P}$ and the columns of $\mathbf{Z}$ span its complement. The covariance matrix of the random effects $\boldsymbol{\alpha}$ is diagonal with elements:

$$\mathbf{G} = \begin{pmatrix} \lambda_u \tilde{\boldsymbol{\Sigma}}_U \otimes \mathbf{I}_d & & \\ & \lambda_v \mathbf{I}_d \otimes \tilde{\boldsymbol{\Sigma}}_V & \\ & & \lambda_u \tilde{\boldsymbol{\Sigma}}_U \otimes \mathbf{I}_{r-d} + \lambda_v \mathbf{I}_{q-d} \otimes \tilde{\boldsymbol{\Sigma}}_V, \end{pmatrix}^{-1}, \tag{16.12}$$

$\tilde{\boldsymbol{\Sigma}}_U$ and $\tilde{\boldsymbol{\Sigma}}_V$ are the nonzero eigenvalues of the marginal penalty matrices, $\mathbf{I}$ is an identity matrix, and d is the dimension of the null space of $\mathbf{P}$. This partition allows the representation of the fitted surface in terms of the sum of three components: one for $\mathbf{u}$ (latitude), one for $\mathbf{v}$ (longitude), and an interaction component which depends on both geographical components simultaneously.

The estimates of the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ follow from standard mixed model theory [39],

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \tag{16.13}$$

$$\hat{\boldsymbol{\alpha}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{16.14}$$

where $\mathbf{V} = \sigma^2\mathbf{I} + \mathbf{Z}\mathbf{G}\mathbf{Z}'$. In the mixed model setting, smoothing parameters λ_u and λ_u become the ratio of variances, therefore, they may be estimated by maximizing the residual log-likelihood (REML) of [40]:

$$\begin{aligned} \ell(\lambda_1, \lambda_2, \sigma^2) = &-\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| \\ &-\frac{1}{2}\mathbf{y}'(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{y}. \end{aligned} \tag{16.15}$$

Recently, Rodriguez-Alvarez et al. [41] presented a fast algorithm for the estimation of smoothing parameter in the context of multidimensional smooth mixed models.

The definition of matrix $\mathbf{Z}$ and covariance matrix $\mathbf{G}$ suggests that the smooth function $f(\mathbf{u}, \mathbf{v})$ accounting for the spatial structure in the data can be decomposed as the sum of three components: one latitude, on for longitude and another for the interaction between them. This decomposition has inspired a new class of smooth models called *P-spline ANOVA models* that have an immediate application in the case of spatial and spatiotemporal data.

16.4 P-spline Smooth ANOVA Models for Spatial and Spatiotemporal data

Sometimes, fitting a multidimensional smooth model of the form $f(\mathbf{x}_1, \dots, \mathbf{x}_k)$ can be restrictive. For example, in the case of spatiotemporal data, using the model

$$\mathbf{y} = f_{st}(\mathbf{u}, \mathbf{v}, \mathbf{t}) + \boldsymbol{\epsilon}\ , \quad \boldsymbol{\epsilon} \sim \mathcal{N}\ (\mathbf{0}, \sigma^2\mathbf{I}) \tag{16.16}$$

can lead to a poor fitting. For example, if there is a strong additive effect of *time*, but the interaction with the geographical location is relatively small, fitting model (16.16) using multidimensional P-splines will impose an interaction model and not an additive one. In order to accommodate all possible settings, we propose the use of the following model as a general approach for the smoothing of space-time data:

$$E[\mathbf{y}] = \gamma + \sum_{i=1}^{k} f_i(\mathbf{x}_i) + \sum_{i<j} f_{ij}(\mathbf{x}_i, \mathbf{x}_j) + \cdots + f_{1,\dots,k}(\mathbf{x}_1, \dots, \mathbf{x}_k), \tag{16.17}$$

where γ is a constant term, f_i are additive univariate functions of the ith covariate, f_{ij} a two-dimensional interaction smooth function of the pair of covariates $(\mathbf{x}_i, \mathbf{x}_j)$, and so on, until a kth order interaction. These types of models can be seen as a functional version of *ANOVA*. Using this terminology, model (16.17) is the sum of smooth functions of *main effects* and *two-way interactions*, *three-way interactions*, and so on. These models have been considered in the literature in the context of Smoothing Splines, as SS-ANOVA models [42, 43]. Since, main effects are *contained* in the higher-order interactions, it is necessary to impose constraints in order to make the model identifiable. This may be complicated and computationally expensive when there are higher-order interactions. As an alternative, we propose a Low-rank S-ANOVA model. P-splines use low-rank bases functions, and so, they are computationally less demanding than other approaches.

In the case of spatiotemporal data, a model such as (16.17) might not be realistic. In general, we will be interested in the spatial effect, the temporal effect, and the interaction between them. Expressing the model in this way, we are, implicitly, giving more flexibility to the spatiospatial structure in the model, and we can gain insight on process behind our data (for example, we can test whether space and time are separable). Therefore, we will consider a *reduced* version of model (16.17) (see [13] for a full description)

$$\mathbf{y} = \gamma + f_s(\mathbf{u}, \mathbf{v}) + f_t(\mathbf{t}) + f_{st}(\mathbf{u}, \mathbf{v}, \mathbf{t}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\ (\mathbf{0}, \sigma^2\mathbf{I}). \tag{16.18}$$

f_s represents the spatial structure common along time, f_t is the common temporal pattern shared by all locations, and f_{st} would account for departures from this overall functions across space and along time. Each of these functions is expressed

again in terms of basis functions and coefficients. The *B*-spline regression basis for this model would be

$$\mathbf{B} = [\mathbf{1}_{nt} : \mathbf{B}_s \otimes \mathbf{1}_t : \mathbf{1}_n \otimes \mathbf{B}_t : \mathbf{B}_s \otimes \mathbf{B}_t], \tag{16.19}$$

where $\mathbf{B}_s$ is the two-dimensional basis defined in (16.5), $\mathbf{B}_t$ is the *B*-spline basis for the time effect and $\mathbf{B}_s \otimes \mathbf{B}_t$ is the basis for the interaction. If data are collected at the same time points for all locations, the 3D-basis is constructed using the Kronecker product, if time points are different, the box-product is used instead.

The vector of regression coefficients is $\mathbf{a} = (\gamma, \mathbf{a}^{(s)\prime}, \mathbf{a}^{(t)\prime}, \mathbf{a}^{(st)\prime})'$ and the penalty matrix is block-diagonal with penalties over $\mathbf{a}$ of the form

$$\mathbf{P} = \text{blockdiag}(0, \mathbf{P}_{(s)}, \mathbf{P}_{(t)}, \mathbf{P}_{(st)}), \tag{16.20}$$

where $\mathbf{P}_{(s)}$ is the two-dimensional penalty matrix for the spatial component, with smoothing parameters λ_u and λ_v as in (16.7), i.e.

$$\mathbf{P}_{(s)} = \lambda_u \mathbf{P}_u \otimes \mathbf{I}_r + \lambda_v \mathbf{I}_q \otimes \mathbf{P}_v, \tag{16.21}$$

$\mathbf{P}_{(t)}$ is the one-dimensional penalty matrix for the time component, with smoothing parameter λ_t, and $\mathbf{P}_{(st)}$ is the three-dimensional penalty matrix for the spatiotemporal component with smoothing parameters τ_u, τ_v, and τ_t:

$$\mathbf{P}_{(st)} = \tau_u \mathbf{I}_t \otimes \mathbf{P}_u \otimes \mathbf{I}_r + \tau_v u \mathbf{I}_t \otimes \mathbf{I}_q \otimes \mathbf{P}_v + \tau_t \mathbf{P}_t \otimes \mathbf{I}_r \otimes \mathbf{I}_q. \tag{16.22}$$

The *B*-spline model matrix for this model is not of full rank since the space spanned by $\mathbf{B}_t$ is contained in the space spanned by $\mathbf{B}_s \otimes \mathbf{B}_t$, and therefore, we encounter the identifiability problems mentioned above. Several approaches have been taken to overcome this problem: (i) add a ridge penalty [44] or (ii) identify and impose the constraints numerically [10]. However, the first alternative may induce to numerical problems, and the second method is difficult to extend in the case of more than two-way interactions. We use here a simpler and more efficient approach based on removing the linearly dependent columns of the basis (identifying the columns to be removed is immediate when the mixed model representation is used). We adapt the transformation given (16.9) to the spatiotemporal model and find that the mixed model matrices are

$$\begin{aligned} \mathbf{X} &= [\overbrace{\mathbf{1}_t \otimes \mathbf{x}_s}^{f_s(\mathbf{u},\mathbf{v})} : \overbrace{\mathbf{t} \otimes \mathbf{1}_n}^{f_t(\mathbf{t})} : \overbrace{\mathbf{t} \otimes \check{\mathbf{x}}}^{f_{s,t}(\mathbf{u},\mathbf{v},\mathbf{t})}] \\ \mathbf{Z} &= [\mathbf{1}_t \otimes \mathbf{Z}_s : \mathbf{Z}_t \otimes \mathbf{1}_n : \mathbf{t} \otimes \mathbf{Z}_s : \mathbf{Z}_t \otimes \check{\mathbf{X}}_s : \mathbf{Z}_t \otimes \mathbf{Z}_s], \end{aligned} \tag{16.23}$$

where $\check{\mathbf{x}} = (\mathbf{u} : \mathbf{v} : \mathbf{x}_s)$, $\mathbf{x}_s = \mathbf{v} \square \mathbf{u}$, and covariance of the random effects is given by

$$\mathbf{G} = \text{blockdiag}\,(\mathbf{F}_{(s)}, \mathbf{F}_{(t)}, \mathbf{F}_{(s,t)})^{-1} \tag{16.24}$$

with blocks

$$\mathbf{F}_{(s)} = \begin{pmatrix} \lambda_u \tilde{\mathbf{\Sigma}}_u \otimes \mathbf{I}_d & & \\ & \lambda_v \mathbf{I}_d \otimes \tilde{\mathbf{\Sigma}}_v & \\ & & \lambda_u \tilde{\mathbf{\Sigma}}_u \otimes \mathbf{I}_{r-d} + \lambda_v \mathbf{I}_{q-d} \otimes \tilde{\mathbf{\Sigma}}_v \end{pmatrix},$$

$$\mathbf{F}_{(t)} = \lambda_t \tilde{\mathbf{\Sigma}}_t \ ,$$

$$\mathbf{F}_{(s,t)} = \text{blockdiag}\left(\mathbf{F}^{(1)}_{(s,t)}, \mathbf{F}^{(2)}_{(s,t)}, \mathbf{F}^{(3)}_{(s,t)}\right).$$

where

$$\mathbf{F}^{(1)}_{(s,t)} = \begin{pmatrix} \tau_u \tilde{\mathbf{\Sigma}}_u \otimes \mathbf{I}_d & & \\ & \tau_v \mathbf{I}_d \otimes \tilde{\mathbf{\Sigma}}_v & \\ & & \tau_u \tilde{\mathbf{\Sigma}}_u \otimes \mathbf{I}_{r-2} + \tau_2 \mathbf{I}_{q-2} \otimes \tilde{\mathbf{\Sigma}}_v \end{pmatrix},$$

$$\mathbf{F}^{(2)}_{(s,t)} = \begin{pmatrix} \tau_t \tilde{\mathbf{\Sigma}}_t \otimes \mathbf{I}_{\mathbf{d}} & & \\ & \tau_u \mathbf{I}_{p-d} \otimes \tilde{\mathbf{\Sigma}}_u + \tau_t \tilde{\mathbf{\Sigma}}_t \otimes \mathbf{I}_{q-d} & \\ & & \tau_v \mathbf{I}_{p-d} \otimes \tilde{\mathbf{\Sigma}}_v + \tau_t \tilde{\mathbf{\Sigma}}_t \otimes \mathbf{I}_{r-d} \end{pmatrix},$$

$$\mathbf{F}^{(3)}_{(s,t)} = \tau_u \mathbf{I}_{p-d} \otimes \tilde{\mathbf{\Sigma}}_u \otimes \mathbf{I}_{r-d} + \tau_v \mathbf{I}_{p-d} \otimes \mathbf{I}_{q-d} \otimes \tilde{\mathbf{\Sigma}}_2 + \tau_t \tilde{\mathbf{\Sigma}}_t \otimes \mathbf{I}_{q-d} \otimes \mathbf{I}_{r-d} \ .$$

Again, estimation of fixed effects coefficients, prediction of random effects, and estimation of smoothing parameters can be done by using standard mixed models methodology. However, the size of the data sets in the spatiotemporal context, makes difficult the use of standard software. We overcome this problem by using the Generalized Linear Array Models (GLAM) algorithms developed by Currie et al. [34] to calculate (16.13), (16.14), and (16.15) (see [13] for details).

16.4.1 Simulation Study

We undertake a small simulation study to show that an ANOVA-type model is preferable to an additive model, or a pure interaction model (which is a common spatiotemporal model with nonseparable covariance structure). For simplicity, we restrict our simulation to the 2D-case and generate data from three possible models:

$$\boldsymbol{\eta}^{(1)} = f_1(\mathbf{x}_1) + f_1(\mathbf{x}_2), \qquad \text{("Two main effects model")}$$

$$\boldsymbol{\eta}^{(2)} = f_{1,2}(\mathbf{x}_1, \mathbf{x}_2), \text{and} \qquad \text{("Interaction model")}$$

$$\boldsymbol{\eta}^{(3)} = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + f_{1,2}(\mathbf{x}_1, \mathbf{x}_2) \qquad \text{("Two main effects and interaction")}$$

and,

$$f_1(\mathbf{x}_1) = \sin(2\pi \mathbf{x}_1),$$

$$f_2(\mathbf{x}_2) = \cos(3\pi \mathbf{x}_2), \text{and}$$

$$f_{1,2}(\mathbf{x}_1, \mathbf{x}_2) = 3 \ \sin(2\pi \mathbf{x}_1)\ (2\mathbf{x}_2 - 1).$$

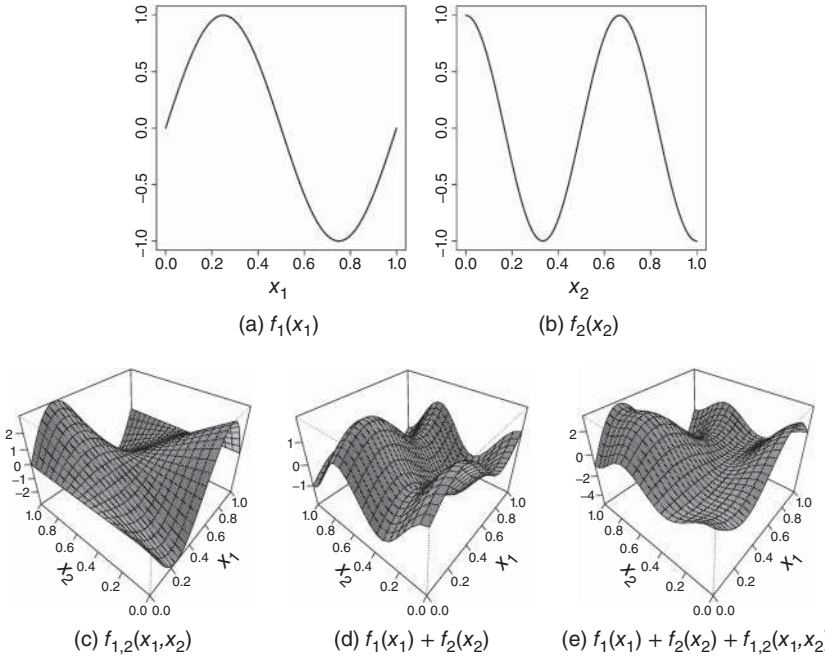


Figure 16.2 Simulated functions: (a) and (b) are the nonlinear main effects of x_1 and x_2; (c) is the additive surface of main effects; (d) is interaction surface; and (e) is the sum of the main effects and the interaction surfaces.

We consider the case of data on a regular grid, the covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ chosen in $[0, 1]$ with dimensions $n_1 = 30$ and $n_2 = 20$, respectively. Figure 16.2 shows the simulated true smooth functions and true surfaces for the proposed scenarios.

Two hundred replicates of three smooth mixed models (*additive*, *anova* and *interaction* models) were fitted for each scenario, with a combination of $\sigma = 0.25$, $\sigma = 0.5$, and $\sigma = 1$. Marginal B-splines bases $\mathbf{B}_1$ and $\mathbf{B}_2$ were calculated with 8 and 6 knots respectively, with cubic splines. Second-order marginal penalties were used in the fitting procedure, and smoothing parameters were chosen by minimizing by REML. To check each model's performance, we computed the mean square error (MSE) for each replicate. Figure 16.3 shows the box-plots of the log(MSE) values for fitted smooth models. The gray shaded box-plot corresponds to the model from which we have simulated each scenario (i.e. in scenario 1, we consider $\boldsymbol{\eta}^{(1)}$ as a function of two main effects, and thus the *additive* model is the favored model). S-ANOVA model clearly gave better results in scenarios 2 and 3 (interaction model and additive plus interaction model). Additive model performed slightly better than the S-ANOVA model in scenario 1. In this case, the S-ANOVA model reduces to an additive model when the smoothing parameters

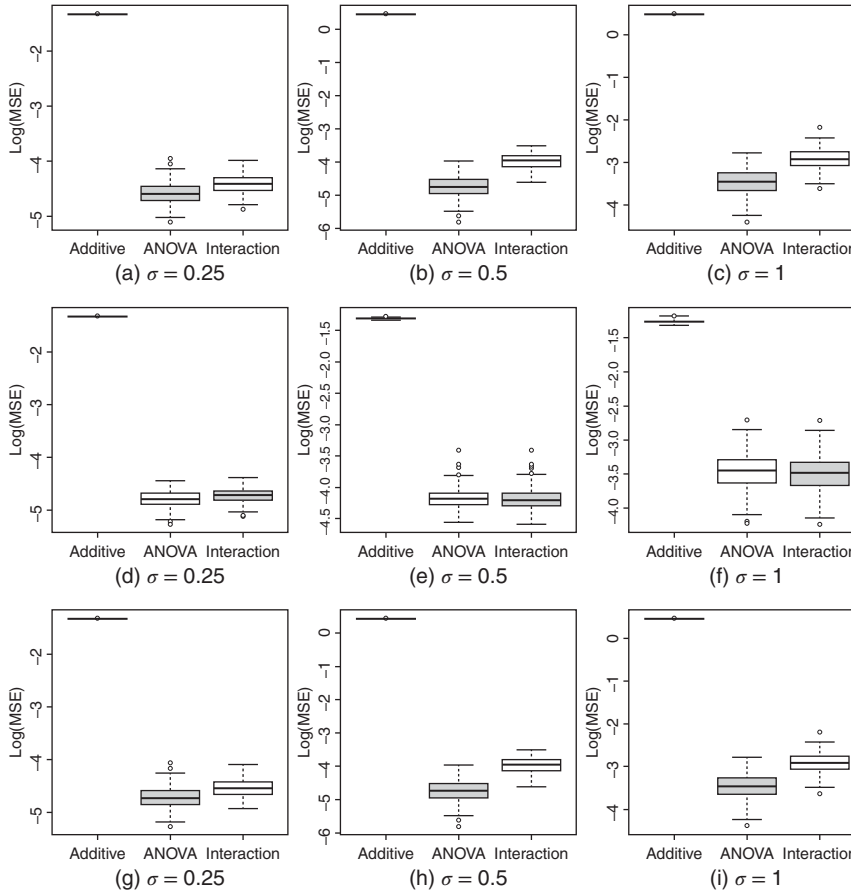


Figure 16.3 `log`(MSE) of fitted smooth model for $R = 200$: scenario 1 (a–c), scenario 2 (d–f), and scenario 3 (g–i).

in the interaction $(\tau_1, \tau_2) \to \infty$. The poor performance of the S-ANOVA model in some replicates might be due to numerical problems, since we have considered an upper bound for the smoothing parameters equal to 10^6.

16.5 P-spline Functional Spatial Regression

An alternative method for modeling and predicting spatiotemporal data is using a FDA based approach as the one developed in [24]. In this case, we have a sample of spatially correlated sample curves $\{y_i(t) : t \in T, i = 1, \ldots, n\}$ which have been observed with error at a finite set of time points $\{t_j : j = 1, \ldots, m\}$ for each

geographical location s_i. Then, the data can be seen as realizations of a spatial functional variable (spatiotemporal stochastic process)

$$\left\{X(s,t) : s \in S \subseteq \mathbb{R}^2, t \in T \subseteq \mathbb{R}\right\}, \tag{16.25}$$

where $s = (u, v)$ is a generic data location in the spatial domain $S = U \times V$ and U, V, and T are real intervals.

Let us assume that the realizations of the functional variable X are square integrable functions in the spatiotemporal domain and belong to the pqr-dimensional tensor function space generated by the three univariate basis of B-splines, so that

$$x(s,t) = \sum_{k=1}^{q}\sum_{l=1}^{r}\sum_{h=1}^{p} a_{klh}\mathbf{B}_k^U(u)\mathbf{B}_l^V(v)\mathbf{B}_h^T(t). \tag{16.26}$$

This means that for all spatial locations, the associated sample curves belong to the finite-dimension space generated by the basis $\left\{\mathbf{B}_h^T : h = 1, \ldots, p\right\}$, so that they admit the basis expansion:

$$x(s,t) = \sum_{h=1}^{p} a_h(s)\mathbf{B}_h^T,$$

where the basis coefficients are realizations of a multivariate spatial process given by

$$a_h(s) = \sum_{k=1}^{q}\sum_{l=1}^{r} a_{klh}\mathbf{B}_k^U(u)\mathbf{B}_l^V(v).$$

For each time point, the associated sample surfaces belong to the tensor function space generated by the basis $\left\{\mathbf{B}_k^U\mathbf{B}_l^V : k = 1, \ldots, q; l = 1, \ldots, r\right\}$ so that can be expressed as

$$x(.,t) = \sum_{k=1}^{q}\sum_{l=1}^{r} a_{kl}(t)\mathbf{B}_k^U\mathbf{B}_l^V,$$

where the basis coefficients are realizations of a multivariate stochastic process given by

$$a_{kl}(t) = \sum_{h=1}^{p} a_{klh}\mathbf{B}_h^T(t).$$

Once the basis coefficients in Eq. (16.26) are estimated from the discrete observation y_{ij}, the spatiotemporal functional variable can be estimated at unobserved locations and times (s_0, t_0) by replacing in such equation. This way we can predict the curve of temporal evolution of the variable across the temporal domain for not sampled geographical locations and the surface of spatial evolution of the variable across the spatial domain for any time point in the temporal domain.

The basis coefficients in Eq. (16.26) can be estimated by introducing the spatial variability through the following functional spatial regression model [24]:

$$\mathbf{y}(t) = \mathbf{B}_s\alpha(t) + \epsilon(t), \quad \forall t \in T, \tag{16.27}$$

where $\mathbf{y}(t) = (y_1(t), \dots, y_n(t))'$ is the vector of response functions, $\mathbf{B}_s$ is the two-dimensional B-spline basis for the geographical position described in Section 16.2, $\alpha(t) = (\alpha_1(t), \dots, \alpha_{qr}(t))'$ is the vector of parameter functions to be estimated and $\epsilon(t) = (\epsilon_1(t), \dots, \epsilon_n(t))'$ the vector of error terms.

Let us assume a basis representation for the functional response $\mathbf{y}(t) = \mathbf{C}B^T(t)$, and a basis representation for the the functional parameter $\alpha(t) = \mathbf{A}B^T(t)$ with $C = (c_{ih})_{n\times p}$ and $A = (a_{(kl)h})_{qr\times p}$ being the corresponding matrices of basis coefficients and $B^T(t) = (B_1^T(t), \dots, B_p^T(t))'$ being the vector of basis functions. Then, the model becomes

$$CB^T(t) = \mathbf{B}_s\mathbf{A}B^T(t) + \epsilon(t), \quad \forall t \in T.$$

The matrix of parameters is estimated by penalized sum of squares, where we have separated regularization for space and time, furthermore, we use a nonisotropic penalty term for space to allow more flexibility, i.e.

$$\begin{aligned}\text{PSSE}(y, \alpha) = &\int (\mathbf{C}B^T(t) - \mathbf{B}_s\mathbf{A}\mathbf{B}^T(t))'(\mathbf{C}B^T(t) - \mathbf{B}_s\mathbf{A}\mathbf{B}^T(t))dt \\ &+ \text{vec}\,(A)'[\text{PEN}^{U,V,T}]\text{vec}(A),\end{aligned} \tag{16.28}$$

where $\text{PEN}^{U,V,T}$ is defined as in (16.22). Interchanging the integration and summation operations implied by the matrix products, calculating the derivatives with respect to $\mathbf{A}$, and using some properties of the Kronecker product, we obtain

$$\text{vec}(\mathbf{A}) = \left[\Psi \otimes (\mathbf{B}_s'\mathbf{B}_s) + \text{PEN}^{U,V,T}\right]^{-1} \text{vec}(\mathbf{B}_s'\mathbf{C}\Psi'),$$

where $\Psi = \int B^T B^T$ is the inner product matrix between the basis functions in the temporal domain.

16.6 Application to Air Pollution Data

In this section, we illustrate both penalized approaches (spatiotemporal smoothing and functional regression) with an application of air pollution data. The data set consists medians over the years 2005 to 2012 of daily ozone levels (O_3) at the 55 monitoring stations in Spain and Portugal. The raw data set together with the map with the geographical locations are shown in Figure 16.4.

The data can be obtained from the `R` package `openair` available at comprehensive R archive network (CRAN). The Openair project is an initiative of the Natural Environment Research Council (NERC) that aims to provide a collection

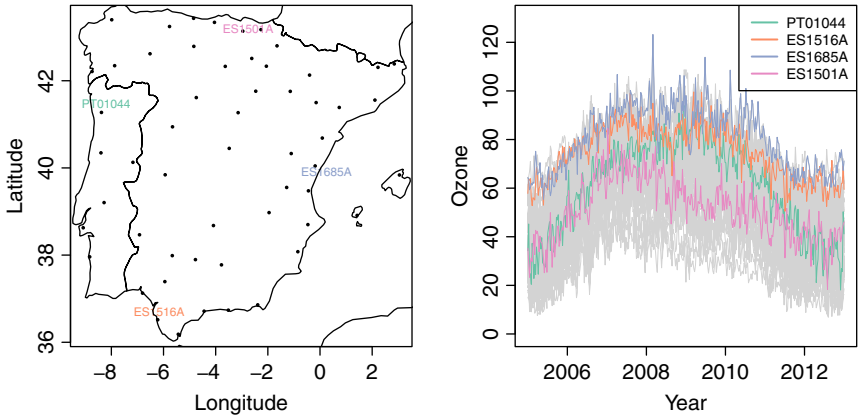


Figure 16.4 Medians of daily ozone curves (from 2002 to 2015) observed at 55 site in Spain and Portugal. Four locations are highlighted.

of open-source tools for the analysis of air pollution data (more details can be found at http://www.openair-project.org/).

16.6.1 Spatiotemporal Smoothing

The P-spline ANOVA model in (16.18) was fitted with 10 basis functions for each longitude and latitude covariates to construct the spatial main effect basis $\mathbf{B}_s$, 18 basis functions for the temporal main effect basis $\mathbf{B}_t$, and the Kronecker product of both matrices for the basis of the interaction effect.

The model is fitted using the mixed model formulation in (16.23) and REML for the estimation of the variance components. The smooth effects of space and time (i.e. $\hat{f}_s$ and $\hat{f}_t$ including the constant terms $\hat{\gamma}$) are shown in Figure 16.5 and represents the main spatial and temporal effects of the ANOVA decomposition. Figure 16.6 shows the space–time interaction estimated by the ANOVA model for four selected days in a year. We can clearly see that the spatial trend is not constant along the days, and it is quite different from the overall spatial trend shown in Figure 16.5, showing the need for a space–time interaction. Approximate F-test also concluded that the interaction term was significant in the model. Finally, Figure 16.7 shows the fitted curves at four selected locations (we will compare them later with the results from the spatial functional approach).

16.6.2 Spatial Functional Regression

To apply the penalized functional spatial regression model (PFSRM) (16.27), we start by constructing a cubic B-spline representation of the curves in terms of 18

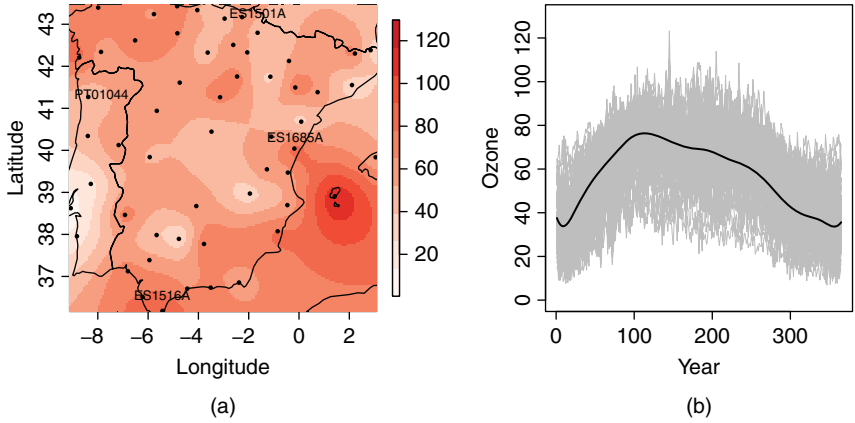


Figure 16.5 Smoothed spatial and temporal main effects for the ANOVA model. (a) Spatial main effect. (b) Temporal main effect.

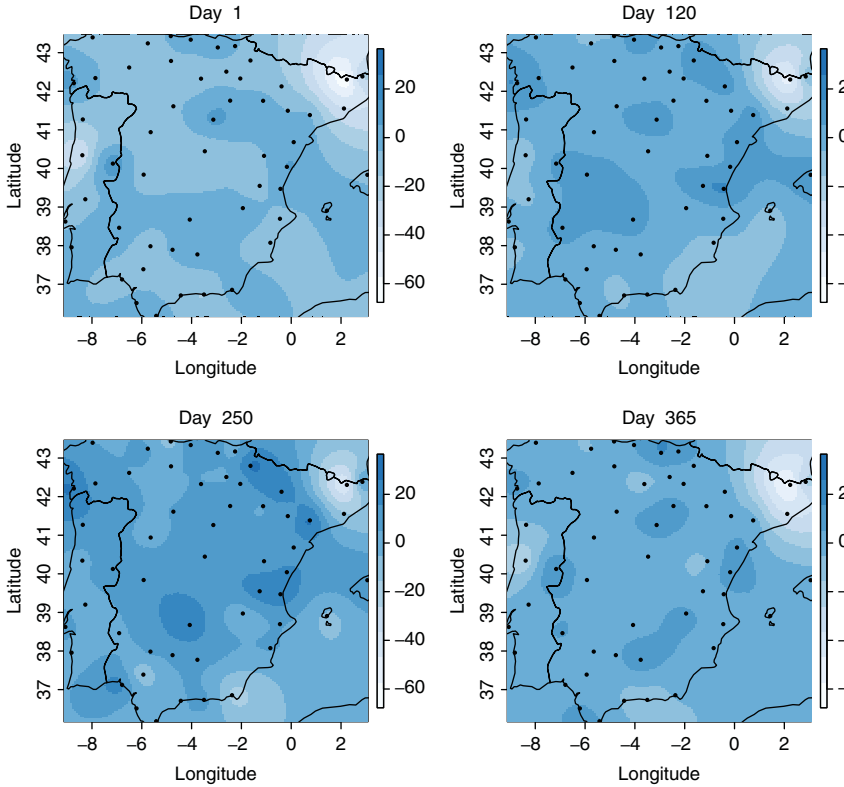


Figure 16.6 Smoothed spatiotemporal interaction for ANOVA model at four selected days.

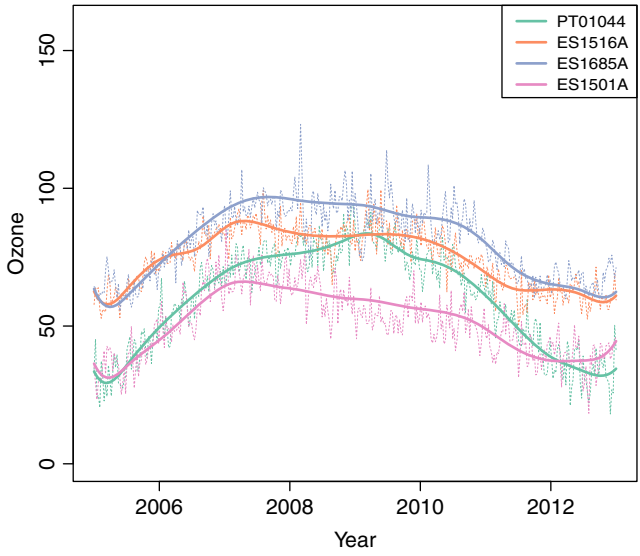


Figure 16.7 Smoothed spatiotemporal fit for ANOVA model at four selected locations.

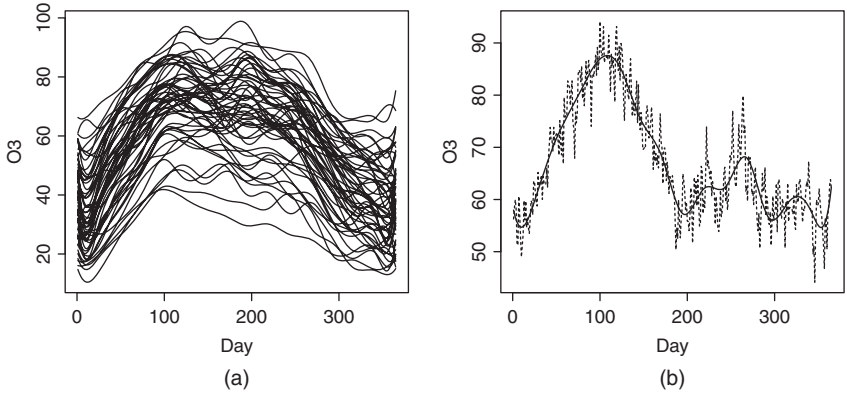


Figure 16.8 Regression splines fitted from the ozone raw data by using a cubic *B*-spline basis with dimension 18 (a). A sample path (dashed line) together with its basis representation (solid line) using 18 *B*-spline basis functions (b).

basis functions. The regression splines fitted this way can be seen in Figure 16.8. The PFSRM model is then estimated by using the Kronecker sum of three second-order P-spline penalties (two for space and one for time) and marginal temporal and spatial basis of dimension 365×18 and 55×100, respectively. The smoothing parameters were selected by generalized cross-validation (see [24] for details).

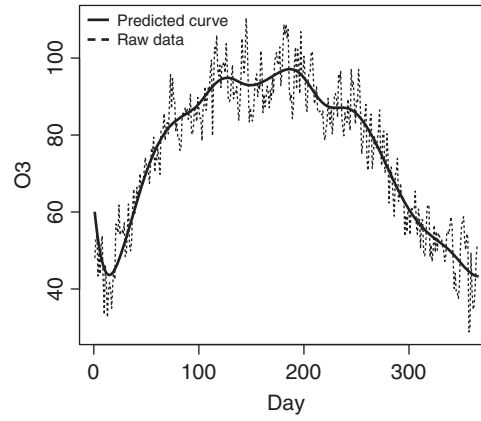


Figure 16.9 Predicted curve from the regression splines of the ozone raw data using 18 cubic *B*-spline basis functions (solid line) and the observed raw data (solid line) in one site.

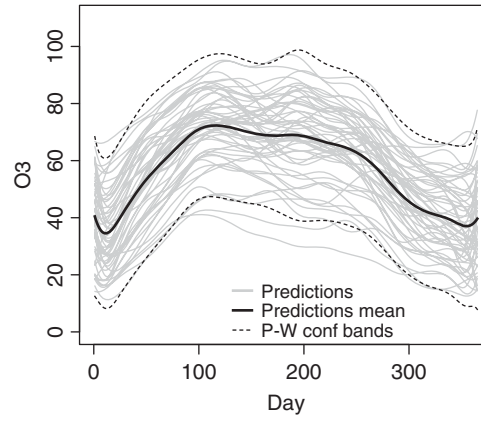


Figure 16.10 Predicted curves (gray) from the regression splines of the ozone raw data using 18 cubic *B*-spline basis functions) join to its mean curve (black and solid line) and the pointwise confidence bands according to the mean ± two times the standard deviation (black and dashed line).

The predicted curves for each of the considered sites were computed by leave-one-out cross-validation. Figure 16.10 displays the predicted curves provided by PFSRM next to their mean curve and point wise confidence bands (computed as the mean ± two times the standard deviation). An example of predicted curve for one site superposed with its raw data is displayed in Figure 16.9.

Finally, we compare the fit of both approaches for four selected locations in Figure 16.11. The curves are quite similar although some discrepancies appear at the beginning and the end of the year.

Inspired by model (16.17), a more flexible functional approach would be

$$\mathbf{y}(t) = \gamma + f(\mathbf{u}, \mathbf{v}) + \boldsymbol{\alpha}(t) + \mathbf{B}_s\beta(t) + \epsilon(t), \quad \forall t \in T, \tag{16.29}$$

and so model (16.27) becomes

$$\mathbf{C}\boldsymbol{\theta}(t) = \mathbf{B}\mathbf{a} + \boldsymbol{\epsilon}(t), \quad \forall t \in T,$$

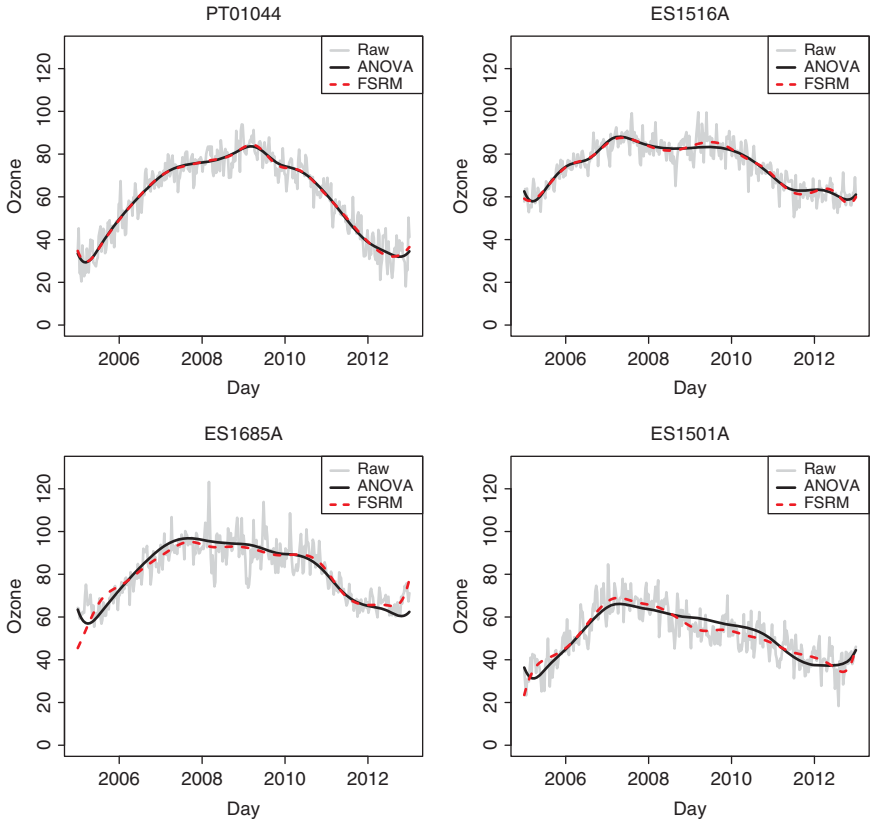


Figure 16.11 Smoothed spatiotemporal fit for ANOVA model at four selected locations.

where $\mathbf{B}$ is given in (16.19), and the vector of regression coefficients is $\mathbf{a} = (\gamma, \mathbf{a}^{(s)\prime}, \mathbf{a}^{(t)\prime}, \mathbf{a}^{(st)\prime})'$. Each component of the vector would be penalized separately as described in (16.20). In order to properly identify the terms in the model, constraints need to be imposed. A possible approach is to constrain the coefficients in the model as follows (see [13]) for details)

$$\sum_h \mathbf{a}_h^{(t)} = \sum_k \mathbf{a}_{kl}^{(s)} = \sum_l \mathbf{a}_{kl}^{(s)} = 0,$$
$$\sum_h \mathbf{a}_{klh}^{(st)} = \sum_k \mathbf{a}_{klh}^{(st)} = \sum_l \mathbf{a}_{klh}^{(st)} = 0.$$

Acknowledgments

This research has been funded by projects MTM2013-47929-P and MTM2014-52184-P from Secretaría de Estado Investigación, Desarrollo e Innovación, Ministerio de Economía y Competitividad, Spain, co-financed by European Regional Development Fund (ERDF).

References

1 Kammann, E. and Wand, M. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52: 1–18.
2 Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70: 209–226.
3 Lee, D.J. and Durbán, M. (2009). Smooth-car mixed models for spatial count data. *Computational Statistics and Data Analysis* 53 (8): 2968–2979.
4 Huang, H.C. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using Kalman filter. *Computational Statistics and Data Analysis* 22: 159–175.
5 Wikle, C.K. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika* 86: 815–829.
6 Gössl, C., Auer, D.P., and Fahrmeir, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics* 57: 554–562.
7 Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, *Monographs on Statistics and Applied Probability*. Chapman and Hall.
8 Lee, D.J. and Durbán, M. (2011). P-spline ANOVA type interaction models for spatio-temporal smoothing. *Statistical Modelling* 11: 49–69.
9 Eilers, P. and Marx, B. (1996). Flexible smoothing with *B*-splines and penalties. *Statistical Science* 11: 89–121.
10 Wood, S. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* 62: 1025–1036.
11 Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* 11: 735–757.
12 Currie, I.D. and Durbán, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling* 2: 333–349.

13 Lee, D.J. (2010). Smoothing mixed models for spatial and spatio-temporal data. PhD thesis. Universidad Carlos III de Madrid.

14 Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2e. Springer-Verlag.

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

16 Aguilera, A.M. and Aguilera-Morillo, M.C. (2013). Penalized PCA approaches for B-spline expansions of smooth functional data. *Applied Mathematics and Computation* 219 (14): 7805–7819.

17 Aguilera-Morillo, M.C., Aguilera, A.M., Escabias, M., and Valderrama, M.J. (2013). Penalized spline approaches for functional logit regression. *Test* 22 (2): 251–277.

18 Aguilera, A.M., Aguilera-Morillo, M.C., and Preda, C. (2016). Penalized versions of functional PLS regression. *Chemometrics and Intelligent Laboratory Systems* 154: 80–92.

19 Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2009). Statistics for spatial functional data: some recent contributions. *Environmetrics* 21: 224–239.

20 Giraldo, R., Delicado, P., and Mateu, J. (2010). Continuous time-varying kriging for spatial prediction of functional data: an environmental application. *Journal of Agricultural, Biological, and Environmental Statistics* 15 (1): 66–82.

21 Giraldo, R., Delicado, P., and Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics* 18 (3): 411–426.

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

23 Menafoglio, A., Secchi, P., and Dalla Rosa, M. (2013). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics* 7: 2209–2240.

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

25 Durbán, M. and Currie, I. (2003). A note on P-spline additive models with correlated errors. *Computational Statistics* 18: 251–262.

26 Diggle, P. and Ribeiro, P. (2007). *Model-Based Geostatistics*, *Springer Series in Statistics*. New York: Springer.

27 Banerjee, S., Carlin, B., and Gelfand, A. (2005). *Hierarchical Modeling and Analysis for Spatial Data*, *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

28 Farhmeir, L. and Kneib, T. (2011). *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*, *Oxford Statistical Science Series*. New York: Oxford University Press.

29 Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*, *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

30 Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society: Series B (Methodological)* 40: 364–372.

31 Dubrule, O. (1994). Comparing splines and kriging. *Computers & Geosciences* 10 (2–3): 327–338.

32 Eilers, P., Marx, B., and Durban, M. (2015). Twenty years of P-splines. *SORT (Statistics and Operations Research Transactions)* 39: 149–186.

33 Wood, N. (2003). Thin plate splines regression. *Journal of the Royal Statistical Society* 65: 95–114.

34 Currie, I.D., Durban, M., and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68: 1–22.

35 Eilers, P., Currie, I., and Durban, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis* 50: 61–76.

36 Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

37 Green, P.J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique* 55: 245–259.

38 Speed, T. (1991). Comment on: that BLUP is a good thing: the estimation of random effects. *Statistical Science* 6: 15–51.

39 Searle, S., Casella, G., and McCulloch, C. (1992). *Variance Components*, *Wiley Series in Probability and Mathematical Statistics*. Wiley.

40 Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58: 545–554.

41 Rodriguez-Alvarez, M.X., Lee, D., Kneib, T. et al. (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing* 25: 941–957.

42 Chen, Z. (1993). Fitting multivariate regression functions by interaction spline models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 55: 473–491.

43 Cu, C. (2013). *Smoothing Spline Anova Models*, *Springer Series in Statistics*. New York: Springer.

44 Eilers, P. and Marx, B. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis* 28: 193–209.

Index

Geostatistical Functional Data Analysis, First Edition. Edited by Jorge Mateu and Ramón Giraldo.