Utah State University

DigitalCommons@USU

All Graduate Theses and Dissertations

Graduate Studies

5-2013

Spatially Indexed Functional Data

Oleksandr Gromenko
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/etd

Part of the Environmental Sciences Commons, and the Statistics and Probability Commons

Recommended Citation

Gromenko, Oleksandr, "Spatially Indexed Functional Data" (2013). *All Graduate Theses and Dissertations*. 1526.
https://digitalcommons.usu.edu/etd/1526

This Dissertation is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.

UtahStateUniversity
MERRILL-CAZIER LIBRARY

SPATIALLY INDEXED FUNCTIONAL DATA

by

Oleksandr Gromenko

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Mathematical and Statistical Sciences

Approved:

Dr. Piotr S. Kokoszka Major Professor	Dr. Jan J. Sojka Committee Member
Dr. Daniel C. Coster Committee Member	Dr. Jürgen Symanzik Committee Member
Dr. Lie Zhu Committee Member	Dr. Mark R. McLellan Vice President for Research and Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2013

Copyright © Oleksandr Gromenko 2013

All Rights Reserved

ABSTRACT

Spatially Indexed Functional Data

by

Oleksandr Gromenko, Doctor of Philosophy

Utah State University, 2013

Major Professor: Dr. Piotr S. Kokoszka
Department: Mathematics and Statistics

The increased concentration of greenhouse gases is associated with the global warming in the lower troposphere. For over twenty years, the space physics community has studied a hypothesis of global cooling in the thermosphere, attributable to greenhouse gases. While the global temperature increase in the lower troposphere has been relatively well established, the existence of global changes in the thermosphere is still under investigation.

A central difficulty in reaching definite conclusions is the absence of data with sufficiently long temporal and sufficiently broad spatial coverage. Time series of data that cover several decades exist only in a few separated regions. The space physics community has struggled to combine the information contained in these data, and often contradictory conclusions have been reported based on the analyses relying on one or a few locations.

To detect global changes in the ionosphere, we present a novel statistical methodology that uses all data, even those with incomplete temporal coverage. It is based on a new functional regression approach that can handle unevenly spaced, partially observed curves. While this research makes a solid contribution to the space physics community, our statistical methodology is very flexible and can be useful in other applied problems.

(142 pages)

PUBLIC ABSTRACT

The increased concentration of greenhouse gases is associated with the global warming in the lower troposphere. For over twenty years, the space physics community has studied a hypothesis of global cooling in the thermosphere, attributable to greenhouse gases. While the global temperature increase in the lower troposphere has been relatively well established, the existence of global changes in the thermosphere is still under investigation.

A central difficulty in reaching definite conclusions is the absence of data with sufficiently long temporal and sufficiently broad spatial coverage. Time series of data that cover several decades exist only in a few separated (industrialized) regions. The space physics community has struggled to combine the information contained in these data, and often contradictory conclusions have been reported based on the analyses relying on one or a few locations.

To detect global changes in the ionosphere, we present a novel statistical methodology that uses all data, even those with incomplete temporal coverage. It is based on a new functional regression approach that can handle unevenly spaced, partially observed curves. While this research makes a solid contribution to the space physics community, our statistical methodology is very flexible and can be useful in other applied problems including spatio-temporal data.

Oleksandr Gromenko

To my Family.

ACKNOWLEDGMENTS

I am sincerely grateful to my advisor, Professor Piotr Kokoszka, for always being an example of academic excellence and professionalism for me. His constant support, guidance, and devotion to research and teaching have been keeping me focused and productive throughout my study despite of all obvious and hidden obstacles. Without Piotr's generous supervision, this dissertation and my whole career as a statistician would be impossible.

I am grateful to Professor Jan J. Sojka for his constant enthusiasm, involvement, and support. To Professor Daniel Coster, Professor Mevin Hooten and many others for offering excellent courses which significantly extended horizons of my scientific interests. To Dr. Lie Zhu and Levan Lomidze for many useful discussions and constant interest to my work. To professor Jüergen Symanzik for many useful suggestions which leaded to substantial improvement of the dissertation.

I am grateful to Professor Vladimir Privman and Professor Valery P. Gusynin for their invaluable help at the beginning of my scientific career. This dissertation would be impossible without a number of random dramatic events in my life...

Work presented in this dissertation was supported in part by NSF Grants DMS-08-04165 and DMS-09-31948.

Oleksandr Gromenko

CONTENTS

Page

ABSTRACT . . . iii

PUBLIC ABSTRACT . . . iv

ACKNOWLEDGMENTS . . . vi

LIST OF TABLES . . . ix

LIST OF FIGURES . . . x

1 INTRODUCTION . . . 1

2 ESTIMATION AND TESTING FOR SPATIALLY INDEXED CURVES 5
- 2.1 Introduction . . . 5
- 2.2 Estimation of the mean function . . . 11
- 2.3 Estimation of the principal components . . . 14
- 2.4 Finite sample performance of the estimators . . . 19
- 2.5 Testing for correlation of two spatial fields . . . 23
- 2.6 Modeling and estimation of the covariance tensor . . . 27
- 2.7 Size and power of the correlation test . . . 29
- 2.8 Application to critical ionospheric frequency and magnetic curves . . . 31

3 TESTING THE EQUALITY OF MEAN FUNCTIONS . . . 41
- 3.1 Introduction . . . 41
- 3.2 Description of the ionosonde data . . . 46
- 3.3 A functional spatio–temporal model for the ionosonde data . . . 48
- 3.4 Estimation of the mean and covariance functions . . . 50
- 3.5 The test procedure . . . 54
- 3.6 Application to the ionosonde data . . . 57
- 3.7 Estimation of the FPC's . . . 61

4 NONPARAMETRIC INFERENCE IN SMALL DATA SETS OF SPATIALLY INDEXED CURVES . . . 63
- 4.1 Introduction . . . 63
- 4.2 Description of the method for scalar data . . . 65
- 4.3 Statistical model for spatially indexed functional data . . . 67
- 4.4 Estimation of the mean function . . . 69
- 4.5 Significance of regression coefficients . . . 71
- 4.6 Application to ionosonde data . . . 73
- 4.7 Validation of the methodology . . . 77
- 4.8 Summary and conclusions . . . 83
- 4.9 Bandwidth selection and the construction of the functional confidence intervals 85

5 EVALUATION OF THE COOLING TREND IN THE IONOSPHERE USING FUNCTIONAL REGRESSION WITH INCOMPLETE CURVES . 88
5.1 Introduction 88
5.2 The data 92
5.3 Statistical model and inference 94
5.4 Application to ionosonde data 102
5.5 Correlation between scores 106
5.6 Estimation of the covariance structure 108
5.7 Covariances of the estimated mean function 111

6 CONCLUSIONS AND FUTURE WORK 114

REFERENCES 115

APPENDICES 123

CURRICULUM VITAE 126

LIST OF TABLES

Table Page

2.1 Models and estimated covariance parameters for the transformed foF2 curves and the magnetic curves. 36

2.2 P–values of the correlation tests applied to the transformed foF2 data. The first column shows the number of FPC's, the second column shows cumulative variances computed as the ratios of the eigenvalues estimated using method CM2. Testing procedures S, SM and T are defined in Section 2.7. The "simple" procedure neglects the spatial dependence of the curves. 36

3.1 Summary information for the ionosonde stations. 47

3.2 Estimates for the parametric covariance models fitted to the ionosonde data. Abbreviation SF states for spatial field. 58

3.3 P-values for different number of the FPC's p, $1 \leq p \leq 7$ for night and day data. Abbreviation CV states for cumulative variance. 59

3.4 Empirical size (in percent) of the tests applied to data generated to resemble the ionosonde data. 60

4.1 P-values for the trend parameter as a function of the number of the FPC's. 76

4.2 Average L_2 distance between the estimated and true mean functions for different estimation methods; the second number represents the standard error. The entries are based on 10^4 replications. 79

4.3 Average L_2 distance between the estimated and true means; the second number represents the standard error. Entries are based on 10^4 replications. . . 81

4.4 Empirical Size of the test of $H_0 : \beta_2 = 0$. 83

4.5 Empirical Size for the "simple" method. 83

5.1 Trends and P-values for different bandwidths. "NM" denotes estimation without magnetic inclination, "M" denotes estimation when magnetic inclination is included. The number of the FPC's, p, was chosen to obtain the cumulative variance closest to but greater than 85% (indicated as Final CV). 104

LIST OF FIGURES

Figure Page

2.1 F2-layer critical frequency curves at three locations. Top to bottom (latitude in parentheses): Yakutsk (62.0), Yamagawa (31.2), Manila (14.7). The functions exhibit a latidudal trend in amplitude. 7

2.2 Locations of 218 ionosonde stations. Circles represent the 32 stations with the longest complete records. 7

2.3 Typical profile of day time ionosphere. The curve shows electron density as a function of height. The right vertical axis indicates the D, E and F regions. 9

2.4 Errors in the estimation of the mean function for sample sizes: 32, 100, 218. The dashed boxes are estimates using the CH variogram, empty are for the MT variogram. The right–most box for each N corresponds to the simple average. The bold line inside each box plot represents the average value of L (2.31). The upper and lover sides of rectangles shows one standard deviation, and horizontal lines show two standard deviations. The right most boxes correspond to the standard method. 22

2.5 Errors in the estimation of the FPC's for sample sizes: 32, 100, 218 . The bold line inside each box plot represents the average value of L. The upper and lover sides of rectangles shows one standard deviation, and horizontal lines show two standard deviations. The right most boxes correspond to the standard method. 23

2.6 Size of the correlation test as a function of p. Solid disks represent method S (based on χ^2 distribution). Circles represent method SM (based on the Monte-Carlo distribution). 31

2.7 Power of the correlation test SM as a function of the population correlation ρ. Each line represents one of the four possible correlated spatial field $\boldsymbol{\xi}_1 - \boldsymbol{\eta}$, $\boldsymbol{\xi}_2 - \boldsymbol{\eta}$, $\boldsymbol{\xi}_3 - \boldsymbol{\eta}$, $\boldsymbol{\xi}_4 - \boldsymbol{\eta}$. The test was performed using $p = 4$, which explains about 85% of variance of the foF2 curves. Since all curves in the graphs are practically the same, we do not specify which curve represents a particular dependent pair $\boldsymbol{\xi}_i - \boldsymbol{\eta}$. 32

2.8 Dots represent the scaling function $G_L(\mathbf{s}_i)$ in the magnetic coordinate system and crosses are same in the geographic coordinate system. Line is the best fit for G_L in the magnetic coordinate system. 33

2.9 Transformed and centered foF2 curves (continuous) and centered magnetic curves (dashed) at 32 locations denoted with circles in Figure 2.2. The scales for the two families of curves are different. The foF2 curves have the same scale, it is shown on the right vertical axes in MHz. The scale of the magnetic curves changes, it is shown on the right vertical axes in each box (unitless). 39

2.10 Scatter plots of the scores $\boldsymbol{\xi}_i, i = 1, 2, 3, 4$ of the foF2 curves, vertical axes, against the scores $\boldsymbol{\eta}$ of the magnetic curves, horizontal axes. 40

3.1 An example of a digital ionogram recorded at Juliusruh ionosonde. The vertical axis shows height in kilometers and the horizontal axis shows frequency in megahertz. The pink and green dots show returned frequencies and their virtual heights. The frequency at which the virtual height tends to infinity is called the electron critical frequency. The black "bell shaped" curve is restored profile of ionosphere, cf. Fig 3.2, and the rightmost point of this curve is the electron critical frequency. 43

3.2 Typical profile of ionosphere. The curve shows electron density as a function of height. The right vertical axis indicates the D, E and F regions of the ionosphere. 44

3.3 Locations of 13 selected ionosonde stations with the corresponding 5 letter codes. Empty circles represent stations with negative magnetic inclination and solid discs represent stations with positive magnetic inclination. 50

3.4 Ionosonde data (gray lines). The black solid line is the mean function for stations with the negative inclination, the black dashed line is the mean functions for stations with the positive inclination. 51

3.5 Normal QQ plots for the estimated scores $\boldsymbol{\xi}_i$, $1 \leq i \leq 7$. 52

3.6 Convergence of the iterative method for the estimation of the means, R_i as a function of the iteration i. 58

3.7 The power function of the test. The lines represent the "exact" gamma test and the empty diamonds represent the MC test for different confidence levels: $\alpha = 10\%$, 5%, 1%. 61

4.1 Illustration of the estimation procedure for scalar data. The true covariance function (dashed line), its estimate (solid line) and the 95% confidence region (dotted lines). 67

4.2 F2-layer critical frequency curves at three locations. Top to bottom (latitude in parentheses): Yakutsk (62.0), Yamagawa (31.2), Manila (14.7). The functions exhibit a latidudal trend in amplitude. 74

4.3 Locations of 28 ionosonde stations in the northern hemisphere. 75

4.4 Normal QQ-plots plots for the estimated scores, ζ_i, $1 \leq i \leq 6$ for the Day data. 82

4.5 Empirical power. Solid line - empirical power for $\alpha = 10\%$, dash–dotted line - empirical power for $\alpha = 5\%$, dashed line - empirical power for $\alpha = 1\%$. . 84

4.6 Convergence of the iterative method for the estimation of the means R_i as a function of iteration i. 84

5.1 Typical profile of day time ionosphere. The curve shows electron density as a function of height. The right vertical axix indicates the D, E and F regions of the ionosphere. 90

5.2 Gray lines represent all foF2 records analysed in this paper 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. 91

5.3 Locations of the 85 ionosonde stations used in this study, black discs. The two circles in northern Canada represent stations located in the auroral zone (dashed line), which were not used. 93

5.4 Estimation of the weights for incomplete records. Left panel: Gray dots represent *all available squared differences* $(X(\mathbf{s}_k;t_i)-X(\mathbf{s}_\ell;t_i))^2$, $1 \leq i \leq T$; black dots represent squared differences $(X(\mathbf{s}_k;t_i) - X(\mathbf{s}_\ell;t_i))^2$, for some fixed t_i. Dashed lines separate regions $P(d_{k\ell})$. Right panel: The thin line shows the estimated variogram, the bold line represents the fitted Gaussian variogram. 99

5.5 Examples of modeling of the foF2 records via (5.1) with $h_\mu = 5$, $h_c = 10$ and $p = 3$. Gray lines represent original records, solid black lines - model, dashed black line - the mean function. Top: An almost complete record, middle: partially observed record, bottom: example of unstable modeling when the number of observations per curve is less than 200. 103

5.6 Normal quantile–quantile plots for scores $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$, $\boldsymbol{\xi}_3$ estimated using $h_\mu = 15$ and $h_c = 15$. 105

5.7 (a) Distribution of the trend parameter as a function of the interval length, (b) the number of stations as a function of the interval length. (c) distribution of the trend parameter as a function of the number of stations. Light gray - full region, dark gray - central 50 percent, dotted line - median, blue solid line - average. Bold black line is the final estimator. 106

5.8 Black lines represent estimated correlograms and gray regions represent the 95% pointwise confidence bounds. 108

CHAPTER 1

INTRODUCTION

The research presented in this dissertation was originally motivated by a very important problem which is directly connected to global warming in the troposphere. The increasing concentration of greenhouse gases has been associated with the global cooling in the ionosphere as opposite to the global warming in the lower troposphere. Roble and Dickinson (1989) using global modeling have shown that the doubling of the mixture of CO_2 and CH_4 will lead to significant cooling of the thermosphere by about 50K. Since then, many researchers in the geophysics community have been working on estimating the linear trend possibly associated with the increasing concentration of greenhouse gases.

To address this problem, we analyze the maximum or the peak frequency of oscillation of free electrons which is directly connected to the temperature of the ionosphere and comparing to the last can be measured in a straightforward way. This frequency has a special nomenclature in geophysics literature, foF2, and we use the same notation everywhere below. The foF2 data represent collections of long (50+ years) mostly equidistant time series with a typical time separation of one hour. These records are collected at different spatial locations all around the globe by ground based instruments called ionosondes.

Global cooling in the ionosphere would cause systematic global decrease of foF2. Behavior of the ionosphere is dominated by several natural factors such as solar activity, Earth's magnetic field, and others. These natural factors have changed over the last decades. Thus, the main challenge in the study was to separate the influence of these natural factors and anthropogenic factors which can be later associated with the increasing concentration of greenhose gases. There exist many different ionospheric models which reproduce the short and middle term behaviour of the ionosphere very well. But modeling of the long term behaviour has always been challenging. Thus, determination of the long term trend should be done from the statistical point of view rather than via physics modeling.

This dissertation consists of four chapters. Each chapter represents a separate article, three of which are already published and one is under review.

The foF2 data are very new to the statistical community. They are freely available from the Space Physics Interactive Data Resource (SPIDR), `http://spidr.ngdc.noaa.gov/spidr/`. Unfortunately, the raw foF2 data are not of good quality and need preprocessing (cleaning) before any statistical study. We developed a more or less automatic cleaning procedure, which besides general cleaning, is also capable of calculating medians and averages for different selected times or time periods. One of the main methodological challenges is that foF2 data have very long gaps of missing observations which, at the early stages of our work, forced us to significantly reduce the analyzed time interval and drop most of the records.

Let $X(\mathbf{s};t)$ be a foF2 measurement at location $\mathbf{s}$ and time t. The main new idea in our research was to use a functional approach instead of applying classical spatio–temporal statistics. Specifically, we treat the whole curve $X(\mathbf{s};\cdot)$ as a single L^2-valued observation. These L^2-valued observations form a spatial random field. Assuming strict stationarity, we represent each curve using the functional space–time model:

$$X(\mathbf{s};t) = \mu(t) + \varepsilon(\mathbf{s};t), \tag{1.1}$$

where $\mu(t) = EX(\mathbf{s};t)$ and $E\varepsilon(\mathbf{s};t) = 0$. The error term is further decomposed using the Karhunen–Loéve expansion: $\varepsilon(\mathbf{s};t) = \sum_{j=1}^{\infty} \xi_j(\mathbf{s})v_j(t)$, where $v_j(t)$ are the functional principal components (FPC's) and $\xi_j(\mathbf{s})$ are the corresponding zero mean scores with some internal spatial correlation structure. In Chapter 2, we proposed several new estimators for the mean function, $\mu(t)$, and the FPC's, $v_j(t)$, which naturally incorporate spatial or another type of dependence. Using numerical simulations, we have shown that our new estimators have better performances compared to the standard ones which assume independence of curves. In the same chapter, we proposed a correlation test for testing if two sets of curves measured at the same locations are uncorrelated. Using this test, we confirmed that the changes in foF2 records are strongly correlated with the changes in the Earth's magnetic field.

After developing a suitable statistical framework for modelling spatially correlated functional data (SCFD), we wanted to test if indeed the mean function, $\mu(t)$, is the same for all spatial locations. In Chapter 3, we proposed a testing procedure for testing if the means of two samples of CFD are the same:

$$H_0 : \mu_1(t) = \mu_2(t),$$
$$H_A : \mu_1(t) \neq \mu_2(t).$$

Using this test, we have shown that the means of foF2 records in western and eastern Europe are statistically different.

One of the reasons why the analysis of SCFD has become so popular is because it can be treated as a fully nonparametric extension of spatio–temporal statistics. By fully nonparametric we mean that neither the mean function $\mu(t)$ nor the covariance $\mathrm{Cov}(X(\mathbf{s}_k; t_i), X(\mathbf{s}_\ell; t_{i'}))$ require any parametric assumptions. Particularly, the Karhunen–Loéve expansion of the error term, mentioned earlier, leads to the following nonparametric covariance function

$$\mathrm{Cov}(X(\mathbf{s}_k; t_i), X(\mathbf{s}_\ell; t_{i'})) = \sum_{j=1}^{\infty} \gamma_j(\mathbf{s}_k, \mathbf{s}_\ell) v_j(t_i) v_j(t_{i'}), \tag{1.2}$$

where $\gamma_j(\mathbf{s}_k, \mathbf{s}_\ell) = E[\xi_j(\mathbf{s}_k)\xi_j(\mathbf{s}_\ell)]$. Model (1.2) is nonseparable, except in some very special cases, and can be constructed without any parametric assumptions, see Chapter 4 for the further details.

In the second part of Chapter 4, we estimated the linear trend not associated with the natural factors such as solar activity and assessed its statistical significance. We proposed a simple procedure and applied it to the noon foF2 records. We found that for the *studied period*, the linear trend is negative and it is statistically significant.

In Chapter 5, we proposed new estimators for the mean function $\mu(t)$ and the FPC's $v_j(t)$ which simultaneously incorporate spatial dependence and handle missing observations. Using this new technique, we were able to analyse 85 foF2 records in the northern hemisphere for the period from 1957 to 2011. We found that the linear ionospheric trend in the northern hemisphere is statistically significant. We believe this result will make a sound contribution to the geophysics community. Our new statistical methodology will also be useful in other research areas which involve the analysis of space–time data with observations and trend determination.

CHAPTER 2

ESTIMATION AND TESTING FOR SPATIALLY INDEXED CURVES WITH APPLICATION TO IONOSPHERIC AND MAGNETIC FIELD TRENDS[1]

Abstract

We develop methodology for the estimation of the functional mean and the functional principal components when the functions form a spatial process. The data consist of curves $X(\mathbf{s}_k; t),\ t \in [0, T]$, observed at spatial locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$. We propose several methods, and evaluate them by means of a simulation study. Next, we develop a significance test for the correlation of two such functional spatial fields. After validating the finite sample performance of this test by means of a simulation study, we apply it to determine if there is correlation between long term trends in the so called critical ionospheric frequency and decadal changes in the direction of the internal magnetic field of the Earth. The test provides conclusive evidence for correlation thus solving a long standing space physics conjecture. This conclusion is not apparent if the spatial dependence of the curves is neglected.

2.1 Introduction

The contribution of this paper to statistics is two–fold: 1) we develop estimation methodology for the functional mean and the functional principal components (FPC's) when the functions form a spatial field; 2) we propose a significance test to determine if two families of curves observed at the same spatial locations are uncorrelated. The contribution to space physics consists in solving a controversy regarding the impact of long term changes in the internal magnetic field of the Earth on long term ionospheric trends. The required physics background is provided later in this section, and in Section 2.8.

---

[1] CO-AUTHORED BY O. GROMENKO, P. KOKOSZKA, L. ZHU, AND J. SOJKA. REPRODUCED FROM THE ANNALS OF APPLIED STATISTICS, VOL. 6, 669-696, 2012. PERMISSION IS NOT REQUIRED.

The data is modeled as curves $X(\mathbf{s}_k; t),\ t \in [0, T]$, observed at spatial locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$. Such functional data structures are quite common, but typically the spatial dependence and the spatial distribution of the points $\mathbf{s}_k$ are not taken into account. A fundamental question is how to estimate the mean function of curves indexed by spatial locations. Clearly, curves located at close by points look similar and must be given smaller weights than curves at points far apart. In addition to the mean function, FPC's play a fundamental role in functional data analysis. Good estimators of FPC's are needed to construct reliable testing and classification procedures, but such issues have been addressed only in the contexts of independent curves, with focus on sparsity and measurement error. The geophysical data that motivate this research are available at fine temporal grids and are measured with errors that are negligible relative to the objectives of the statistical analysis. A focus of recent geophysical research is on the detection and estimation of global and/or regional long term trends (the global warming paradigm), so before a statistical analysis is undertaken, the data are typically smoothed to remove daily or even annual periodicity. The question we address is how to combine the temporal trajectories available at many spatial locations to obtain meaningful summary trends. We argue that one can do better than using simple averaging. The focus of this paper is thus on combining information from spatially dependent curves, which are smooth and available at all time points.

Many environmental and geophysical data sets fall into the framework considered in this paper. The data set that motivated this research consists of the curves of the ionospheric F2-layer critical frequency, foF2. Three such curves are shown in Figure 2.1. In principle, foF2 curves are available at over 200 locations throughout the globe, see Figure 2.2, but sufficiently complete data are available at only 30-40 locations which are very unevenly spread; for example, there is a dense network of observatories over Europe and practically no data over the oceans. The study of this data set has been motivated by the hypothesis of Roble and Dickinson (1989) who suggested that the increasing amounts of (radiative) greenhouse gases should lead to global cooling in mesosphere and thermosphere, as opposed to the global warming in lower troposphere, cf. Figure 2.3. Rishbeth (1990) pointed out

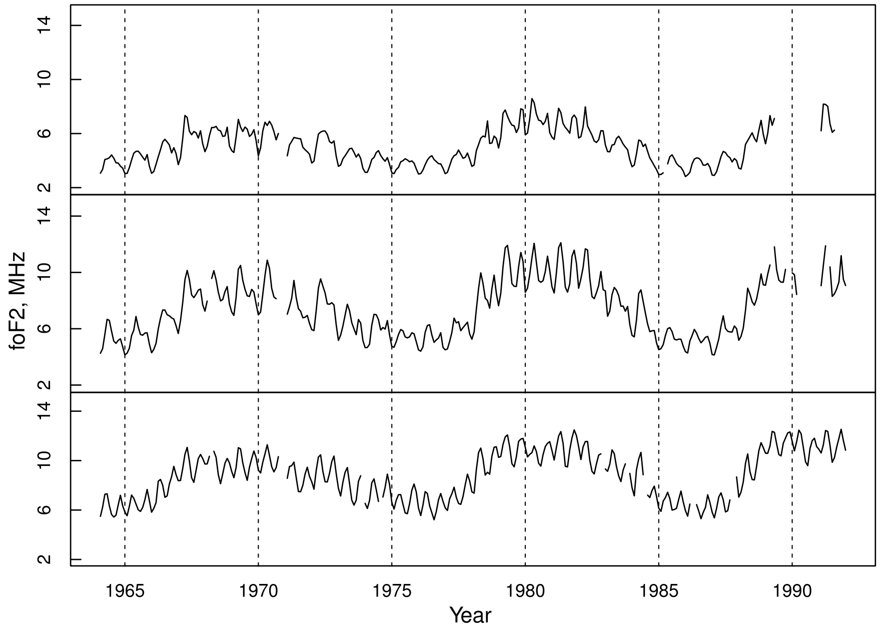

Figure 2.1: F2-layer critical frequency curves at three locations. Top to bottom (latitude in parentheses): Yakutsk (62.0), Yamagawa (31.2), Manila (14.7). The functions exhibit a latidudal trend in amplitude.

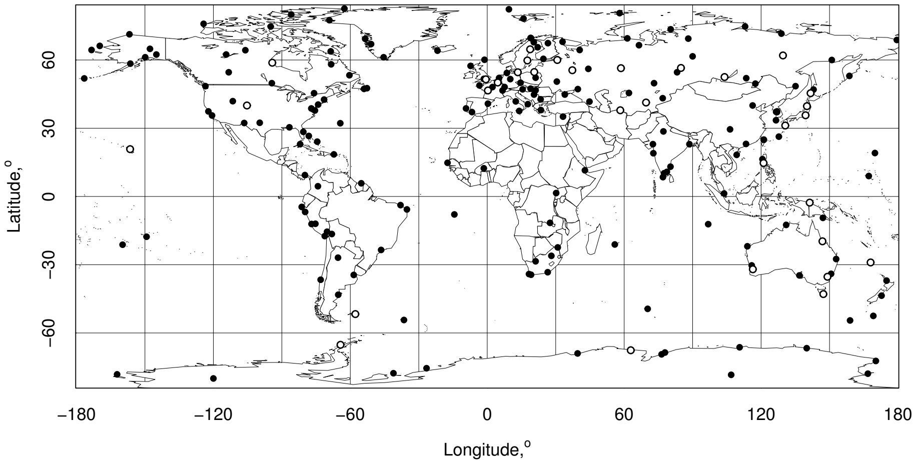

Figure 2.2: Locations of 218 ionosonde stations. Circles represent the 32 stations with the longest complete records.

that such cooling would result in a thermal contraction and the global lowering of the ionospheric peak densities, which can be computed from the critical frequency foF2. The last twenty years have seen very extensive research in this area, see Lastovicka *et al.* (2008) for a partial overview. One of the difficulties in determining a global trend is that the foF2 curves appear to exhibit trends in opposing directions over various regions. A possible explanation suggests that these trends are caused by long term trends in the magnetic field of the Earth. There is however currently not agreement in the space physics community if this is indeed the case. In general, to make any trends believable, a suitable statistical modeling, and a proper treatment of "errors and uncertainties" is called for Ulich *et al.* (2003). This paper makes a contribution in this direction. Space physics data measured at terrestrial observatories always come in the form of temporal curves at fixed spatial locations. In Maslova *et al.* (2009), Maslova *et al.* (2010a), and Maslova *et al.* (2010b) the tools of functional data analysis were used to study such data, but the spatial dependence of the curves was not fully exploited.

Spatio-temporal modeling has received a great deal of attention of late, see Part V of Gelfand *et al.* (2010) and Chapters 3, 4 and 6 of Gneiting *et al.* (2007) which discuss spatio–temporal models for geostatistical data. There has however not been much research specifically on spatially indexed functional data; Delicado *et al.* (2010) review recent contributions. For geostatistical functional data, several approaches to kriging have been proposed, see Yamanishi and Tanaka (2003), Nerini *et al.* (2010), Giraldo *et al.* (2011) and Bel *et al.* (2011).

Throughout the paper, $\{X(\mathbf{s})\}$ denotes a random field defined on a spatial domain and taking values in the Hilbert space $L^2 = L^2([0,1])$ with the inner product

$$\langle f, g\rangle = \int_0^1 f(t)g(t)dt, \quad f, g \in L^2.$$

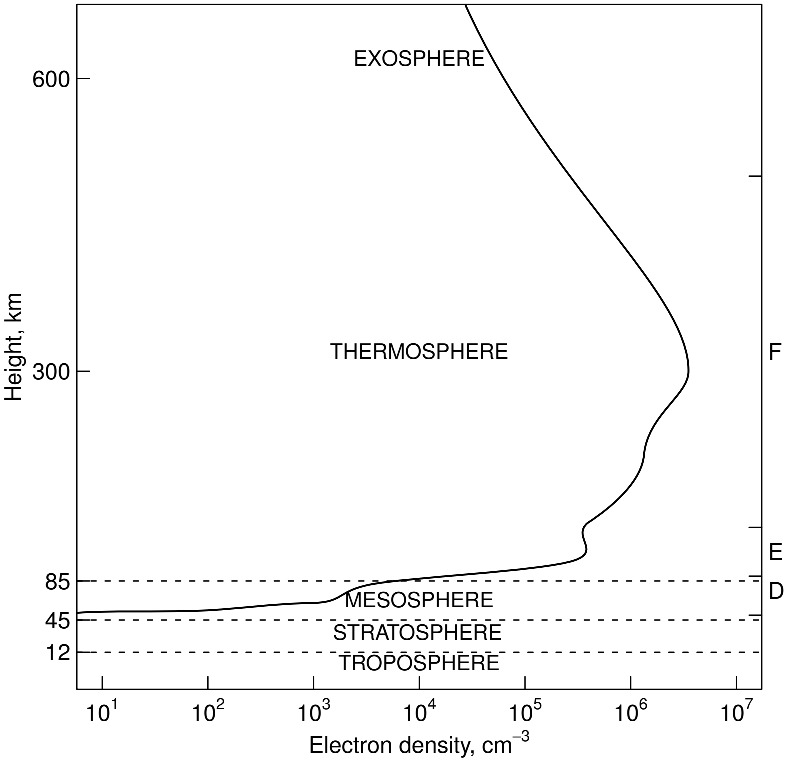

Figure 2.3: Typical profile of day time ionosphere. The curve shows electron density as a function of height. The right vertical axis indicates the D, E and F regions.

The value of the function $X(\mathbf{s}) \in L^2$ at time $t \in [0,1]$ is denoted by $X(\mathbf{s};t)$. We postulate the model

$$X(\mathbf{s};t) = \mu(t) + \sum_{i=1}^{\infty} \xi_i(\mathbf{s}) e_i(t), \quad \xi_i(\mathbf{s}) = \langle X(\mathbf{s}) - \mu,\ e_i \rangle, \tag{2.1}$$

where the e_i form a complete orthonormal system. Note that the mean function μ and the FPC's e_i do not depend on $\mathbf{s}$. A sufficient condition for this is that the distribution in L^2 of the function $X(\mathbf{s})$ does not depend on the location $\mathbf{s}$. A stronger sufficient condition is the strict stationarity of the field $\{X(\mathbf{s})\}$.

For the applications we have in mind, it is enough to assume that the spatial domain is a subset of the plane or a two–dimensional sphere. On the plane, the distance between points is the usual Euclidean distance; on the sphere, we use the chordal distance defined as the Euclidean distance in the three–dimensional space. The reason for using the chordal distance is that any spatial covariance functions in $\mathbb{R}^3$ restricted to the unit sphere is then also a covariance function on the sphere. Denoting the latitude by L and the longitude by l, the chordal distance, $0 \leq d_{k,\ell} \leq 2$, between two points, $\mathbf{s}_k, \mathbf{s}_\ell$, on the unit sphere is given by

$$d_{k,\ell} = 2\left[\sin^2\left(\frac{L_k - L_\ell}{2}\right) + \cos L_k \cos L_\ell \sin^2\left(\frac{l_k - l_\ell}{2}\right)\right]^{1/2}. \tag{2.2}$$

For arbitrary (not necessarily spatially indexed) functions, $X_1, X_2, \ldots, X_N$, the sample mean is defined as $\bar{X}_N = N^{-1}\sum_{n=1}^{N} X_n$, and the sample covariance operator as

$$\widehat{C}(x) = N^{-1}\sum_{n=1}^{N}\left[\langle (X_n - \bar{X}_N), x\rangle (X_n - \bar{X}_N)\right], \quad x \in L^2.$$

The sample FPC's are computed as the eigenfunctions of $\widehat{C}$. These are the estimates produced by several software packages, including the popular `R` package `fda`, see Ramsay *et al.* (2009). The consistency of the sample mean and the sample FPC's relies on the assumption that the functional observations form a simple random sample. If the functions $X_k = X(\mathbf{s}_k)$ are spatially distributed, the sample mean and the sample FPC's need not even be consistent, see Hörmann and Kokoszka (2013). This happens if the spatial dependence is strong or if there are clusters of the points $\mathbf{s}_k$. We will demonstrate that better estimators are available and we will use them as part of the procedure for testing the independence of two functional fields $\{X(\mathbf{s}), \mathbf{s} \in \mathbf{S}\}$ and $\{Y(\mathbf{s}), \mathbf{s} \in \mathbf{S}\}$. The procedure is based on the observed pairs of functions $(X(\mathbf{s}_k), Y(\mathbf{s}_k)),\ 1 \leq k \leq N$. The test we propose is applied to ionosonde (X) and magnetic (Y) curves, and conclusively shows that the temporal evolution of these two families is strongly correlated.

The remainder of the paper is organized as follows. Sections 2.2 and 2.3 focus, respectively, on the estimation of the mean function and the FPC's in a spatial setting. Section 2.4 demonstrates by means of a simulation study that the methods we propose improve on the standard approach, and discusses their relative performance and computational cost. In Section 2.5, we develop a test for the correlation of two functional spatial fields. This test requires estimation of a covariance tensor. After addressing this issue in Section 2.6, we study in Section 2.7 the finite sample properties of several implementations of the test. Finally, in Section 2.8, we apply the methodology developed in the previous section to test for the correlation between the ionospheric critical frequency and magnetic curves.

2.2 Estimation of the mean function

We propose three methods of estimating the mean function μ, which we call M1, M2, M3. As will become apparent in this section, several further variants, not discussed here, are conceivable. But the results of Section 2.4 show that while all these methods offer an improvement over the simple sample mean, their performance is comparable. We represent the observed functions as

$$X(\mathbf{s}_k;t) = \mu(t) + \varepsilon(\mathbf{s}_k;t), \quad k = 1, 2, \ldots, N, \tag{2.3}$$

where ε is an unobservable field with $E\varepsilon(\mathbf{s};t) = 0$. All methods assume that the function valued field $\varepsilon(\cdot)$ is strictly stationary and isotropic, even though weaker, more technical assumptions could be made for the specific methods. Methods M1 and M2 are akin to the kriging technique advocated by Giraldo *et al.* (2011) in that they treat the curves as single entities and seek to minimize the integrated mean squared error. Method M3 is similar in spirit to the approach to functional kriging developed by Nerini *et al.* (2010) and Giraldo *et al.* (2010) who use cokriging of basis coefficients.

Methods M1 and M2 estimate μ by the weighted average

$$\hat{\mu}_N = \sum_{n=1}^{N} w_n X(\mathbf{s}_n). \tag{2.4}$$

The optimal weights w_k are defined to minimize $E\|\sum_{n=1}^{N} w_n X(\mathbf{s}_n) - \mu\|^2$ subject to the condition $\sum_{n=1}^{N} w_n = 1$ ($\|x\|^2 = \int_0^1 x^2(t)dt$). Using the method of the Lagrange multiplier, we see that the unknowns $w_1, w_2, \ldots, w_N, r$ are solutions to the system of $N+1$ equations

$$\sum_{n=1}^{N} w_n = 1, \quad \sum_{k=1}^{N} w_k C_{kn} - r = 0, \quad n = 1, 2, \ldots, N, \tag{2.5}$$

where

$$C_{k\ell} = E[\langle \varepsilon(\mathbf{s}_k), \varepsilon(\mathbf{s}_\ell) \rangle]. \tag{2.6}$$

Set $\mathbf{w} = (w_1, \ldots, w_N)^T$. An easy way to solve equations (2.5) is to compute $\mathbf{v} = \mathbf{C}^{-1}\mathbf{1}$, where $\mathbf{C} = [C_{k\ell}, 1 \le k, \ell \le N]$, and then set $\mathbf{w} = a\mathbf{v}$, where a is a constant such that $\mathbf{1}^T\mathbf{w} = 1$.

Method M1. At each time point t_j, we fit a parametric spatial model to the scalar field $X(\mathbf{s}; t_j)$. To focus attention, we provide formulas for the exponential model

$$\mathrm{Cov}(X(\mathbf{s}_k; t_j), X(\mathbf{s}_\ell; t_j)) = \sigma^2(t_j) \exp\left\{ -\frac{d(\mathbf{s}_k, \mathbf{s}_\ell)}{\rho(t_j)} \right\}. \tag{2.7}$$

It is clear how they can be modified for other popular models.

Observe that under model (2.7),

$$\begin{aligned} C_{k\ell} &= E \int \left(X(\mathbf{s}_k; t) - \mu(t)\right)\left(X(\mathbf{s}_\ell; t) - \mu(t)\right) dt \\ &= \int \mathrm{Cov}(X(\mathbf{s}_k; t_j), X(\mathbf{s}_\ell; t_j)) dt \\ &= \int \sigma^2(t) \exp\left\{ -\frac{d(\mathbf{s}_k, \mathbf{s}_\ell)}{\rho(t)} \right\} dt. \end{aligned}$$

One way to estimate $C_{k\ell}$ is to set

$$\widehat{C}_{k\ell} = \int \hat{\sigma}^2(t) \exp\left\{ -\frac{d(\mathbf{s}_k, \mathbf{s}_\ell)}{\hat{\rho}(t)} \right\} dt, \tag{2.8}$$

with the estimates $\hat{\sigma}^2(t_j)$ and $\hat{\rho}(t_j)$ obtained using some version of empirical variogram, (2.29) or (2.30) in this study.

Another way to proceed, is to replace the $\hat{\rho}(t_j)$ by their average $\hat{\rho} = m^{-1} \sum_{j=1}^{m} \hat{\rho}(t_j)$, where m is the count of the t_j at which the variogram is estimated successfully. Then, the $C_{k\ell}$ are approximated by

$$\widehat{C}_{k\ell} = \left(\int \hat{\sigma}^2(t) dt \right) \exp\left\{ -\frac{d(\mathbf{s}_k, \mathbf{s}_\ell)}{\hat{\rho}} \right\}.$$

As explained above, in order to compute the weights w_j in (2.5), it is enough to know the matrix $\mathbf{C}$ only up to a multiplicative constant. Thus we may set

$$\widehat{C}_{k\ell} = \exp\left\{-\frac{d(\mathbf{s}_k,\mathbf{s}_\ell)}{\hat{\rho}}\right\}. \tag{2.9}$$

Once the matrix $\mathbf{C}$ has been estimated, we compute the weights w_j, and estimate the mean via (2.4).

If (2.8) is used, we refer to this method as M1a, if (2.9) is used, we call it M1b.

Method M1 relies on the estimation of the variograms at every point t_j. Method M2, described below, requires only one optimization, so it is much faster than M1.

Method M2. We define the *functional* variogram

$$\begin{aligned} 2\gamma(\mathbf{s}_k,\mathbf{s}_\ell) &= E||X(\mathbf{s}_k) - X(\mathbf{s}_\ell)||^2 \\ &= 2E\|X(\mathbf{s}_k) - \mu\|^2 - 2E\left[\langle X(\mathbf{s}_k) - \mu, X(\mathbf{s}_\ell) - \mu\rangle\right] \\ &= 2E\|X(\mathbf{s}) - \mu\|^2 - 2C_{k\ell}. \end{aligned} \tag{2.10}$$

The variogram (2.10) can be estimated by its empirical counterparts, like (2.29) or (2.30), with the $|X(\mathbf{s}_k) - X(\mathbf{s}_\ell)|$ replaced by

$$||X(\mathbf{s}_k) - X(\mathbf{s}_\ell)|| = \left\{\int \left(X(\mathbf{s}_k;t) - X(\mathbf{s}_\ell;t)\right)^2 dt\right\}^{1/2}.$$

Next, we fit a parametric model, for example we postulate that

$$\gamma(\mathbf{s}_k,\mathbf{s}_\ell) = \sigma_f^2\left(1 - \exp\left\{-\frac{d(\mathbf{s}_k,\mathbf{s}_\ell)}{\rho_f}\right\}\right). \tag{2.11}$$

The subscript f is used to emphasize the *functional* variogram. Denoting by $\hat{\rho}_f$ the resulting NLS estimate, we estimate the C_{kl} by (2.9) with $\hat{\rho}$ replaced by $\hat{\rho}_f$.

Method M3. This method uses a basis expansion of the functional data, it does not use the weighted sum (2.4). Suppose $B_j, 1 \le j \le K$, are elements of a functional basis with

K so large that for each k

$$X(\mathbf{s}_k) \approx \sum_{j\le K} \langle B_j, X(\mathbf{s}_k)\rangle B_j \tag{2.12}$$

to a good approximation. By (2.3), we obtain for every j

$$\langle B_j, X(\mathbf{s}_k)\rangle = \langle B_j, \mu\rangle + \langle B_j, \varepsilon(\mathbf{s}_k)\rangle, \quad k = 1, 2, \ldots, N. \tag{2.13}$$

For every fixed j, the $\langle B_j, X(\mathbf{s}_k)\rangle$ form a stationary and isotropic scalar spatial field with a constant unknown mean $\langle B_j, \mu\rangle$. This mean can be estimated by postulating a covariance structure for each $\langle B_j, X(\mathbf{s}_k)\rangle$, for example

$$\mathrm{Cov}\left(\langle B_j, X(\mathbf{s}_k)\rangle, \langle B_j, X(\mathbf{s}_\ell)\rangle\right) = \sigma_j^2 \exp\left\{-\frac{d(\mathbf{s}_k, \mathbf{s}_\ell)}{\rho_j}\right\}.$$

The mean $\langle B_j, \mu\rangle$ is estimated by a weighted average of the $\langle B_j, X(\mathbf{s}_k)\rangle$ (the weights depend on j). Denote the resulting estimate by $\hat{\mu}_j$. The mean function μ is then estimated by $\hat{\mu}(t) = \sum_{j\le K} \hat{\mu}_j B_j(t)$.

2.3 Estimation of the principal components

Assume now that the mean function μ has been estimated, and this estimate is subtracted from the data. To simplify the formulas, in the following we thus assume that $EX(\mathbf{s}) = 0$.

We consider analogs of methods M2 and M3. Extending Method M1 is possible, but presents a computational challenge because a parametric spatial model would need to be estimated for every pair (t_i, t_j). For the ionosonde data studied in Section 2.8, there are 336 points t_j. Estimation on a single data set would be feasible, but not a simulation study based on thousands of replications.

In both approaches, which we term CM2 and CM3, the FPC's are estimated by expansions of the form

$$v_j(t) = \sum_{\alpha=1}^{K} x_\alpha^{(j)} B_\alpha(t), \tag{2.14}$$

where the B_α are elements of an *orthonormal* basis. We first describe an analog of method M3, which is conceptually and computationally simpler.

Method CM3. The starting point is the expansion

$$X(\mathbf{s}; t) = \sum_{j=1}^{\infty} \xi_j(\mathbf{s}) B_j(t),$$

where, by the orthonormality of the B_j, the $\xi_j(\mathbf{s})$ form an observable field $\xi_j(\mathbf{s}_k) = \langle B_j, X(\mathbf{s}_k)\rangle$. Using the orthonormality of the B_j again, we obtain

$$\begin{aligned} C(B_j) &= E\left[\langle \sum_{\alpha=1}^{\infty} \xi_\alpha(\mathbf{s}) B_\alpha, B_j\rangle \sum_{i=1}^{\infty} \xi_i(\mathbf{s}) B_i\right] \\ &= E\left[\xi_j(\mathbf{s}) \sum_{i=1}^{\infty} \xi_i(\mathbf{s}) B_i\right] \\ &= \sum_{i=1}^{\infty} E[\xi_i(\mathbf{s})\xi_j(\mathbf{s})] B_i. \end{aligned} \tag{2.15}$$

Thus, to estimate C, we must estimate the means $E[\xi_i(\mathbf{s})\xi_j(\mathbf{s})]$.

Fix i and j, and define the scalar field z by $z(\mathbf{s}) = \xi_i(\mathbf{s})\xi_j(\mathbf{s})$. We can postulate a parametric model for the covariance structure of the field $z(\cdot)$, and use an empirical variogram to estimate $\mu_z = Ez(\mathbf{s})$ as a weighted average of the $z(\mathbf{s}_k)$. Denote the resulting estimate by $\hat{r}_{ij}$.

The empirical version of (2.15) is then

$$\widehat{C}(B_j) = \sum_{i=1}^{K} \hat{r}_{ij} B_i. \tag{2.16}$$

Relation (2.16) defines the estimator $\widehat{C}$ which acts on the span of $B_j, 1 \le j \le K$.

Its eigenfunctions are of the form $x = \sum_{1\le\alpha\le K} x_\alpha B_\alpha$. Observe that

$$\widehat{C}(x) = \sum_{\alpha} x_\alpha \sum_{i} \hat{r}_{i\alpha} B_i = \sum_{i} \left(\sum_{\alpha} \hat{r}_{i\alpha} x_\alpha\right) B_i.$$

On the other hand,

$$\lambda x = \sum_i \lambda x_i B_i.$$

Since the B_i form an orthonormal basis, we obtain

$$\sum_\alpha \hat{r}_{i\alpha} x_\alpha = \lambda x_i.$$

Setting

$$\mathbf{x} = [x_1, x_2, \ldots, x_K]^T, \quad \widehat{\mathbf{R}} = [\hat{r}_{ij},\ 1 \le i, j \le K],$$

we can write the above as a matrix equation

$$\widehat{\mathbf{R}}\mathbf{x} = \lambda \mathbf{x}. \tag{2.17}$$

Denote the solutions to (2.17) by

$$\hat{\mathbf{x}}^{(j)} = [\hat{x}_1^{(j)}, \hat{x}_2^{(j)}, \ldots, \hat{x}_k^{(j)}]^T, \quad \hat{\lambda}_j, \qquad 1 \le j \le K. \tag{2.18}$$

The $\hat{\mathbf{x}}^{(j)}$ satisfy $\sum_{\alpha=1}^K \hat{x}_\alpha^{(j)} \hat{x}_\alpha^{(i)} = \delta_{ij}$. Therefore the $\hat{v}_j$ defined by

$$\hat{v}_j = \sum_{\alpha=1}^K \hat{x}_\alpha^{(j)} B_\alpha \tag{2.19}$$

are also orthonormal (because the B_j are orthonormal). The $\hat{v}_j$ given by (2.19) are the estimators of the FPC's, and the $\hat{\lambda}_j$ in (2.18) of the corresponding eigenvalues.

As in method M3, the value of K can be taken to the number of basis functions used to create the functional objects in `R`, so it can be a relatively large number, e.g. $K = 49$. Even though the range of j in (2.18) and (2.19) runs up to K, only the first few estimated FPC's $\hat{v}_j$ would be used in further work.

Method CM2. Recall that under the assumption of zero mean function, the covariance operator is defined by $C(x) = E[\langle X(\mathbf{s}), x\rangle X(\mathbf{s})]$. It can be estimated by the simple average

$$\frac{1}{N}\sum_{n=1}^{N}\langle X(\mathbf{s}_n), \cdot\rangle X(\mathbf{s}_n) = \frac{1}{N}\sum_{n=1}^{N} C_k, \tag{2.20}$$

where C_k is the operator defined by

$$C_k(x) = \langle X(\mathbf{s}_k), x\rangle X(\mathbf{s}_k).$$

As for the mean, more precise estimates can be obtained by using the weighted average

$$\widehat{C} = \sum_{k=1}^{N} w_k C_k. \tag{2.21}$$

Before discussing the estimation of the weights w_k, we comment that the FPC's v_j and their eigenvalues λ_j can be estimated using (2.21) and the representation (2.14). As in method CM3, set $x = \sum_{1\le\alpha\le K} x_\alpha B_\alpha$, and observe that

$$\widehat{C}(x) = \sum_{j=1}^{K}\left(\sum_{\alpha=1}^{K} s_{j\alpha}x_\alpha\right) B_j,$$

where

$$s_{j\alpha} = \sum_{k=1}^{N} w_k \langle X_k, B_j\rangle\langle X_k, B_\alpha\rangle.$$

Thus, analogously to (2.17), we obtain a matrix equation $\mathbf{S}\mathbf{x} = \lambda\mathbf{x}$, from which the estimates of the v_j, λ_j can be found as in (2.18) and (2.19).

We now return to the estimation of the weights w_k in (2.21). One way to define the optimal weights is to require that they minimize the expected Hilbert–Schmidt norm of $\widehat{C} - C$. Recall that the Hilbert–Schmidt norm of an operator K is defined by

$$||K||_{\mathcal{S}}^2 = \sum_{i=1}^{\infty} ||K(e_i)||^2 = \sum_{i=1}^{\infty}\int |K(e_i)(t)|^2 dt,$$

where $\{e_i, i \geq 1\}$ is any orthonormal basis in L^2. Since $|| \cdot ||_{\mathcal{S}}$ is a norm in the the Hilbert space $\mathcal{S}$ of the Hilbert–Schmidt operators with the inner product

$$\langle K_1, K_2 \rangle_{\mathcal{S}} = \sum_{i=1}^{\infty} \langle K_1(e_i), K_2(e_i) \rangle,$$

we can repeat all algebraic manipulations needed to obtain the weight w_i in (2.4). The optimal weights in (2.21) thus satisfy

$$\sum_{n=1}^{N} w_n = 1, \quad \sum_{k=1}^{N} w_k \kappa_{kn} - r = 0, \quad n = 1, 2, \ldots, N, \tag{2.22}$$

where

$$\kappa_{k\ell} = E[\langle C_k - C, C_\ell - C \rangle_{\mathcal{S}}].$$

Finding the weights thus reduces to estimating the expected inner products $\kappa_{k\ell}$.

Since method M2 of Section 2.2 relies only on estimating inner product in the Hilbert space L^2, it can be extended to the Hilbert space $\mathcal{S}$. First observe that, analogously to (2.10),

$$E||C_k - C_\ell||_{\mathcal{S}}^2 = 2E||C_k - C||_{\mathcal{S}}^2 - 2\kappa_{k\ell}.$$

We can estimate the variogram

$$\gamma_C(d) = E||\langle X(\mathbf{s}), \cdot \rangle X(\mathbf{s}) - \langle X(\mathbf{s}+\mathbf{d}), \cdot \rangle X(\mathbf{s}+\mathbf{d})||_{\mathcal{S}}^2, \quad d = ||\mathbf{d}||$$

by fitting a parametric model. In formulas (2.29) and (2.30), the squared distances $(X(\mathbf{s}_k) - X(\mathbf{s}_\ell))^2$ must be replaced by the squared norms $||C_k - C_\ell||_{\mathcal{S}}^2$. These norms are equal to

$$||C_k - C_\ell||_{\mathcal{S}}^2 = \sum_{i=1}^{\infty} \int (f_{ik} X_k(t) - f_{i\ell} X_\ell(t))^2 \, dt,$$

where

$$f_{ik} = \int X_k(t) e_i(t) dt.$$

The inner products f_{ik} can be computed using the `R` package `fda`.

2.4 Finite sample performance of the estimators

In this section, we report the results of a simulation study designed to compare the performance of the methods proposed in Sections 2.2 and 2.3 in a realistic setting motivated by the ionosonde data. It is difficult to design an exhaustive simulation study due to the number of possible combinations of the point distributions, dependence structures, shapes of mean functions and the FPC's and ways of implementing the methods (choice of spatial models, variogram estimation etc.). We do however think that our study provides useful information and guidance for practical application of the proposed methodology.

Data generating processes. We generate functional data at location $\mathbf{s}_k$ as

$$X(\mathbf{s}_k; t) = \mu(t) + \sum_{i=1}^{p} \xi_i(\mathbf{s}_k) e_i(t), \tag{2.23}$$

where the e_i are orthonormal functions, cf. model (2.1).

To evaluate the estimators of the mean, we use $p = 2$ and

$$e_1(t) = \sqrt{2}\sin(2\pi t \cdot 6), \quad e_2(t) = \sqrt{2}\sin(2\pi t/2). \tag{2.24}$$

We use two mean functions

$$\mu(t) = a\sqrt{2}\sin(2\pi t \cdot 3), \quad a = 2 \tag{2.25}$$

and

$$\mu(t) = a\sqrt{t}\sin(2\pi t \cdot 3), \quad a = 1 \tag{2.26}$$

The mean function (2.25) resembles the mean shape for the ionosonde data. It is however a member of the Fourier basis, and can be isolated using only one basis function, what could possibly artificially enhance the performance of method M3. We therefore also consider the mean function (2.26). Combining the mean function (2.25) and the FPC's

(2.24), we obtain functions which very closely resemble the shapes of the ionosonde curves. In the above formulas, time is rescaled so that $t \in [0, 1]$.

To evaluate the estimators of the FPC's, we use $p = 3$ and

$$X(\mathbf{s}_k; t) = \xi_1(\mathbf{s}_k) \frac{e_1(t) + e_2(t)}{\sqrt{2}} + \xi_2(\mathbf{s}_k) e_3(t), \tag{2.27}$$

where $e_1(t) = \sqrt{2} \sin(2\pi t \cdot 7)$, $e_2(t) = \sqrt{2} \sin(2\pi t \cdot 2)$, $e_3(t) = \sqrt{2} \sin(3\pi t \cdot 3)$. Direct verification, which uses the independence of the fields ξ_1 and ξ_2, shows that the FPC's are $v_1 = 2^{-1/2}(e_1 + e_2)$ and $v_2 = 2_3$ (for the parameters of the ξ_i specified below).

To complete the description of the data generating processes, we must specify the dependence structure of the scalar spatial fields ξ_1 and ξ_2. A common assumption for the Karhunen-Loéve expansions used in statistical inference is that the score processes ξ_i are independent, and this is what we assume. We use the exponential and Gaussian models (2.46) with chordal distances (2.2) between the locations described below. To make simulated data look similar to the real foF2 data we chose $\sigma_1 = 1$, $\rho_1 = \pi/6$ for $\xi_1(\mathbf{s})$ field and $\sigma_2 = 0.1$, $\rho_2 = \pi/4$ for $\xi_2(\mathbf{s})$ field.

The locations $\mathbf{s}_k$ are selected to match the locations of the real ionosonde stations. For the sample size 218 we use all available locations, shown in Figure 2.2. Size 32 corresponds to the ionosondes with the longest record history. We also consider a sample of size 100; the 100 stations were selected randomly out of the 218 stations.

Details of implementation. All methods require the specification of a parametric spatial model for the variogram. Even though for some methods the variograms are defined for L^2– or $\mathcal{S}$–valued objects, only a *scalar* model is required. In this simulation study we use the exponential and Gaussian models.

Methods M3, CM2 and CM3 require the specification of a basis $\{B_j\}$ and the number K of the basis functions. We use the Fourier basis and $K = 1 + 4[\sqrt{\#\{t_j\}}]$, where $\#\{t_j\}$ is the count of the points at which the curves are observed. For our real and simulated data $K = 1 + 4[\sqrt{336}] = 73$, a number that falls between the recommended values of 49 and 99 for the number of basis functions. Specifically, the basis functions B_j are

$$\{1, \sqrt{2}\sin(2\pi t \cdot i), \sqrt{2}\cos(2\pi t \cdot i); \quad i = 1, 2, \ldots, 36\}. \tag{2.28}$$

All methods require the estimation of a parametric model on an empirical variogram. There are several versions of the empirical variogram for scalar fields. The classical estimator proposed by Matheron is given by

$$\hat{\gamma}(d) = \frac{1}{|N(d)|} \sum_{N(d)} (X(\mathbf{s}_k) - X(\mathbf{s}_l))^2, \tag{2.29}$$

where $N(d) = \big\{(\mathbf{s}_i, \mathbf{s}_j) : d_{\mathbf{s}_i,\mathbf{s}_j} = d; i, j = 1, ..., N\big\}$ and $|N(d)|$ is the number of distinct pairs in $N(d)$. A robust estimator proposed by Cressie and Hawkins is defined as

$$\hat{\gamma}(d) = \left(\frac{1}{|N(d)|} \sum_{N(d)} |X(\mathbf{s}_k) - X(\mathbf{s}_l)|^{1/2}\right)^4 / \left(0.457 + \frac{0.494}{|N(d)|}\right). \tag{2.30}$$

For details, we refer to Section 4.4 of Schabenberger and Gotway (2005), where other ways of variogram estimation are also discussed. In our study we use only estimators (2.29) and (2.30), and refer to them, respectively, as MT and CH.

Results of the simulation study. For comparison of different methods we introduce the quantity L which is the average of the integrated absolute differences between real and estimated mean functions or FPC's. For the mean function, L is defined by

$$L = \frac{1}{R} \sum_{r=1}^{R} \int |\hat{\mu}_r(t) - \mu(t)| dt, \tag{2.31}$$

where R is the number of replications, we use $R = 10^3$. For the FPC's the definition is fully analogous. We also compute the standard deviation for L, based on the normal approximation for R independent runs.

The results of the simulations for the mean function (2.26) are shown in Figure 2.4. The data generating processes have exponential covariance functions. If the ξ_i in (2.23) have

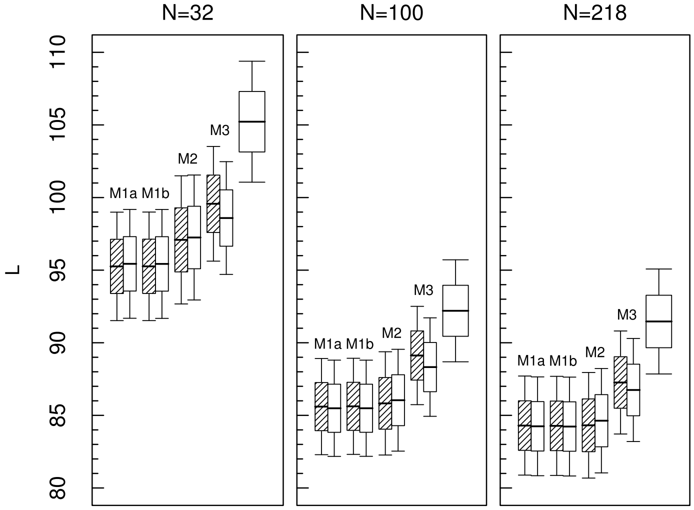

Figure 2.4: Errors in the estimation of the mean function for sample sizes: 32, 100, 218. The dashed boxes are estimates using the CH variogram, empty are for the MT variogram. The right–most box for each N corresponds to the simple average. The bold line inside each box plot represents the average value of L (2.31). The upper and lover sides of rectangles shows one standard deviation, and horizontal lines show two standard deviations. The right most boxes correspond to the standard method.

Gaussian covariances, the results are not visually distinguishable. The errors values for mean (2.25) are slightly different, but the relative position of the box plots practically does not change. All methods M1, M2 and M3 are significantly better than the sample average. Method M2 strikes the best balance between the computational cost and the precision of estimation. Note that methods M1 and M2 were designed to minimizes the expected L^2 distance, and all three methods are compared using the L^1 distance, so this comparison does not *a priori* favor them. In the context of forecasting, using different loss functions to evaluate the forecasts than to design them can lead to spurious conclusions, see Gneiting (2011). In our context, if the L^2 distance is used to compare the methods, the ranking and conclusions are the same.

Errors in the estimation of the FPC's in model (2.27) are shown in Figure 2.5. The displayed errors are those for the ξ_i with exponential covariances and the CH variogram. The results for Gaussian covariances and the MT variogram are practically the same. The

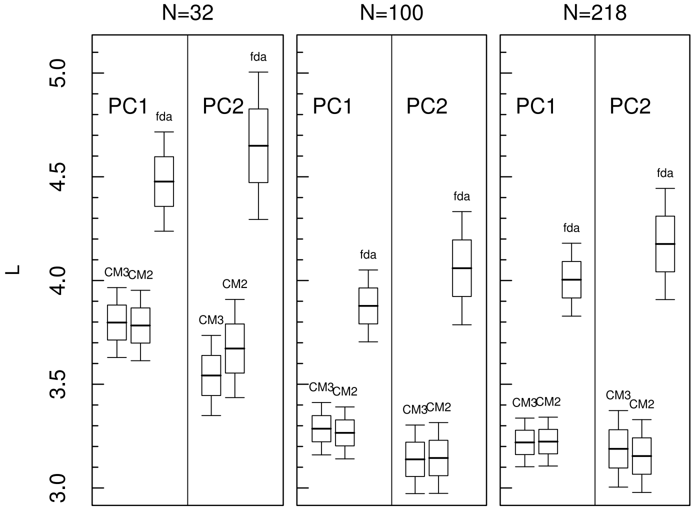

Figure 2.5: Errors in the estimation of the FPC's for sample sizes: 32, 100, 218 . The bold line inside each box plot represents the average value of L. The upper and lover sides of rectangles shows one standard deviation, and horizontal lines show two standard deviations. The right most boxes correspond to the standard method.

performance of methods CM2 and CM3 is comparable, and they are both much better than using the eigenfunctions of the empirical covariance operator (2.20), which is the standard method implemented in the `fda` package. The computational complexity of methods CM2 and CM3 is the same.

Conclusions. For simulated data generated to resemble the ionosonde data, all methods inroduced in Sections 2.2 and 2.3 have integrated absolute deviations (away from a true curve) statistically significantly smaller than the standard methods designed for iid curves. Methods M2 and CM2, based on weighted averages estimated using functional variograms, offer a computationally efficient and unified approach to the estimation of the mean function and of the FPC's in this spatial setting.

2.5 Testing for correlation of two spatial fields

Motivated by the problem of testing for correlation between foF2 and magnetic curves, described in detail in Section 2.8, we now propose a relevant statistical significance test.

There are N spatial locations: $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$. At location $\mathbf{s}_k$, we have two curves:

$$X_k = X(\mathbf{s}_k) = X(\mathbf{s}_k; t),\ t \in [0,1],$$

and

$$Y_k = Y(\mathbf{s}_k) = Y(\mathbf{s}_k; t),\ t \in [0,1].$$

We want to test the null hypothesis that the collections of curves $\{X_k,\ 1 \le k \le N\}$ and $\{Y_k,\ 1 \le k \le N\}$ are uncorrelated in a sense defined below. The null distribution is derived assuming a stronger condition that these two families are independent. Szekely *et al.* (2007) and Szekely and Rizzo (2009) introduced measures of dependence based on the distance of characteristic functions which allow to test independence (rather than just lack of correlation) of random variables X and Y given a sample of iid observations (X_k, Y_k). The extension of their theory to the case of spatially dependent observations (X_k, Y_k) is not obvious, so we consider only a test for linear dependence.

The idea of the test is as follows. To lighten the notation, assume that

$$EX_k(t) = 0 \quad \text{and} \quad EY_n(t) = 0.$$

The mean functions will be estimated and subtracted using one of the methods of Section 2.2. We approximate the curves X_n and Y_n by the expansions

$$X_n(t) \approx \sum_{i=1}^{p} \langle X_n, v_i \rangle v_i(t), \quad Y_n(t) \approx \sum_{j=1}^{q} \langle Y_n, u_j \rangle u_j(t),$$

where the v_i and the u_j are the corresponding FPC's. At this point, the functions $v_i,\ 1 \le i \le p$ and $u_j,\ 1 \le j \le q$ are deterministic, so the independence of the curves X_n of the curves Y_n implies the independence of the vectors

$$[\langle X_n, v_1 \rangle, \langle X_n, v_2 \rangle, \ldots, \langle X_n, v_p \rangle]^T, \quad 1 \le n \le N$$

and

$$[\langle Y_n, u_1\rangle, \langle Y_n, u_2\rangle, \ldots, \langle Y_n, u_q\rangle]^T, \quad 1 \le n \le N.$$

Then, under H_0, the expected value of the sample covariances

$$A_N(i,j) = \frac{1}{N}\sum_{n=1}^{N}\langle X_n, v_i\rangle\langle Y_n, u_j\rangle \tag{2.32}$$

is zero. If their estimated versions are large as a group, i.e. if some of the estimated $A_N(i,j)$ are too large, we reject the null hypothesis.

To construct a test statistic, we introduce the quantities

$$V_{k\ell}(i,i') = E[\langle v_i, X_k\rangle\langle v_{i'}, X_\ell\rangle], \qquad U_{k\ell}(j,j') = E[\langle u_j, Y_k\rangle\langle u'_j, Y_\ell\rangle].$$

Note that $V_{k\ell}(i,i') = 0$ and $U_{k\ell}(j,j') = 0$, if the observations in each sample are independent (and have mean zero). Thus, the $V_{k\ell}(i,i')$ and the $U_{k\ell}(j,j')$ are specific to dependent data, they do not occur in the currently available testing procedures developed for independent curves. Setting $X_{ik} = \langle v_i, X_k\rangle,\ Y_{jk} = \langle u_j, Y_k\rangle$, observe that if the X_{ik} are uncorrelated with the Y_{jk}, then

$$E[\sqrt{N}A_N(i,j)\sqrt{N}A_N(i',j')] = \frac{1}{N}E\left[\sum_{k=1}^{N}X_{ik}Y_{jk}\sum_{\ell=1}^{N}X_{i'\ell}Y_{j'\ell}\right]$$

$$= \frac{1}{N}\sum_{k=1}^{N}\sum_{\ell=1}^{N}E[X_{ik}X_{i'\ell}]E[Y_{jk}Y_{j'\ell}] = \frac{1}{N}\sum_{k=1}^{N}\sum_{\ell=1}^{N}V_{k\ell}(i,i')U_{k\ell}(j,j').$$

The normalized covariance tensor of the $\sqrt{N}A_N(i,j)$ thus has entries

$$\sigma_N(i,j;\ i',j') = \frac{1}{N}\sum_{k,\ell=1}^{N}V_{k\ell}(i,i')U_{k\ell}(j,j'). \tag{2.33}$$

The idea of the test, is to approximate the distribution of the matrix

$$\mathbf{A}_N = [A_N(i,j),\ 1 \le i \le p,\ 1 \le j \le q]$$

via $\sqrt{N}\mathbf{A}_N \approx \mathbf{Z}$, where $\mathbf{Z}$ is a $p \times q$ Gaussian matrix whose elements have covariances $E[Z(i,j)Z(i',j')] = \sigma_N(i,j;\ i',j')$.

We now explain how to implement this idea. Denote by $\hat{\lambda}_i, \hat{\gamma}_j$ and $\hat{v}_i, \hat{u}_j$ the eigenvalues and the eigenfunctions estimated either by method CM2 or CM3. The covariances $A_N(i,j)$ are then estimated by

$$\hat{A}_N(i,j) = \frac{1}{N}\sum_{n=1}^{N}\langle X_n, \hat{v}_i\rangle\langle Y_n, \hat{u}_j\rangle.$$

If the observations within each sample are independent, an appropriate test statistic is

$$N\sum_{i=1}^{p}\sum_{j=1}^{q}\hat{\lambda}_i^{-1}\hat{\gamma}_j^{-1}\hat{A}_N^2(i,j).$$

Since $\lambda_i = E[\langle v_i, X\rangle^2]$, this is essentially the sum of all correlations, and it tends to a chi–squared distribution with pq degrees of freedom, as shown in Kokoszka *et al.* (2008). This is however not that case for dependent data. To explain, set

$$\mathbf{a}_N = \mathrm{vec}(\mathbf{A}_N),$$

i.e. $\mathbf{a}_N$ is a column vector of length pq consisting of the columns of $\mathbf{A}_N$ stacked on top of each other, starting with the first column. Then $\sqrt{N}\mathbf{a}_N$ is approximated by a Gaussian vector $\mathbf{z}$ with covariance matrix $\mathbf{\Sigma}$ constructed from the entries (2.33). It follows that

$$\hat{S}_N = N\hat{\mathbf{a}}_N^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{a}}_N \approx \chi^2_{pq}, \tag{2.34}$$

where $\hat{\mathbf{a}}_N = \mathrm{vec}(\hat{\mathbf{A}}_N)$. The entries of the matrix $\hat{\mathbf{\Sigma}}$ are

$$\hat{\sigma}_N(i,j;\ i',j') = \frac{1}{N}\sum_{k,\ell=1}^{N}\hat{V}_{k\ell}(i,i')\hat{U}_{k\ell}(j,j'), \tag{2.35}$$

where $\hat{V}_{k\ell}(i,i')$ and $\hat{U}_{k\ell}(j,j')$ are estimators of $V_{k\ell}(i,i')$ and $U_{k\ell}(j,j')$, respectively. The test rejects H_0 if $\hat{S}_N > \chi^2_{pq}(1-\alpha)$, where $\chi^2_{pq}(1-\alpha)$ is the $100(1-\alpha)$th percentile of the chi–squared distribution with pq degrees of freedom. One can use Monte Carlo versions of the above test, for example the test is based on the approximation

$$\hat{T}_N := N\hat{\mathbf{a}}_N^T\hat{\mathbf{a}}_N \approx \mathbf{w}^T\hat{\mathbf{\Sigma}}\mathbf{w}, \tag{2.36}$$

where the components of $\mathbf{w}$ are are iid standard normal.

The test procedure can be summarized as follows:

1. Subtract the mean functions, estimated by one of the methods of section 2.2, from both samples.
2. Estimate the FPC's by method CM2 or CM3.
3. Using a model for the covariance tensor (2.35), see Section 2.6, compute the test statistic $\hat{S}_N$. (This tensor is not needed to compute $\hat{T}_N$, but it is needed to find its Monte Carlo distribution.)
4. Find the P–value using either a Monte-Carlo distribution or the χ^2 approximation.

We now turn to the important issue of modeling and estimation of the matrix $\mathbf{\Sigma}$.

2.6 Modeling and estimation of the covariance tensor

The estimation of the $V_{k\ell}(i,i')$ involves only the X_n, and the estimation of the $U_{k\ell}(j,j')$ only the Y_n, so we describe only the procedure for the $V_{k\ell}(i,i')$. We assume that the mean has been estimated and subtracted, so that we can define

$$C_h(x) = E[\langle X(\mathbf{s}), x\rangle X(\mathbf{s}+\mathbf{h})], \quad h = ||\mathbf{h}||. \tag{2.37}$$

The estimation of the $V_{k\ell}(i,i')$ relies on the identity

$$V_{k\ell}(i,i') = \langle C_h(v_i), v_{i'}\rangle, \qquad h = d(\mathbf{s}_k, \mathbf{s}_\ell).$$

To propose a practical approach to the estimation of $\mathbf{\Sigma}$, we consider an extension of the multivariate intrinsic model, see e.g. Chapter 22 of Wackernagel (2003). A most direct extension is to assume that

$$C_h = Cr(h), \tag{2.38}$$

where C is a covariance operator, i.e. a symmetric positive definite operator with summable eigenvalues, and $r(h)$ is a correlation function of a scalar random field. Since $r(0) = 1$, we have $C = C_0$, so C in (2.38) must be the the covariance operator of each $X(\mathbf{s})$. If we assume the intrinsic model (2.38), then

$$V_{k\ell}(i,j) = \langle r(h)C(v_i), v_j\rangle = \lambda_i \delta_{ij} r(d(\mathbf{s}_k, \mathbf{s}_l)). \tag{2.39}$$

To allow more modeling flexibility, we postulate that

$$V_{k\ell}(i,j) = \lambda_i \delta_{ij} r_i(d(\mathbf{s}_k, \mathbf{s}_l)). \tag{2.40}$$

Under (2.39) (equivalently, under (2.38)), each scalar field $\langle X(\mathbf{s}), v_i\rangle$ has the same correlation function, only their variances are different. Under (2.40), the fields $\langle X(\mathbf{s}), v_i\rangle$ can have different correlation functions. As will be seen below, model (2.40) also leads to valid covariance matrix.

The correlations $r_i(d(\mathbf{s}_k, \mathbf{s}_l))$ and the variances λ_i can be estimated using a parametric model for the scalar field $\xi_i(\mathbf{s}) = \langle X(\mathbf{s}), v_i\rangle$. The resulting estimates $\hat{r}_i(d(\mathbf{s}_k, \mathbf{s}_l))$ and $\hat{\lambda}_i$ lead to the estimates $\hat{V}_{k\ell}(i,j)$ via (2.40). Analogous estimates of the functional field Y are $\hat{\gamma}_j(d(\mathbf{s}_k, \mathbf{s}_l))$, $\hat{\tau}_j$ and $\hat{U}_{k\ell}(i,j)$.

For ease of reference, we note that under model (2.40) and H_0, the covariance tensor,

$$\left[\frac{1}{N}\sum_{k=1}^{N}\sum_{\ell=1}^{N} \hat{V}_{k\ell}(i,i')\hat{U}_{k\ell}(j,j'),\ 1 \le i,i' \le p,\ 1 \le j,j' \le q\right],$$

has the following matrix representation

$$\hat{\mathbf{\Sigma}} = \mathrm{diag}\left(\sum_{k=1}^{N} \sum_{\ell=1}^{N} \hat{\mathbf{\Sigma}}_{\xi_1}(k,\ell) \hat{\mathbf{\Sigma}}_{\eta_1}(k,\ell), \ \dots, \sum_{k=1}^{N} \sum_{\ell=1}^{N} \hat{\mathbf{\Sigma}}_{\xi_p}(k,\ell) \hat{\mathbf{\Sigma}}_{\eta_q}(k,\ell) \right), \tag{2.41}$$

where

$$\hat{\mathbf{\Sigma}}_{\xi_i}(k,\ell) = \frac{1}{\sqrt{N}} \hat{\lambda}_i \hat{r}_i(d(\mathbf{s}_k, \mathbf{s}_\ell))$$

and

$$\hat{\mathbf{\Sigma}}_{\eta_j}(k,\ell) = \frac{1}{\sqrt{N}} \hat{\gamma}_j \hat{\tau}_j(d(\mathbf{s}_k, \mathbf{s}_\ell)).$$

This form is used to construct the Monte Carlo tests discussed in Section 2.7.

The matrices $\mathbf{\Sigma}$ and $\widehat{\mathbf{\Sigma}}$ are positive definite. See Horváth and Kokoszka (2012) for the verification.

2.7 Size and power of the correlation test

As in Section 2.4, our objective is to evaluate the finite sample performance of the test introduced in Section 2.5 in a realistic setting geared toward the application presented in Section 2.8.

Data generating processes. We generate samples of zero mean Gaussian processes

$$X(\mathbf{s};t) = \sum_{i=1}^{p} \xi_i(\mathbf{s}) v_i(t); \qquad Y(\mathbf{s};t) = \sum_{j=1}^{q} \eta_j(\mathbf{s}) u_j(t). \tag{2.42}$$

The process X is designed to resemble in distribution appropriately transformed and centered foF2 curves; the process Y the centered magnetic curves. Following the derivation presented in Section 2.8, we use $p = 7$ and $q = 1$. The curves v_i and u_1 are the estimated FPC's of the real data. The scalar Gaussian spatial fields ξ_i and η_1 follow parametric models estimated for real data, details of the models are presented in Table 2.1. The ξ_i are independent. Under H_0, the ξ_i are independent of η_1. The dependence under H_A can be generated in many ways. We considered the following scenarios: ξ_1 and η_1 are dependent, ξ_i and η_1 are independent for $i \neq 1$, then ξ_2 and η_1 are dependent, ξ_i and η_1 are independent for $i \neq 2$, etc. To produce two dependent spatial fields ξ_i and η, we generated N iid pairs $\mathbf{x}_i = [x_{1i}, x_{2i}]^T$, $1 \leq i \leq N$, where

$$\mathbf{x}_i \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Then we merged all x_{1i} into vector $\mathbf{y}_1 = [x_{11}, ..., x_{1N}]^T$ and all x_{2i} into vector $\mathbf{y}_2 = [x_{21}, ..., x_{2N}]^T$. Performing the Cholesky rotation, we obtain correlated spatial vectors:

$$\boldsymbol{\xi}_i = \mathbf{V}\mathbf{y}_1, \ (\boldsymbol{\Sigma}_{\xi_i} = \mathbf{V}\mathbf{V}^T), \qquad \boldsymbol{\eta} = \mathbf{U}\mathbf{y}_2, \ (\boldsymbol{\Sigma}_{\eta} = \mathbf{U}\mathbf{U}^T).$$

We used sample sizes $N = 32$ and $N = 100$ corresponding to the locations determined as in Section 2.4.

Testing procedures. We studied the finite sample behavior of three methods, which we call S, SM and T. Method S rejects H_0 if the statistic $\hat{S}_N$ (2.34) exceeds a chi–square critical value. Method SM uses a Monte Carlo distribution of the statistic $\hat{S}_N$: after estimating all parameters from the data and assuming the Gaussian distribution of the fields ξ_i and η_1, we can replicate the values of the statistic $\hat{S}_N$ under H_0 using the covariance matrix (2.41). Method T uses the statistics $\hat{T}_N$ (2.36), and approximates its distribution by the Monte Carlo distribution of $\mathbf{w}^T\hat{\boldsymbol{\Sigma}}\mathbf{w}$, as explained in Section 2.5. For determining the critical values in methods SM and T, we used 10^7 Monte Carlo replications. The empirical size and power are based on 10^5 independent runs.

Conclusions. As Figure 2.6 shows, the empirical size is higher than the nominal size, and it tends to increase with the number p of principal components used to construct the test, especially for $N = 32$. The usual recommendation is to use p which explains about 85% of the variance. For the foF2 data with $N = 32$, this corresponds to $p = 4$. Applied to real data in Section 2.8, all tests (S, SM and T) lead to extremely strong rejections, so the inflated empirical size is not a problem. Figure 2.6 also shows that the Monte Carlo approximation is useful for $N = 32$, this is the sample size we must use in Section 2.8. The

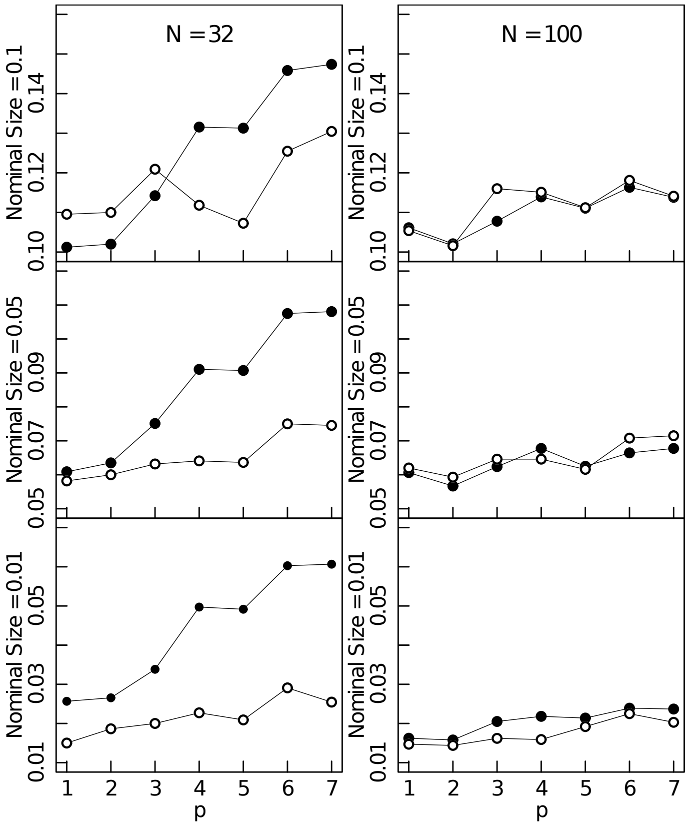

Figure 2.6: Size of the correlation test as a function of p. Solid disks represent method S (based on χ^2 distribution). Circles represent method SM (based on the Monte-Carlo distribution).

size of test T is practically indistinguishable for that of test SM. Figure 2.7 shows the power of method SM; power curves for method T are practically the same, method S has higher power. The simulation study shows that a strong rejection when the test is applied to real data can be viewed as a reliable evidence of dependence.

2.8 Application to critical ionospheric frequency and magnetic curves

In this section, we apply the correlation test, which uses the estimation methodology of Sections 2.2 and 2.3, to foF2 and magnetic curves.

Description of the data. The F2 layer of the ionosphere is the upper part of the F layer shown in Figure 2.3. The F2 layer electron critical frequency, foF2, is measured

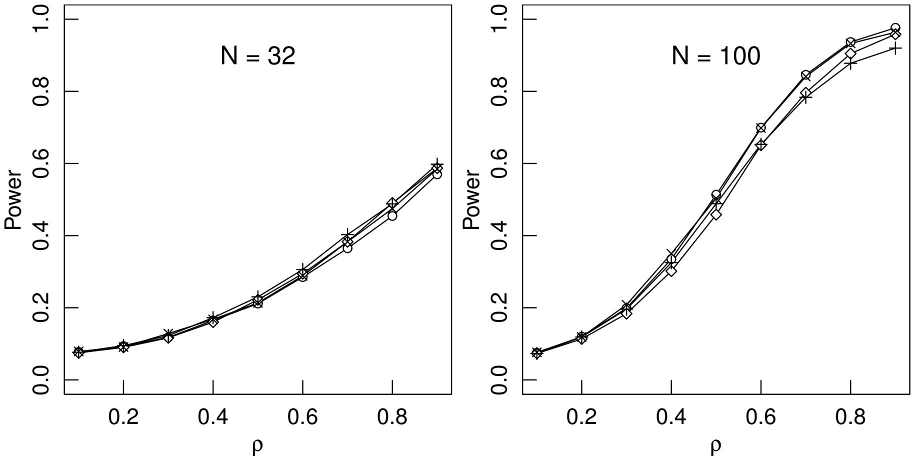

Figure 2.7: Power of the correlation test SM as a function of the population correlation ρ. Each line represents one of the four possible correlated spatial field $\boldsymbol{\xi}_1 - \boldsymbol{\eta}$, $\boldsymbol{\xi}_2 - \boldsymbol{\eta}$, $\boldsymbol{\xi}_3 - \boldsymbol{\eta}$, $\boldsymbol{\xi}_4 - \boldsymbol{\eta}$. The test was performed using $p = 4$, which explains about 85% of variance of the foF2 curves. Since all curves in the graphs are practically the same, we do not specify which curve represents a particular dependent pair $\boldsymbol{\xi}_i - \boldsymbol{\eta}$.

using an instrument called the ionosonde, a type of radar. The foF2 frequency is used to estimate the location of the peak electron density, so an foF2 trend corresponds to a trend in the average height of the ionosphere over a spatial location. The foF2 data have therefore been used to test the hypothesis of ionospheric global cooling discussed in the introduction. Hourly values of foF2 are available from the SPIDR database `http://spidr.ngdc.noaa.gov/spidr/` for more than 200 ionosondes. We use monthly averages for 32 selected ionosondes, with sufficiently complete records, for the period 1964 – 1992. Their locations are shown in Figure 2.2. Three typical foF2 curves are shown in Figure 2.1. We omit the details of the procedure for obtaining curves like those shown in Figure 2.1, but we emphasize that it requires a great deal of work. In particular, the SPIDR data suffer from two problems. First, for some data, the amplitude is artificially magnified ten times, and needs to be converted into standard units (MHz). Second, in many cases, missing observations are not replaced by the standard notation 9999, but rather just skipped. Thus if one wants to use equally-spaced time series, skipped data must be found and replaced by missing values. For filling in missing values, we perform linear interpolation. We developed

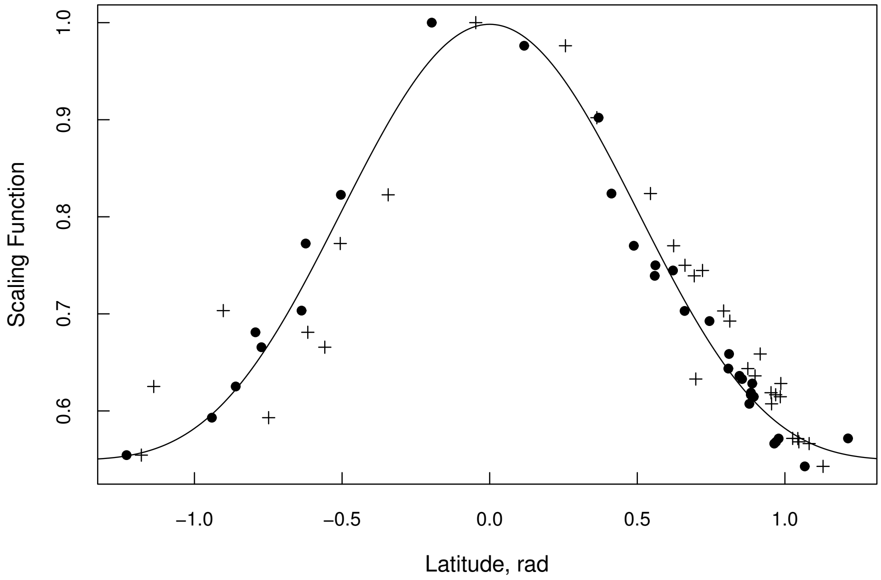

Figure 2.8: Dots represent the scaling function $G_L(\mathbf{s}_i)$ in the magnetic coordinate system and crosses are same in the geographic coordinate system. Line is the best fit for G_L in the magnetic coordinate system.

a customized C++ code to handle these issues. We emphasize that one of the reasons why this global data set has not been analyzed so far is that useable data have been derived only over relatively small regions, e.g. Western Europe, and more often for a single location.

As explained in Section 2.1, the foF2 data are used to test hypotheses on long term ionospheric trends. We thus removed annual and higher frequency variations using 16 month averaging with MODWT filter, see Chapter 5 of Percival and Walden (2000). This leads to 32 time series at different locations, each containing 336 equally-spaced temporal observations. The amplitude of the foF2 curves exhibits a nonlinear latitudinal trend; it decreases as the latitude increases, see Figure 2.1. To remove this trend, which may potentially bias the test, we assume that the foF2 signal, $F(\mathbf{s};t)$, at location $\mathbf{s}$ follows the model

$$F(\mathbf{s};t) = G(L(\mathbf{s}))X(\mathbf{s};t), \tag{2.43}$$

where $X(\mathbf{s};t)$ is a constant amplitude field, and $G(\cdot)$ is a scaling function which depends only on the *magnetic* latitude L (in radians). Since the trend in the amplitude of $F(\mathbf{s};t)$

is caused by the solar radiation which is nonlinearly proportional to the zenith angle, we postulate that the function $G(\cdot)$ has the form

$$G(L) = a + b\cos^c(L). \tag{2.44}$$

The parameters a, b, c are estimated as follows. Let $\mathbf{s}_0$ be the position of the ionosonde closest to the magnetic equator. For identifiability , we set $G(L(\mathbf{s}_0)) = 1$. For the remaining locations $\mathbf{s}_k$, we compute $\hat{G}(L(\mathbf{s}_k))$ as the average, over all 336 time points t_j of the ratio $F(\mathbf{s}_k; t_j)/F(\mathbf{s}_0; t_j)$. Figure 2.8 shows these ratios as a function of the magnetic and geographic latitude. The ratios in the magnetic latitude show much less spread, and this is another reason why we work with the magnetic latitude. The curve $G(L)$ (2.44) is fitted to the $\hat{G}(L(\mathbf{s}_k))$ in magnetic latitude by nonlinear least squares. The fitted values are $a = 0.5495$, $b = 0.4488$, $c = 4.2631$.

We now describe how we construct the curves that reflect the relevant long term changes in the internal magnetic field of the earth. The height of the F2 layer (and so the foF2 frequency) can be affected by a vertical plasma drift which responds to the magnetic field. The vertical plasma drift is due to the wind effect, and is given by (we use the same notations as in Mikhailov and Marin (2001))

$$W = (V_{nx}\cos D - V_{ny}\sin D)\sin I\cos I + V_{nz}\sin^2 I.$$

In the above formula, V_{nx} , V_{ny} and V_{nz} are, respectively, meridional (parallel to constant longitude lines), zonal (parallel to constant latitude lines) and vertical components of the thermospheric neutral wind; I and D are inclination and declination of the earth magnetic field. Detailed figures are provided in Chapter 13 of Kivelson and Russell (1997). Usually $V_{nz} \ll V_{nx}, V_{ny}$, and assuming that the difference between magnetic and geographic coordinates, D, is small (at least for low- and mid-latitude regions) we can simplify the above formula to $W = V_{nx}\sin I\cos I$. Thus, only the meridional thermospheric wind is significant. Measuring neutral wind components (V_{nx}, V_{ny}, V_{nz}) is difficult, and long term wind records are not available. We therefore replace V_{nx} by its average. For the test of correlation, the specific value of this average plays no role, so we define the magnetic curves as

$$Y(\mathbf{s};t) = \sin I(\mathbf{s};t)\cos I(\mathbf{s};t). \tag{2.45}$$

The curves $I(\mathbf{s};t)$ are computed using the international geomagnetic reference field (IGRF); the software is available at `http://www.ngdc.noaa.gov/IAGA/vmod/`.

The test is applied to the curves $X(\mathbf{s}_k;t)$ defined by (2.43) and (2.44), and to the curves $Y(\mathbf{s}_k;t)$ defined by (2.45).

Application of the correlation test. We first estimate and subtract the mean functions of the fields $X(\mathbf{s}_k)$ and $Y(\mathbf{s}_k)$ using method M2 (the other spatial methods give practically the same estimates). The principal components v_i and u_i are estimated using method CM2 (method CM3 gives practically the same curves).

We apply the test, for all $1 \le p \le 7$ and $q = 1$. The first seven eigenvalues of the field X (computed per (2.17) or its analog for method CM2) explain about 95% of the variance. The first eigenvalue of the field Y explains about 99% of the variance. The eigenfunction u_1 is approximately equal to the linear function: $u_1(t) \sim t$. This means that at any location, after removing the average, the magnetic field either linearly increases or decreases, with slopes depending on the location, see Figure 2.9. To lighten the notation, we drop the "hats" from the estimated scores and denote the zero mean vector $[\xi_i(\mathbf{s}_1), ..., \xi_i(\mathbf{s}_N)]^T$ by $\boldsymbol{\xi}_i$, and $[\eta_1(\mathbf{s}_1), ..., \eta_1(\mathbf{s}_N)]^T$ by $\boldsymbol{\eta}$ The covariances $\boldsymbol{\Sigma}_{\xi_i}$ and $\boldsymbol{\Sigma}_{\eta}$ are estimated using parametric spatial models determined by the inspection of the empirical variograms. In this application, it is sufficient to use two covariance models:

$$\begin{aligned} &\text{Gaussian}: && c(\mathbf{s}_k,\mathbf{s}_\ell) = c_0 + \sigma^2\exp\{-d^2(k,\ell)/\rho^2\},\\ &\text{Exponential}: && c(\mathbf{s}_k,\mathbf{s}_\ell) = c_0 + \sigma^2\exp\{-d(k,\ell)/\rho\}. \end{aligned} \tag{2.46}$$

When the scores do not have a spatial structure, we use the sample variance (flat variogram). The estimated models and their parameters are listed in Table 2.1.

The P–values for different number of FPC's $1 \le p \le 7$ are summarized in Table 2.2.

Table 2.1: Models and estimated covariance parameters for the transformed foF2 curves and the magnetic curves.

Spatial field	Model	Parameters		
		c_0	σ^2	ρ
$\boldsymbol{\eta}$	Gaussian	–	5.99 ± 0.48	0.32 ± 0.04
$\boldsymbol{\xi}_1$	Gaussian	–	20.05 ± 2.20	0.12 ± 0.03
$\boldsymbol{\xi}_2$	–	–	3.30 ± 0.43	–
$\boldsymbol{\xi}_3$	Exponential	–	2.63 ± 0.52	0.16 ± 0.07
$\boldsymbol{\xi}_4$	Gaussian	–	2.66 ± 0.39	0.18 ± 0.05
$\boldsymbol{\xi}_5$	–	–	2.74 ± 0.32	–
$\boldsymbol{\xi}_6$	Gaussian	0.16 ± 0.02	0.85 ± 0.24	0.17 ± 0.06
$\boldsymbol{\xi}_7$	–	–	1.22 ± 0.18	–

Table 2.2: P–values of the correlation tests applied to the transformed foF2 data. The first column shows the number of FPC's, the second column shows cumulative variances computed as the ratios of the eigenvalues estimated using method CM2. Testing procedures S, SM and T are defined in Section 2.7. The "simple" procedure neglects the spatial dependence of the curves.

p	CV, %	Spatial			Simple
		S	SM	T	
1	47.88	$6.22 \cdot 10^{-5}$	$3.05 \cdot 10^{-4}$	$3.05 \cdot 10^{-4}$	0.035
2	62.59	$3.26 \cdot 10^{-6}$	$2.91 \cdot 10^{-4}$	$2.99 \cdot 10^{-4}$	0.095
3	73.67	$4.53 \cdot 10^{-8}$	$2.43 \cdot 10^{-4}$	$2.32 \cdot 10^{-4}$	0.043
4	84.40	$1.47 \cdot 10^{-26}$	$1.6 \cdot 10^{-7}$	$2.24 \cdot 10^{-5}$	0.039
5	88.70	$4.95 \cdot 10^{-26}$	$2.6 \cdot 10^{-7}$	$2.27 \cdot 10^{-5}$	0.046
6	92.21	$6.73 \cdot 10^{-27}$	$5.9 \cdot 10^{-7}$	$2.21 \cdot 10^{-5}$	0.060
7	94.57	$2.12 \cdot 10^{-32}$	$1.6 \cdot 10^{-7}$	$1.92 \cdot 10^{-5}$	0.030

Independent of p and a specific implementation of the test, all P–values are very small, and so the rejection of the null hypothesis is conclusive; we conclude that there is a statistically significant correlation between the foF2 curves $X(\mathbf{s}_k)$ and the magnetic curves $Y(\mathbf{s}_k)$. We also applied a version of our test which neglects any spatial dependence, this is the test proposed by Kokoszka *et al.* (2008). The P-values hover around the 5% level, but still point toward rejection. The evidence is however much less clear cut. This may partially explain why this issue has been a matter of much debate in the space physics community. The correlation between the foF2 and magnetic curves is far from obvious. Figure 2.9 shows these pairs at all 32 locations. It is hard to conclude by eye that the direction of the magnetic field change impacts the foF2 curves.

Discussion. A very important role in our analysis is played by the transformation (2.44). Applying the test to the original foF2 curves $F(\mathbf{s}_k; t)$, gives the P–values 0.209 ($p = 1$) and 0.011 ($p = 2$) for the spatial S test, and 0.707 ($p = 1$), 0.185 ($p = 2$), 0.139 ($p = 3$) for the "simple" test. As explained above, the amplitude of the field $F(\mathbf{s}_k; t)$ evolves with the latitude. This invalidates the assumption of a mean function which is independent of the spatial location. Thus even for the spatial test, the mean function confounds the first FPC. However, the spatial estimation of the mean function and of the FPC's "quickly corrects" for the violation of assumptions, and the null hypothesis is rejected for $p \geq 2$. When the spatial structure is neglected (and no latitudal transformation is applied) no correlation between the foF2 curves and magnetic curves is found.

The rejection of the null hypothesis means that after adjusting the foF2 curves for the latitude and the global mean, their regional variability is correlated with the regional changes in the magnetic field. This conclusion agrees with recent space physics research, see Cnossen and Richmond (2008) and Lastovicka (2009), and can, to some extent, be visually confirmed, post–analysis, by the examination of the scatter plots shown in Figure 2.10. It implies that long term magnetic trends must be considered as additional covariates in testing for long term trends in the foF2 curves. The main covariate is the solar activity which drives the shape of the mean function, but, as explained in the introduction, the impact of the concentration of the greenhouse gases is of particular interest, see Qian *et al.* (2009) among many other contributions.

A broader conclusion of the work presented in this paper is that methods of functional data analysis must be applied with care to curves obtained at spatial locations. Neglecting the spatial dependence can lead to incorrect conclusions and biased estimates. The same applies to space physics research. If trends or models are estimated separately at each spatial location, one should not rely on results obtained by some form of a simple averaging. This is however the prevailing approach. Interestingly, the results related to global ionospheric trends are often on the borderline of statistical significance. Standard t-tests lead either to rejection or acceptance, depending on a specific method used (a similar phenomenon is observed in the last column of Table 2.2). It is hoped that the methodology developed in this paper will be useful in addressing such issues.

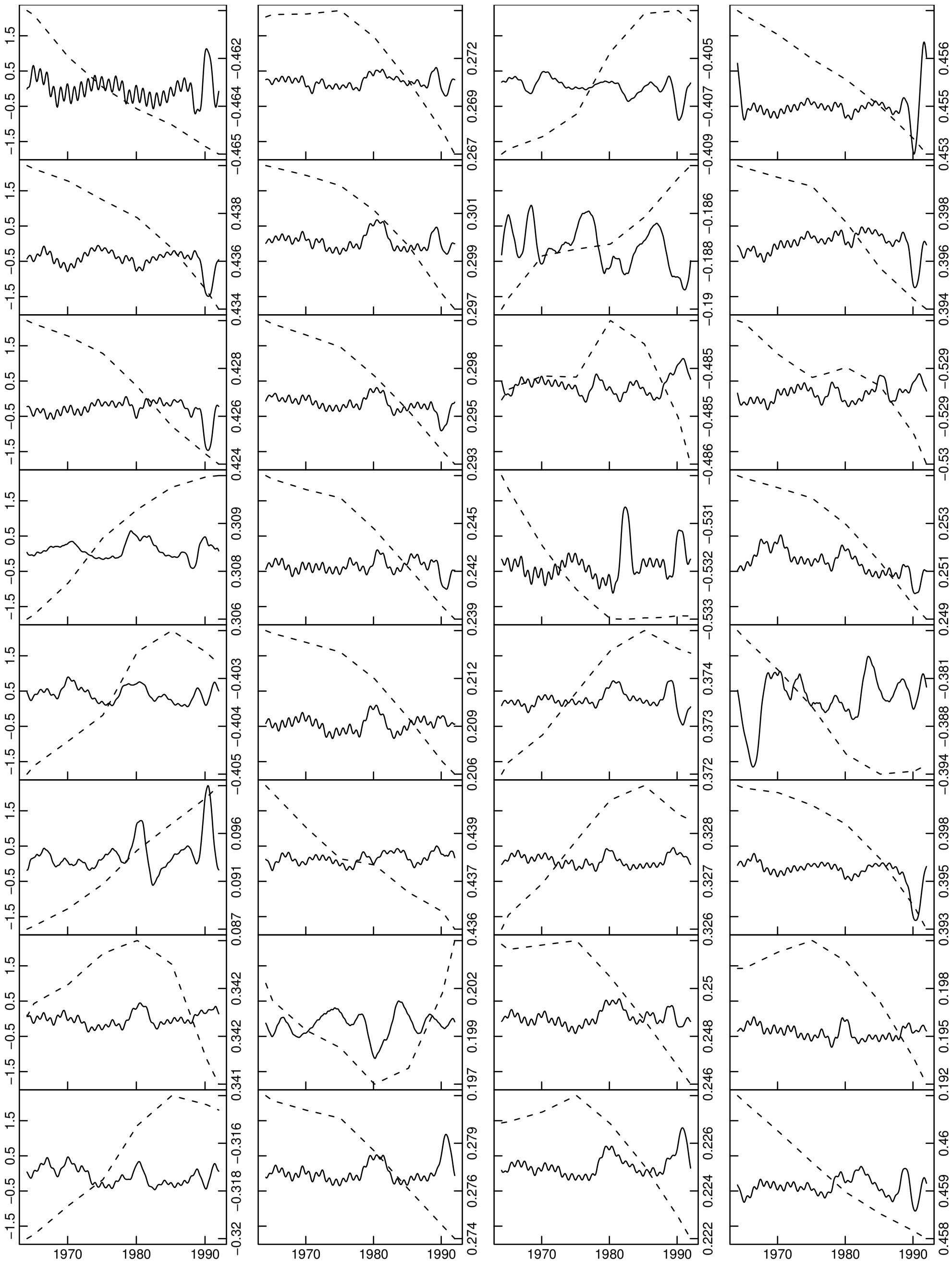

Figure 2.9: Transformed and centered foF2 curves (continuous) and centered magnetic curves (dashed) at 32 locations denoted with circles in Figure 2.2. The scales for the two families of curves are different. The foF2 curves have the same scale, it is shown on the right vertical axes in MHz. The scale of the magnetic curves changes, it is shown on the right vertical axes in each box (unitless).

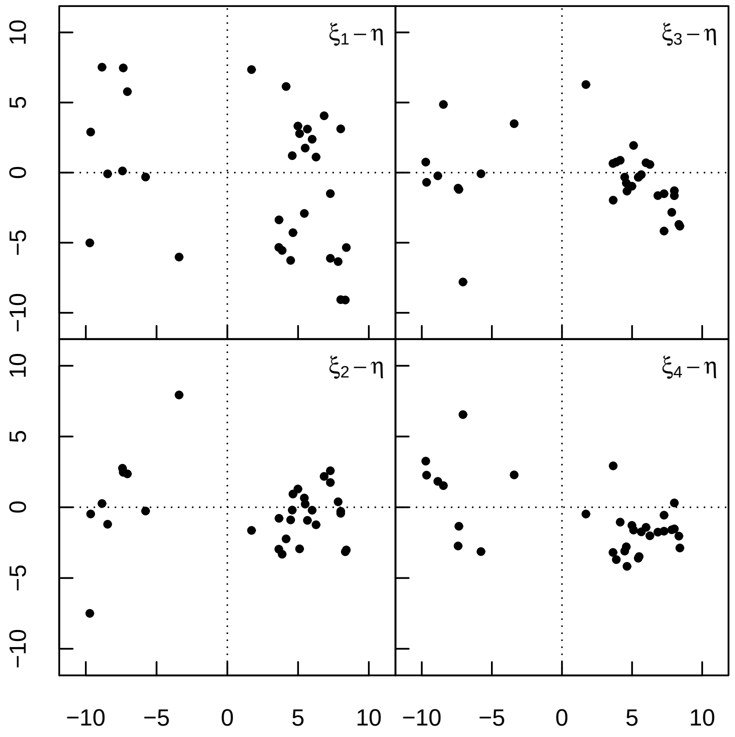

Figure 2.10: Scatter plots of the scores $\boldsymbol{\xi}_i, i = 1, 2, 3, 4$ of the foF2 curves, vertical axes, against the scores $\boldsymbol{\eta}$ of the magnetic curves, horizontal axes.

CHAPTER 3

TESTING THE EQUALITY OF MEAN FUNCTIONS OF IONOSPHERIC CRITICAL FREQUENCY CURVES[1]

Abstract

This paper develops a significance test for evaluating the equality of the mean functions in two samples of spatially indexed functional data. The problem is motivated by an important question in space physics research which is related to the hypothesis of ionospheric global cooling (as opposed to the conjectured global warming of near surface atmosphere). The critical electron frequency of the ionosphere's F2 region, foF2, can be used to empirically test conjectures about the trends in the ionosphere. We apply the proposed test to the foF2 records over eastern and western Europe to verify if there exists a conjectured difference between first order behavior of these records over regions with different magnetic inclinations. It is found that the difference between the means is statistically significant for night time records. The implications of this result are discussed. Finite sample performance of the proposed test is validated via numerical simulations.

3.1 Introduction

Over the last two decades, functional data analysis has established itself as an important and dynamic area of statistics which has provided intuitive and computationally feasible approaches to many applied problems. The monograph of Ramsay and Silverman (2005) offers an excellent and accessible introduction to the central ideas of the field. Ramsay *et al.* (2009) provide a concise introduction which focusses on computational issues. Many recent developments are studied in the books of Ferraty and Vieu (2006), Bosq and Blanke (2007), Ferraty and Romain (2011), Shi and Choi (2011) and Horváth and Kokoszka

---

[1] CO-AUTHORED BY O. GROMENKO AND P. KOKOZKA. REPRODUCED FROM THE JOURNAL OF THE ROYAL STATISTICAL SOCIETY: SERIES C, VOL. 61, 715–731, 2012.

(2012). Relatively little attention has however focused on inferential methods for spatially indexed curves, even though data of this type are quite common; a data object is a curve $X(\mathbf{s}_k)$ observed at location $\mathbf{s}_k$. In many cases, the curves are functions of time so that $X(\mathbf{s}_k;t)$ is the value of the function $X(\mathbf{s}_k)$ at time t. Such a data structure is a special case of a spatio–temporal process which calls for statistical tools specifically designed for it. Many environmental and geophysical data sets are of this type. A well-known example is the Canadian weather data set which consists of temperature and precipitation curves at 35 locations, see Ramsay and Silverman (2005). Another example is the Australian rainfall data set, recently studied by Delaigle and Hall (2010), which consists of daily rainfall measurements from 1840 to 1990 at 191 Australian weather stations. Snow water curves measured at several dozen locations in Western US states over many decades provide climatic information which is of importance for urban and agricultural development planning in the Western states with little summer time rainfall, see e.g. Carroll *et al.* (1995) and Carroll and Cressie (1996). Another important example is pollution curves: $X(\mathbf{s}_k;t)$ is the concentration of a pollutant at time t at location $\mathbf{s}_k$. Data of this type were studied by Kaiser *et al.* (2002). In many studies, $X(\mathbf{s}_k;t)$ is the count at time t of infectious disease cases, where $\mathbf{s}_k$ is a representative location, e.g. a “middle point” of a county. Still another example arises in modeling brain activity based on continuous time records obtained from probes placed at different locations in the brain, see Aston and Kirch (2012) and references therein. Delicado *et al.* (2010) review other examples and recent contributions to the methodology for spatially indexed functional data.

The data which motivate the research presented in this paper are not well–known in the statistical community, but have played a central role in space physics research for many decades. Since the 1930’s the ionosphere has been studied by an instrument called the ionosonde, which is a type of radar vertically emiting a frequency spectrum and recording the profile of the returned signal. The returned profile is called the ionogram, an example is given in Fig. 3.1. The ionogram contains implicit information about the physical properties of the ionosphere directly above the location of the ionosonde. The specific data set we

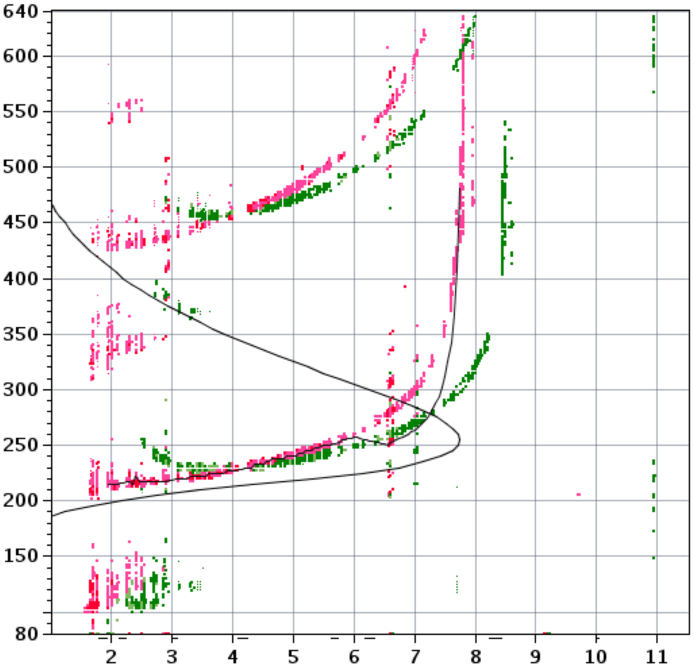

Figure 3.1: An example of a digital ionogram recorded at Juliusruh ionosonde. The vertical axis shows height in kilometers and the horizontal axis shows frequency in megahertz. The pink and green dots show returned frequencies and their virtual heights. The frequency at which the virtual height tends to infinity is called the electron critical frequency. The black "bell shaped" curve is restored profile of ionosphere, cf. Fig 3.2, and the rightmost point of this curve is the electron critical frequency.

study, derived from ionograms, consists of hourly records of the so-called electron critical frequency in the F2 ionosphere region, foF2, at 13 locations in Europe, we describe this data set in detail in Section 3.2. The F2 region is the main part of the ionosphere's F region in the range of heights 250–350 km above the sea level, see Fig. 3.2. For a brief overview of the structure of the ionosphere and its properties see, for example, Chapter 1 in Kelly (2009).

To describe the specific problem studied in this paper, we must provide some broader background. The increased concentration of greenhouse gases in the upper atmosphere has been associated with the global warming in the lower troposphere. Roble and Dickinson (1989) suggested that the increasing amounts of these radiatively active gases would lead

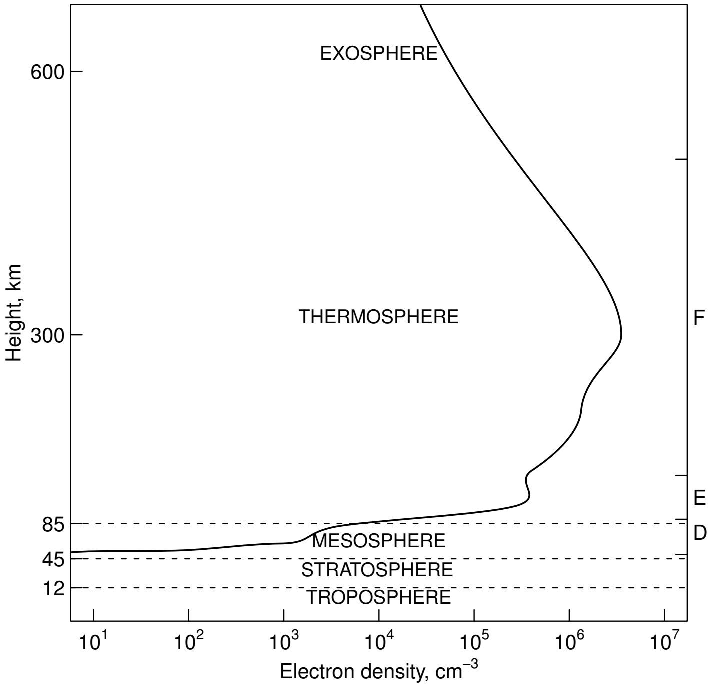

Figure 3.2: Typical profile of ionosphere. The curve shows electron density as a function of height. The right vertical axis indicates the D, E and F regions of the ionosphere.

to global cooling in the mesosphere and thermosphere. Shortly afterwards, Rishbeth (1990) pointed out that this would result in a thermal contraction and the global lowering of the ionospheric peak densities. The peak density height of the F2 region can be approximately computed using the critical frequency foF2. Thus, if the hypothesis of Roble and Dickinson (1989) were true, cooling of the ionosphere would results in a systematic global change of foF2 which, in space physics research, is referred to as a global foF2 trend. It initially appeared that the evaluation of such a trend in the ionosphere might be easier than the evaluation of a global warming trend in the near surface atmosphere which exhibits a very strong local variability, making a definition of a global warming trend much less obvious. Consequently, during the last two decades, trends in foF2 (defined in several ways) have been extensively studied by many authors, Lastovicka *et al.* (2006) and Lastovicka (2009) offer reviews of the relevant literature. In spite of this great interest and extensive research, the nature of the trends in the upper ionosphere is not yet fully understood, and there is no universal agreement on the existence of a global trend. A central issue is that trends over some regions appear to be upward, and over other regions to be downward. These trend estimates have been obtained by regression methods not fully justified and validated for the spatially distributed time series data, but they prompted explanations different from the global contraction hypothesis of Rishbeth (1990). In particular, Danilov and Mikhailov (1999) and Mikhailov and Marin (2000) argued that the trends in upper ionosphere can be related to changes in geomagnetic activity. A different explanation is that the trends could be due to long term changes in the internal magnetic field of the Earth, see Foppiano *et al.* (1999). Recently Elias (2009) pointed out that it is more likely that observed trends in foF2 are a combined result of the influence of several factors. The precise understanding of the influence of different factors is needed to make a reliable conclusion about the origin and significance of the observed trends, but because of the extremely complex nature of the Earth's ionosphere, separation of the influence of different factors from physical perspective is almost impossible. For this reason, a reliable statistical analysis is needed. In particular, it is important to determine if the trends over two regions are indeed different, or if the apparent differences are spurious and stem from using regression methods not suitable for the foF2 data. The present paper makes a contribution in this direction by proposing a new methodology.

Dependence makes the study of spatially indexed functional data different from the more common analysis of functional objects. Most methodological and theoretical developments in functional data analysis have been motivated by data obtained from designed experiments in which functional observations on subjects can be treated as independent. In this paper, we carefully take into account the spatial dependence between the curves to develop a significance test for testing if the mean curves over two disjoint regions are different. We apply the new test to establish if there is a difference in the mean functions of foF2, due to the Earth's magnetic field. We analyze foF2 data recorded over eastern and western Europe. The inclination of the Earth's magnetic field is positive in the eastern part and negative in the western part of Europe. We use this property to separate Europe into two regions.

While there is extensive literature on comparison of spatial fields in different settings such as medical imaging and forecasts, see for example Hering and Genton (2011) and Gilleland *et al.* (2009) and references therein, in functional data analysis spatial dependence is usually ignored. A test for the equality of the mean functions in two independent samples of independent curves is proposed in Horváth and Kokoszka (2012). It is extended to time series data in Horváth *et al.* (2013). These are asymptotic tests whose test statistic is a quadratic form constructed from estimated variance operators (long–run variances for time series data). The test statistic is asymptotically chi–square distributed under the null hypothesis of equal mean functions. An approach of this type is in principle possible for spatially correlated curves, but it gives very poor results in small samples. We therefore pursue a different approach. Very small sample sizes combined with spatial dependence are two main concerns we must tackle.

The paper is organized as follows. In Section 3.2, we provide a detailed description of the foF2 data. Section 3.3 introduces the required notation, formalizes the problem, and proposes a statistical model for the data. In Section 3.4, we introduce an iterative method for the estimation of mean functions and of the covariance matrix. With these preliminaries, we construct in Section 3.5 the test statistic and study its properties. This allows us to apply the test to the foF2 data in Section 3.6 and arrive at well supported conclusions. Section 3.7 contains an algorithmic description of one of the estimation procedures used in the paper.

3.2 Description of the ionosonde data

The foF2 data are collected by a global network of ionosonde stations which consist of over two hundred observatories, but only a few dozen have operated for sufficiently long periods of time to provide data useful to study long term ionospheric trends. The complete raw records are stored at the National Oceanic and Atmospheric Administration web site `http://spidr.ngdc.noaa.gov/spidr/`. The quality of the raw data from most stations is poor. The main problem is that even if the measurements are listed as equidistant, in

Table 3.1: Summary information for the ionosonde stations.

Code	Location	Latitude,^{0}	Longitude,^{0}	Magnetic inclination
AZ136	Arkhangelsk	64.4	40.5	+
DB049	Dourbes	50.1	4.6	−
JR055	Juliusruh	54.6	13.4	+
KI167	Kiruna	67.8	20.4	+
KL154	Kaliningrad	54.7	20.6	+
LN047	Lannion	48.8	−3.4	−
LY164	Lycksele	64.7	18.8	+
MO155	Moscow	55.5	37.3	+
PQ052	Pruhonice	50.0	14.6	+
PT046	Poitiers	46.6	0.3	−
SL051	Slough	51.5	−0.6	−
UP158	Uppsala	59.8	17.6	+
US057	South Uist	57.4	−7.3	−

reality they are not. Some absent data are flagged as missing, and some are not. The gaps range from several hours to several months, and must be identified and filled. The second problem is that in December the foF2 signal is artificially increased ten times. Because of the amount of observations (over seventy thousand in eight years per curve (station)), cleaning must be done algorithmically. We developed a customized C++ code which cleans the foF2 data and performs interpolation; the code is available upon request. We suppose that the absence of sufficiently long records of clean data at many locations has been a practical major obstacle in studying the global foF2 trend. We hope that our code will be useful to the space physics community.

We selected 13 ionosonde stations with hourly records of the foF2 starting from January 1972 and ending in December 1980. All stations are located in Europe. The main criterion used for ionosonde station selection was the lack of long periods of missing observations. The selected ionosonde stations are listed in Table 3.1. For ease of reference, we also provide a map of Europe with the locations of the ionosonde stations, see Fig. 3.3. The time interval was selected to match two criteria: maximum data availability and coverage of a solar minimum period. The last criterion is needed to reduce latitudinal variability of the foF2 signal due to the solar activity. For further reduction of the effects due to the solar activity, we calculated nightly sample averages of data recorded between 22 and 2 LT at each day. All such averages are calculated based on 3-5 consecutive temporal observations. This results in one data point per day. Finally, we used the maximum overlap discrete wavelet transform to smooth the noisy curves, see Chapter 5 of Percival and Walden (2000). We used filter length roughly corresponding to averaging over 32 days. The data curves thus prepared and transformed are shown in Fig. 3.4, which also shows the estimated mean functions for the negative and positive inclination samples. We call such smooth nightly averages ionosonde data. The question we answer in this paper is whether the two mean curves (for positive and negative inclination), whose estimates are shown in Fig. 3.4, are significantly different.

3.3 A functional spatio–temporal model for the ionosonde data

This section introduces the requisite Hilbert space framework within which the testing problem can be clearly stated. Following the usual approach adopted in functional data analysis, we assume that the smoothed foF2 curves belong to the Hilbert space of square integrable functions on the interval $[0, 1]$, which is the rescaled time interval from January 1, 1972 to December 31, 1980. This space is denoted $L^2 = L^2([0, 1])$, and is equipped with the inner product $\langle f, g\rangle = \int f(t)g(t)dt$ and the norm $\|f\|^2 = \int f^2(t)dt$. The curve at location $\mathbf{s}$ is denoted $X(\mathbf{s})$, and the value at time $t \in [0, 1]$ is denoted by $X(\mathbf{s}; t)$. We treat the function $X(\mathbf{s})$ as a random element of the space L^2.

We consider a spatial domain $\mathcal{D}$ which is separated into two disjoint regions $\mathcal{D}_1$ and $\mathcal{D}_2$; $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$. Let $\mathbf{s}^{(1)} \in \mathcal{D}_1$ and $\mathbf{s}^{(2)} \in \mathcal{D}_2$ denote the generic locations in these two regions, and $\mathbf{s}_k^{(1)}$ and $\mathbf{s}_\ell^{(2)}$ denote the spatial locations in regions $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively, at which data curves are available. The number of observation in region $\mathcal{D}_j$ is N_j. We reserve indexes $j, j' = 1, 2$ to specify the region.

We assume that the random functions in the two regions may differ only in the mean function, so the spatio–temporal model for the data is

$$X(\mathbf{s}^{(j)}; t) = \mu_j(t) + \varepsilon(\mathbf{s}^{(j)}; t), \quad j = 1, 2, \tag{3.1}$$

where $\varepsilon(\cdot)$ is a strictly stationary zero mean random field in L^2. To define the covariance function, we must assume that

$$E\|\varepsilon(s)\|^2 = E\int \varepsilon^2(\mathbf{s};t)dt < \infty.$$

It follows that the expectations $E[\langle\varepsilon(\mathbf{s}),\varepsilon(\mathbf{s}')\rangle]$ exist, depend only on the distance between $\mathbf{s}$ and $\mathbf{s}'$, and the distribution of $\varepsilon(\mathbf{s})$ does not depend on $\mathbf{s}$.

In this paper, the distance $d_{k,\ell}$ between two points $\mathbf{s}_k, \mathbf{s}_\ell$ is a chordal distance defined as

$$d_{k,\ell} = 2\left[\sin^2\left(\frac{L_k - L_\ell}{2}\right) + \cos L_k \cos L_\ell \sin^2\left(\frac{l_k - l_\ell}{2}\right)\right]^{1/2}, \tag{3.2}$$

where L denotes the latitude and l the longitude. The reason for using the chordal distance is that any spatial covariance functions in $\mathbb{R}^3$ restricted to the unit sphere is then also a covariance function on the sphere.

In this framework, the testing problem can be formulated as

$$H_0 : \mu_1(t) = \mu_2(t),\ (\|\mu_1 - \mu_2\|^2 = 0);$$
$$H_A : \mu_1(t) \neq \mu_2(t),\ (\|\mu_1 - \mu_2\|^2 > 0).$$

In the remainder of this section we tighten our assumptions in a way that makes a construction of a test possible. The functional error field admits the Karhunen-Loève expansion

$$\varepsilon(\mathbf{s};t) = \sum_{i=1}^{\infty} \xi_i(\mathbf{s})v_i(t),$$

in which the functional principal components (FPC's) v_i are unknown L^2 valued parameters. Later in the paper we truncate the infinite sum by taking into account the first p summands which capture the desired level of the variance. Let $\boldsymbol{\xi}_i$ be a column vector comprised of the scores $\xi_i(\mathbf{s}_k)$, $\boldsymbol{\xi}_i = [\xi_i(\mathbf{s}_1^{(1)}),\ldots,\xi_i(\mathbf{s}_{N_1}^{(1)}),\xi_i(\mathbf{s}_1^{(2)}),\ldots,\xi_i(\mathbf{s}_{N_2}^{(2)})]^T$. Below we use two assumptions: the vector fields $\boldsymbol{\xi}_i$ are normal, $\boldsymbol{\xi}_i \sim N(\mathbf{0},\boldsymbol{\Gamma}_i)$, where $\boldsymbol{\Gamma}_i = \mathrm{Var}(\boldsymbol{\xi}_i)$ and the fields $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_{i'}$ are independent if $i' \neq i$.

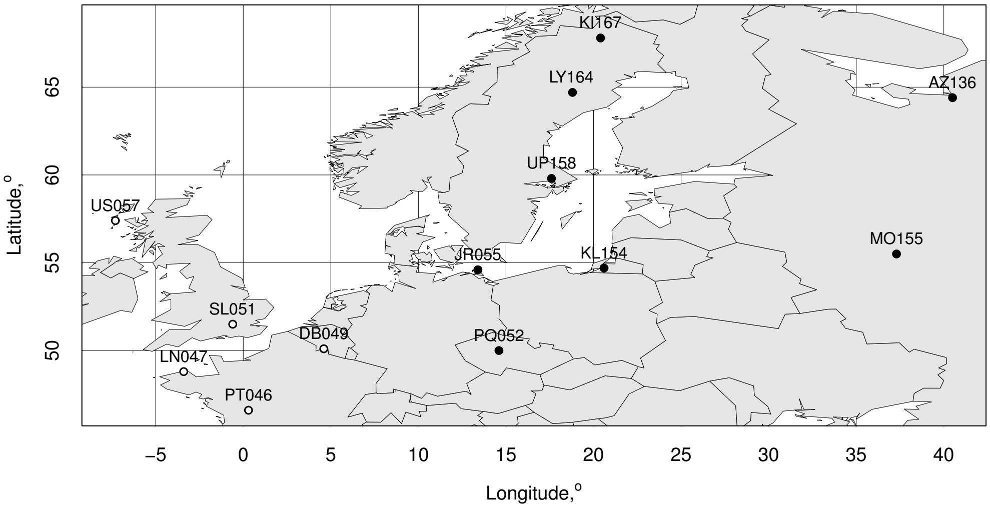

Figure 3.3: Locations of 13 selected ionosonde stations with the corresponding 5 letter codes. Empty circles represent stations with negative magnetic inclination and solid discs represent stations with positive magnetic inclination.

The assumption of normality can be verified empirically. The QQ plots in Fig. 3.5 do not contradict it in any obvious way. The assumption of the independence of score functions is required to construct a computable test statistic, and is needed for most inferential procedures, Delaigle and Hall (2010) discuss this point. The scores at the same location are always uncorrelated, i.e. $E[\xi_i(\mathbf{s})\xi_{i'}(\mathbf{s})] = 0$, if $i' \neq i$. Our assumptions imply that $E[\xi_i(\mathbf{s})\xi_{i'}(\mathbf{s}')] = 0$, for arbitrary locations $\mathbf{s}$ and $\mathbf{s}'$.

The above development shows that in order to construct a useful test statistic, we must address the estimation of the mean function and of the functional principal components in a spatial setting. We turn to these issues in Section 3.4.

3.4 Estimation of the mean and covariance functions

In this section, we propose an extension of the methodology for the estimation of the mean function and the FPC's introduced in Chapter 2. There are several new elements: the modifications required to deal with two samples, with possibly different means, the introduction of an iterative estimation process, and the choice of the bin size in variogram

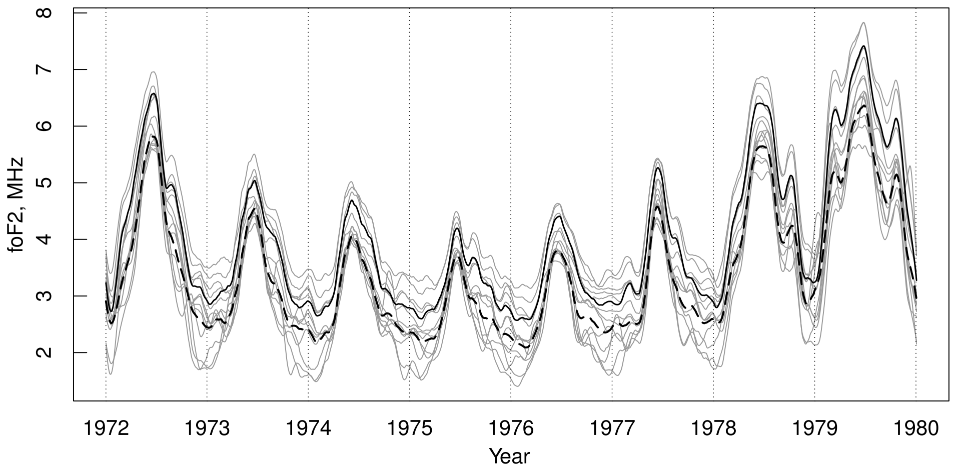

Figure 3.4: Ionosonde data (gray lines). The black solid line is the mean function for stations with the negative inclination, the black dashed line is the mean functions for stations with the positive inclination.

estimation.

Recall that $\mathbf{s}_k^{(j)}$ refers to a location in region $j = 1, 2$. The mean functions μ_j are estimated by weighted sums of observations:

$$\hat{\mu}_j(t) = \sum_{k=1}^{N_j} w_k^{(j)} X(\mathbf{s}_k^{(j)}; t), \quad \sum_{k=1}^{N_j} w_k^{(j)} = 1, \quad j = 1, 2. \tag{3.3}$$

To find the weights $w_k^{(j)}$, it is convenient to use matrix notation. We introduce the following vectors and matrices:

$$\mathbf{1}_1 = \begin{bmatrix} \mathbf{1}_{N_1} \\ \mathbf{0}_{N_2} \end{bmatrix}, \ \mathbf{1}_2 = \begin{bmatrix} \mathbf{0}_{N_1} \\ \mathbf{1}_{N_2} \end{bmatrix}, \ \mathbf{I}_1 = \begin{bmatrix} I_{N_1 \times N_1} & \mathbf{0}_{N_1 \times N_2} \\ \mathbf{0}_{N_2 \times N_1} & \mathbf{0}_{N_2 \times N_2} \end{bmatrix}, \ \mathbf{I}_2 = \begin{bmatrix} \mathbf{0}_{N_1 \times N_1} & \mathbf{0}_{N_1 \times N_2} \\ \mathbf{0}_{N_2 \times N_1} & I_{N_2 \times N_2} \end{bmatrix}$$

and

$$\mathbf{w}_1 = [w_1^{(1)}, \ldots, w_{N_1}^{(1)}, 0, \ldots, 0], \ \ \mathbf{w}_2 = [0, \ldots, 0, w_1^{(2)}, \ldots, w_{N_2}^{(2)}],$$

$$\mathbf{X}(t) = [X(\mathbf{s}_1^{(1)}; t), \ldots, X(\mathbf{s}_{N_2}^{(2)}; t)]^T.$$

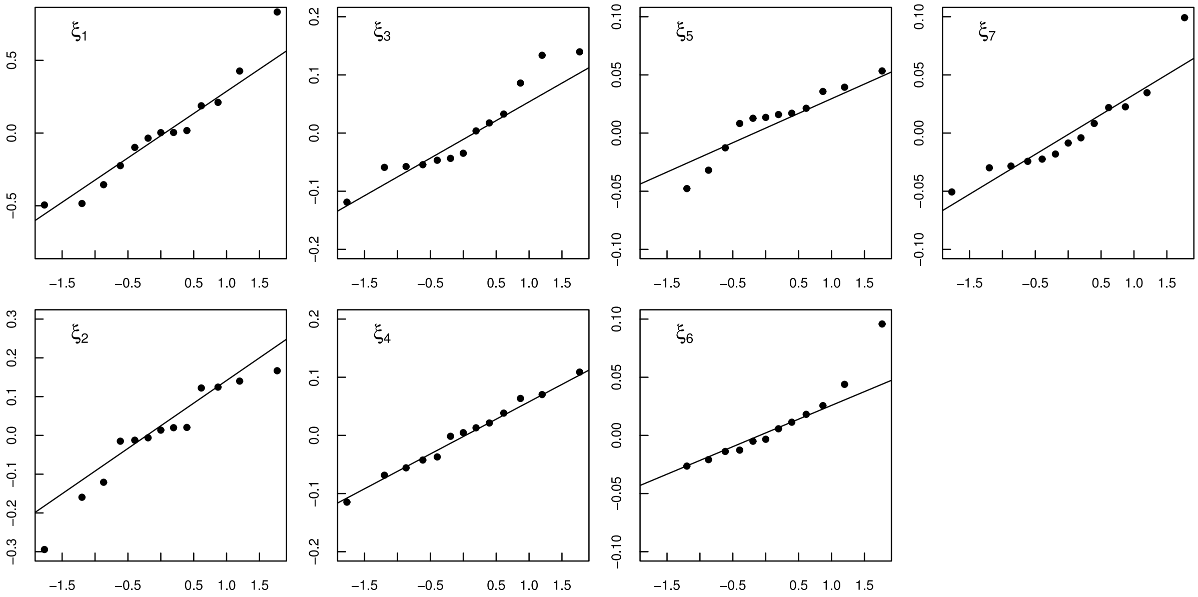

Figure 3.5: Normal QQ plots for the estimated scores $\boldsymbol{\xi}_i,\ 1 \leq i \leq 7$.

Then, the estimates are given by

$$\hat{\mu}_j(t) = \mathbf{w}_j \mathbf{X}(t), \quad \mathbf{w}^{(j)} \mathbf{1}_j = 1, \quad j = 1, 2.$$

To give closed form expressions for the weights which minimize the constrained least squared errors, we define the covariance matrix $\mathbf{C}$, which has the block form

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}.$$

The elements of the sub-matrices $\mathbf{C}_{jj'}$ are given by

$$\begin{aligned} c_{jj'}(k,\ell) =& E\left[\left\langle X(\mathbf{s}_k^{(j)}) - \mu_j, X(\mathbf{s}_\ell^{(j')}) - \mu_{j'} \right\rangle\right] \\ =& \sum_{i=1}^{p} E[\xi_i(\mathbf{s}_k^{(j)}) \xi_i(\mathbf{s}_\ell^{(j')})],\ \ 1 \leq k \leq N_j,\, 1 \leq \ell \leq N_{j'}\,. \end{aligned} \tag{3.4}$$

(The estimation of the $c_{jj'}(k, \ell)$ is discussed at the end of this section.) Using the method of Lagrange multipliers, we find the optimal weights:

$$\mathbf{w}_1 = (\mathbf{1}_1^T \mathbf{C}^{-1} \mathbf{1}_1)^{-1} \mathbf{1}_1^T \mathbf{C}^{-1} \mathbf{I}_1, \;\; \mathbf{w}_2 = (\mathbf{1}_2^T \mathbf{C}^{-1} \mathbf{1}_2)^{-1} \mathbf{1}_2^T \mathbf{C}^{-1} \mathbf{I}_2. \tag{3.5}$$

The covariance matrix $\mathbf{C}$ requires the estimation of the FPC's, $v_i(t)$, which, in turn, requires the estimation of the mean functions, $\mu_j(t)$, and centering the data. Thus, the following iterative approach is used:

1. Compute the simple mean (average) for each sample and subtract it from the curves in each sample.
2. Estimate the FPC's and their number required to capture desired level of variability using the method CM3 proposed in Chapter 2 and calculate the scores. We review the method CM3 in Section 3.7.
3. Estimate the covariance matrix $\mathbf{C}$ using (3.4), find the weights using (3.5) and estimate the mean functions using the weighted sums, (3.3).
4. Subtract the mean functions from the curves in each sample and repeat steps (b)–(d) until a suitably defined convergence of the mean function estimates is reached. We discuss the convergence in Section 3.6.

To calculate the expectations $E[\xi_i(\mathbf{s}_k^{(j)})\xi_i(\mathbf{s}_\ell^{(j')})]$ appearing in the definition of the matrix $\mathbf{C}$, we use parametric modeling, common in spatial statistics:

$$E[\xi_i(\mathbf{s}_k^{(j)})\xi_i(\mathbf{s}_\ell^{(j')})] = \mathrm{Cov}(\xi_i(\mathbf{s}_k^{(j)}), \xi_i(\mathbf{s}_\ell^{(j')})) = f_i(d_{k,\ell}, \sigma^2, \rho, \ldots),$$

where $d_{k,\ell}$ is a chordal distance between $\mathbf{s}_k$ and $\mathbf{s}_\ell$ defined in Section 3.3. The selected parametric models for scores as well as estimated parameters are summarized in Section 3.6. They are estimated using weighted least square fitting of a robust variogram estimator of Hawkins and Cressie (1984). We emphasize that there is no guarantee that the iterative algorithm converges for an arbitrary dataset.

When sample size is small, estimation of the empirical variogram can be a challenge. In fact, improper selection of the bin parameter could potentially cause significant overestimation of the sill and the range, which may affect the conclusion of the test. To achieve stability of the estimator of the empirical variogram the bin parameter should be selected in such a way that each lag interval contains at least 30 distinct distances, see, for example, Section 2.4 in Cressie (1993). In the case of small sample size this recommendation is hard to fulfill, thus, visual validation both for variogram estimation and its fitting is essential. In our study, we use two criteria for validation. The first criterion is that the empirical variogram should have a flat-shaped region for large lags. The second criterion is that the estimated range should be less than the characteristic size of the spatial domain. We also recommend to use a simple estimator of the variance and FPC's to approximate the magnitude order of the sill. Usually, the maximum lag is taken to be half of the characteristic size of the spatial domain. Finally, let us draw attention to the fact that overestimation of the sill can cause under-rejection of the null hypothesis which in practice is more preferable than over-rejection. We discuss bin size selection for the ionosonde data in Section 3.6

3.5 The test procedure

To test the null hypothesis formulated in Section 3.3, we use the statistic

$$\hat{S} = \|\hat{\mu}_1 - \hat{\mu}_2\|^2 = \int \left(\hat{\mu}_1(t) - \hat{\mu}_2(t)\right)^2 dt, \tag{3.6}$$

where the mean functions $\hat{\mu}_1$ and $\hat{\mu}_2$ are estimated using the iterative procedure described in Section 3.4. A natural approach for the small sample is to use a Monte Carlo (MC) test based on (3.6), which is feasible under the assumptions stated in Section 3.3 (normality and the independence of score processes). We describe this approach first. Then we introduce tests based on gamma approximation to the distribution of (3.6) under H_0.

Both the MC test and the gamma test use the expansion under H_0

$$\hat{S}_0 = \sum_{i=1}^{p} \left[(\mathbf{w}_1 - \mathbf{w}_2)\, \hat{\boldsymbol{\xi}}_i \right]^2 \tag{3.7}$$

in which the $\hat{\boldsymbol{\xi}}_i$ are the scores estimated as explained in Section 3.4. The weights $\mathbf{w}_j$ are also estimated, but are treated as deterministic in both procedures. This is because the variability of the mean function estimates is quantified by the variability of the scores.

Monte Carlo approximation. This method relies on the following procedure.

1. Estimate the mean functions μ_j, the weights $w_k^{(j)}$ and the FPC's v_i as described in Section 3.4 and Section 3.7.
2. Calculate the test statistics $\hat{S}$ using (3.6).
3. Generate scores $\tilde{\boldsymbol{\xi}}_i \sim N(\mathbf{0}, \boldsymbol{\Gamma}_i)$, where $\boldsymbol{\Gamma}_i$ is the estimated covariance matrix. (Recall that the variances $\boldsymbol{\Gamma}_i = E[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T]$ are estimated using a parametric model.)
4. Calculate the MC statistic, $\tilde{S}$, by plugging the weights $w_k^{(j)}$ and the generated scores $\tilde{\boldsymbol{\xi}}_i$ into (3.7).
5. Compare the MC statistic and the observed test statistics.
6. Repeat steps (c)–(e) M times to reach a required precision. (We use $M = 10^6$ which provides precision of 3 orders of magnitude.)
7. The ratio of the Monte-Carlo statistics which are larger than the test statistics, $\hat{S}$, and the number of iterations, M, is the Monte-Carlo P–value.

The MC procedure is time consuming. We now describe a procedure based on gamma approximation of the test statistics. This procedure is much faster and has comparable finite sample properties in synthetic data sets generated to resemble the ionosonde data.

Gamma approximation. The test statistics (3.7) can be written as follows

$$\hat{S}_0 = \sum_{k=1}^{p} \left[(\mathbf{w}_1 - \mathbf{w}_2)\, \hat{\boldsymbol{\xi}}_k \right]^2 = \sum_{k=1}^{p} \hat{\sigma}_k^2 z_k^2, \tag{3.8}$$

where the z_i are standard normal and $\hat{\sigma}_i^2 = (\mathbf{w}_1 - \mathbf{w}_2)\hat{\mathbf{\Gamma}}_i(\mathbf{w}_1 - \mathbf{w}_2)^T$. It is not difficult to verify that $\sigma_k^2 = O(N^{-1})$, and so the characteristic function of $\hat{S}$ admits the approximation

$$\prod_{k=1}^{p} \left(1 - 2i\sigma_k^2 t\right)^{-1/2} \approx \left(1 - 2i\sum_{k=1}^{p} \sigma_k^2 t\right)^{-1/2}. \tag{3.9}$$

The right-hand side of (3.9) is the characteristic function of the gamma density

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0,$$

with

$$\alpha = 1/2, \qquad \beta = 2\sum_{k=1}^{p} \sigma_k^2. \tag{3.10}$$

Instead of using (3.10), a better approximation is obtained by using exact values given by

$$\alpha = \frac{\left(E\hat{S}_0\right)^2}{\mathrm{Var}(\hat{S}_0)}, \qquad \beta = \frac{\mathrm{Var}(\hat{S}_0)}{E\hat{S}_0}. \tag{3.11}$$

This requires estimates of $E\hat{S}$ and $\mathrm{Var}(\hat{S})$ valid under H_0. Treating the weights $w_k^{(j)}$ as deterministic and using (3.8), we obtain

$$E\hat{S}_0 = \sum_{i=1}^{p} \hat{\sigma}_i^2 = (\mathbf{w}_1 - \mathbf{w}_2)\hat{\mathbf{C}}(\mathbf{w}_1 - \mathbf{w}_2)^T \tag{3.12}$$

and

$$E\hat{S}_0^2 = 3\sum_{i=1}^{p} \hat{\sigma}_i^4 + 2\sum_{i>i'}^{p} \hat{\sigma}_i^2\hat{\sigma}_{i'}^2 \text{ and } (E\hat{S}_0)^2 = \sum_{i=1}^{p} \hat{\sigma}_i^4 + 2\sum_{i>i'}^{p} \hat{\sigma}_i^2\hat{\sigma}_{i'}^2,$$

so that

$$\mathrm{Var}(\hat{S}_0) = E\hat{S}_0^2 - (E\hat{S}_0)^2 = 2\sum_{i=1}^{p} \hat{\sigma}_i^4. \tag{3.13}$$

For the ionosonde data the difference between the "approximate" and the "exact" values given by (3.10) and (3.11) respectively is about 20%, which is the main cause of the difference between P-values in Table 3.3.

We can now summarize the testing procedure based on the gamma approximation.

1. Estimate the mean functions μ_j, the weights $w_k^{(j)}$ and the FPC's v_i as described in Section 3.4.
2. Compute the test statistics $\hat{S}$ using (3.6).
3. Estimate the variances $\boldsymbol{\Gamma}_i = E[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T]$ using a spatial parametric model for each $1 \leq i \leq p$.
4. Estimate the parameters α and β using either (3.10) or (3.11).
5. Compute the P–value:

$$\text{P-value} = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_{\hat{S}}^{\infty} x^{\alpha-1} e^{-x/\beta} dx \, .$$

3.6 Application to the ionosonde data

In this section we describe the application of the methodology we propose to the ionosonde data. We begin with the details of implementation.

We estimated the variograms using the robust estimator of Hawkins and Cressie (1984) and fit parametric spatial models using weighted nonlinear least squares. As explained at the end of Section 3.4, the bin parameter needs to be selected with great care in order to obtain a reasonable estimate of the spatial parametric model. The number of the distinct distances is $C_2^{13} = 78$ and the maximum lag is about 20^0 or 0.3 rad, see Fig. 3.3. By letting each lag contain $N(h_i) = 20$ distinct distances we conclude that the bin parameter should be 5^0 or 0.06 rad. Because $N(h_i) < 30$ visual validation is needed in each step of the iterative algorithm.

Visual examination revealed that for some score processes $\boldsymbol{\xi}_i$ the Exponential model with zero nugget offers the best fit. Under this model, the covariances are given by

$$\text{Cov}(\xi_i(\mathbf{s}_k^{(j)}), \xi_i(\mathbf{s}_\ell^{(j')})) = \sigma_i^2 \exp\{-d_{k,\ell}/\rho_i\} \, .$$

The estimated sills, σ_i^2, and ranges, ρ_i are summarized in Table 3.2, “Simple” stands for the ordinary variance estimation.

Table 3.2: Estimates for the parametric covariance models fitted to the ionosonde data. Abbreviation SF states for spatial field.

SF	Model	σ^2	ρ	SF	Model	σ^2	ρ
$\boldsymbol{\xi}_1$	Exp.	$1.57 \cdot 10^{-1}$	0.056	$\boldsymbol{\xi}_5$	Exp.	$2.17 \cdot 10^{-3}$	0.028
$\boldsymbol{\xi}_2$	Simple	$1.72 \cdot 10^{-2}$	–	$\boldsymbol{\xi}_6$	Exp.	$2.76 \cdot 10^{-3}$	0.025
$\boldsymbol{\xi}_3$	Exp.	$6.56 \cdot 10^{-3}$	0.036	$\boldsymbol{\xi}_7$	Simple	$1.49 \cdot 10^{-3}$	–
$\boldsymbol{\xi}_4$	Simple	$3.92 \cdot 10^{-3}$	–				

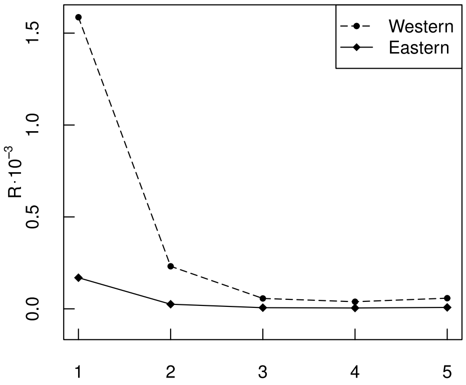

Figure 3.6: Convergence of the iterative method for the estimation of the means, R_i as a function of the iteration i.

Using the iterative approach introduced in the Section 3.4, we estimate the mean functions for the two regions described in Section 3.2. The convergence is evaluated by means of the quantity

$$R_i = \int_0^1 |\mu^{(i+1)}(t) - \mu^{(i)}(t)| dt,$$

where the index i denotes the number of iterations. For the ionosonde data, the convergence is reached after four iterations, as shown in Fig. 3.6.

Table 3.3: P-values for different number of the FPC's p, $1 \leq p \leq 7$ for night and day data. Abbreviation CV states for cumulative variance.

	Spatial					Simple		
Target CV,%	Final p	Final CV,%	MC	Exact	Approx	Final p	Final CV,%	Exact
75	1	75.45	0.023	0.0244	0.0244	–	–	–
80	2	85.14	0.024	0.0250	0.0296	1	82.49	$1.42 \cdot 10^{-3}$
90	4	91.81	0.024	0.0254	0.0334	2	91.53	$1.23 \cdot 10^{-3}$
95	7	95.17	0.025	0.0257	0.0357	3	95.32	$1.15 \cdot 10^{-3}$

These estimates allow us to construct the following decomposition of variance:

$$\sum_{k=1}^{N_1} \|X(\mathbf{s}_k^{(1)}) - \hat{\mu}_1\|^2 + \sum_{\ell=1}^{N_2} \|X(\mathbf{s}_\ell^{(2)}) - \hat{\mu}_2\|^2 = \sum_{i=1}^{N_1+N_2} \left\{ \sum_{k=1}^{N_1} \hat{\xi}_i^2(\mathbf{s}_k^{(1)}) + \sum_{\ell=1}^{N_2} \hat{\xi}_i^2(\mathbf{s}_\ell^{(2)}) \right\}.$$

The sum on the right–hand side is replaced by the sum

$$V(p) := \sum_{i=1}^{p} \left\{ \sum_{k=1}^{N_1} \hat{\xi}_i^2(\mathbf{s}_k^{(1)}) + \sum_{\ell=1}^{N_2} \hat{\xi}_i^2(\mathbf{s}_\ell^{(2)}) \right\},$$

with p so large that $V(p)$ exceeds a specified percentage of $V(N_1 + N_2)$. This procedure is fairly standard for independent functional data. For spatially indexed functions, the estimated score processes $\hat{\boldsymbol{\xi}}_i$ take into account the spatial correlations.

In some cases, the results of a test for functional data depend on the selection of the cut–off value p, with the usual recommendation being to use the value of p which explains between 80 and 90 percent of the variance. For the ionosonde data, $p = 1$ already explains over 75% of the variance, and the results of the test, for both MC and gamma implementations, do not significantly depend on p. Table 3.3 shows the P-values for the MC method, gamma method with the exact parameters α and β calculated using (3.11) (Exact), gamma method which uses (3.10) (Approx) and the P–values obtained by ignoring the spatial dependence (Simple). In the last case we used estimation procedure and the gamma method which neglect spatial dependence. We conclude that there is enough evidence to support the alternative hypothesis. For the ionosonde data the conclusion of our analysis is

Table 3.4: Empirical size (in percent) of the tests applied to data generated to resemble the ionosonde data.

p	Spatial									Simple		
	MC			Exact			Approx			Exact		
	10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%
1	9.96	4.89	1.02	9.57	4.66	0.93	9.57	4.66	0.93	20.17	11.12	1.66
2	9.94	5.05	1.01	9.34	4.70	1.01	8.89	4.20	0.76	19.51	10.91	1.86
3	10.03	4.97	1.00	9.29	4.76	0.98	8.60	4.02	0.64	19.51	11.14	1.99
4	9.96	4.96	1.02	9.25	4.77	1.02	8.51	3.91	0.62	19.46	11.09	2.01
5	10.04	5.02	1.01	9.23	4.69	1.02	8.34	3.80	0.61	19.41	11.07	2.08
6	9.97	4.86	0.92	9.23	4.65	0.97	8.26	3.67	0.54	19.31	11.03	2.02
7	9.88	4.95	0.93	9.12	4.63	0.96	8.15	3.60	0.53	19.17	10.99	2.04

that there is statistically significant evidence to claim that the mean function over Europe changes due to magnetic inclination. This conclusion agrees with the findings of Cnossen and Richmond (2008) who used a physical ionospheric model.

The remainder of this section is devoted to the statistical validation of the conclusions presented above. Due to very small sample sizes, the only feasible way of assessing the final sample performance of the tests is through a simulation study. The results of any such study depend on the model for the data generating process (DGP). We have taken great care to ensure that the stochastic structure of the simulated data resembles that of the real data. The key assumptions made to simulate the data are the normality and the independence of the score processes $\boldsymbol{\xi}_i$.

The DGP is given by

$$X(\mathbf{s}^{(1)};t) = m_1(t) + \sum_{i=1}^{p} \xi_i(\mathbf{s}^{(1)})v_i(t), \quad X(\mathbf{s}^{(2)};t) = m_2(t) + \sum_{i=1}^{p} \xi_i(\mathbf{s}^{(2)})v_i(t). \tag{3.14}$$

To evaluate the empirical size, we set $m_1(t) = m_2(t) = 0$. To evaluate the empirical power, we set $m_1(t) = \mu_1(t)$, the estimated mean function for the western region, and $m_2 = (1-r)\mu_1 + r\mu_2$, where μ_2 is the estimated mean function for the eastern region. The power is then a function of the "separation" $0 \le r \le 1$. The FPC's v_i are those estimated from the real data. The scores are generated from zero mean normal distribution

$\boldsymbol{\xi}_i \sim N(\mathbf{0}, \boldsymbol{\Gamma}_i)$, using the Cholesky decomposition. For the empirical size we select $1 \le p \le 7$ and for the power $p = 7$. The number of replication of the process (3.14) is 10^5. The results for the empirical size and the power are summarized in Table 3.4 and Fig. 3.7, respectively. Starting with Table 3.4, we see that MC test has more accurate empirical size, very close to the nominal size. The Gamma test (Exact and Approx) is slightly too conservative. Method Exact without taking into account spatial correlation has inflated size. This agrees with the very small P-values reported for real data, and shows that conclusions based on this method must be treated as tentative.

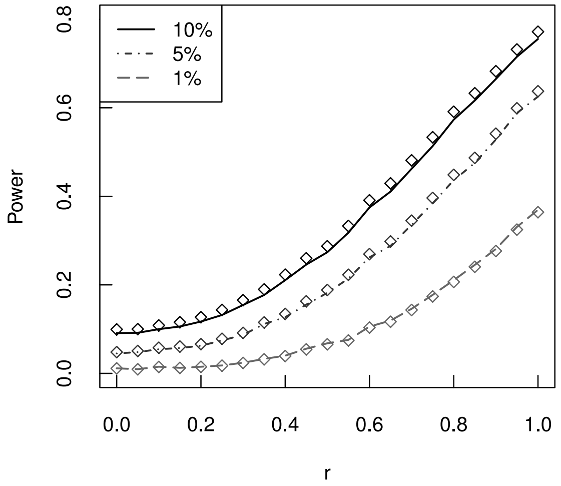

Figure 3.7: The power function of the test. The lines represent the "exact" gamma test and the empty diamonds represent the MC test for different confidence levels: $\alpha = 10\%, 5\%, 1\%$.

3.7 Estimation of the FPC's

In this section we summarize a procedure for the estimation of the FPC's when the curves are spatially correlated. The method we outline is called MC3 in Chapter 2. It is easy to implement and has good size and power in finite samples.

1. Assuming that $EX(\mathbf{s}) = 0$ postulate the expansion

$$X(\mathbf{s}) \approx \sum_{m=1}^{K} \eta_m(\mathbf{s}) B_m,$$

   where, B_m is an orthonormal basis, and $\eta_m(\mathbf{s})$ form an observable field $\eta_m(\mathbf{s}_k) = \langle B_m, X(\mathbf{s}_k)\rangle$. The value of K can be taken to the number of basis functions used to create the functional objects in `R`, for example $K = 49$.

2. Fix n and m, and define the scalar field z by $z(\mathbf{s}_k) = \eta_n(\mathbf{s}_k)\eta_m(\mathbf{s}_k)$.

3. Estimate $\mu_z = Ez(\mathbf{s})$ as a weighted average of the $z(\mathbf{s}_k)$ for all n, m and denote resulting estimate by $\hat{r}_{nm}$.

4. Find the solution of the following matrix equation:

$$\widehat{\mathbf{R}}\mathbf{x} = \lambda\mathbf{x}, \tag{3.15}$$

   where

$$\mathbf{x} = [x_1, x_2, \ldots, x_K]^T, \quad \widehat{\mathbf{R}} = [\hat{r}_{nm},\ 1 \le n, m \le K].$$

   We denote the solutions to (3.15) by

$$\hat{\mathbf{x}}^{(n)} = [\hat{x}_1^{(n)}, \hat{x}_2^{(n)}, \ldots, \hat{x}_k^{(n)}]^T, \quad \hat{\lambda}_j, \qquad 1 \le n \le K. \tag{3.16}$$

5. The FPC's v_n, $1 \le n \le K$ are estimated by

$$\hat{v}_n = \sum_{\alpha=1}^{K} \hat{x}_\alpha^{(n)} B_\alpha. \tag{3.17}$$

   Because the B_n are orthonormal estimated FPC's are also orthonormal. The $\hat{\lambda}_n$ in (3.16) are estimators of the corresponding eigenvalues.

CHAPTER 4

NONPARAMETRIC INFERENCE IN SMALL DATA SETS OF SPATIALLY INDEXED CURVES WITH APPLICATION TO IONOSPHERIC TREND DETERMINATION[1]

Abstract

This paper is concerned with estimation and testing in data sets consisting of a small number (about 20–30) of curves observed at unevenly distributed spatial locations. Such data structures may be referred to as spatially indexed functional data. Motivated by an important space physics problem, we model such data as a mean function plus spatially dependent error functions. Given a small number of spatial locations, the parametric methods for the estimation of the spatial covariance structure of the error functions are not satisfactory. We propose a fully nonparametric estimator for the mean function. We also derive a test to determine the significance of the regression coefficients if the mean function is a linear combination of known covariate functions. In particular, we develop methodology for the estimation a trend in spatially indexed functional data, and for assessing its statistical significance. We apply the new tools to global ionosonde records to test the hypothesis of ionospheric cooling. Nonparametric modeling of the space–time covariances is surprisingly simple, much faster than those previously proposed, and less sensitive to computational errors. In simulated data, the new estimator and test uniformly dominate those based on parametric modeling.

4.1 Introduction

Models for data which exhibit both space and time dependence have attracted increasing attention in geophysical and environmental research. This is a fast growing branch of statistics, for a general overview see Cressie and Wikle (2011) and Sherman (2011), for

---

[1] CO-AUTHORED BY O. GROMENKO AND P. KOKOSZKA. REPRODUCED FROM COMPUTATIONAL STATISTICS AND DATA ANALYSIS, VOL. 59, 82–94, 2013.

a fast, accessible introduction, we recommend Gneiting *et al.* (2007). Space–time data could be roughly separated into several categories according to the amount of information contained, respectively, in their spatial and temporal components. One category are the data which have a very rich spatial component and relatively limited temporal component. Such data usually come from satellites, see e.g. Jun and Stein (2009), Cressie *et al.* (2010) and Katzfuss and Cressie (2011), among many others. Another category are data which have a rich temporal component and a relatively simple spatial structure. Such data come typically as collections of long time series recorded at different spatial locations by ground based instruments. For example, the Irish wind data studied by Haslett and Raftery (1989) and consequently used in many other papers, the Canadian weather data extensively used in Ramsay and Silverman (2005) and Ramsay *et al.* (2009), pollution data studied by Bowman *et al.* (2009), and many others.

In this paper, we propose a flexible, fully nonparametric methodology for data of the latter type. It includes estimation of the mean function and is applied to testing the statistical significance of a linear trend. Our methodology builds on the theory of Hall *et al.* (1994) and Hall and Patil (1994) by 1) developing a practically applicable tool set for the estimation and testing in the spatial context with few data locations, 2) extending it to the framework of spatially indexed functional data, 3) developing suitable confidence bounds, and 4) applying it to an important space physics problem. The work presented in this paper is a direct result of our attempts to solve this important space physics problem in a fairly conclusive manner that would be satisfactory to the space physics community. Since the problem concerns the detection of a long term (many decades) trend, we hope that out methodology is general and useful enough to be applicable to other similar data sets and problems. Spatially indexed functional data have been the focus of several recent studies, see Delicado *et al.* (2010), Giraldo *et al.* (2011), Nerini *et al.* (2010), and Chapter 2 and 3 in this dissertation. Existing approaches however often fail when the number of spatial locations is small because in such cases the numerical optimization required to fit a parametric spatial model may fail, or the fit may be poor. The research we report is, to a large extent, a result of computational difficulties we encountered with standard approaches. The resulting new methodology is computationally faster and the algorithms never fail to converge (in our data sets and simulations).

The paper is organized as follows. In Section 4.2, we develop a nonparametric covariance estimation procedure for scalar data. Next, in Section 4.3, a statistical model for spatially indexed functional data is introduced. Section 4.4 presents the estimation procedure for this model. In Section 4.5 , we derive a test for assessing the significance of regression coefficients when the mean function is a linear combination of known covariate functions. The application of this test to the assessment of a long term cooling trend in the ionosphere is presented in Section 4.6. Section 4.7 presents the results of simulation studies that validate the methodology we propose and its application to the ionosonde data.

4.2 Description of the method for scalar data

In this section, we assume that ζ is a mean zero stationary and isotropic *scalar* random field observed at locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$, and $\mathbf{\Gamma}$ is the $N \times N$ matrix of covariances

$$\gamma(d_{k\ell}) = \mathrm{Cov}(\zeta(\mathbf{s}_k), \zeta(\mathbf{s}_\ell)) = E[\zeta(\mathbf{s}_k)\zeta(\mathbf{s}_\ell)],$$

where $d_{k\ell}$ is the distance between $\mathbf{s}_k$ and $\mathbf{s}_\ell$. Estimation of $\mathbf{\Gamma}$ is not trivial for small samples. Standard variogram based estimator for small spatial data sets is generally unstable, and the the optimization often fails to converge. It is recommended that every lag interval should contain at least 30 distinct distances, but for small sample sizes, it is difficult to meet this condition without reducing the number of intervals to a level which makes fitting a parametric model difficult. We therefore develop nonparametric methodology, based on the work of Hall *et al.* (1994) and Hall and Patil (1994), which is suitable for small data sets. It forms the basis of the estimation and testing procedures for functional spatially indexed data, but can also be used for different spatio–temporal models, as illustrated in Example 4.7.1.

Recall that $d_{k\ell}$ is the distance between $\mathbf{s}_k$ and $\mathbf{s}_\ell$, and consider the preliminary estimator

$$\tilde{\gamma}(d_{k\ell}) = \zeta(\mathbf{s}_k)\zeta(\mathbf{s}_\ell). \tag{4.1}$$

It is possible that for some distances there exist several distinct estimators $\tilde{\gamma}(d_{k\ell})$, in fact for $d_{k\ell} = 0$ there are always N different preliminary estimators. The estimated covariances are ordered so that the corresponding distances do not decrease: denoting the $d_{k\ell}$ by d_i, we thus have $d_i \le d_{i+1}$, $1 \le i \le N(N+1)/2$. The resulting sequence $\{\tilde{\gamma}(d_i) : 1 \le i \le N(N+1)/2\}$ is very noisy and must be smoothed. We use local linear regression, see Fan and Gijbels (1996), rather than the kernel smoother suggested by Hall *et al.* (1994). The reason for using the local linear regression is that it introduces a slightly smaller bias for small and large distances d_i. Let $\kappa(x)$ be a compactly supported symmetric probability density function. The smoothed value of $\gamma(d)$ is thus estimated by $\hat{m}(d)$ computed by minimizing

$$(\hat{m}(d), \hat{m}_1) = \arg\min_{m,m_1} \sum_{i=1}^{N(N+1)/2} \kappa\left(\frac{d - d_i}{h}\right) \left\{\tilde{\gamma}(d_i) - m(d) - m_1(d - d_i)\right\}^2 . \tag{4.2}$$

We performed simulations using several popular kernels (triangular, quadratic, Epanechnikov, triweight, tricube), and found that they produce practically the same estimates. The results reported in this paper are based on the Epanechnikov kernel. As with all problems of this type, the most difficult issue is the selection of the bandwidth h; Hall *et al.* (1994) do not recommend any specific procedure. They developed an interactive software which allows the user to choose the bandwidth and visually compare the resulting estimates. We describe our method of bandwidth selection in Section 4.9.

To construct a positive definite covariance function, we use Bochner's theorem: We compute the Fourier transform of $\hat{m}$ and delete all negative lobes. The inverse Fourier transform is then our final estimator $\hat{\gamma}(d)$. We enhance the idea of Hall *et al.* (1994) by providing a procedure to construct functional confidence intervals for $\hat{\gamma}(\cdot)$, see Section 4.9. The application of the procedure to simulated data is illustrated in Figure 4.1.

Hall *et al.* (1994) showed that to achieve consistency in the estimation of $\gamma(d)$, the

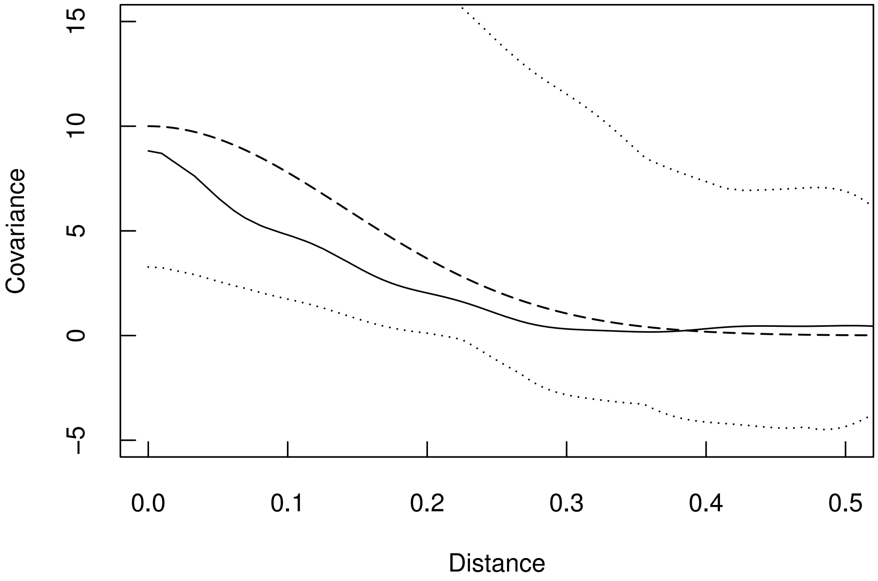

Figure 4.1: Illustration of the estimation procedure for scalar data. The true covariance function (dashed line), its estimate (solid line) and the 95% confidence region (dotted lines).

distance between $\min(d_i)$ and $\max(d_i)$ (the range) must grow much slower than the number of the d_i. This condition is naturally satisfied in the spatial setting because adding one more $\mathbf{s}_k$ roughly increases the range at most by a unit, but increases the number of the d_i by N.

4.3 Statistical model for spatially indexed functional data

The methodology developed in this paper is motivated by the problem of the estimation and modeling of the mean function $\mu(\cdot)$ in the model

$$X(\mathbf{s};t) = \mu(t) + \varepsilon(\mathbf{s};t). \tag{4.3}$$

The data are curves $X(\mathbf{s}_k;t),\ t \in [0,T]$, observed at spatial locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$. Such functional data structures are quite common: examples are discussed in Delicado *et al.* (2010), Nerini *et al.* (2010), Giraldo *et al.* (2011), Hörmann and Kokoszka (2013), Chapters 2 and 3 of this dissertation and Chapters 17 and 18 of Horváth and Kokoszka (2012).

In model (4.3), the error curves $\varepsilon(\mathbf{s})$ are assumed to form a mean zero strictly stationary and isotropic spatial random field taking values in the Hilbert space of square integrable functions with the usual inner product, see e.g. Chapter 2 of Horváth and Kokoszka (2012). We assume that

$$E\|\varepsilon(\mathbf{s})\|^2 = E\int \varepsilon^2(\mathbf{s};t)dt < \infty.$$

If the above assumptions are satisfied, the error term can be represented using the Karhunen-Loève expansion $\varepsilon(\mathbf{s};t) = \sum_{j=1}^{\infty} \zeta_j(\mathbf{s})v_j(t)$, where v_j is the jth functional principal component (FPC), and $\zeta_j(\mathbf{s}) = \langle X(\mathbf{s}) - \mu, v_j\rangle$ is the score of $\varepsilon(\mathbf{s})$ with respect to it. Recall that the v_i are the eigenfunctions of the covariance operator $E[\langle X(\mathbf{s}) - \mu, \cdot\rangle (X(\mathbf{s}) - \mu)]$. This leads to the model

$$X(\mathbf{s};t) = \mu(t) + \sum_{j=1}^{\infty} \zeta_j(\mathbf{s})v_j(t). \tag{4.4}$$

In the above model, the mean function $\mu(\cdot)$ does not depend on the spatial location. It represents the mean temporal evolution of the spatially distributed curves. The inference for $\mu(\cdot)$, when then number of the spatial locations $\mathbf{s}_k$ is small, is the focus of this paper.

The fields $\zeta_j(\cdot)$ are mean zero purely spatial random fields. Set

$$\boldsymbol{\zeta}_j = [\zeta_j(\mathbf{s}_1), \ldots, \zeta_j(\mathbf{s}_N)]^T, \qquad \boldsymbol{\Gamma}_j = \mathrm{Var}[\boldsymbol{\zeta}_j].$$

The matrix $\boldsymbol{\Gamma}_j$ is a positive definite $N \times N$ matrix with elements

$$\gamma_j(\mathbf{s}_k - \mathbf{s}_\ell) = E\left[\zeta_j(\mathbf{s}_k)\zeta_j(\mathbf{s}_\ell)\right].$$

Our modeling framework requires that for every k, l

$$E[\zeta_j(\mathbf{s}_k)\zeta_{j'}(\mathbf{s}_\ell)] = 0 \quad \text{ if } j \neq j'. \tag{4.5}$$

Note that (4.5) is always true for $k = \ell$, but for $k \neq \ell$ it does not follow from any mathematical argument. One can show that the separability of the spatio–temporal covariance function implies (4.5), and we need assumption (4.5) to ensure that the spatio–temporal covariance function is positive definite. Observe that under (4.5) the covariances are given by

$$\mathrm{Cov}(X(\mathbf{s}_k;t),X(\mathbf{s}_\ell;t')) = c(\mathbf{s}_k,\mathbf{s}_\ell;t,t') = \sum_{j=1}^{\infty}\gamma_j(\mathbf{s}_k-\mathbf{s}_\ell)v_j(t)v_j(t'). \tag{4.6}$$

It is not difficult to verify that if each $\mathbf{\Gamma}_j$ is positive definite, then

$$\sum_{k,\ell=1}^{N}\sum_{t,t'} a_k(t)a_\ell(t')c(\mathbf{s}_k,\mathbf{s}_\ell;t,t') \geq 0.$$

Model (4.6) is obviously nonseparable and enjoys the property of full symmetry. Its *spatial* component is (strictly) stationary and isotropic and temporal component is nonstationary. Model (4.6) is thus more general than those proposed in Cressie and Huang (1999) and Gneiting (2002).

4.4 Estimation of the mean function

In this section, we put together the developments of Sections 4.2 and 4.3, and propose a complete nonparametric methodology for the estimation of the mean in model (4.4). In Chapter 2 we considered several approaches to the estimation of the function μ and found that the smallest integrated mean square and absolute errors are obtained by using a weighted sum of functional observations:

$$\hat{\mu}(t) = \sum_{k=1}^{N} w_k X(\mathbf{s}_k;t), \qquad \sum_{k=1}^{N} w_k = 1. \tag{4.7}$$

The weights w_k are found by minimizing the expected value of the L^2 norm of the difference $\hat{\mu}(t)-\mu(t)$, and are given by

$$\mathbf{w} = \mathbf{C}^{-1}\mathbf{1}/(\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}), \tag{4.8}$$

where

$$\mathbf{C} = E\left[\langle \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^T\rangle\right] = \sum_{j=1}^{\infty}\mathbf{\Gamma}_j. \tag{4.9}$$

The estimation of μ is summarized in the following algorithm.

(a) Compute the sample mean $N^{-1} \sum_{k=1}^{N} X(\mathbf{s}_k; t)$. This yields the first estimate $\hat{\mu}(t)$, which will be improved in subsequent iterations.

(b) Calculate the functional version of (4.1), i.e.

$$\tilde{\gamma}(d_{k\ell}) = \langle X(\mathbf{s}_k) - \hat{\mu}, X(\mathbf{s}_\ell) - \hat{\mu} \rangle = \int \left(X(\mathbf{s}_k; t) - \hat{\mu}(t)\right) \left(X(\mathbf{s}_\ell; t) - \hat{\mu}(t)\right) dt$$

and estimate the covariance matrix $\mathbf{C}$ as described in Section 4.2.

(c) Compute the weights using (4.8), and the mean function, $\hat{\mu}(t)$, using (4.7).

(d) Repeat steps (b)-(c) until suitable convergence is reached.

The convergence of the algorithm is evaluated by means of the quantity

$$R_i = \int_0^1 |\mu^{(i+1)}(t) - \mu^{(i)}(t)| dt,$$

where the index i denotes the number of iteration. When the R_i do not decrease with i, we stop the algorithm. The graphs of the R_i for real data are shown in Fig. 4.6.

A similar estimation procedure for model (4.4) was considered in Chapter 2 (Method M2) and Chapter 3, but the nonparametric covariance estimation was not used. We note that we do not estimate the covariances $\mathbf{\Gamma}_j$ of the scores processes separately. This is a very computationally expensive process, and our approach avoids it.

A simulation study that validates the above method and shows its superiority (in small samples) relative to current approaches is presented in Section 4.7.

4.5 Significance of regression coefficients

In this section, we assume that the mean function $\mu(\cdot)$ is a linear combination of q known functions, so that model (4.4) takes the form

$$X(\mathbf{s};t) = \sum_{i=1}^{q} \beta_i z_i(t) + \sum_{j=1}^{\infty} \zeta_j(\mathbf{s}) v_j(t). \tag{4.10}$$

Model (4.10) is motivated by the application to ionosonde data described in Section 4.6, where there is a dominant explanatory function $z(t)$ which quantifies the solar activity. Model (4.10) can include a linear trend $z(t) = t$, and can be used to test the significance of the coefficient of this trend. Problems related to testing the significance of long term trends abound in geophysical and ecological sciences. We now describe how this can be done if the relevant time series are measured at several spatial locations.

Introduce the following vectors

$$\boldsymbol{\beta} = [\beta_1, \dots, \beta_q]^T, \quad \mathbf{z}(t) = [z_1(t), \dots, z_q(t)]^T,$$

and matrices

$$\mathbf{Q} = [\langle z_i, z_{i'} \rangle, 1 \leq i, i' \leq q], \quad \boldsymbol{\Omega} = [\langle z_i, v_j \rangle, 1 \leq i \leq q, 1 \leq j \leq p].$$

The number p of the FPC's is typically selected using the cumulative variance criterion, see e.g. Ramsay and Silverman (2005) or Horváth and Kokoszka (2012). A general recommendation is to use p such that the first p components explain about 85-90% of the variance. It is often useful to perform the inference for several values of p. If the conclusions do not depend on p, we can place more confidence in them.

To estimate the parameter vector $\boldsymbol{\beta}$, we minimize

$$\left\| \sum_{n=1}^{N} w_n X(\mathbf{s}_n) - \sum_{i=1}^{q} \beta_i z_i \right\|^2, \tag{4.11}$$

which lead to the solution

$$\hat{\boldsymbol{\beta}} = \mathbf{Q}^{-1} \left\langle \mathbf{z}, \mathbf{w}^T \mathbf{X} \right\rangle . \tag{4.12}$$

The quantity $\left\langle \mathbf{z}, \mathbf{w}^T \mathbf{X} \right\rangle$ is a $q \times 1$ vector with the ith entry $\left\langle z_i, \sum_{k=1}^{N} w_k X(\mathbf{s}_k) \right\rangle$. Since $\mathbf{w}^T \mathbf{X}$ is an estimate of the mean function $\mu(\cdot)$, (4.12) can be written as

$$\hat{\boldsymbol{\beta}} = \mathbf{Q}^{-1} \left\langle \mathbf{z}, \hat{\mu} \right\rangle . \tag{4.13}$$

It is clear now that the estimation of the regression coefficients is a two step procedure: first we estimate $\mu(\cdot)$ using the methodology of Section 4.4, then we use equation (4.13).

The variance of (4.12), $\mathrm{Var}[\hat{\boldsymbol{\beta}}]$, can be calculated in two different ways. The first way is by substituting the bootstrap sample of $\hat{\mu}$ into (4.13) and using the sample variance. This approach is computationally very expensive and we do not recommend it. Long run times are due to the spatial estimation and the estimation of the FPC's. Instead, we propose an approach based on the following calculations. Observe that

$$\begin{aligned} \left\langle \mathbf{z}, \mathbf{w}^T \boldsymbol{\varepsilon} \right\rangle &= \left[\sum_{k=1}^{N} \sum_{j=1}^{p} w_k \zeta_j(\mathbf{s}_k) \left\langle v_j, z_1 \right\rangle , \ldots , \sum_{k=1}^{N} \sum_{j=1}^{p} w_k \zeta_j(\mathbf{s}_k) \left\langle v_j, z_q \right\rangle \right]^T \\ &= \left[\sum_{j=1}^{p} \tilde{\zeta}_j \left\langle v_j, z_1 \right\rangle , \ldots , \sum_{j=1}^{p} \tilde{\zeta}_j \left\langle v_j, z_q \right\rangle \right]^T , \end{aligned} \tag{4.14}$$

where $\tilde{\zeta}_j$ is the weighted sum of the scores: $\tilde{\zeta}_j = \sum\limits_{k=1}^{N} w_k \zeta_j(\mathbf{s}_k)$. Let $\tilde{\boldsymbol{\zeta}} = [\tilde{\zeta}_1, \ldots, \tilde{\zeta}_p]^T$ then $\left\langle \mathbf{z}, \mathbf{w}^T \boldsymbol{\varepsilon} \right\rangle = \boldsymbol{\Omega} \tilde{\boldsymbol{\zeta}}$ and hence

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{Q}^{-1} \boldsymbol{\Omega} \tilde{\boldsymbol{\zeta}}.$$

Thus

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \mathbf{Q}^{-1} \boldsymbol{\Omega} \mathrm{Var}[\tilde{\boldsymbol{\zeta}}] \boldsymbol{\Omega}^T \mathbf{Q}^{-1}. \tag{4.15}$$

Assuming that the weights are known constants, we have

$$\mathrm{Var}[\tilde{\boldsymbol{\zeta}}] = \mathrm{diag} \left(\mathbf{w}^T \boldsymbol{\Gamma}_1 \mathbf{w}, \ldots , \mathbf{w}^T \boldsymbol{\Gamma}_p \mathbf{w} \right) , \tag{4.16}$$

which is a $p \times p$ diagonal matrix. The diagonal form of $\mathrm{Var}[\tilde{\boldsymbol{\zeta}}]$ is a consequence of assumption (4.5). All quantities appearing in (4.15) and (4.16) can be estimated using the methodology presented in the previous sections.

If the functions $X(\mathbf{s}_k)$ are normally distributed, then, by (4.12), the estimator $\hat{\boldsymbol{\beta}}$ is approximately normal. Thus to test $\beta_i = 0$, for a fixed i, we assume that the statistic $\hat{\beta}_i / \sqrt{\mathrm{Var}[\hat{\beta}_i]}$ has the standard normal distribution. We note that the spatial quasi–bootstrap method also requires assumptions on the distribution of the spatial processes ξ_j, and the only practical assumption is that these processes are normal (we must use the Cholesky decomposition). In Section 4.6 we verify that the assumption of normality is approximately satisfied by the ionosonde data.

A simulation study that validates our method is presented in Section 4.7.

4.6 Application to ionosonde data

We first provide some background and motivation. The problem we study is related to the the hypothesis of Roble and Dickinson (1989) who argued that the increasing amounts of greenhouse gases in the upper atmosphere should lead to its global cooling because these gases radiate heat into space. Rishbeth (1990) pointed out that such a cooling would result in a thermal contraction and the global lowering of the highest or peak electron density. The ionospheric layer which contain the peak electron density is known as the F2 region. The peak electron density above any location on the Earth can be measured indirectly. Data obtained from a type of radar called the ionosonde allow to compute the a critical frequency, denoted foF2. This frequency is related to the peak electron density by an equation based partially on laws of physics and partially on empirical corrections. There are, in fact, several versions of this equation, but the main point is that a lowering in the peak electron density corresponds to a decrease in the foF2 frequency. There has consequently been extensive space physics research aimed at determining if a decreasing temporal trend in the foF2 frequency indeed exists. An interested reader is referred to Lastovicka *et al.* (2008), as a starting point. We note that a long term change in the upper atmosphere can impact space–based navigation, short-wave (3-30Mhz) radio communication and the

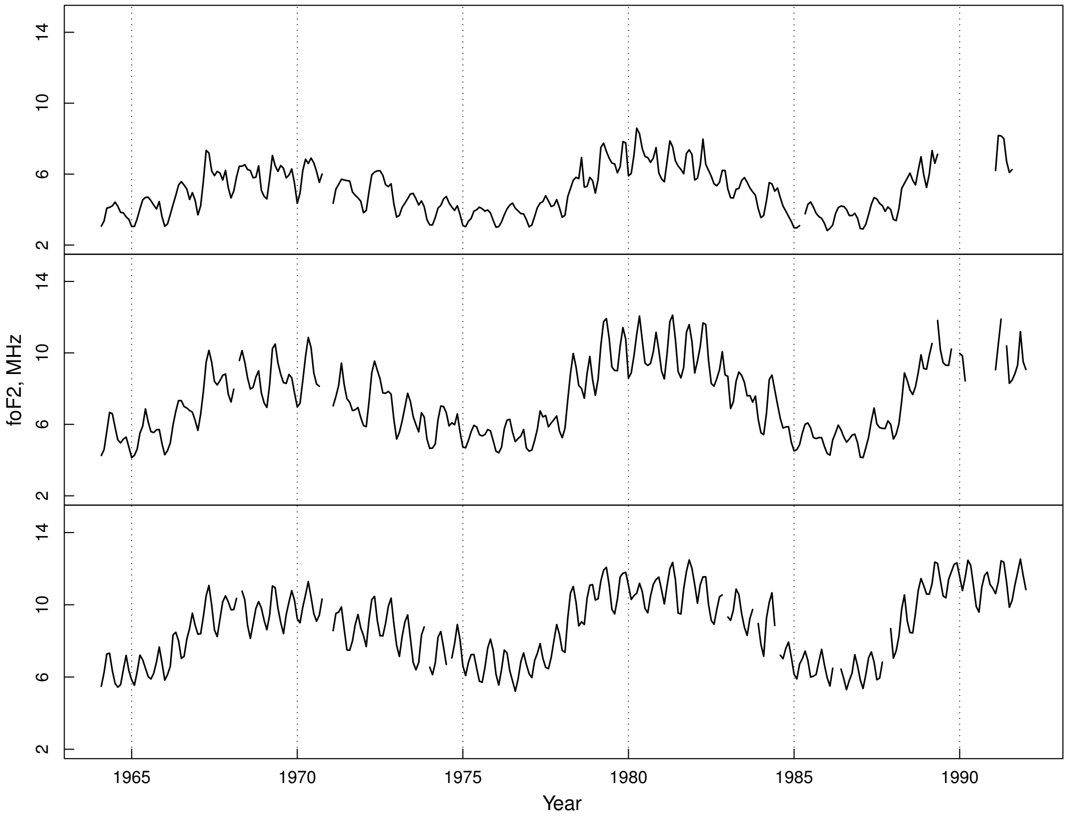

Figure 4.2: F2-layer critical frequency curves at three locations. Top to bottom (latitude in parentheses): Yakutsk (62.0), Yamagawa (31.2), Manila (14.7). The functions exhibit a latidudal trend in amplitude.

operation of low orbit satellites. But perhaps even more importantly, long–term changes in the upper atmosphere and in the lower atmosphere (troposphere) can be governed by the same factors such as solar activity, changes the Earth's magnetic field and greenhouse gases concentration. Consequently, understanding of the influence of these factors and their combinations in the upper atmosphere can provide additional information on long–term changes in the troposphere.

Long-term changes in the upper atmosphere are usually described using a linear approximation which is called the ionospheric trend. The main problem in its determination is the separation of the solar activity and other factors, like the long term changes in the internal magnetic field of the Earth, see Clilverd *et al.* (2003). The solar cycle however

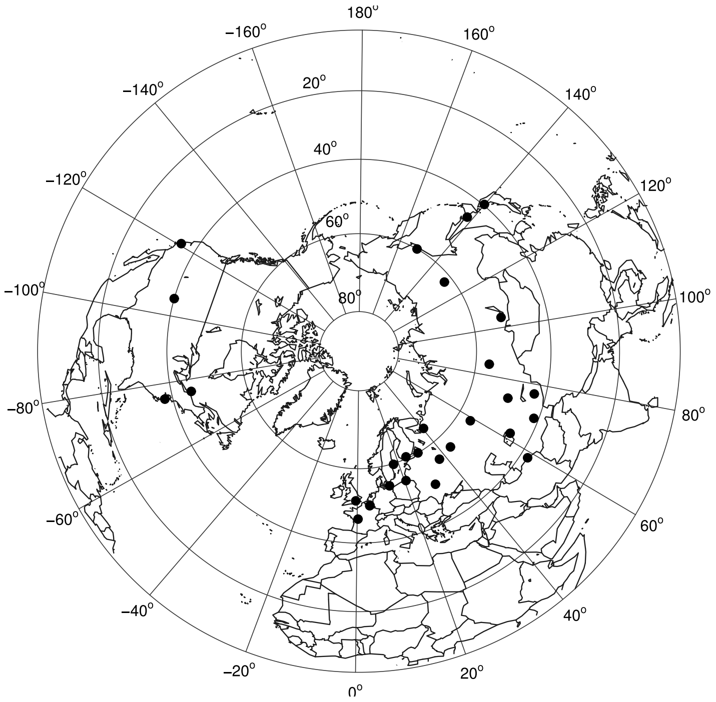

Figure 4.3: Locations of 28 ionosonde stations in the northern hemisphere.

clearly dominates the shape of the foF2 curves, as shown in Fig. 4.2. A comprehensive overview of statistical methods proposed in the space physics community is given in Lastovicka *et al.* (2006). The main problem from which they suffer is their inability to combine the information from many spatial locations. The model and the testing approach we propose is an attempt to overcome this difficulty. We consider the following specialization of model (4.10):

$$\mathrm{foF2}(\mathbf{s};t) = \beta_1 + \beta_2 t + \beta_3 \mathrm{SRF}(t) + \sum_{j=1}^{p} \zeta_j(\mathbf{s}) v_j(t). \tag{4.17}$$

The covariate SRF is the solar radio flux measured in $\mathrm{W/m^2Hz}$. It is a proxy for the solar activity. Another possible proxy is the sunspot number. Our primary interest in this section is in testing the hypothesis $H_0: \ \beta_2 = 0$.

In this paper, we use data from 28 ionosonde stations, see Fig. 4.3, located in the mid–latitude region form 30^0 to 60^0 in the magnetic coordinate system. To study the solar influence on trends we split data into Night data from 22 to 2 LT, Noon Data from 10 to

Table 4.1: P-values for the trend parameter as a function of the number of the FPC's.

Target CV,%	Day Final p	Final CV,%	P-value	Noon Final p	Final CV,%	P-value	Night Final p	Final CV,%	P-value
80	2	80.81	$3.3 \cdot 10^{-03}$	2	84.08	$5.58 \cdot 10^{-3}$	2	81.82	0.3025
85	–	–	–	–	–	–	3	89.08	0.3197
90	3	90.81	0.0101	3	92.32	$5.79 \cdot 10^{-3}$	4	92.40	0.3302
–	4	93.10	0.0102	4	94.70	$5.79 \cdot 10^{-3}$	–	–	–
–	5	94.62	0.0103	–	–	–	–	–	–
95	6	95.87	0.0103	5	96.26	$6.04 \cdot 10^{-3}$	5	95.24	0.3304

14 LT and Day data (no time filter is applied). Here and below LT means local (latitudal) time. The data are available from 1967-08-01 to 1989-08-01. This interval covers 22 years or two solar 11-year cycles. We do not discuss the details of creating the functional objects from the raw data, as these are fairly complex. An interested reader is referred to Chapter 2 and Chapter 3 herein. The distance between the locations on the globe is measured using the chordal distance.

Table 4.1 shows the P–values obtained for several values of p. These values were selected by targeting a specific percentage of variance explained by the first p FPC's. It is seen that the trend coefficient is significant for the Day and Noon data, but not for the Night data. A more comprehensive discussion of this finding will require a more detailed space physics research, but we note that our result is consistent with discussions published in space physics literature. Due to different physical processes, the behavior of the upper atmosphere is different at different times of a day. Our finding, in a sense, confirms that the problem of trend determination is complex, an a clear cut answer may not be available. The problem must be formulated in a more precise way. One might clearly wonder if the acceptance of $H_0 : \beta_2 = 0$ is not a type I error. As seen in Fig. 4.5, the test has the power of about 80%, so a type one error is a possibility. One way to increase power, would be to consider more ionosonde locations. There are however two problems that must first be solved. 1) As seen in Fig. 4.3, the amplitude of the curves depends on the latitude. Thus extending the latitudinal coverage would violate te assumption of stationarity that underlines our methodology. 2) Most stations outside the northern hemisphere have incomplete records. These are not a few missing observations, but missing stretches of data of length 5–20 years. A suitable methodology to accommodate such incomplete records would need to be developed.

4.7 Validation of the methodology

In this section, we present the results of several simulation studies that validate the estimation on testing methods introduced in Sections 4.4 and 4.5.

Methodology of Section 4.4. To demonstrate the superior performance (in small samples) of the new nonparametric estimation procedure, we designed the following simulation study. We generate data using model (4.4) with two FPC's, i.e.

$$X(\mathbf{s};t) = \mu(t) + \zeta_1(\mathbf{s})v_1(t) + \zeta_2(\mathbf{s})v_2(t). \tag{4.18}$$

We set

$$v_1(t) = \sin(2\pi t \cdot 7) + \sin(2\pi t \cdot 2), \quad v_2(t) = \sqrt{2}\sin(2\pi t \cdot 3);$$

$$\boldsymbol{\zeta}_1 \sim N(\mathbf{0}, \boldsymbol{\Gamma}_1), \quad \boldsymbol{\zeta}_2 \sim N(\mathbf{0}, \boldsymbol{\Gamma}_2),$$

where the elements of the covariance matrices have the following parametric form:

$$\gamma_1(d_{k\ell}) = \exp(-d_{k\ell}/0.1), \quad \gamma_2(d_{k\ell}) = 0.2\exp(-(d_{k\ell}/0.3)^2).$$

The spatial locations $\mathbf{s}_k$ are uniformly distributed on the unit square (different locations for every MC replication).

We compare four estimation procedures:

- S Simple average, which totally ignores the spatio–temporal dependence;
- T The infeasible method which uses the true covariance matrix $\mathbf{C} = \boldsymbol{\Gamma}_1 + \boldsymbol{\Gamma}_2$;

P The method that uses a *misspecified* parametric spatio–temporal covariance function $\sigma^2 C(\mathbf{h}; u)$, with $C(\mathbf{h}; u)$ given by

$$C(\mathbf{h}; u) = \frac{1}{1 + a|u|^{2\alpha}} \exp\left(-\frac{c\|\mathbf{h}\|}{(1 + a|u|^{2\alpha})^{\beta/2}}\right). \tag{4.19}$$

(This covariance function is discussed in detail in Example 4.7.1 below in this section.)

N The new nonparametric method.

Method S corresponds to using only step (a) in the algorithm presented earlier in this section. It would be the default method if standard R or Matlab software were to be used. The remaining methods are iterative, and differ by the way in which the covariance matrix in step (b) is estimated. The parametric model used in method P is discussed in greater detail in Example 4.7.1.

Table 4.2 compares the performance of the four methods for sample sizes ranging from 15 to 40. We report Monte Carlo averages and standard deviations of the L_2 distance defined as

$$L_2 = \left\{ \int (\hat{\mu}_i(t) - \mu(t))^2 \, dt \right\}^{1/2}. \tag{4.20}$$

The most important conclusion is that method N is significantly better than all other methods at the 5% confidence level (a difference of more than two standard deviations, which will be the benchmark in the discussion that follows). It is even better than the parametric method T which assumes the known true covariances. This illustrates a relatively well–known fact that a flexible nonparametric model can approximate the stochastic structure of the data better than a true parametric model, when the number of data points is small. Method P, based on a more flexible parametric family is significantly worse than method N, but "almost" significantly better than Method T: the differences between N and P are significant at about 10% level. The standard method S is significantly worse than any other methods. This means that taking into account spatial dependence of the curves even in a suboptimal way significantly improves the estimates. We emphasize that for method P only the cases in which the variogram optimization converged were considered; it did not

Table 4.2: Average L_2 distance between the estimated and true mean functions for different estimation methods; the second number represents the standard error. The entries are based on 10^4 replications.

Sample	S	T	P	N
15	0.164 ± 0.006	0.151 ± 0.005	0.145 ± 0.006	0.069 ± 0.004
20	0.137 ± 0.005	0.123 ± 0.004	0.116 ± 0.005	0.061 ± 0.003
25	0.122 ± 0.005	0.108 ± 0.004	0.098 ± 0.004	0.059 ± 0.004
30	0.115 ± 0.004	0.105 ± 0.004	0.098 ± 0.004	0.060 ± 0.004
35	0.109 ± 0.004	0.096 ± 0.004	0.089 ± 0.004	0.060 ± 0.004
40	0.104 ± 0.005	0.089 ± 0.004	0.082 ± 0.004	0.059 ± 0.005

converge in over 10% of replications. The conclusions reached from the analysis of Table 4.2 do not change if the data generating process (4.18) is modified by adding more principal components which however do not account for more than 10% of variability, as is the case for the ionosonde data.

The results reported in Table 4.2 reveal good performance of Method P based on the misspecified covariances (4.19). The example below considers this spatio–temporal model a bit closer. It is not a model for spatially indexed functional data that drives our methodology, so it serves to underline the flexibility and extendability of our approach, which proposes to estimate the spatial components nonparametrically, when the number of spatial locations is small. It turns out that our nonparametric approach can improve on the current estimation methodology for model (4.19) as well.

EXAMPLE 4.7.1 We now consider the model $Z(\mathbf{s};t) = \mu + \varepsilon(\mathbf{s};t)$, with the true mean $\mu = 0$. The spatial locations $\mathbf{s}_k$ are uniformly distributed on the unit square (different locations for every MC replication). The errors are normal with the space–time correlation function given by (4.19), i.e. by Eq. (14) in Gneiting (2002). The scale parameters a and c are nonnegative, the smoothness parameter α, and the space-time interaction parameter β take values in $[0, 1]$. The goal is to study performance of the nonparametric method in three regimes: no space-time interaction $\beta = 0$, moderate space-time interaction $\beta = 0.5$, and strong space-time interaction $\beta = 1$. The parameter a is also of importance; if it is small, it induces long range temporal correlation, if it is large the temporal correlation decays fast. The space scale parameter and the smoothness parameter are fixed ($c = 5$, $\alpha = 0.7$). Again, to estimate the mean, we must estimate the covariance structure. The details are outlined at the end of this example. The methods we study are:

S Simple average, which totally ignores the spatio–temporal dependence;

T The infeasible method which uses the *correct* model and *true* parameter values;

P The parametric method that uses the *correct* model and *estimated* parameters;

N The new nonparametric method.

The results of our simulations are displayed in Table 4.3. In all but four cases, method N is significantly better than method P at 5% level of significance. In the remaining four cases, the average L_2 distance for method N is smaller, but the difference is smaller than two standard deviations. The difference is much more significant than 5% for $N = 15$ and $N = 20$. The marginally insignificant result for $N = 15$ and $\beta = 1, a = 1$ could be a type II error. In all cases considered the infeasible method (T) is not significantly better than N at 5% level. Somewhat surprisingly, in a few cases with $N = 15$ and $N = 20$ method P is not significantly better than the trivial method S.

We conclude this example by outlining the covariance estimation procedure for model (4.19), see Section 4 in Gneiting (2002) for more details. By plugging in (4.19) into (4.9) one can see that for mean estimation only a purely spatial covariance is needed:

$$C(\mathbf{h}) = \int_0^1 C(\mathbf{h}; 0)dt = \exp\left(-c\|\mathbf{h}\|\right).$$

After determining the parameter c via nonlinear fitting we calculate the weights using (4.8). For method N, we estimate the covariance function nonparametrically, as described in Section 4.2.

Methodology of Section 4.5. There are many reasonable data generating processes that could be used to evaluate the finite sample performance of the normal test based on

Table 4.3: Average L_2 distance between the estimated and true means; the second number represents the standard error. Entries are based on 10^4 replications.

Parameters	Sample	S	T	P	N
$\beta = 0,\ a = 0.1,$	15	0.201 ± 0.002	0.185 ± 0.003	0.192 ± 0.004	0.181 ± 0.001
	20	0.188 ± 0.002	0.167 ± 0.002	0.183 ± 0.007	0.166 ± 0.002
	25	0.178 ± 0.002	0.158 ± 0.003	0.158 ± 0.004	0.155 ± 0.002
	30	0.173 ± 0.003	0.154 ± 0.003	0.158 ± 0.006	0.150 ± 0.003
$\beta = 0,\ a = 1,$	15	0.199 ± 0.002	0.183 ± 0.002	0.191 ± 0.003	0.181 ± 0.002
	20	0.189 ± 0.002	0.171 ± 0.003	0.178 ± 0.006	0.166 ± 0.002
	25	0.179 ± 0.002	0.159 ± 0.002	0.163 ± 0.003	0.156 ± 0.002
	30	0.172 ± 0.002	0.150 ± 0.002	0.154 ± 0.003	0.149 ± 0.001
$\beta = 0.5,\ a = 0.1,$	15	0.200 ± 0.002	0.182 ± 0.002	0.196 ± 0.004	0.182 ± 0.002
	20	0.189 ± 0.002	0.166 ± 0.002	0.171 ± 0.002	0.166 ± 0.002
	25	0.179 ± 0.003	0.157 ± 0.005	0.167 ± 0.006	0.154 ± 0.003
	30	0.173 ± 0.002	0.152 ± 0.002	0.158 ± 0.003	0.151 ± 0.002
$\beta = 0.5,\ a = 1,$	15	0.199 ± 0.002	0.181 ± 0.002	0.188 ± 0.002	0.180 ± 0.002
	20	0.182 ± 0.002	0.161 ± 0.002	0.171 ± 0.004	0.161 ± 0.002
	25	0.179 ± 0.002	0.157 ± 0.002	0.167 ± 0.004	0.156 ± 0.002
	30	0.174 ± 0.002	0.162 ± 0.006	0.168 ± 0.005	0.151 ± 0.001
$\beta = 1,\ a = 0.1,$	15	0.199 ± 0.001	0.185 ± 0.003	0.196 ± 0.004	0.181 ± 0.002
	20	0.189 ± 0.002	0.171 ± 0.003	0.180 ± 0.007	0.166 ± 0.002
	25	0.177 ± 0.002	0.157 ± 0.002	0.166 ± 0.004	0.156 ± 0.002
	30	0.172 ± 0.002	0.150 ± 0.001	0.158 ± 0.003	0.148 ± 0.001
$\beta = 1,\ a = 1,$	15	0.201 ± 0.001	0.185 ± 0.002	0.189 ± 0.002	0.186 ± 0.002
	20	0.187 ± 0.002	0.172 ± 0.003	0.176 ± 0.003	0.166 ± 0.002
	25	0.181 ± 0.002	0.159 ± 0.002	0.165 ± 0.003	0.157 ± 0.002
	30	0.172 ± 0.002	0.150 ± 0.001	0.152 ± 0.003	0.149 ± 0.001

$\hat{\beta}_i/\sqrt{\mathrm{Var}[\hat{\beta}_i]}$. A number of spatial designs, shapes of the FPC's and the covariance functions for the scores could be employed. To avoid producing a large number of tables, we focus on a simulation study relevant to the science problem we consider in Section 4.6. The data generating processes are designed to resample true data.

To evaluate the empirical size and power, we generate the data using model (4.17) with $p = 3$ because the first three estimated FPC's explain about 91–92% of the variance for each of the three types of data. The coefficients β_1 and β_3 are equal to these coefficients estimated from the real data. To evaluate the size, we set $\beta_2 = 0$, to study the power,

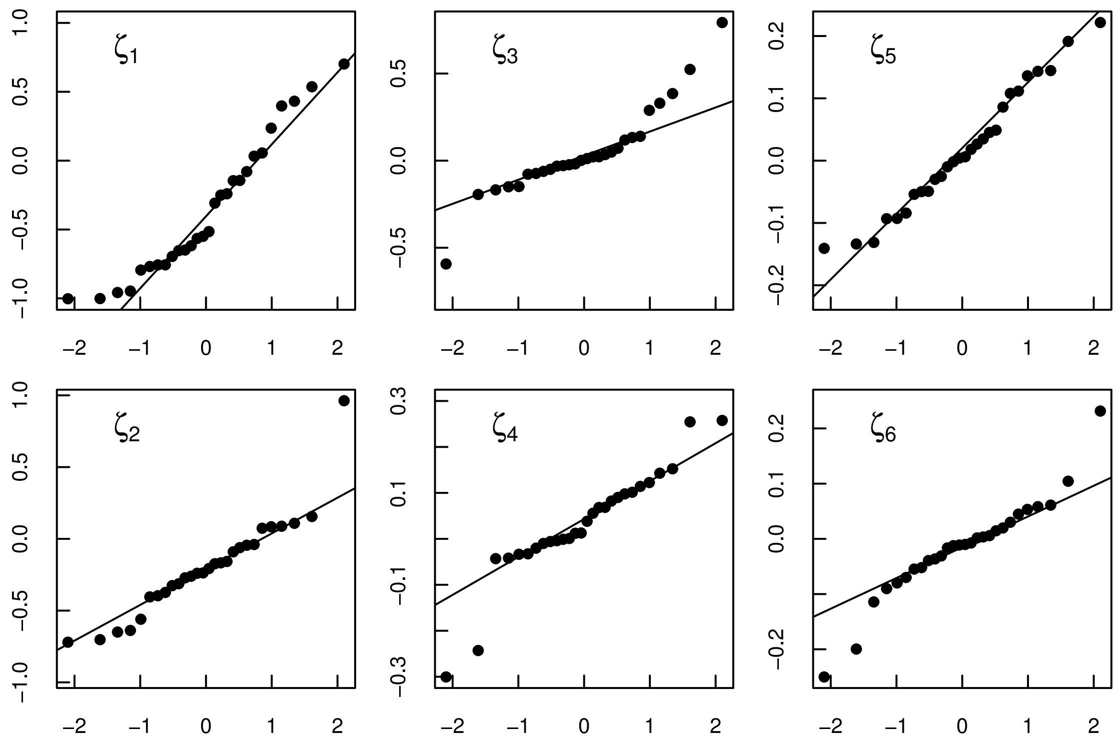

Figure 4.4: Normal QQ-plots plots for the estimated scores, ζ_i, $1 \leq i \leq 6$ for the Day data.

we consider $0 < \beta_2 \leq 0.5$. The FPC's v_j are equal to those estimated from the real data. The vectors $\boldsymbol{\zeta}_j$ are $N(\mathbf{0}, \hat{\boldsymbol{\Gamma}}_j)$ with the $\hat{\boldsymbol{\Gamma}}_j$ equal to the covariance matrices estimated from the data (using the nonparametric method). Monte Carlo replications are generated by repeated simulations of the vectors $\boldsymbol{\zeta}_j$. The assumption of normality holds to a reasonable approximation as shown in Fig. 4.4 for the Day data. The plots for the Noon and Night data look similar. There is one outlying point in the QQ–plot of the $\zeta_2(\mathbf{s}_k)$, and the plot for the $\zeta_3(\mathbf{s}_k)$ indicates some departure from normality. The third FPC contributes however less than 10% to the variance, so its impact on our conclusions is small. In fact, the after removing this point, the trend parameter practically did not change.

The empirical size of the test developed in Section 4.5 is reported in Table 4.4 as a function of p. It is remarkably close to the nominal size and does not depend on p, as long as it remains in a reasonable range. This remains true if the scores are not normal, but

Table 4.4: Empirical Size of the test of $H_0 : \beta_2 = 0$.

	Day			Noon			Night		
p	10%	5%	1%	10%	5%	1%	10%	5%	1%
2	9.06	4.24	0.90	10.28	5.52	1.14	10.47	5.34	1.17
3	9.94	4.75	1.23	9.82	5.15	1.01	10.51	5.24	1.04
4	9.80	4.76	1.03	10.01	5.08	1.00	9.78	4.77	1.07
5	10.21	4.93	1.16	9.52	4.91	1.06	9.95	4.88	1.01

Table 4.5: Empirical Size for the "simple" method.

	Day			Noon			Night		
p	10%	5%	1%	10%	5%	1%	10%	5%	1%
2	65.97	60.82	49.91	27.34	19.21	8.35	54.70	47.59	34.79
3	61.95	55.42	43.23	69.34	64.02	53.98	55.51	47.27	35.65
4	60.22	53.49	41.40	68.20	63.73	54.63	55.52	48.73	35.94
5	62.18	55.31	43.51	70.51	62.15	53.24	54.75	47.08	34.24

severe departures from normality may distort the size for some values of p. For comparison, Table 4.5 shows the sizes for the method which can be termed "simple." It also relies on (4.13) and (4.15), but it uses the standard estimation procedure implemented in software packages discussed in Ramsay *et al.* (2009). (In the "simple" method, we set $w_k = 1/N$ and the v_j in $\mathbf{\Omega}$ are estimated as the eigenfunctions of the usual empirical covariance operator.) The "simple" method does not take into account the spatial dependence of the curves. This method severely overrejects. The empirical power of the new method is displayed in Fig. 4.5. The power curves for the Day data rise less steeply than those for the Noon and Night data. We will return to these curves when we discuss the results of the application of the test to the real data.

4.8 Summary and conclusions

The research reported in the paper is the outcome of our attempts to solve the hypothesis of global ionospheric cooling in a manner that would survive a rigorous scientific scrutiny. Our initial approaches failed because the standard parametric spatial model fitting

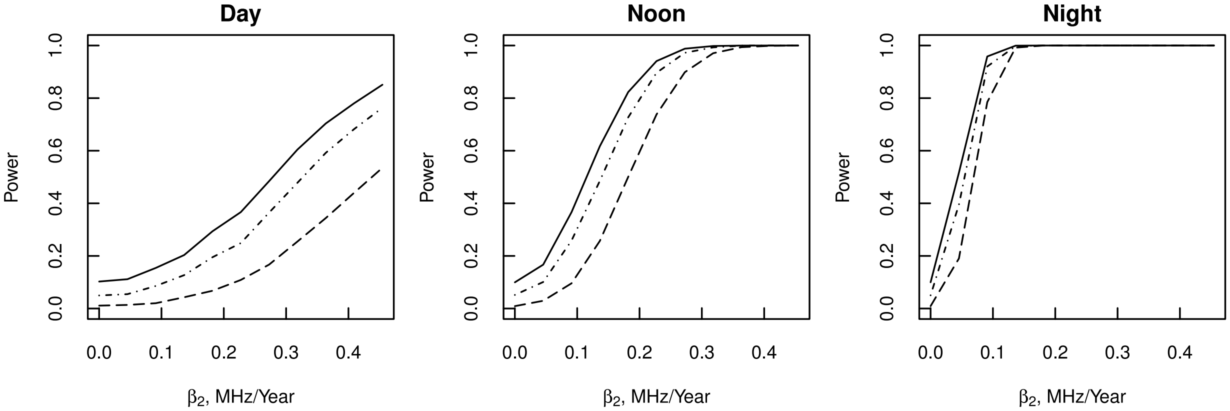

Figure 4.5: Empirical power. Solid line - empirical power for $\alpha = 10\%$, dash–dotted line - empirical power for $\alpha = 5\%$, dashed line - empirical power for $\alpha = 1\%$.

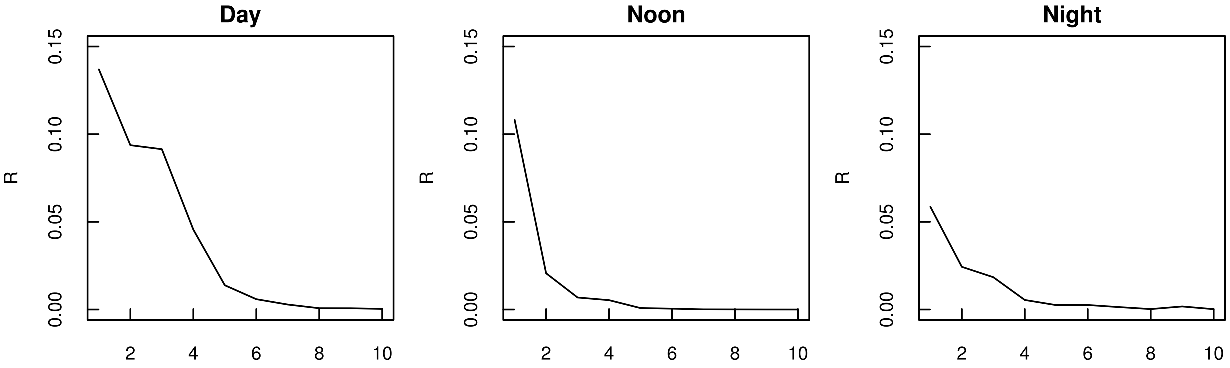

Figure 4.6: Convergence of the iterative method for the estimation of the means R_i as a function of iteration i.

produces very poor results if a small sample of spatial locations is available. To address this issue, we built on the nonparametric approach and developed a set of tools that can be used with confidence in small samples of spatially distributed curves. The main ingredients of the new methodology are the following. 1) Nonparametric estimation procedure for the mean function which produces significantly smaller mean squared errors than any of the existing procedures. 2) Estimation procedure for the mean function expressed as a linear combination of known functions. 3) A test to determine the significance of the coefficient of any of the known functions in 2). As explained in the introduction, we hope that our methodology will be used in other problems of inference for functional data available at a small number of spatial locations.

4.9 Bandwidth selection and the construction of the functional confidence intervals

In this section, we describe the procedures for the selection of the bandwidth h and for the construction of the functional confidence intervals, the two important ingredients of the methodology introduced in Section 4.2.

Regarding the choice of h, we performed a very extensive simulation study to evaluate the performance of several potential methods. It is important to note that the nonparametric covariance function estimation described in Section 4.2 is only an ingredient of a broader methodology for the estimation and testing in the functional spatio–temporal framework. The choice of the bandwidth must thus be be tailored to the problems we want to solve. These problems revolve around the estimation of the mean function $\mu(t)$ in model (4.4). The procedure employed in all numerical work reported in this paper is the following: We first center the functions by their average (their possible spatial dependence is not taken into account). This step transforms the functions to a set of *approximately* mean zero functions. We then estimate the covariance functions using a several choices of h, the same h for every functional principal component. We select h for which the estimated mean function is closest to zero. In other words, we select h which minimizes

$$\|\hat{\mu}_h - 0\|^2 = \int \hat{\mu}_h^2(t)dt$$

where $\hat{\mu}_h$ is the final estimated mean function (of the centered curves) obtained using bandwidth h.

The other approaches we experimented with included: 1) cross–validation to minimize the integrated mean squared error of $\hat{m}$ given by (4.2) or of $\hat{\gamma}$ described in the next paragraph. 2) Cross–validation to minimize the integrated mean squared error of the estimated function $\hat{\mu}_h$. 3) Spatial versions of the cross validations in 1) and 2), in which the removed observation is replaced by a spatial prediction obtained using kriging with various values of h. None of these approaches yielded uniformly satisfactory results.

We now turn to the construction of functional confidence bounds for the covariance function $\gamma(\cdot)$. The idea is as follows. First, using the quasi–bootstrap procedure proposed by Solow (1985) and further improved by Clark and Allingham (2011), we produce a collection of M independent covariance curves, $\hat{\gamma}_i(d)$, $1 \le i \le M$. Then using the concept of a functional depth we construct the confidence bounds. The concept of functional depth has been extensively used lately, see Fraiman and Muniz (2001), Febrero *et al.* (2008), López-Pintado and Romo (2009) and Sun and Genton (2011), but not in the context of covariances of spatial data.

We start with outlining the quasi–bootstrap procedure. The estimated covariance matrix $\hat{\mathbf{\Gamma}}$ is decomposed using the Cholesky decomposition as $\hat{\mathbf{\Gamma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^T$, where $\hat{\mathbf{L}}$ is the lower triangular matrix. Using $\hat{\mathbf{L}}$, the spatial field $\boldsymbol{\zeta}$ can be decorrelated as $\boldsymbol{\zeta}_0 = \hat{\mathbf{L}}^{-1}\boldsymbol{\zeta}$. Next, $\boldsymbol{\zeta}_0$ is resampled with replacement and recorrelated $\boldsymbol{\zeta}^i = \hat{\mathbf{L}}\boldsymbol{\zeta}_0$. The superscript $1 \le i \le M$ refers to the iteration step. Based on the bootstrap sample $\boldsymbol{\zeta}^i$, the correlation is estimated by $\hat{\gamma}_i(d)$ using the nonparametric method. These steps are repeated sufficiently many, say $M = 1,000$, times. This leads to a collection of independent covariance curves $\hat{\gamma}_i(d)$.

A functional depth can be defined in many different ways. Here we use the definition presented in Chapter 1 of Horváth and Kokoszka (2012). Let

$$F_{N,d}(\gamma) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{I}\left\{\hat{\gamma}_i(d) \le \gamma\right\},$$

be the empirical distribution function at point d. We define the functional depth (FD) of the curve $\hat{\gamma}_i$ by integrating the univariate depth:

$$FD_N(\hat{\gamma}_i) = \int (1 - |1/2 - F_{N,d}(\hat{\gamma}(s))|)\, ds.$$

Next, we select K curves with the largest depth. Then, for each d we find $\min(\hat{\gamma}_i(d))$ and $\max(\hat{\gamma}_i(d))$ from among these K curves. This produces the lower and upper bounds which form the functional analogue of the univariate $1 - \alpha = K/M$ confidence interval. An Example of the functional confidence bounds is shown in Fig. 4.1.

CHAPTER 5

EVALUATION OF THE COOLING TREND IN THE IONOSPHERE USING FUNCTIONAL REGRESSION WITH INCOMPLETE CURVES[1]

Abstract

Long–term trends in the ionosphere can impact the operation of space–based civilian and defense systems. The ionospheric cooling trend studied in this paper is also related to the global warming hypothesis; both are attributed to the same drivers. The hypothesis that such a trend exists has been an important focus of space physics research for two decades. A central difficulty in reaching broadly agreed on conclusions was the absence of data with sufficiently long temporal and sufficiently broad spatial coverage. Time series of data that cover several decades exist only in several separated (industrialized) regions. The space physics community has struggled to combine the information contained in these data, and often contradictory conclusions have been reported based on the analyses relying on one or a few locations. We present a statistical analysis that uses all data, even those with incomplete temporal coverage. It is based on a new functional regression approach that can handle unevenly spaced, partially observed curves. We conclude that a statistically significant cooling trend exists in the Northern Hemisphere. This confirms the hypothesis put forward in the space physics community over two decades ago. We also define the minimum requirements on the number of curves and their temporal extent in order for statistical significance of trend to be determined.

5.1 Introduction

---

[1] CO-AUTHORED BY O. GROMENKO, P. KOKOSZKA, AND J. SOJKA. THIS PAPER IS SUBMITTED TO THE JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION: APPLICATIONS AND CASE STUDIES.

This paper is concerned with a long standing problem of space physics research. The increased concentration of greenhouse gases in the upper atmosphere is associated with global warming in the lower troposphere. Roble and Dickinson (1989) suggested that the increasing amounts of these radiatively active gases, mostly CO_2 and CH_4, would lead to a global cooling in the thermosphere by about 50K. Rishbeth (1990) pointed out that this would result in a thermal contraction of the atmosphere and the global lowering of the ionospheric peak height and the decrease of the inospheric peak density, see Fig. 5.1. The F region peak has been observed for many decades by globally distributed ground–based ionosondes. The ionosonde is a type of radar projecting a spectrum of high-frequencies (HF) vertically into the ionosphere. In principle, these observations could be used to quantitatively test the hypothesis of Roble and Dickinson (1989). A long term change in the upper atmosphere can impact space–based navigation (including GPS systems), HF (2-30MHz) radio communication and the operation of low orbit satellites. It is associated with the global warming hypothesis because a physical mechanism for the conjectured cooling trend is also attributable to greenhouse gases.

The ionospheric layer which contains the peak electron density is known as the F2 region (the right–most peak in Fig. 5.1). Ionosonde measurements allow us to observe a critical frequency, denoted foF2. The ionosonde observes the frequency associated with the peak plasma density known as the plasma frequency or the critical frequency of the ordinary ray propagation of the transmitted radio wave. This frequency decreases as the peak density proportional to the square root of the peak density. There has consequently been extensive space physics research aimed at determining if a decreasing temporal trend in the foF2 frequency indeed exists. Lastovicka *et al.* (2008) review some of the relevant literature.

Long-term changes in the upper atmosphere are usually described using a linear approximation referred to as the ionospheric trend. The main problem in its determination is the separation of the solar activity and other factors, like the long term changes in the internal magnetic field of the Earth; the solar cycle dominates the shape of the foF2 curves,

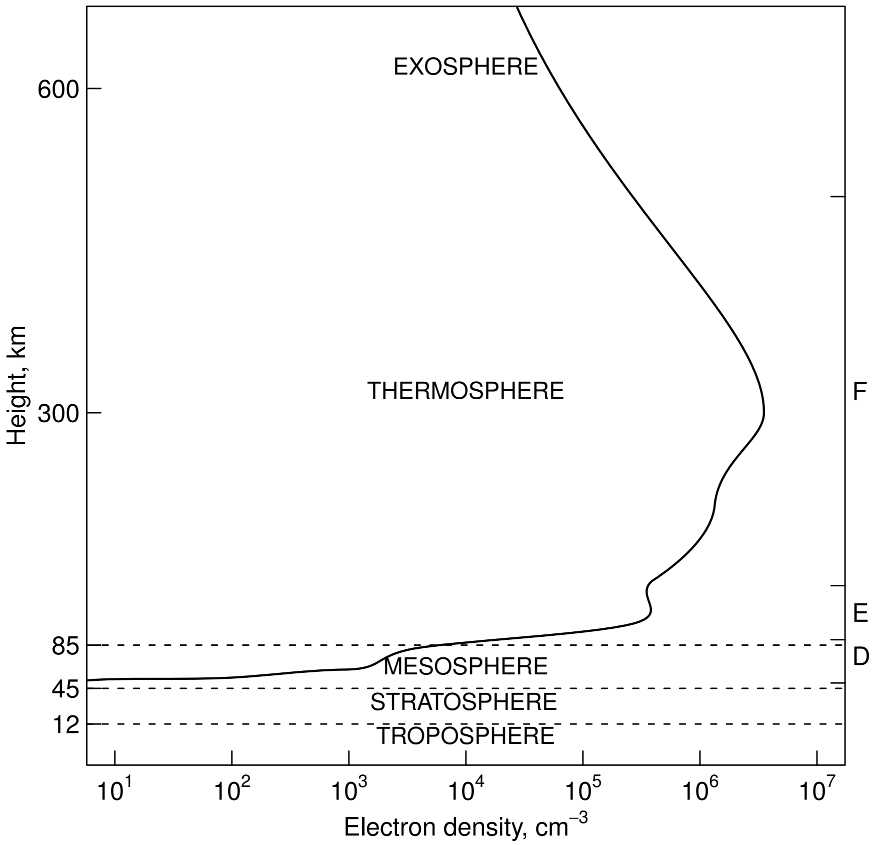

Figure 5.1: Typical profile of day time ionosphere. The curve shows electron density as a function of height. The right vertical axix indicates the D, E and F regions of the ionosphere.

see Fig. 5.2. A comprehensive overview of statistical methods proposed in the space physics community is given in (Lastovicka *et al.*, 2006). The main problem from which they suffer is their inability to combine the information from many spatial locations. The usual approach is to calculate trends separately at a number of locations, often using different time periods, and then average these trends to obtain a sense of a global trend, see Bremer *et al.* (2012) for a recent contribution and a discussion of previous work. There has, however, long been a sentiment in the ionospheric physics community, that, in addition to informative exploratory analyses, an inferential statistical framework should be developed to address the question of the existence of long term ionospheric trends; Ulich *et al.* (2003) stress that to make any trends believable, a suitable statistical modeling, a proper treatment of “errors and uncertainties” is called for.

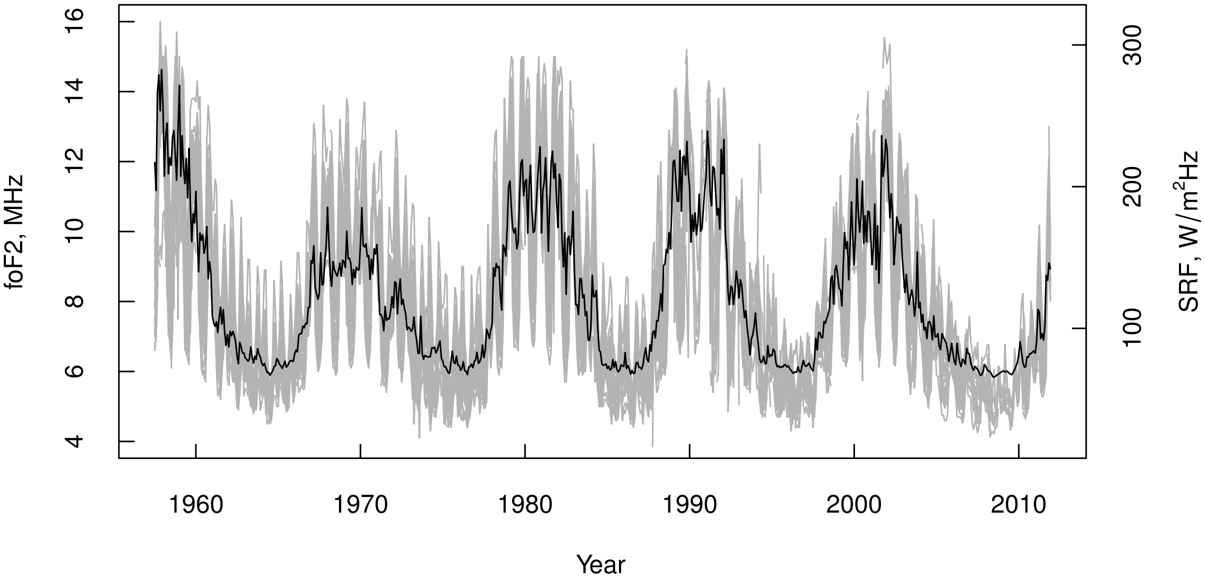

Figure 5.2: Gray lines represent all foF2 records analysed in this paper with the scale on the left-hand side. The black line represents the observed solar radio flux with the scale on the right-hand side.

Our objective is to make a contribution in this direction which establishes the existence of the negative foF2 trend over the mid–latitude Northern Hemisphere with statistical significance. This is achieved by developing an inferential framework which allows us to combine incomplete ionosonde records from globally distributed locations and take their spatial dependence into account. The absence of complete records has been a major stumbling block in space physics research to date. Our approach is developed in the framework of functional data analysis: the ionosonde records are viewed as spatially indexed curves which are only partially observed.

There has been an increasing interest in correlated (in particular spatially dependent) functional data. Such data occur in many settings of practical relevance: meteorological and pollution variables at many locations measured over long periods of time, records of brain activity at a number of locations within the brain, economic or health variables indexed by counties, etc. An interested reader is referred to Delicado *et al.* (2010), Giraldo *et al.*

(2009, 2010, 2011, 2012), Nerini *et al.* (2010), Secchi *et al.* (2011, 2012), Jiang and Serban (2012), Crainiceanu *et al.* (2012), Staicu *et al.* (2010, 2012), and Chapters 2, 3 and 4 herein. Even though our new functional regression technique has been developed to solve a specific science problem, it is hoped that it will be received with interest as a more broadly applicable tool of functional data analysis.

The remainder of the paper is organized as follows. In Section 5.2, we introduce the space physics data we work with. Section 5.3 is devoted to the new statistical methodology we had to develop to solve the problem outlined above. Some technical aspects of this methodology are explained in Sections 5.5, 5.6 and 5.7. In Section 5.4, we apply these tools to establish, with statistical significance, the existence of a negative foF2 trend in the mid–latitude Northern Hemisphere.

5.2 The data

The main data used in our study are the foF2 values calculated from ionosonde radio wave echoes from the F2 layer. The raw data are available at the Space Physics Interactive Data Resource (SPIDR), `http://spidr.ngdc.noaa.gov/spidr/`. In principle, these are equidistant time series at over 200 locations with the typical separation between the observations of one hour. In practice, these raw data contain huge gaps, often over a decade long, as well as a large number of “sporadic” missing observations, most likely due to equipment failure or maintenance. Missing data are often not indicated and not plugged in a standard way. Even at the same location, the foF2 values are sometimes reported in different units at various times in the pat six decades. We developed a C++ program which converts the raw data to standard units and to regularly spaced time series with missing values flagged. Due to rounding of geographical coordinates, some stations appear to have the same location. When this happens, we use exact locations provided by external sources. Also, some stations are listed twice with different 5-digit SPIDR codes. For example, HAJ43 and HAJ45 both are Hanscom AFB, MA. In this case, we use the record with the 5-digit code which has the lower numeric part. In some cases, we merge such records to obtain longer temporal coverage.

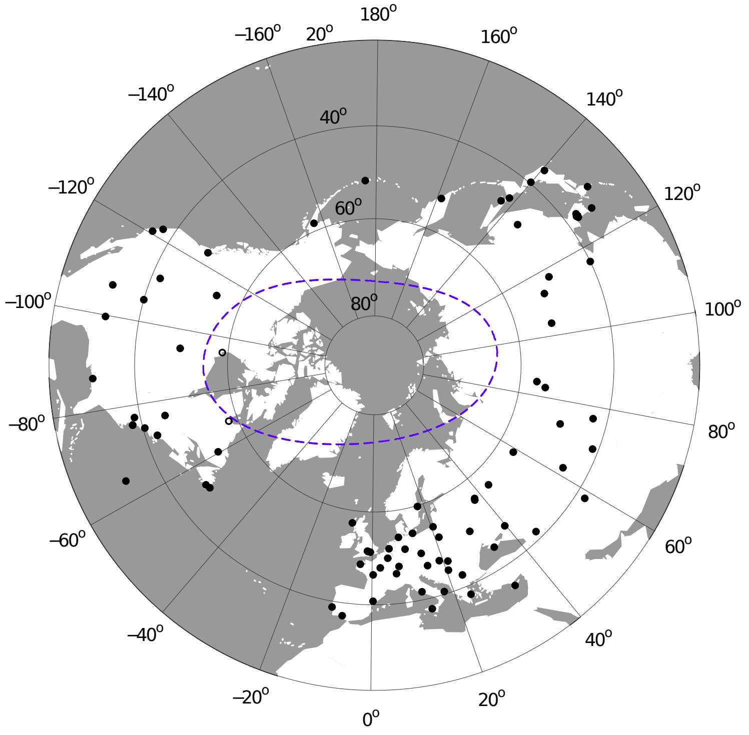

Figure 5.3: Locations of the 85 ionosonde stations used in this study, black discs. The two circles in northern Canada represent stations located in the auroral zone (dashed line), which were not used.

For the study reported in the paper, we calculated monthly medians using only near noon observations, 10-14 LT (LT denotes local solar time). At noon, the behavior of the ionosphere is completely dominated by the solar radiation, see Fig. 5.2, which can be removed using our regression model. At night, the behavior of the ionosphere is complicated, and we postpone the study of the night time data to a more specialized space physics paper. Our statistical study requires the assumption of spatial stationarity. To make this assumption reasonable, we focus only on the mid–latitude region located between 30°N and 60°N geographic latitude. The ionosphere can be divided into three regions, equatorial, mid–latitude and auroral. It exhibits different electron density profiles in each of these regions, with the profile shown in Fig. 5.1 typical of the mid–latitude region. The reason for choosing the northern hemisphere is that it contains the longest records with the most extensive spatial coverage, see Fig. 5.3. We dropped two Canadian stations located between 30°N and 60°N which are however in the auroral zone (determined by the magnetic coordinates). Visual examination shows that these two records indeed appear to be outliers. The total number of selected stations is 85. The majority of the ionosondes started to operate in 1957, the international geophysical year. We selected the time interval from July 1957 to December 2011, so that the total number of months is 654. While the total number of selected stations is 85, the number of stations available at any specific month never exceeds 50.

The foF2 curves are used as responses in our functional regression. The main explanatory variable is the observed solar radio flux (SRF), also available at SPIDR, which is a well established proxy for the solar activity. We also use another regressor which is a function of the direction of the internal magnetic field of the Earth, which has changed at every ionosonde location over the time span of the data. These directions are computed using the international geomagnetic reference field (IGRF); the software is available at `http://www.ngdc.noaa.gov/IAGA/vmod/`.

5.3 Statistical model and inference

In order to develop an inferential procedure, a statistical model for the data must be postulated. Denote by $\{\mathbf{s}_k, 1 \leq k \leq N\}$ the locations at which the functional field is observed. We assume that each curve $X(\mathbf{s}_k, \cdot)$ is an element of a *strictly stationary* spatial random field taking values in the space L^2 of square integrable functions. This assumption implies that all curves $X(\mathbf{s}_k, \cdot)$ have the same distribution in L^2, in particular, they have the same mean function $\mu(t) = E[X(\mathbf{s}, t)]$ and the same functional principal components (FPC's), which we denote by $v_j(t)$. The inference on the mean function $\mu(\cdot)$ is the main objective of this research; this function may or may not contain the conjectured foF2 trend. The main difficulty arising in the work that follows is that the curves $X(\mathbf{s}_k, \cdot)$ are often not available over long periods of time, these periods being different at different locations $\mathbf{s}_k$. There is also a small measurement error, which we denote by $\theta(\mathbf{s}; t)$. We assume that $\theta(\mathbf{s}, t)$ are mean zero iid random variables with variance σ_θ^2. We also assume that the random fields X and θ are independent. Under these assumptions, the model for the data is

$$X(\mathbf{s};t) = \mu(t) + \sum_{j=1}^{\infty} \zeta_j(\mathbf{s}) v_j(t) + \theta(\mathbf{s};t), \tag{5.1}$$

where the second term on the right-hand side is the Karhunen–Loéve expansion, see e.g. Chapter 17 of Horváth and Kokoszka (2012). Each ζ_j is a strictly stationary mean zero scalar random field. For every $\mathbf{s}$, $E[\zeta_{j'}(\mathbf{s})\zeta_j(\mathbf{s})] = 0,\ j' \neq j$ (this is a general property of the scores). We impose a stronger assumption that for any $\mathbf{s}, \mathbf{s}'$,

$$E[\zeta_{j'}(\mathbf{s}')\zeta_j(\mathbf{s})] = 0, \qquad j' \neq j, \tag{5.2}$$

which is needed to derive a test statistic whose distribution can be approximated. In Section 5.5, we show that (5.2) is a very reasonable assumption for the foF2 data. Using a mathematical argument, we also show in Section 5.5 that (5.2) holds for every separable spatio–temporal random field, i.e. a field for which

$$\mathrm{Cov}\left(X(\mathbf{s},t), X(\mathbf{s}',t')\right) = \Sigma(\mathbf{s},\mathbf{s}')c(t,t'). \tag{5.3}$$

We however do not assume separability. The covariance structure of our model, which can be viewed as a spatio–temporal field, is

$$\mathrm{Cov}\left(X(\mathbf{s},t), X(\mathbf{s}',t')\right) = \sum_{j=1}^{\infty} \Sigma_j(\mathbf{s},\mathbf{s}') v_j(t) v_j(t'),$$

where

$$\Sigma_j(\mathbf{s},\mathbf{s}') = E[\zeta_j(\mathbf{s})\zeta_j(\mathbf{s}')].$$

In our estimation procedure, we assume that the fields ζ_j are isotropic, an assumption that holds reasonably well for the foF2 data.

5.3.1 Estimation in the presence of incomplete records

In this section, we introduce a new method for the estimation of the mean function $\mu(\cdot)$ and the FPC's $v_j(\cdot)$ in model (5.1). This new approach is called for by the need to deal with incomplete data.

Estimation of the mean function. For complete records, Chapter 2 proposed several approaches. The most straightforward method is to estimate the mean by the weighted sum:

$$\hat{\mu}(t) = \sum_{k=1}^{N} w_k X(\mathbf{s}_k; t), \quad \sum_{k=1}^{N} w_k = 1. \tag{5.4}$$

The optimal weights are found by minimizing the expected value of the L^2 distance between the mean and its estimator (5.4), subject to constrain $\mathbf{w}^T\mathbf{1} = 1$. This leads to the following expression for the weights:

$$\mathbf{w} = \mathbf{\Sigma}^{-1}\mathbf{1}/(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1}), \tag{5.5}$$

where $\mathbf{\Sigma}$ is an $N \times N$ positive definite matrix with entries

$$\Sigma(\mathbf{s}_k, \mathbf{s}_\ell) = E \int \{X(\mathbf{s}_k; t) - \mu(t)\} \{X(\mathbf{s}_\ell; t) - \mu(t)\} \, dt. \tag{5.6}$$

It is clear that (5.4) and (5.6) require the curves $X(\mathbf{s}_k; t)$ to be complete. This is particularly important in (5.6); for incomplete records, it may happen that the segments over which $X(\mathbf{s}_k; t)$ and $X(\mathbf{s}_\ell; t)$ are available are practically disjoint. A similar problem occurs if one attempts to use the methods of the estimation of the FPC's v_j developed in Chapter 2 in situations when large segments of data are missing. These difficulties motivate us to propose an different approach which we now describe.

Let N be the total number of curves, i.e. the number of ionosonde stations used in the study. By t_i, $1 \le i \le T$, we denote all possible times at which the ionosonde records can, in principle, be evaluated. By N_i, $0 \le N_i \le N$, we denote the number of observations actually available at time t_i, and by T_k the actual number of observations of the curve $X(\mathbf{s}_k)$.

The new method also uses weights to account for spatial dependence. To handle missing observation we use smoothing, common in longitudinal data analysis. The mean function is estimated by the local linear indexed regression:

$$\begin{aligned}&(\hat{m}_0(t), \hat{m}_1(t)) \\ &= \arg\min_{m_0,m_1} \sum_{i=1}^{T} \kappa_\mu \left(\frac{t-t_i}{h_\mu}\right) \left\{ \sum_{k=1}^{N_i} w_k(t_i) X(\mathbf{s}_k; t_i) - m_0 - m_1(t-t_i) \right\}^2 .\end{aligned} \tag{5.7}$$

The curve $\hat{m}_0(\cdot)$ is the estimate of the mean function $\mu(\cdot)$. We report the results obtained by using the Epanechnikov kernel $\kappa_\mu(t) = 3/4(1-t^2)\mathbf{1}_{[-1,1]}(t)$ because it has several desirable properties, see e.g. Theorem 3.4 in Fan and Gijbels (1996). Simulations and application to foF2 data show that the choice of kernel plays practically no role. The conclusions for the foF2 data do not depend on the choice of the bandwidth h_μ either, as long as it is in a reasonable range, so that the smoothed curves visually follow the raw data. Specific values are given in Section 5.4.

The main idea encapsulated in formula (5.7) is that at each time t_i we use only the N_i available curves; the weights $w_k(t_i)$, which capture the spatial structure, depend on t_i. Their calculation is discussed in the following. This is a novel aspect because smoothing methodology developed to date, see Yao *et al.* (2005), Yao and Lee (2006), Müller and Yao (2008), assumes independence of the curves.

Calculation of the weights. In Chapter 2 proposed the so-called functional variogram:

$$2\gamma(d_{k\ell}) = E\left\{ \int (X(\mathbf{s}_k; t) - X(\mathbf{s}_\ell; t))^2 dt \right\}. \tag{5.8}$$

A natural estimator of $2\gamma(d_{k\ell})$ for complete records is

$$2\tilde{\gamma}(d_{k\ell}) = \frac{1}{p_{k\ell}} \sum_{P(d_{k\ell})} \frac{1}{T} \sum_{i=1}^{T} (X(\mathbf{s}_k; t_i) - X(\mathbf{s}_\ell; t_i))^2, \tag{5.9}$$

where $P(d_{k\ell}) = \{(\mathbf{s}_k, \mathbf{s}_\ell) : \|\mathbf{s}_k - \mathbf{s}_\ell\| = d_{k\ell}\}$ and $p_{k\ell}$ is the cardinality of $P(d_{k\ell})$. *The points with $d = 0$ are not included.* When the records are incomplete, averaging over time can be a source of a severe bias especially for short records. Thus, preaveraging over time should be avoided. Instead, we perform averaging for *all available squared differences* $(X(\mathbf{s}_k;t_i) - X(\mathbf{s}_\ell;t_i))^2$, $1 \le i \le T$, for locations which fall into $P(d_{k\ell})$, see Fig. 5.4. The resulting estimator is noisy and the corresponding spatial covariance is not necessarily positive definite. We thus fit a valid parametric variogram model to the $\tilde{\gamma}(d_{k\ell})$, using nonlinear least squares, and restore the covariance. We use the Gaussian model

$$\gamma(d) = (\sigma^2 - \sigma_\nu^2)(1 - \exp(-d^2/\rho^2)) + \sigma_\nu^2 \mathbf{1}_{(0,\infty)}(d) \tag{5.10}$$

because it fits the estimated variogram for the real data well, see Figure 5.4.

Once the parameters σ^2, σ_ν^2 and ρ^2 have been estimated, calculation of the weights $w_k(t_i)$ is straightforward: we first estimate the covariance matrix $\mathbf{\Sigma}(t_i)$ by plugging the distances between locations with available observations into (5.10) and then use formula

$$\mathbf{w}(t_i) = \mathbf{\Sigma}(t_i)^{-1}\mathbf{1}/(\mathbf{1}^T\mathbf{\Sigma}(t_i)^{-1}\mathbf{1}).$$

Note that above $\mathbf{\Sigma}(t_i)$ is the $N_i \times N_i$ dimensional matrix and $\mathbf{1}$ is the $N_i \times 1$ dimensional vector.

We will also work with “universal” weights $\mathbf{w}$ obtained by plugging in distances between all locations into (5.10) and using (5.5). We need the weights $\mathbf{w}$ for the trend estimation in section 5.3.2.

Estimation of the covariance structure. To determine the statistical significance of the conjectured cooling trend, we need to estimate several elements of the second order structure of the incomplete functional field X. We will see in Section 5.3.2 that what is needed are estimates of the FPC's v_j and of the matrices $\mathbf{\Sigma}_j$ whose entries are

$$\Sigma_j(k,\ell) = E[\zeta_j(\mathbf{s}_k)\zeta_j(\mathbf{s}_\ell)], \qquad 1 \le k,\ell \le N. \tag{5.11}$$

To calculate these estimates, we extended the ideas used in the estimation of the mean function to the estimation of the second order structure by using bivariate smoothing. The

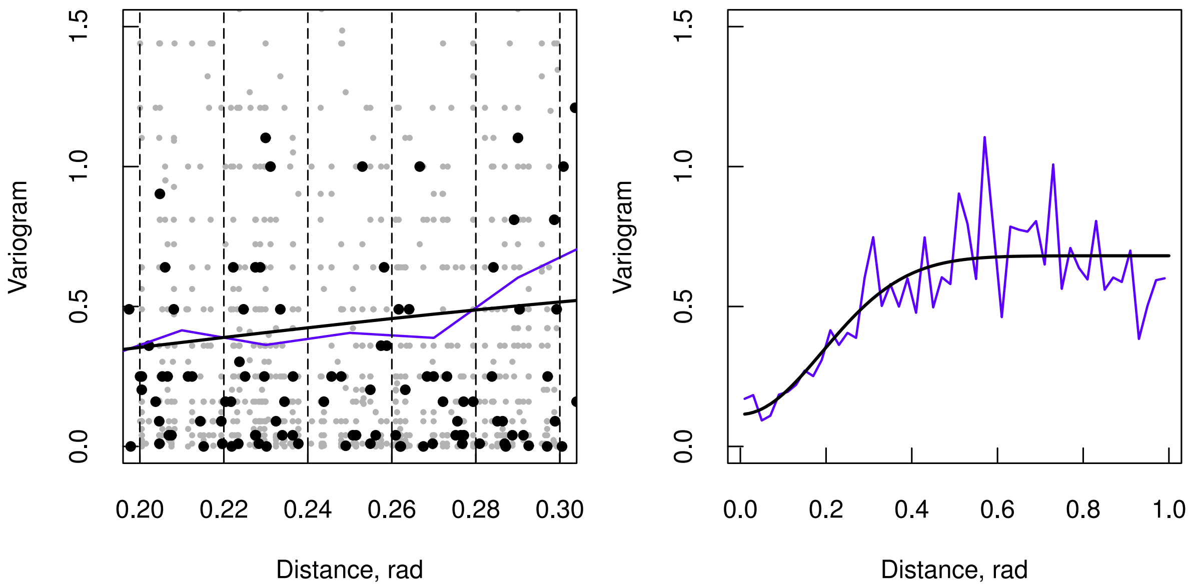

Figure 5.4: Estimation of the weights for incomplete records. Left panel: Gray dots represent *all available squared differences* $(X(\mathbf{s}_k;t_i) - X(\mathbf{s}_\ell;t_i))^2$, $1 \le i \le T$; black dots represent squared differences $(X(\mathbf{s}_k;t_i) - X(\mathbf{s}_\ell;t_i))^2$, for some fixed t_i. Dashed lines separate regions $P(d_{k\ell})$. Right panel: The thin line shows the estimated variogram, the bold line represents the fitted Gaussian variogram.

details are however fairly technical, and are presented in Section 5.6.

5.3.2 Functional regression

Estimation of the trend. In Chapter 4 we proposed a procedure for determining the linear trend for complete records when all covariates are global, like the SRF which does not depend on the spatial location. Here we generalize that approach to the case of incomplete curves and covariates which may depend on the spatial location.

We assume that the mean function $\mu(t)$ is a linear combination of q known functions (covariates) $z_i(t;\mathbf{s})$, so that model (5.1) takes the form

$$X(\mathbf{s};t) = \sum_{i=1}^{q} \beta_i z_i(t;\mathbf{s}) + \sum_{j=1}^{\infty} \zeta_j(\mathbf{s}) v_j(t) + \theta(\mathbf{s};t). \tag{5.12}$$

Some covariates are global, but we use the notation $z_i(t;\mathbf{s})$ for all of them. All covariates are fully observed and are treated as deterministic regressors.

For an arbitrary weight vector $\mathbf{w} = [w_1, \dots, w_N]^T$, set

$$z_{wi}(t) = \sum_{k=1}^{N} w_k z_i(t;\mathbf{s}_k).$$

Next, introduce the following vectors

$$\boldsymbol{\beta} = [\beta_1, \dots, \beta_q]^T, \quad \mathbf{z}(t) = [z_{w1}(t), \dots, z_{wq}(t)]^T,$$

and matrices

$$\mathbf{Q} = \left[\langle z_{wi}, z_{wi'} \rangle, \, 1 \le i, i' \le q\right], \quad \boldsymbol{\Omega} = \left[\langle z_{wi}, v_j \rangle, \, 1 \le i \le q, 1 \le j \le p\right].$$

The number p of the FPC's in the definition of the matrix $\boldsymbol{\Omega}$ is selected using the cumulative variance criterion, see e.g. Ramsay and Silverman (2005) or Horváth and Kokoszka (2012). A general recommendation is to use p such that the first p components explain about 85-90% of the variance. It is often useful to perform inference for several values of p. If the conclusions do not depend on p, we can place more confidence in them.

We now explain how the parameter vector $\boldsymbol{\beta}$ is estimated. If the responses $X(\mathbf{s}_k)$ are fully observed, we minimize

$$\left\| \sum_{k=1}^{N} w_k \left\{ X(\mathbf{s}_k) - \sum_{i=1}^{q} z_i(\mathbf{s}_k)\beta_i \right\} \right\|^2, \quad \sum_{k=1}^{N} w_k = 1. \tag{5.13}$$

This leads to the solution

$$\hat{\boldsymbol{\beta}} = \mathbf{Q}^{-1} \left\langle \mathbf{z}, \mathbf{w}^T\mathbf{X} \right\rangle, \qquad \mathbf{z} = [z_{w1}, \dots, z_{wq}]^T, \tag{5.14}$$

with the weights $\mathbf{w}$ given by (5.5). The quantity $\left\langle \mathbf{z}, \mathbf{w}^T\mathbf{X} \right\rangle$ is the $q \times 1$ vector with the ith entry $\left\langle z_{wi}, \sum_{k=1}^{N} w_k X(\mathbf{s}_k) \right\rangle$.

Notice that $\mathbf{w}^T\mathbf{X}$ is the estimate of the mean function μ for full records. Solution (5.14) can thus be written as

$$\hat{\boldsymbol{\beta}} = \mathbf{Q}^{-1} \langle \mathbf{z}, \hat{\mu} \rangle . \tag{5.15}$$

When the record are partially observed, μ is estimated using the indexed regression discussed above. Thus (5.15) is suitable for estimating the parameter vector when data contain missing observations. This is the approach we take.

Significance of regression coefficients. The variance of the estimator (5.15) is

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \mathbf{Q}^{-1} E \left[\langle \mathbf{z}, \hat{\mu} - \mu \rangle \langle \mathbf{z}, \hat{\mu} - \mu \rangle^T \right] \mathbf{Q}^{-1}, \tag{5.16}$$

where the middle term is a $q \times q$ matrix whose (i, j) element is

$$\iint z_{wi}(t) z_{wj}(t') \mathrm{Cov}(\hat{\mu})(t, t') dt dt'. \tag{5.17}$$

The formula for $\mathrm{Cov}(\hat{\mu})(t, t')$ is given in Section 5.7, where we also derive the approximations:

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \mathbf{Q}^{-1} \boldsymbol{\Omega} \mathrm{Var}[\boldsymbol{\zeta}_w] \boldsymbol{\Omega}^T \mathbf{Q}^{-1} + \sigma_\theta^2 \mathbf{Q}^{-1} \mathbf{w}^T \mathbf{w}, \tag{5.18}$$

where, assuming that the weights are known constants,

$$\mathrm{Var}[\boldsymbol{\zeta}_w] = \mathrm{diag} \left(\mathbf{w}^T \boldsymbol{\Sigma}_1 \mathbf{w}, \ldots, \mathbf{w}^T \boldsymbol{\Sigma}_p \mathbf{w} \right) . \tag{5.19}$$

The last expression is a $p \times p$ diagonal matrix. The diagonal form of $\mathrm{Var}[\boldsymbol{\zeta}_w]$ is a consequence of assumption (5.2). Quantities appearing in (5.18) and (5.19) can be estimated using the methodology presented in the previous sections.

Using numerical simulations, we found that the estimator $\hat{\boldsymbol{\beta}}$ is approximately normal even if the functions $X(\mathbf{s}_k)$ are not normally distributed. The ionosonde data do not show alarming departures from normality, see Fig. 5.6. Thus to test $\beta_i = 0$, for a fixed i, we assume that the statistic $\hat{\beta}_i / \sqrt{\mathrm{Var}[\hat{\beta}_i]}$ has the standard normal distribution, the P-value is calculated as

$$\text{P-value} = 2\Phi(\hat{\beta}_i/\sqrt{\text{Var}[\hat{\beta}_i]}). \tag{5.20}$$

5.4 Application to ionosonde data

The specific form of regression (5.12) used in this section is

$$X(\mathbf{s}_k;t) = \beta_1 + \beta_2 t + \beta_3 \text{SRF}(t) + \beta_4 M(\mathbf{s}_k;t) + \sum_{j=1}^{p} \zeta_j(\mathbf{s}_k) v_j(t), \quad 1 \le k \le 85. \tag{5.21}$$

As explained in Section 5.6, the estimated noise variance is extremely small, so the noise term is not included in the final model. The global functional covariate SRF is the observed solar radio flux. The local covariates, $M(\mathbf{s}_k)$, reflect the potential impact of decadal changes in the direction of the internal magnetic field of the Earth, and are given by

$$M(\mathbf{s}_k;t) = \sin I(\mathbf{s}_k;t) \cos I(\mathbf{s}_k;t), \tag{5.22}$$

where $I(\mathbf{s}_k;t)$ is the inclination of the Earth's magnetic field, see e.g. Chapter 13 of Kivelson and Russell (1997). The space physics background justifying formula (5.22) is complicated. An interested reader is referred to Elias (2009) and references therein. We also consider a restricted model with $\beta_4 = 0$. Using the full and the restricted models will allow us to evaluate the impact of the decadal changes in the internal field on the conjectured trend. Our interest is in estimating the coefficient β_2, which we call the "trend," and evaluating its statistical significance.

Examples of modeling of the foF2 curves using model (5.21) are shown in Fig. 5.5. Estimates of the linear trend as well as their statistical significance for different smoothing bandwidths are summarized in Table. 5.1. We found that inclusion of the changes in the Earth's magnetic field changes the estimated values only slightly. A small impact of this covariate is however clear, as it decreases the value of the estimated trend. Our main conclusion is that *the trend is negative and it is statistically significant.* The estimated trend value practically does not depend on the bandwidth h_μ. When $p = 3, 4$ the estimated

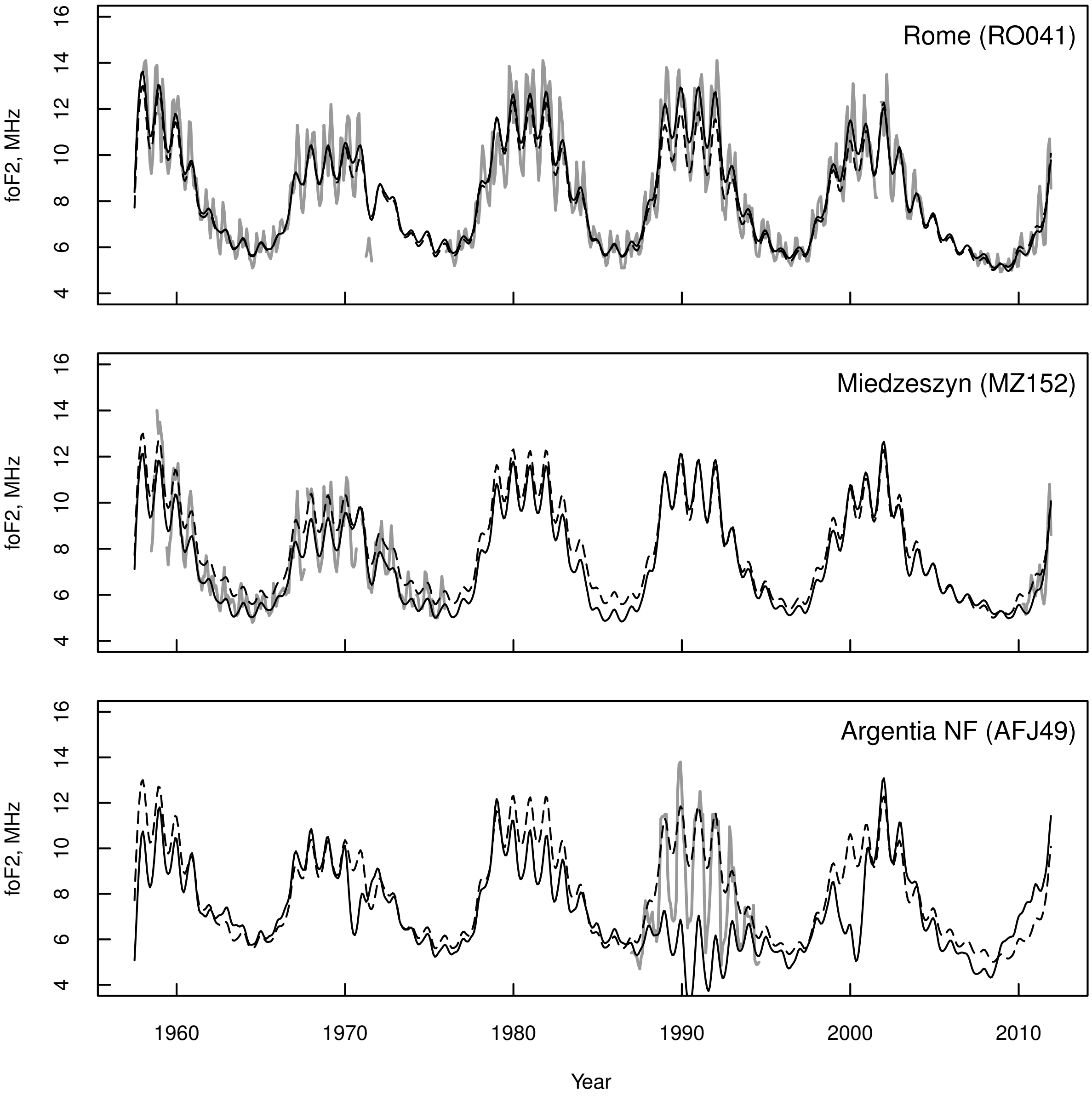

Figure 5.5: Examples of modeling of the foF2 records via (5.1) with $h_\mu = 5$, $h_c = 10$ and $p = 3$. Gray lines represent original records, solid black lines - model, dashed black line - the mean function. Top: An almost complete record, middle: partially observed record, bottom: example of unstable modeling when the number of observations per curve is less than 200.

standard deviation and P-values do not depend on the bandwidth h_c. But when the number of FPC's is 2, the estimated standard deviations and P-values are much smaller than those

Table 5.1: Trends and P-values for different bandwidths. "NM" denotes estimation without magnetic inclination, "M" denotes estimation when magnetic inclination is included. The number of the FPC's, p, was chosen to obtain the cumulative variance closest to but greater than 85% (indicated as Final CV).

		Final		NM		M	
h_μ	h_C	CV,%	p	$\beta_2, 10^{-3}$MHz/Year	P-value	$\beta_2, 10^{-3}$MHz/Year	P-value
5	10	86.77	3	-5.22 ± 2.25	0.020	-4.91 ± 2.24	0.028
	15	90.44	3	-5.22 ± 2.30	0.023	-4.91 ± 2.29	0.032
	20	87.07	2	-5.22 ± 1.48	$4.36 \cdot 10^{-4}$	-4.91 ± 1.47	$8.62 \cdot 10^{-4}$
10	10	88.75	4	-5.18 ± 2.15	0.016	-4.88 ± 2.14	0.023
	15	87.57	3	-5.18 ± 2.11	0.014	-4.88 ± 2.10	0.020
	20	86.50	2	-5.18 ± 1.83	$4.67 \cdot 10^{-3}$	-4.88 ± 1.82	$7.52 \cdot 10^{-3}$
15	10	88.55	4	-5.28 ± 2.16	0.014	-4.99 ± 2.15	0.020
	15	87.09	3	-5.28 ± 2.11	0.012	-4.99 ± 2.11	0.018
	20	91.36	3	-5.28 ± 2.20	0.016	-4.99 ± 2.19	0.023
20	10	88.58	4	-5.07 ± 2.13	0.017	-4.79 ± 2.12	0.024
	15	87.02	3	-5.07 ± 1.99	0.011	-4.79 ± 1.98	0.015
	20	91.04	3	-5.07 ± 2.04	0.013	-4.79 ± 2.03	0.018

for $p = 3, 4$. We believe that this happens due to oversmoothing of the covariance surface and the resulting underestimation of the variance. The normal quantile-quantile plots are shown in Fig. 5.6 which suggests that there is a slight deviation from normality, but according to our simulations (not reported here) the t–statistics in (5.20) is robust enough to such departures.

Our conclusion (significant negative trend) agrees with the hypothesis of Roble, Dickinson and Rishbeth discussed in Section 5.1. The exploratory analysis in Bremer *et al.* (2012) applied to 37 stations located worldwide and various time periods yields however mixed evidence. The average trend (simple average of individual trends) is either negative or practically zero depending on the time period and the number of stations. It is therefore important to determine if our conclusion depends on the choice of locations and the time interval.

To assess the robustness of our conclusion we performed two experiments, which could be called temporal and spatial subsampling. In the first experiment, we study the depen-

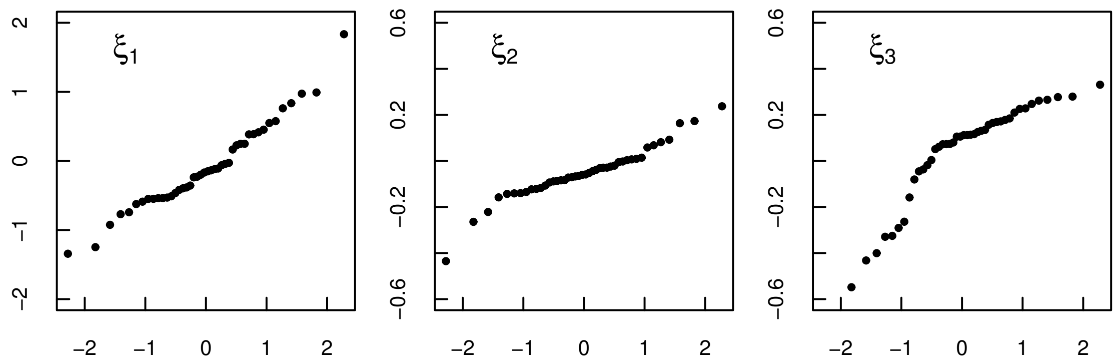

Figure 5.6: Normal quantile–quantile plots for scores $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$, $\boldsymbol{\xi}_3$ estimated using $h_\mu = 15$ and $h_c = 15$.

dence of the trend on the time interval length. To do this, we fix the interval length, L, and determine the trend for all possible intervals. Since the observations must be sequential, the number of possible intervals is limited, $M = T - L + 1$. The variability of the trend parameter is shown in Fig. 5.7 (a). While the average trend parameter is uniformly less than the final estimator, the variability is very high if $L \leq 400$. This indicates, that if the time interval is less than about 30 years, the sample may not contain enough information to allow us to reach definite conclusions. In the second experiment, we use the whole time interval (654 months), but we randomly select subsamples of locations of a specified size. For each subsample size (e.g. 60), we draw 10^3 subsamples without replacement. The distribution of the values of β_2 as a function of the subsample size is given in Fig. 5.7 (c). While the average trend parameter is consistent with our final estimator, the variability is high for small samples. If the sample size is greater than 65, we practically always obtain a negative trend. However, for sample size smaller than 50, there is a good chance that a positive trend will be obtained, if some specific locations are chosen. This phenomenon has been a source of much debate in the space physics community. Studies of this type have often been based on a handful of stations from a specific region because most of the foF2 series are of poor quality and have limited temporal coverage. In particular, studies based on individual records from the Asian part of Russia (east of 30° longitude) indicated a positive trend, Bremer (1998). Finally, we note that there is a clear relationship between the time interval length and the number of stations that can be used. The distribution of the number of stations as a function of the length is shown in Fig. 5.7 (b). We see that if the time interval is about 400 months, about 70 stations are available. This establishing a convincing connection between the two experiments. Our analysis indicates that a negative trend in the Northern Hemisphere will be obtained provided the time interval is longer than 35 years (about 3 solar cycles) and at least 70 stations are used.

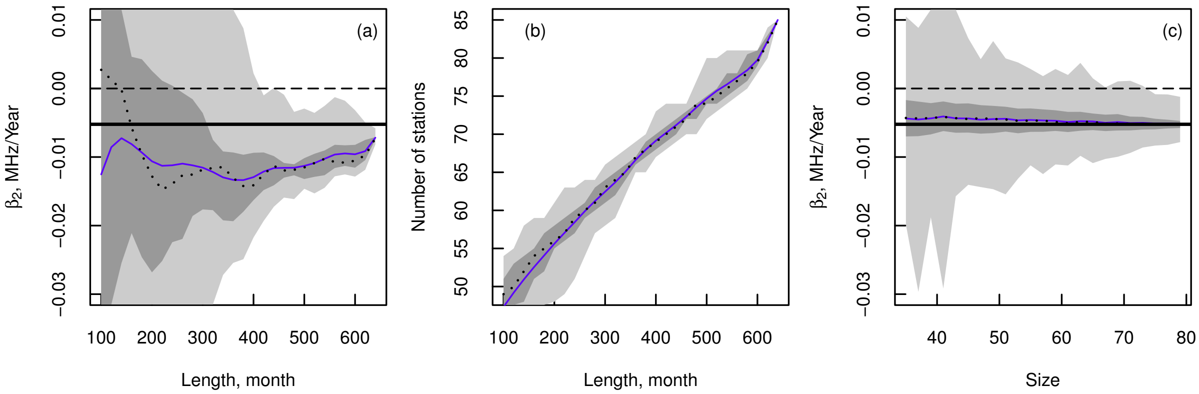

Figure 5.7: (a) Distribution of the trend parameter as a function of the interval length, (b) the number of stations as a function of the interval length. (c) distribution of the trend parameter as a function of the number of stations. Light gray - full region, dark gray - central 50 percent, dotted line - median, blue solid line - average. Bold black line is the final estimator.

5.5 Correlation between scores

We first verify that if the covariance function is separable, i.e. if (5.3) holds, then assumption (5.2) holds. Indeed, since the FPC's v_i are the eigenfunctions of the covariance kernel $c(\cdot,\cdot)$, we have

$$\begin{aligned} E[\zeta_j(\mathbf{s})\zeta_i(\mathbf{s}')] &= E\left[\int (X(\mathbf{s};t)-\mu(t))\, v_j(t)dt \int \big(X(\mathbf{s}';t')-\mu(t')\big)\, v_i(t')dt'\right] \\ &= \iint \mathrm{Cov}\,\big(X(\mathbf{s},t), X(\mathbf{s}',t')\big)\, v_j(t)v_i(t')dtdt' \\ &= \iint \Sigma(\mathbf{s},\mathbf{s}')c(t,t')v_j(t)v_i(t')dtdt' \\ &= \Sigma(\mathbf{s},\mathbf{s}')\int\left\{\int c(t,t')v_i(t')dt'\right\} v_j(t)dt \\ &= \Sigma(\mathbf{s},\mathbf{s}')\int \lambda_i v_i(t)v_j(t)dt \\ &= \lambda_i\delta_{ij}\Sigma(\mathbf{s},\mathbf{s}'), \end{aligned}$$

where δ_{ij} is Kronecker's delta.

Now we explain a data driven approach to checking assumption (5.2). Recall that $\boldsymbol{\xi}_j = [\xi_j(\mathbf{s}_1),\ldots,\xi_j(\mathbf{s}_N)]^T$ is a zero mean random vector with the covariance matrix $\boldsymbol{\Sigma}_j$. Consider the $N\times N$ cross–covariance matrix $\boldsymbol{\Sigma}_{jj'} = E[\boldsymbol{\xi}_j\boldsymbol{\xi}_{j'}]$. The matrix $\boldsymbol{\Sigma}_{jj'}$ does not need to be positive definite. We want to test

$$H_0 : \boldsymbol{\Sigma}_{jj'} = \mathbf{0}, \ \text{ vs. } \ H_A : \boldsymbol{\Sigma}_{jj'} \neq \mathbf{0}.$$

This can be done assuming that $\boldsymbol{\Sigma}_{jj'}$ is isotropic, i.e that $\Sigma_{jj'}(\mathbf{s}_k,\mathbf{s}_\ell) = \Sigma_{jj'}(d_{k\ell})$, where $d_{k\ell}$ is the chordal distance between locations $\mathbf{s}_k$ and $\mathbf{s}_\ell$. To estimate $\boldsymbol{\Sigma}_{jj'}$ we use the standard binning approach:

$$\hat{\Sigma}_{jj'}(d) = \frac{1}{p(d)}\sum_{P(d)} \xi_j(\mathbf{s}_k)\xi_{j'}(\mathbf{s}_\ell), \tag{5.23}$$

where $P(d) = \{(\mathbf{s}_k,\mathbf{s}_\ell) : \|\mathbf{s}_k-\mathbf{s}_\ell\| = d; k,\ell = 1,\ldots,N\}$ and $p(d)$ is the number of distinct pairs. We call $\hat{\Sigma}_{jj'}(d)$ the correlogram. Precise estimation of the confidence intervals for $\hat{\Sigma}_{jj'}(d)$ is a difficult task. Thus, we take a simplified approach which nevertheless provides useful information. It can be argued that under suitable mixing conditions $\hat{\Sigma}_{jj'}(d)$ is approximately normally distributed, see for example chapter 2.4.1 in Cressie (1993) and references therein. Let $\hat{\boldsymbol{\Sigma}}_{jj'} = [\hat{\Sigma}_{jj'}(d_1),\ldots\hat{\Sigma}_{jj'}(d_H)]^T$, be a vector of length H. Then

$\mathrm{Var}(\hat{\mathbf{\Sigma}}_{jj'})$ is a $H \times H$ positive definite covariance matrix. To find the confidence bounds we need to estimate the diagonal elements of $\mathrm{Var}(\hat{\mathbf{\Sigma}}_{jj'})$. We estimate the diagonal elements using the sample variance estimator:

$$\mathrm{Var}(\hat{\Sigma}_{jj'}(d_i)) \approx \frac{1}{p(d_i)-1} \sum_{P(d_i)} \Big(\xi_j(\mathbf{s}_k)\xi_{j'}(\mathbf{s}_\ell) - \hat{\Sigma}_{jj'}(d_i)\Big)^2,$$

Now the pointwise confidence bounds can be constructed in a standard way. The estimated correlograms and the corresponding pointwise 95% confidence bounds for different pairs j, j' are shown in Fig. 5.8.

The pointwise confidence intervals cover the zero line almost entirely which shows that the difference between the estimated correlograms and zero is not statistically significant.

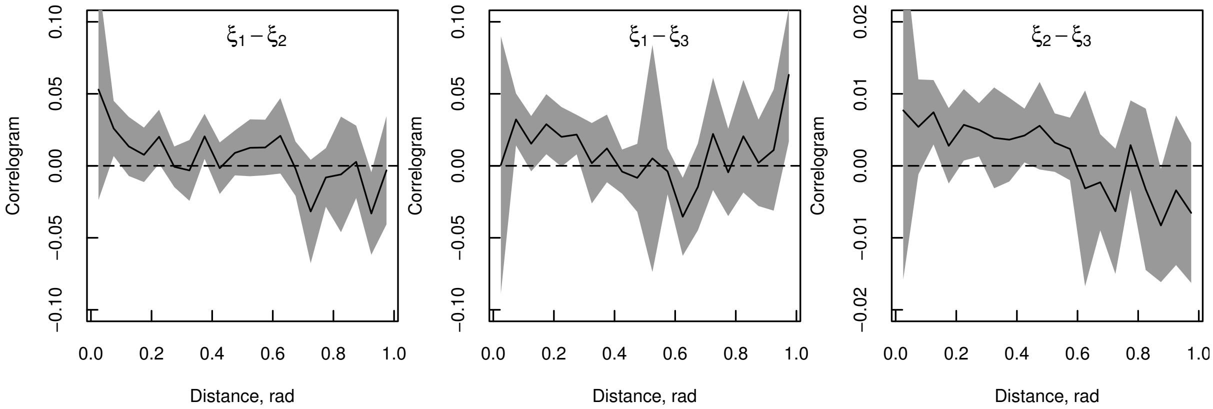

Figure 5.8: Black lines represent estimated correlograms and gray regions represent the 95% pointwise confidence bounds.

5.6 Estimation of the covariance structure

We begin with the estimation of the $v_j(t)$. Since the $v_j(t)$ are the eigenfunctions of the covariance operator, it is enough to obtain an estimate of the covariance surface, and then numerically solve the eigenfunction equations. We emphasize that since many functions are not observed over long temporal segments, the approaches developed in Chapter 2 cannot be used, as they involve various integrals over the whole temporal domain. Our objective is thus to estimate

$$c(t, t') = \mathrm{Cov}(X(t), X(t')).$$

The elements of the covariance function $c(t, t')$ for spatially correlated functional data are estimated in two steps. In the first first step, which takes into account spatial correlations, we obtain a preliminary estimator $\tilde{c}(t_i, t_{i'})$ which is noisy and contains missing values. The second step is smoothing.

To obtain the preliminary estimator, we perform the following procedure. For fixed i and i' define the scalar field

$$\psi(\mathbf{s}) = \left[X(\mathbf{s}; t_i) - \hat{\mu}(t_i)\right]\left[X(\mathbf{s}; t_{i'}) - \hat{\mu}(t_{i'})\right].$$

The estimation of $\tilde{c}(t_i, t_{i'})$ is thus reduced to the estimation of the mean (independent of $\mathbf{s}$) of the scalar spatial process $\psi(\cdot)$ based on the pseudo–observations $\psi(\mathbf{s}_1), \ldots, \psi(\mathbf{s}_{N(i,i')})$, where $N(i, i')$ is the number of records available simultaneously at times t_i and $t_{i'}$. To estimate $\mu_\psi = E\psi(\mathbf{s})$ as a weighted average of the $\psi(\mathbf{s}_k)$, the covariance matrix of the $\psi(\mathbf{s}_k)$ must be estimated. This can be accomplished using either parametric modeling for large samples, or the nonparametric approach developed in Chapter 4, for small samples ($N(i, j) \leq 20$). In this paper we use only parametric modeling with the exponential covariance. If the number of observations in the the $\psi(\mathbf{s}_k)$ is less than 20, we do not perform estimation. This is admissible because of the second step. We point out that the first step is computationally very expensive, but possible on a parallel machine.

Now we explain smoothing of the preliminary estimator. The core of the covariance estimation methodology is discussed in Staniswalis and Lee (1998), Yao *et al.* (2003) and Yao *et al.* (2005). In the presence of measurement error, the diagonal elements are contaminated by the noise variance and should not be included as input for the smoothing step. The estimator of $c(t, t')$ thus is

$$\hat{c}(t, t') = \arg\min_{\mathbf{u}} \sum_{1 \leq i \neq i' \leq T} \kappa_c\left(\frac{t - t_i}{h_c}, \frac{t' - t_{i'}}{h_c}\right)\left\{\tilde{c}(t_i, t_{i'}) - f(\mathbf{u}, t, t', t_i, t_{i'})\right\}^2, \qquad (5.24)$$

where $\tilde{c}(t_i, t_{i'})$ is the estimator obtained in step one, and

$$f(\mathbf{u}, t, t', t_i, t_{i'}) = u_0 + u_1(t - t_i) + u_2(t' - t_{i'}), \quad \mathbf{u} = [u_0, u_1, u_2]^T.$$

The choice of the kernel κ_c is not crucial, but the selection of the bandwidth h_c requires attention. The natural way to choose h_c is by using leave one curve out cross–validation. Unfortunately, due to extreme computational cost of the first step this type of cross validation is not possible at this time. We recommend to simply try several different bandwidths. Notice that preliminary estimator does not depend on h_c, thus this procedure is computationally fast. The conclusions reported in Section 5.4 do not depend on the choice of h_c as long as this bandwidth is reasonable. We used bandwidths corresponding to effective averaging over periods from half a year to two years, which is unlikely to effect decadal trends.

To asses the significance of the linear trend, the estimation of the score covariances (5.11) is required. Estimation of the matrices $\mathbf{\Sigma}_j$ for the ionosonde data is not straightforward. When a curve is complete or almost complete numerical integration in $\zeta_j(\mathbf{s}) = \langle X(\mathbf{s}) - \mu, v_j \rangle$ works very well. But when huge parts of a curve are missing, numerical integration leads to a very poor estimate of a score. One possible remedy is to use the conditional estimation advocated by Yao *et al.* (2005). Unfortunately, for ionosonde data this approach performs poorly as well. We believe that the conditional estimation does not work because observations are not missing at random. For the ionosonde data, using visual validation, we found that if the number of observation per curve is bigger than 200 (about 1/3 of the maximum number of monthly observations), then the score estimates, both obtained by numerical integration and the conditional estimation lead to predicted curves $\hat{\mu}(t) + \sum_{j=1}^{p} \xi_j(\mathbf{s}) v_j(t)$ which are close to the observed curves over the segments for which the latter are available. Fig. 5.5 (bottom) is an example of a situation when this is not the case. Thus, for accurate estimation of the score covariances we use only scores estimated for curves with more 200 observations per curve. The size of the selected spatial fields ζ_j is only 55 among 85 possible locations.

Once the locations have been selected, we perform estimation of the covariance using the semivariogram method. Namely, we estimate the semi–variogram

$$2\gamma_j(d) = \frac{1}{p(d)} \sum_{P(d)} (\zeta_j(\mathbf{s}_k) - \zeta_j(\mathbf{s}_\ell))^2,$$

where $P(d) = \{(\mathbf{s}_k, \mathbf{s}_\ell) : \|\mathbf{s}_k - \mathbf{s}_\ell\| = d\}$ and $p(d)$ is the cardinality of $P(d)$. *The points with* $d = 0$ *are not included.* We emphasize that not all N locations $\mathbf{s}_k$ are used, only those which have more than 200 temporal measurements. We then fit the empirical semivariogram to some valid parametric model, $\tilde{\gamma}_j(d)$. The elements of the score covariances are calculated as

$$\Sigma_j(k, \ell) = \tilde{\gamma}_j(0) - \tilde{\gamma}_j(d_{k\ell}), \;\; 1 \le k, \ell \le N.$$

Finally, we note that the measurement noise variance can be estimated using Eq. (2) in Yao *et al.* (2005). We found that for the ionosonde data the contribution of the measurement noise is completely negligible. We thus omit it in numerical calculations, but preserve the noise variance in formulas to enhance their broader applicability.

5.7 Covariances of the estimated mean function

Everywhere below we assume that the sampling variability of the weights $w_k(t)$ can be neglected. Without this assumptions it is difficult to derive usable expressions.

We introduce the following vectors and matrices, with their dimensions given in parenthesis to facilitate the understanding:

$$\boldsymbol{\mu}_w = \left[\sum_{k=1}^{N_1} w_k(t_1) X(\mathbf{s}_k; t_1), \ldots, \sum_{k=1}^{N_T} w_k(t_T) X(\mathbf{s}_k; t_T) \right]^T, \qquad (T \times 1);$$

$$\mathbf{m}(t) = [m_0(t), m_1(t)]^T, \qquad (2 \times 1)$$

$$\mathbf{Z}(t) = \begin{bmatrix} 1 & t - t_1 \\ \vdots & \vdots \\ 1 & t - t_T \end{bmatrix}, \qquad (T \times 2);$$

$$\mathbf{K}(t) = \mathrm{diag}\left[\kappa((t-t_1)/h), \ldots, \kappa((t-t_T)/h)\right], \qquad (T \times T).$$

Note that $\mathbf{K}(t)$, $\mathbf{Z}(t)$ and $\mathbf{m}(t)$ depend on continuous t, while $\boldsymbol{\mu}_w$ is a vector of a fixed length.

The solution to (5.7) has the form

$$\hat{\mathbf{m}}(t) = [\mathbf{Z}^T(t)\mathbf{K}(t)\mathbf{Z}(t)]^{-1}\mathbf{Z}^T(t)\mathbf{K}(t)\boldsymbol{\mu}_w, \tag{5.25}$$

with the covariance matrix

$$\mathrm{Cov}(\hat{\mathbf{m}})(t,t') = [\mathbf{Z}^T(t)\mathbf{K}(t)\mathbf{Z}(t)]^{-1}\mathbf{Z}^T(t)\mathbf{K}(t)\mathrm{Var}(\boldsymbol{\mu}_w)\mathbf{K}(t')\mathbf{Z}(t')[\mathbf{Z}^T(t')\mathbf{K}(t')\mathbf{Z}(t')]^{-1}. \tag{5.26}$$

Thus the covariances of the mean function are

$$\mathrm{Cov}(\hat{\mu})(t,t') = e_1^T \mathrm{Cov}(\hat{\mathbf{m}})(t,t')e_1, \tag{5.27}$$

where $e_1 = [1, 0]^T$.

The only factor in (5.26) that requires attention is $\mathrm{Var}(\boldsymbol{\mu}_w)$ which is a $T \times T$ matrix. To calculate it in a general setting with measurement error we use model (5.1). Treating the weights w_k as fixed, we obtain

$$\begin{aligned}
&\mathrm{Cov}(\boldsymbol{\mu}_w(t_i), \boldsymbol{\mu}_w(t_{i'})) \\
&= \mathrm{Cov}\left(\sum_{k=1}^{N_i} w_k(t_i)X(\mathbf{s}_k;t_i)\ ,\ \sum_{\ell=1}^{N_{i'}} w_\ell(t_{i'})X(\mathbf{s}_\ell;t_{i'})\right) \\
&= \sum_{k=1}^{N_i}\sum_{\ell=1}^{N_{i'}} w_k(t_i)w_\ell(t_{i'})E\left[\left\{\sum_{j=1}^{\infty}\zeta_j(\mathbf{s}_k)v_j(t_i)+\theta(t_i)\right\}\left\{\sum_{j'=1}^{\infty}\zeta_{j'}(\mathbf{s}_\ell)v_{j'}(t_i)+\theta(t_{i'})\right\}\right] \\
&= \sum_{k=1}^{N_i}\sum_{\ell=1}^{N_{i'}} w_k(t_i)w_\ell(t_{i'})\left\{\sum_{j=1}^{\infty} E\left[\zeta_j(\mathbf{s}_k)\zeta_j(\mathbf{s}_\ell)\right]v_j(t_i)v_j(t_{i'})\right\} + \delta_{ii'}\sigma_\theta^2\sum_{k=1}^{N_i} w_k^2(t_i),
\end{aligned} \tag{5.28}$$

where the last equality follows from (5.2).

Formulas (5.28) and (5.26) are computationally intensive. By comparing the values for selected times t, t', we found that practically the same values are obtained by replacing the smoothing matrices by identity matrices and replacing the weights $w_k(t_i)$ by the universal weights w_k given by (5.5). Such a simplification leads to the approximations

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{\mu}})(t,t') &= \sum_{j=1}^{\infty} \mathbf{w}^T \boldsymbol{\Sigma}_j \mathbf{w} v_j(t) v_j(t'), \quad t \neq t' \\ \mathrm{Var}(\hat{\boldsymbol{\mu}})(t,t) &= \sum_{j=1}^{\infty} \mathbf{w}^T \boldsymbol{\Sigma}_j \mathbf{w} v_j^2(t) + \sigma_\theta^2 \mathbf{w}^T \mathbf{w}, \qquad (5.29)\end{aligned}$$

Only the first few, p, FPC's in the infinite sums in (5.28) and (5.29) are used in practice. Usually this number is selected to capture the desired level of variability. Inserting (5.29) so truncated into (5.17), leads to (5.18).

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

Results reported in this dissertation make contributions both into statistical methodology and space physics applications. Specifically, we introduced a new statistical framework for analysis and testing of complete and incomplete spatially correlated functional data. Using this framework we answered a very important space physics question regarding global changes in the ionosphere. We found that the critical frequency, foF2, is decreasing which means that the temperature of the inosphere is also decreasing.

Now I describe several projects which I would like to address in the very near future.

1. In Chapter 3 we introduced the test for testing the equality of two functional means when curves are spatially correlated. One possible extension of this work would be to generalize testing procedure to test multiple hypothesis:

$$H_0 : \mu_1(t) = \mu_2(t) = \ldots = \mu_N(t),$$
$$H_A : \mu_i(t) \neq \mu_j(t), \text{ for some } \; i, j.$$

2. We are planning to generalize our estimation procedure to let the mean function be space dependent. This problem is not very difficult when there are repeated measurements (curves) at each location and it could potentially became a project for a master thesis. But when at each location there is only one curve this problem become much more complicated, one possible remedy is to restrict $\mu(\mathbf{s};t)$ to some parametric form. Other possibilities still need to be explored.

3. The core of the third project is to create an automatic procedure for selecting spatial regions where the mean functions or other quantities such as linear trends are the same. This project is rather long and perhaps in combination with important applied problem can become a core of a PhD dissertation.

REFERENCES

Aston, J. A. D. and Kirch, C. (2012). Estimation of the distribution of change-points with application to fMRI data. *Annals of Applied Statistics*, **6,** 1906–1948.

Bel, L., Bar-Hen, A., Cheddadi, R. and Petit, R. (2011). Spatio-temporal functional regression on paleo–ecological data. *Journal of Applied Statistics*, **38,** 695–704.

Bosq, D. and Blanke, D. (2007). *Inference and Prediction in Large Dimensions.* Wiley, Hoboken.

Bowman, A.W., Giannitrapani, M. and Scott, E.M. (2009). Spatiotemporal smoothing and sulphur dioxide trends over Europe. *Journal of the Royal Statistical Society:Series C*, **58,** 737–752.

Bremer, J. (1998). Trends in the ionospheric E and F regions over Europe. *Annales Geophysicae*, **16,** 986–996.

Bremer, J., Damboldt, T., Mielich, J. and Suessmann, P. (2012). Comparing long–term trends in the ionospheric F2–region with two different methods. *Journal of Atmospheric and Solar–Terrestrial Physics*, **77,** 174–185.

Carroll, S. S. and Cressie, N. (1996). A comparison of geostatistical methodologies used to estimate snow water equivalent. *Water Resources Bulletin*, **32,** 267–278.

Carroll, S. S., Day, G. N., Cressie, N. and Carroll, T. R. (1995). Spatial modeling of snow water equivalent using airborne and ground–based snow data. *Environmetrics*, **6,** 127–139.

Clark, R.G. and Allingham, S. (2011). Robust resampling confidence intervals for empirical variograms. *Mathematical Geosciences*, **43,** 243–259.

Clilverd, M. A., Ulich, T. and Jarvis, M. J. (2003). Residual solar cycle influence on trends in ionospheric F2-layer peak height. *Journal of Geophysical Research*, **108,** A12, SIA15.1–SIA15.8.

Cnossen, I. and Richmond, A.D. (2008). Modelling the effects of changes in the Earth's magnetic field from 1957 to 1997 on the ionospheric hmf2 and fof2 parameters. *Journal of Atmospheric and Solar-Terrestrial Physics*, **70,** 1512 – 1524.

Crainiceanu, C. M., Staicu, A. M., Ray, S. and Punjabi, N. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Statistics in Medicine*, **31,** 3223–3240.

Cressie, N. and Huang, H-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, **94,** 1330–1340.

Cressie, N., Shi, T. and Kang, E. L. (2010). Fixed rank filtering for spatio–temporal data. *Journal of Computational and Graphical Statistics*, **19,** 724–745.

Cressie, N. and Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data.* Wiley, Hoboken.

Cressie, N. A. C. (1993). *Statistics for Spatial Data.* Wiley, Hoboken.

Danilov, A.D. and Mikhailov, A.V. (1999). Long-term trends in the parameters of the f2-region: a new approach. *Geomagnetism and Aeronomy*, **39,** 473–479.

Delaigle, A. and Hall, P. (2010). Defining probability density function for a distribution of random functions. *The Annals of Statistics*, **38,** 1171–1193.

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

Elias, A.G. (2009). Trends in the F2 ionospheric layer due to long-term variations in the Earth's magnetic field. *Journal of Atmospheric and Solar-Terrestrial Physics*, **71,** 1602–1609.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman & Hall/CRC, London.

Febrero, M., Galeano, P. and W.González-Manteiga (2008). Outlier detection in functional data by depth measures with application to identify abnormal NO_x levels. *Environmetrics*, **19,** 331–345.

Ferraty, F. and Romain, Y. (2011) (eds). *The Oxford Handbook of Functional Data Analysis.* Oxford University Press, Oxford.

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

Foppiano, A.J., Cid, L. and Jara, V. (1999). Ionospheric long-term trends for South American mid-latitudes. *Journal of Atmospheric and Solar-Terrestrial Physics*, **61,** 717–723.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, **10,** 419–440.

Gelfand, A. E., Diggle, P. J., Fuentes, M. and Guttorp, P. (2010) (eds). *Handbook of Spatial Statistics.* Chapman & Hall/CRC, London.

Gilleland, E., Ahijevych, D., Brown, B.G., Casati, B. and Ebert, E.E. (2009). Intercomparison of Spatial Forecast Verification Methods. *Weather and Forecasting*, **24,** 1416–1130.

Giraldo, R., Delicado, P. and Mateu, J. (2009). Continuous time–varying kriging for spatial prediction of functional data: an environmental application. *Journal of Agricultural, Biological, and Environmental Statistics*, **15,** 66–82.

Giraldo, R., Delicado, P. and Mateu, J. (2010). Ordinary kriging for function–valued spatial data. *Environmental and Ecological Statistics*, **18,** 411–426.

Giraldo, R., Delicado, P. and Mateu, J. (2011). A generalization of cokriging and multivariable spatial prediction for functional data. Technical report. Universitat Politécnica de Catalunya, Barcelona.

Giraldo, R., Delicado, P. and Mateu, J. (2012). Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica*, **66,** 403–421.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, **97,** 590–600.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of American Statistical Association*, **106,** 746–762.

Gneiting, T., Genton, M. and Guttorp, P. (2007). Geostatistical space–time models, stationarity, separability and full symmetry. In *Statistical Methods for Spatio–Temporal Systems* (eds B. Finkenstadt, L. Held and V. Isham), pp. 151–175. Chapman & Hall/CRC, London.

Hall, P., Fisher, N. I. and Hoffmann, B. (1994). On the Nonparametric Estimation of Covariance Functions. *The Annals of Statistics*, **22,** 2115–2134.

Hall, P. and Patil, P. (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields*, **99,** 399–424.

Haslett, J. and Raftery, A. E. (1989). Space–time modelling with long–memory dependence: assesing Ireland's wind power resource. *Applied Statistics*, **38,** 1–50.

Hawkins, D. M. and Cressie, N. (1984). Robust kriging – a proposal. *Journal of the International Association for Mathematical Geology*, **16,** 3–18.

Hering, A.S. and Genton, M.G. (2011). Comparing Spatial Predictions. *Technometrics*, **53,** 414–425.

Hörmann, S. and Kokoszka, P. (2013). Consistency of the mean and the principal components of spatially distributed functional data. *Bernoulli*; Forthcoming.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications.* Springer, New York.

Horváth, L., Kokoszka, P. and Reeder, R. (2013). Estimation of the mean of functional time series and a two sample problem. *Journal of the Royal Statistical Society: Series* B, **75,** 103–122.

Jiang, H. and Serban, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics*, **54,** 108–137; With discussion.

Jun, M. and Stein, M. L. (2009). Nonstationary covariance models for global data. *The Annals of Applied Statistics*, **2,** 1271–1289.

Kaiser, M. S., Daniels, M. J., Furakawa, K. and Dixon, P. (2002). Analysis of particulate matter air pollution using Markov random field models of spatial dependence. *Environmetrics*, **13,** 615–628.

Katzfuss, M. and Cressie, N. (2011). Spatio–temporal smoothing and EM estimation for massive remote–sensing data sets. *Journal of Time Series Analysis*, **32,** 430–446.

Kelly, M. C. (2009). *The Earth's Ionosphere*, 2nd edn. Academic Press, San Diego.

Kivelson, M. G. and Russell, C. T. (1997) (eds). *Introduction to Space Physics.* Cambridge University Press, Cambridge.

Kokoszka, P., Maslova, I., Sojka, J. and Zhu, L. (2008). Testing for lack of dependence in the functional linear model. *Canadian Journal of Statistics*, **36,** 207–222.

Lastovicka, J. (2009). Global pattern of trends in the upper atmosphere and ionosphere: recent progress. *Journal of Atmospheric and Solar-Terrestrial Physics*, **71,** 1514–1528.

Lastovicka, J., A, V. Mikhailov, Ulich, T., Bremer, J., Elias, A., Ortiz de Adler, N., Jara, V., Abbarca del Rio, R., Foppiano, A., Ovalle, E. and Danilov, A. (2006). Long term trends in foF2: a comparison of various methods. *Journal of Atmospheric and Solar-Terrestrial Physics*, **68,** 1854–1870.

Lastovicka, J., Akmaev, R. A., Beig, G., Bremer, J., Emmert, J. T., Jacobi, C., Jarvis, J. M., Nedoluha, G., Portnyagin, Yu. I. and Ulich, T. (2008). Emerging pattern of global change in the upper atmosphere and ionosphere. *Annales Geophysicae*, **26,** 1255–1268.

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

Maslova, I., Kokoszka, P., Sojka, J. and Zhu, L. (2009). Removal of nonconstant daily variation by means of wavelet and functional data analysis. *Journal of Geophysical Research*, **114,** A03202.

Maslova, I., Kokoszka, P., Sojka, J. and Zhu, L. (2010a). Estimation of Sq variation by means of multiresolution and principal component analyses. *Journal of Atmospheric and Solar–Terrestial Physics*, **72,** 625–632.

Maslova, I., Kokoszka, P., Sojka, J. and Zhu, L. (2010b). Statistical significance testing for the association of magnetometer records at high–, mid– and low latitudes during substorm days. *Planetary and Space Science*, **58,** 437–445.

Mikhailov, A. V. and Marin, D. (2001). An interpretation of the foF2 and hmF2 long-term trends in the framework of the geomagnetic control concept. *Annales Geophysicae*, **19,** 733–748.

Mikhailov, A.V. and Marin, D. (2000). Geomagnetic control of the fof2 long-term trends. *Annales Geophysicae*, **18,** 653–665.

Müller, H-G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association*, **103,** 1534–1544.

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

Percival, D. B. and Walden, A. T. (2000). *Wavelet Methods for Time Series Analysis.* Cambridge University Press, Cambridge.

Qian, L., Burns, A. G., Solomon, S. C. and Roble, R. G. (2009). The eect of carbon dioxide cooling on trends in the f2-layer ionosphere. *Journal of Atmospheric and Solar-Terrestrial Physics*, **71,** 1592–1601.

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

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd edn. Springer, New York.

Rishbeth, H. (1990). A greenhouse effect in the ionosphere? *Planetary and Space Science*, **38,** 945–948.

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.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis.* Chapman & Hall/CRC, Boca Raton.

Secchi, P., Vantini, S. and Vitelli, V. (2011). A clustering algorithm for spatially dependent functional data. *Procedia Environmental Sciences*, **7,** 176–181.

Secchi, P., Vantini, S. and Vitelli, V. (2012). Bagging Voronoi classifiers for clustering spatial functional data. *International Journal of Applied Earth Observation and Geoinformation*, **22,** 53–64.

Sherman, M. (2011). *Spatial Statistics and Spatio–Temporal data: Covariance Functions and Directional Properties.* Wiley, New York.

Shi, J. Q. and Choi, T. (2011). *Gaussian Process Regression Analysis for Functional Data.* Chapman & Hall/CRC, London.

Solow, A. (1985). Bootstrapping correlated data. *Mathematical Geosciences*, **17,** 769–775.

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

Staicu, A. M., Crainiceanu, C. M., Reich, D. S. and Ruppert, D. (2012). Modeling functional data with spatially heterogeneous shape characteristics. *Biometrics*, **68,** 331–343.

Staniswalis, J. G. and Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, **93,** 1403–1418.

Sun, Y. and Genton, M.G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, **20,** 316–334.

Szekely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *Annals of Applied Statistics*, **3,** 1236–1265.

Szekely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, **35,** 2769–2794.

Ulich, T., Clilverd, M. A. and Rishbeth, H. (2003). Determining long-term change in the ionosphere. *Eos, Transactions American Geophysical Union*, **84,** 581–585.

Wackernagel, H. (2003). *Multivariate Geostatistics.* Springer, Berlin.

Yamanishi, Y. and Tanaka, Y. (2003). Geographically weighted functional multiple regression analysis: A numerical investigation. *Journal of the Japanese Society of Computational Statistics*, **15,** 307–317.

Yao, F. and Lee, T. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society: Series* B, **68,** 3–25.

Yao, F., Müller, H-G. and Wang, J-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100,** 577–590.

Yao, F., Müller, H.G., Clifford, A.J., Dueker, S.R., Follett, J., Lin, Y., Buchholz, B. and Vogel, J.S. (2003). Shrinkage estimation for functional principal component scores, with application to the population kinetics of plasma folate. *Biometrics*, **59,** 676–685.

APPENDICES

License Number	3130360404630
License date	Apr 15, 2013
Licensed content publisher	John Wiley and Sons
Licensed content publication	Journal of the Royal Statistical Society: Series C (Applied Statistics)
Licensed content title	Testing the equality of mean functions of ionospheric critical frequency curves
Licensed copyright line	© 2012 Royal Statistical Society
Licensed content author	Oleksandr Gromenko, Piotr Kokoszka
Licensed content date	Apr 24, 2012
Start page	715
End page	731
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Oleksandr Gromenko
Customer address	Department of Mathematics and Statistics Utah State University 3900 Old Main Hill Logan, UT 84322
License number	3130361093211
License date	Apr 15, 2013
Licensed content publisher	Elsevier
Licensed content publication	Computational Statistics & Data Analysis
Licensed content title	Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination
Licensed content author	Oleksandr Gromenko, Piotr Kokoszka
Licensed content date	March 2013
Licensed content volume number	59
Licensed content issue number	
Number of pages	13
Start Page	82
End Page	94
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Spatially Indexed Functional Data
Expected completion	date May 2013
Estimated size (number of pages)	140
Elsevier VAT number	GB 494 6272 12

CURRICULUM VITAE

Oleksandr Gromenko

Education

2009–2013 (Expected)

PhD in Statistics, Utah State University, Department of Math&Stat, Logan, UT.

Dissertation: Spatially Indexed Functional Data. *Advisor:* Professor Piotr Kokoszka

2007–2009

M.Sc in Physics, Clarkson University, Department of Physics, Potsdam, NY.

Specialization: Statistical Physics. *Advisor:* Professor Vladimir Privman

2002–2007

B.Sc and M.Sc in Physics, National Taras Shevchenko University, Department of Physics, Kiev, Ukraine. *Thesis title:* Dynamical chiral symmetry breaking in SU(N_C) gauge theories with large number of fermion flavors. *Advisor:* Professor Valeriy P. Gusynin

Honors and Awards

- The School of Graduate Studies Dissertation Fellowship, USU (2012-2013)
- Multiple travel awards from the Graduate Student Senate, Dean's office and the Department of Math&Stat, USU
- Winner in personal and team competition in All-Ukrainian Student's Tournaments in Physics (2003, 2004, 2005)

Research Interests

Computational Statistics, Functional Data Analysis, Spatio–Temporal Statistics, Nonparametric Statistics, Statistical Physics, Applications to Geosciences

COMPUTER SKILLS

Experience and Proficiency in C/C++ and R. Active user of MPI and OpenMP

TEACHING AND WORKING EXPERIENCE

Graduate Teaching Assistant at Utah State University

STAT 1040, Introduction to Statistics (Fall 2009, Spring 2010, Fall 2010)

STAT 2000, Statistical Methods (Spring 2012)

STAT 3000, Statistics for Scientists (Fall 2011)

Graduate Research Assistant at Utah State University

(Summer 2012, Spring, Summer 2011, Summer 2010)

Graduate Teaching Assistant at Clarkson University

PH131, Classical Mechanics (Fall 2008, Fall 2009)

PH132, Electricity & Magnetism (Spring 2007, Spring 2009)

Research Staff, JINR/CERN (Spring, Summer 2006)

ACADEMIC SERVICE

Referee: Statistical Modelling, Computational Statistics, Journal of Computational and Graphical Statistics, Communications in Statistics - Simulation and Computation

Membership: American Statistical Association, Institute of Mathematical Statistics

PERSONAL

Languages: Russian (native), English (fluent), Ukrainian (fluent)
US status: F-1 Visa

PAPERS IN JOURNALS

1. **Gromenko O.**, Kokoszka P., and Sojka J. (2013) Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves, under revision.

2. **Gromenko O.**, Kokoszka P. (2013) Nonparametric estimation in small data sets of spatially indexed curves with application to temporal trend determination. *Computational Statistics and Data Analysis* **59**, 82–94.

3. **Gromenko O.**, Kokoszka P. (2012) Testing the equality of mean functions of ionospheric critical frequency curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **61**, 715–731.

4. **Gromenko O.**, Kokoszka P., Zhu L. and Sojka J. (2012) Estimation and testing for spatially distributed curves with application to ionospheric and magnetic field trends. *The Annals of Applied Statistics* **6**, 669–696; `http://arxiv.org/abs/1206.6655`

5. ben-Avraham D., **Gromenko O.** and Politi P. (2009) Deterministic reaction models with power-law forces. *Journal of Physics A: Mathematical and Theoretical* **42**, 495004; `http://arxiv.org/abs/0909.0183`

6. **Gromenko O.** and Privman V. (2009) Random sequential adsorption of oriented superdisks. *Physical Review E* **79**, 042103; `http://arxiv.org/abs/0902.3089`

7. **Gromenko O.** and Privman V. (2009) Random sequential adsorption of objects of decreasing size. *Physical Review E* **79**, 011104; `http://arxiv.org/abs/0809.5061`

8. **Gromenko O.**, Privman V. and Glasser M. L. (2008) Random sequential adsorption model of damage and crack accumulation: exact one–dimensional results. *Journal of Computational and Theoretical Nanoscience* **5**, 2119–2123; `http://arxiv.org/abs/0712.3567`

PAPERS IN PROCEEDINGS AND PREPRINTS

1. **Gromenko O.**, Kokoszka P. (2011) Estimation and testing for geostatistical functional data. In *Recent Advances in Functional Data Analysis and Related Topics*, edited by F. Ferraty, Springer.

2. **Gromenko O.** (2007) Dynamical chiral symmetry breaking in SU(N_C) gauge theories with large number of fermion flavors; `http://arxiv.org/abs/0710.1591`

PROFESSIONAL CONFERENCES AND WORKSHOPS

1. *Contributed talk:* "Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination", Joint Statistical Meetings, San Diego, CA (2012/07)

2. *Poster presentation:* "Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination", Workshop on Analysis of High-Dimensional and Functional Data in Honor of Peter Hall, Davis, (CA 2012/05)

3. *Contributed talk:* "Testing the equality of mean functions of ionospheric critical frequency curves", Intermountain Graduate Research Symposium, Utah State University, Logan, UT (2012/04)

4. *Poster presentation:* "Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends ", Climate Modeling Opening Workshop, Pleasanton, CA (2011/08)

5. *Poster presentation:* "Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends", Joint CEDAR/GEM Workshop, Santa Fe, NM (2011/06)

6. *Contributed talk:* "Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends", Intermountain Graduate Research Symposium, Utah State University, Logan, UT (2011/04)

Summer Schools and Educational Workshops

1. Computational Advertising, SAMSI, NC (2012/08)
   Project/Talk: "Application of Penalized Matrix Factorization to Yahoo! Music Data"

2. Industrial Math/Stat Modeling Workshop for Graduate Students, NC (2011/07)
   Project/Talk: "Robust Optimal Design of Heliostat Arrays for Concentrating Solar Power Plants"

3. Nuclear theory and astrophysical applications. Helmholtz International Summer School, Joint Institute for Nuclear Researches, Dubna, Moscow region, Russian Federation (2005/08)
4. Summer School on Particle Physics. Abdus Salam International Centre for Theoretical Physics, Trieste, Italy (2005/06)