Universidade do Minho
Escola de Ciências

Luís Manuel dos Santos de Melo Margalho

Spatio-temporal modelling of environmental data

Maio de 2015

Universidade do Minho
Escola de Ciências

Luís Manuel dos Santos de Melo Margalho

Spatio-temporal modelling of environmental data

Doutoramento
Programa Doutoral em Matemática e Aplicações

Trabalho efectuado sob a orientação do
Professora Doutora Raquel Menezes Mota Leite e
Professora Doutora Inês Pereira Silva Cunha Sousa

Maio de 2015

STATEMENT OF INTEGRITY

I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, _____

Full name: _____

Signature: _____

Acknowledgements

First of all, my sincere appreciation must be assigned to my supervisors, professor Raquel Menezes and professor Inês Sousa. Without their helpful suggestions, constant support and availability to guide me during the period of preparation of this work, it would become an impossible task to complete.

My acknowledgments must also be addressed to my co-workers at Instituto Superior de Engenharia de Coimbra, for their words of encouragement and the support they gave me for the past years.

One word of acknowledgement, as a member of the project "Modelos conjuntos para processos espaço-temporais e respectivo desenho amostral, em Ciências do Ambiente e Saúde", FCT PTDC/MAT/112338/2009, should be given to Centre of Mathematics of Minho University, and to Centre for Environmental Biology of Lisbon University, for sharing the data base used in this work.

I wish also to thank the partial financial support from Coimbra Institute of Engineering / Politechnic of Coimbra, and from Minho University.

At last, but not at least, I am so grateful to whom I can call, above all, my friends: my first mention must go to my sisters, brother and my favorite nieces, but also to Carlos and Hermann, Carla, Ricardo, Paula e Jorge. To all of you, my deepest gratitude.

To my parents memory.

To my family.

Abstract

Spatio-temporal modelling of environmental data

Environmental monitoring may be defined as a description of processes and activities performed to characterize and monitor the quality of the environment. Monitoring schemes may differ greatly in their spatial and temporal extent, but as an outcome of any environmental monitoring process, data are gathered exhibiting both a spatial and a temporal dimension.

With this work, we aim to analyze the predictive accuracy when characterizing the spatio-temporal patterns of heavy metal deposition in mainland Portugal. The data set in use consists of measurements of heavy metal deposition in mosses, resulting from three nationwide surveys performed in 1992, 1996 and 2002.

Firstly, we begin with an exploratory descriptive analysis and an exploratory spatial analysis of the data, using well known techniques of spatial interpolation. After, we propose to make a spatio-temporal prediction of heavy metal concentration for the most recent survey, allowing to incorporate georeferenced explanatory covariates of the process under observation, calling on an existing spatio-temporal prediction model. This model focuses on the spatial dimension by defining random fields for the mean, the scale and the residuals components, and incorporates the time dimension by means of strictly temporal random fields, which work as corrections for the temporal evolution of the process.

Motivated by the fact that the data set in use in dense in the spatial dimension but sparse in the temporal one, a novel model-based approach is proposed for Gaussian data, corresponding to a saturated correlation model

in the time dimension. The proposed model is derived in order to accommodate not exclusively geo-referenced covariates, but also covariates associated to the temporal behavior of the process.
Regarding the results obtained in terms of predictive accuracy, a comparison of predictions from a purely spatial model with the ones from a spatio-temporal model showed that the latter improve the accuracy of predicted value. Moreover, if the comparison is restricted to the two spatio-temporal models, the new model proposal provides better results.

Keywords: geostatistics, spatio-temporal modelling, sparse time dimension, environmental biomonitoring, mosses.

Resumo

Modelação espaço-temporal de dados ambientais

Por monitorização ambiental entende-se uma descrição dos processos e atividades realizadas para caracterizar e monitorizar a qualidade do meio ambiente. Apesar de diferentes estudos de monitorização ambiental poderem diferir em termos de extensão espacial e temporal, de qualquer processo de monitorização resultam dados que apresentam tanto uma dimensão espacial como uma dimensão temporal.

Com este trabalho, pretende-se analisar a precisão das predições efectuadas ao caracterizar os padrões espaço-temporais de deposição de metais pesados em Portugal continental. A base de dados utilizada neste estudo consiste em medidas de deposição de metais pesados em musgos, resultante de três campanhas de amostragem a nível nacional, realizados em 1992, 1996 e 2002. Inicialmente será efectuada uma análise exploratória descritiva e uma análise exploratória espacial dos dados, utilizando técnicas bem conhecidas de interpolação espacial. De seguida, será desenvolvida uma previsão espaço-temporal da concentração de metais pesados para a campanha mais recente, permitindo incorporar variáveis geo-referenciadas explicativas do processo sob observação. Para isso, iremos recorrer a um modelo de previsão espaço-temporal existente. Este modelo incide sobre a dimensão espacial do processo através da definição de campos aleatórios para a média, para a escala e para os resíduos, e incorporando a dimensão temporal por meio de campos aleatórios estritamente temporais, que funcionam como correcções para a evolução temporal do processo.

Motivados pelo fato de o conjunto de dados em uso ser denso na dimensão espacial, mas escasso em termos temporais, é proposta uma abordagem

model-based para dados Gaussianos, e que corresponde a um modelo de correlação saturado na dimensão temporal. O modelo proposto é deduzido de forma a acomodar não somente covariáveis geo-referenciadas, mas também covariáveis associadas ao comportamento temporal do processo.
No que respeita à precisão dos valores de concentração de metais pesados, a comparação das previsões obtidas por meio de modelos puramente espaciais com as obtidas por modelos espaço-temporais revelou um melhor desempenho por parte destes últimos. É de realçar ainda que, se a comparação for restringida aos dois modelos espaço-temporais, a abordagem *model-based* proporciona melhores resultados.

Palavras-Chave: geoestatística, modelação espaço-temporal, dimensão temporal reduzida, bio-monitorização ambiental, musgos.

Contents

Glossary of Terms

APE	Absolute Prediction Error
AR	Auto Regressive
BLUE	Best Linear Unbiased Estimator
Exp	Exponential
GIS	Geographic Information System
MAPE	Mean Absolute Prediction Error
Max	Maximum
Min	Minimum
Mn	Manganese
MVN	Multivariate Normal Distribution
NO_3	Nitrate
OK	Ordinary Kriging
Pb	Lead
PM_x	Particulate Matter with diameter of x micrometres or less
Quart 1	First quartile
Quart 3	Third quartile
St. dev.	Standard Deviation
UK	Universal Kriging

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

1

Introduction

1.1 Understanding the problem

Although pollution had been known to exist for a very long time, it was after the industrial revolution in the 19^{th} century that its growth started to have global proportions. Over the last decades the interest in the impacts over the public health attributed to environmental pollution, namely the pollution caused by heavy metals, has increased. Heavy metals are metallic elements that are present in natural environments, where they occur at low concentrations, but also in contaminated environments, where high concentrations are observed. Heavy metals may be released into the environment as a consequence of human activities, *e.g.* from metal smelting and refining industries, plastic and rubber industries, or from burning of waste containing these elements. Once release to the air, the elements are deposited onto the soil, vegetation and water, and may persist in the environment for many years poisoning humans through inhalation, ingestion and skin absorption.

Worldwide, several projects have emerged on the subject of environmental pollution and in the assessment of its impact on humans. Some examples are the Atmospheric Heavy Metal Deposition in Europe, which has the objective of characterize qualitatively and quantitatively the atmospheric deposition of heavy metals in northern Europe (Rühling & Steinnes (1998)), the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air), a cohort study aiming at assessing the relationship between chronic exposure to air pollution and the progression of sub-clinical cardiovascular disease (Lindstrom *et al.* (2011)), the National Morbidity, Mortality, and Air Pollution Study (NMMAPS), aim-

ing to characterize the effects of airborne particles less than 10 μm in aerodynamic diameter (PM_{10}), alone and in combination with gaseous air pollutants (Samet *et al.* (2000)).

Biomonitoring can be defined as the process in which plants are used to assess changes in the environment, generally changes due to anthropogenic causes, and biomonitors are organisms that contain information on the quantitative aspects of the quality of the environment (Markert *et al.* (2003)).

Biomonitoring projects like the ones mentioned before, rely on the use of biomonitors, as they have proven to be excellent tools providing information which can be used to assess environmental quality, but also to investigate trends by monitoring systems, repeating measurements in time (Markert *et al.* (2003)). The use of plants as biomonitors is frequent for ecosystem quality assessment due to their sensitivity to chemical changes in environmental composition. The advantages of this use include, among others, low costs, the possibility of long-term sampling, and high availability. Lower plant organisms, like mosses, are often used in analysis of atmospheric depositions, soil quality and water purity, due not only to the mentioned sensitivity to chemical changes but also to their capacity to accumulate and store heavy metals and other toxins (Gadzała-Kopciuch *et al.* (2004)).

In Europe, mosses and lichens as biomonitors are widely used, where surveys that have been performed led to geographical and longitudinal descriptive studies of airborne metals (Sarmento (2012)).

1.2 Statistics and environmental studies

Statistical techniques are commonly applied in several areas in order to interpret, analyze, and understand data which may involve more than one type of measurement. Multivariate statistics deals with problems where more than one dependent variable is analyzed simultaneously with other variables.

Geostatistics is a branch of multivariate statistics that takes into account the spatial distribution information to accurately predict and display correlations present in the data. Within the field of geostatistics, interpolation methods are used to provide accurate estimations of the variable of interest by using the correlation that results from

known sample points and their geographic location relative to the point of estimation (Milillo (2009)).

One interpolation method broadly used in geostatistics is Kriging, a technique aiming at estimating values of the variable of interest at locations which have not been sampled, weighting the surrounding measured values based on the distance between the measured location and the not sampled one. If this interpolation technique refers to data which is primarily transformed from continuous values to binary, it is designated as Indicator Kriging.

In the literature one may find several studies applying geostatistics techniques to analyze biomonitoring data. Cocchi *et al.* (2007) use data of PM_{10} measurements from 11 spatial locations collected over 1096 days. Bruno *et al.* (2003) use a data set consisting of daily ozone measurements made at 32 monitoring locations, for the period 1998-2002. Mitchell *et al.* (2005) study the effect of high levels of CO_2 on rice, using data from 13 spatial locations and 112 time points. These few examples share the common feature of the number of time observations is larger than the number of spatial locations.

However, despite the easiness on gathering data enabled by modern technologies, there are cases where, due to the intrinsic nature of the process of data acquisition, data are collected over a large number of spatial locations but only a reduced number of time periods. One of such cases is the one resulting from the Portuguese participation on the Atmospheric Heavy Metal Deposition in Europe project, which yielded measurements of heavy metal concentration in mosses collected at 146 spatial locations on three nationwide surveys.

1.3 Main objectives

The last mentioned project illustrates the existence of an increasing interest in problems dealing simultaneously with spatial and temporal relationships between observations. To understand data collected not only across space at a given moment, but also along time at each location, there is a growing need for models that can accommodate both the spatial and the temporal dimension of data, usually known as spatio-temporal models.

With this work, strongly motivated by environmental monitoring studies, we aim not

only to understand how important is to the prediction process to consider data from the past, but also how the inclusion of variables explaining the process under observation can improve the accuracy of predictions. Therefore, our main goals are (*i*) to propose an extension of an existing geostatistical spatio-temporal model, allowing for considering explanatory covariates relevant to the process under observation, and (*ii*) to propose a simple spatio-temporal model, suitable for studies with a reduced number of time observations and also accommodating the possible existence of explanatory covariates. Moreover, for the latter model, the spatio-temporal covariance function will be prepared to take into account different scale parameters for the spatial and the temporal components, opposite to the most traditional interpretation of this function as proposed in Rodriguez-Iturbe & Mejia (1974). These two models are applied to the environmental biomonitoring data set resulting from the three available surveys of the Portuguese participation on the Atmospheric Heavy Metal Deposition in Europe project, in order to create prediction maps and error maps of heavy metal concentrations all over the Portuguese mainland territory and for the most recent survey.

1.4 Geostatistical software

Geostatistics provides a set of mathematical tools that have been used to data analysis, and to generate prediction maps from point observations together with the associated uncertainty maps. To perform this task, an important piece is certainly the computer program that implements the (geo)statistical algorithm that has been selected to predict the target variable (Hengl (2007), Fischer & Getis (2009)).

The increasing popularity of geostatistics has originated a substantial expansion of software suitable to that purpose, providing several solutions in terms of price, operating systems, user-friendliness, functionalities, graphical and visualization capabilities (Fischer & Getis (2009)). Among all the available software, R (the open source version of the S language for statistical computing, available at `http://www.r-project.org/`) is today identified as one of the fastest growing and most comprehensive statistical computing tools/communities (Hengl (2007)). It become even more attractive for geostatistical analysis after the integration of the geostatistical tools *geoR*, which implements model-based, likelihood-based and Bayesian geostatistical methods (Ribeiro & Diggle

(2001)), and *gstat*, which offers univariate and multivariate geostatistical methods for estimation and simulation, namely variogram modelling, simple, ordinary and universal kriging, and spatio-temporal kriging (Pebesma (2004)).

When dealing with spatial data, these are typically classified according to three classes: point-referenced data, which corresponds to the measurement of certain characteristic over a finite set of known spatial locations, areal (or lattice) data, which corresponds to measurements of certain characteristic over regions from a partition of some bounded spatial domain, and point processes data, where the data identifies the random spatial location where the measurement was made. Besides the aforementioned *geoR* and *gstat*, suitable for point referenced data, some other available packages in R are *geoRglm*, which extends *geoR* for Binomial and Poisson processes (Christensen & Ribeiro (2002)), *RandomFields*, offering tools for the simulation of different kinds of random fields, model estimation and inference for regionalized variables and data analysis, and model estimation for (geostatistical) linear (mixed) models (Schlather *et al.* (2013)), *geoCount*, providing functions to analyze and model geostatistical count data with generalized linear spatial models (Jing & de Oliveira (2015)), *spatial*, which provides functions for kriging and point pattern analysis (Venables & Ripley (2002)), *DCluster*, for the detection of spatial clusters of diseases (Gómez-Rubio *et al.* (2005)), or the R-inla packages, to solve models using Integrated Nested Laplace Approximation (INLA), which is a new approach to statistical inference for latent Gaussian Markov random fields (Rue *et al.* (2009)).

The available resources for geostatistical spatial or spatio-temporal analysis are not limited to R. For example *mGstat*, a geostatistical toolbox for Matlab, provides an interface to R's gstat, using Matlab as a scripting language. *SGeMS*, provides state of the art geostatistical simulation algorithms. The web-page Geospatial Analysis, `http://www.spatialanalysisonline.com/software.html`, contains a large number of examples from different GIS and related software packages and tool sets. The recent paper of E. Pebesma and co-authors, Pebesma *et al.* (2015), give an overview of published works using geostatistical software, covering spatial analysis topics such as visualization (micromaps, links to Google Maps or Google Earth), point pattern analysis, geostatistics, analysis of areal aggregated or lattice data, spatio-temporal statistics, Bayesian spatial statistics, and Laplace approximations.

1.5 Thesis outline

This thesis is organized as follows. The initial concepts on the area of spatial geostatistics are introduced in Chapter 2. A practical application of these concepts, focusing on the Kriging and Indicator Kriging techniques, is proposed based on a data set consisting of measurements of nitrate (NO_3) concentration in water samples, collected in four different moments independently of each other in a Portuguese river basin. Chapter 3 reveals the principal biomonitoring data set to be used to illustrate the usefulness of the geostatistical spatio-temporal models to be proposed. First, an exploratory data analysis of manganese (Mn) concentration and lead (Pb) concentration in mosses is conducted and, after, the spatial behavior of this data is also studied. The inclusion of covariates relevant to the process under observation is considered. The generalization of the spatial geostatistics framework to spatio-temporal geostatistics is addressed in Chapter 4. One particular spatio-temporal model, proposed to analyze interpolation errors when considering data from repeated observations of monitoring networks, is introduced. This model allows for the sampled number of spatial location to be larger than the number of temporal observations. The concepts are illustrated by means of the application of the mentioned model to the Mn and Pb data. In Chapter 5, a generalization of the previous model is proposed, aiming to take into account the existence of relevant covariates to the process under observation.

Departing from the particular characteristics of studies like the one originating the data set revealed in Chapter 3, namely the fact that the time dimension is much smaller than the spatial one, a new spatio-temporal geostatistical model is proposed in Chapter 6. This model proposal assumes the separability between the spatial and the temporal components. Moreover, complementing the most common interpretation of the covariance structure, different spatial and temporal sources of variability are allowed and, as a consequence of the reduced number of observations in time, the temporal correlation corresponds to a saturated model. The model is applied to the Mn data. The model's good performance in predicting Mn concentration values is assessed by comparing the predicted values with results obtained by using the spatio-temporal model introduced earlier, which also assumes spatial and temporal separability and enables the use of covariates.

Some conclusions and directions for future work are in Chapter 7.

2

Spatial geostatistics

2.1 Introduction

One type of information that distinguishes environmental data from most other types of data, is that the former always belong to some location in space and always was collected at a given time.

In case this information of *where* and *when* takes part of the data to be interpreted, we fall in the scope of geostatistics, a set of of numerical techniques that deal with the characterization of spatial attributes (Olea (2012)). Geostatistics offers a way of describe the spatial continuity of natural phenomena, adapting classical regression techniques to take advantage of this continuity (Isaacs & Srivastava (1989)).

The development of this branch of statistics started in the beginning of 1950's, with the work of D. Krige (Krige (1951)), in the mining and petroleum industries. Latter on, in a more formal way, the problem of spatial prediction was also addressed by Matheron (1971). Geostatistics has since then been applied to many other fields, in or related to the earth sciences. Some examples are applications related to environmental sciences (Høst *et al.* (1995), Guttorp & Loperfido (2008), Cocchi *et al.* (2007)), meteorology (Cressie & Huang (1999), Kyriakidis *et al.* (2001)) or hydrology (Rouhani & Wackernagel (1990), Goovaerts (2000)).

Although initially the main focus of geostatistics was on spatial variables, over the past years the conceptual viewpoint accommodated also a temporal dimension, addressing variables that change both in space and time. Several books have been written focusing on the subject of geostatistics, both in development of the mathematical and statistical

theory (*e.g.*, Cressie (1993), Cressie & Wikle (2011), Diggle & Ribeiro (2007)) and in applications (*e.g.*, Goovaerts (1997), Journel & Huijbregts (1997)).
In this chapter, our main objective is to present a concise review of the main characteristics defining the concepts of spatial geostatistics.

2.2 Spatial geostatistics

Let us consider a continuous spatial process (or random field)

$$\{Z(\mathbf{s}),\, \mathbf{s} \in D\} \tag{2.1}$$

where $\mathbf{s}$ are locations within some spatial region $D \subset \mathbb{R}^d$. Typically, $d = 1, 2$ or 3. The characterization of the spatial process (2.1) is usually made by means of the cumulative distribution function

$$F_{\mathbf{s}_1,\ldots,\mathbf{s}_n}(z_1,\ldots,z_n) = P\big(Z(\mathbf{s}_1) \leq z_1,\ldots,Z(\mathbf{s}_n) \leq z_n\big)\ ,\ n \geq 1 \tag{2.2}$$

which must observe the usual conditions of being symmetric and consistent:

- symmetry:

$$F_{\mathbf{s}_{i_1},\ldots,\mathbf{s}_{i_n}}(z_{i_1},\ldots,z_{i_n}) = F_{\mathbf{s}_1,\ldots,\mathbf{s}_n}(z_1,\ldots,z_n)$$

  for any permutation $\mathbf{s}_{i_1},\ldots,\mathbf{s}_{i_n}$ of the indexes $1,\ldots,n$

- consistency:

$$F_{\mathbf{s}_1,\ldots,\mathbf{s}_n,\mathbf{s}_{n+1},\ldots,\mathbf{s}_{n+k}}(z_1,\ldots,z_n,\infty,\ldots,\infty) = F_{\mathbf{s}_1,\ldots,\mathbf{s}_n}(z_1,\ldots,z_n)$$

When the joint cumulative distribution (2.2) is a multivariate Gaussian distribution, the random field (2.1) is denoted as a Gaussian random field.
In general, we observe only one (partial) realization of the random field. That is, we have a sample of size one which is a collection of n observations $z(\mathbf{s}_1),\ldots,z(\mathbf{s}_n)$ at the known locations $(\mathbf{s}_1,\ldots,\mathbf{s}_n)$. In order to make possible to perform inferences about the spatial process, some assumptions on the regularity of the process ought to be made, which will now be presented.

Moments

The moment of order k of the random field $Z(\mathbf{s})$, defined at any location $\mathbf{s} \in D$ is

$$\mathrm{E}\Big[\big(Z(\mathbf{s})\big)^k\Big] = \int x^k dF_{\mathbf{s}}(x) \tag{2.3}$$

provided this integral exists.

Expected value

The expected value of a random field $Z(\mathbf{s})$ is defined as the order-one moment,

$$\mu(\mathbf{s}) = \mathrm{E}\big[Z(\mathbf{s})\big] \tag{2.4}$$

for any location $\mathbf{s} \in D$. In general, the expected value is allowed to depend on the location $\mathbf{s}$. In geostatistical applications, $\mu(\mathbf{s})$ is often referred to as the trend.

Variance and Covariance

The variance of a random field $Z(\mathbf{s})$ is defined as the second-order moment about the expected value $\mu(\mathbf{s})$,

$$\mathrm{Var}\big[Z(\mathbf{s})\big] = \mathrm{E}\Big[\big(Z(\mathbf{s}) - \mu(\mathbf{s})\big)^2\Big] \tag{2.5}$$

for any location $\mathbf{s} \in D$. As for the expected value, the variance is generally dependent on the location $\mathbf{s}$. The covariance is defined by

$$\mathrm{Cov}\big(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\big) = \mathrm{E}\Big[\big(Z(\mathbf{s}_i) - \mu(\mathbf{s}_i)\big)\big(Z(\mathbf{s}_j) - \mu(\mathbf{s}_j)\big)\Big] \tag{2.6}$$

for any locations $\mathbf{s}_i$ and $\mathbf{s}_j$ in D, $i, j = 1, \ldots, n$.

Strict stationarity

Given the set of $n \geq 1$ spatial locations $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$ in D and the vector $\mathbf{h} \in \mathbb{R}^d$ such that $\mathbf{s}_i + \mathbf{h} \in D$, $i = 1, \ldots, n$, $Z(\mathbf{s})$ is said to be strictly stationary if the distributions of $\big(Z(\mathbf{s}_1 + \mathbf{h}), \ldots, Z(\mathbf{s}_n + \mathbf{h})\big)$ and $\big(Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\big)$ are identical, that is, any translation of a set of locations does not alter the joint distribution,

$$F_{\mathbf{s}_1+\mathbf{h},\ldots,\mathbf{s}_n+\mathbf{h}}(z_1, \ldots, z_n) = F_{\mathbf{s}_1,\ldots,\mathbf{s}_n}(z_1, \ldots, z_n), \ \forall \mathbf{h} \in \mathbb{R}^d.$$

It is usual to consider, however, less restrictive conditions to caracterize a stationary random field.

Second-order (or weak) stationarity

The random field $Z(\mathbf{s})$ is said to be second-order stationary if it shows a constant first order moment and the covariance between two given locations only depends on the separation vector,

- $\mathrm{E}\big[Z(\mathbf{s})\big] = \mu,\ \forall \mathbf{s} \in D$
- $\mathrm{Cov}\big(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\big) = \mathrm{C}(\mathbf{s}_i - \mathbf{s}_j),\ \mathbf{s}_i,\ \mathbf{s}_j \in D$

The function C(.) is usually known as the *stationary covariance function* or *covariogram*.
In particular, if the spatial process $Z(\mathbf{s})$ is such that $\mathrm{C}(\mathbf{0}) > 0$, the second-order stationarity can also be stated by means of the *correlation function* or *correlogram*, denoted by $\rho(.)$,

$$\mathrm{Corr}\big(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\big) = \frac{\mathrm{Cov}\big(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\big)}{\mathrm{C}(\mathbf{0})} = \rho(\mathbf{s}_i - \mathbf{s}_j),\ \mathbf{s}_i, \mathbf{s}_j \in D \tag{2.7}$$

Notice that $\rho(\mathbf{s}_i - \mathbf{s}_j) = \rho(\mathbf{s}_j - \mathbf{s}_i)$ and $\rho(\mathbf{0}) = 1$.

Intrinsic stationarity

If the spatial process $Z(\mathbf{s})$ verifies that the first order moment is constant and the variance of the difference between the observation at two locations only depends on the difference between those locations, it is said to be intrinsically stationary,

- $\mathrm{E}\big[Z(\mathbf{s})\big] = \mu,\ \forall \mathbf{s} \in D$
- $\mathrm{Var}\big[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\big] = 2\gamma(\mathbf{s}_i - \mathbf{s}_j),\ \mathbf{s}_i,\ \mathbf{s}_j \in D$

The functions $2\gamma(.)$ and $\gamma(.)$ are usually known as the *variogram* and the *semi-variogram*, respectively, although some authors use for the latter also the *variogram*. According to Cressie (1993), the variogram is a model-based measure of the spatial statistical

dependence in a geostatistical process.

Relationship between different definitions of stationarity

Variance's properties state that

$$\begin{aligned} 2\gamma(\mathbf{s}_i - \mathbf{s}_j) &= \mathrm{Var}\big[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\big] \\ &= \mathrm{Var}\big[Z(\mathbf{s}_i)\big] + \mathrm{Var}\big[Z(\mathbf{s}_j)\big] - 2\mathrm{Cov}\big(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\big) \end{aligned} \tag{2.8}$$

meaning that the knowledge of the variance of the random field $Z(\mathbf{s})$ enables the identification of the variogram and reciprocally.
If $Z(\mathbf{s})$ is a second-order stationary process, its variance is known and constant, say $\sigma^2(\mathbf{s}) = \mathrm{C}(\mathbf{0})$, and consequently

$$\begin{aligned} 2\gamma(\mathbf{s}_i - \mathbf{s}_j) &= 2\Big(\mathrm{C}(\mathbf{0}) - \mathrm{Cov}\big(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\big)\Big) \\ \gamma(\mathbf{s}_i - \mathbf{s}_j) &= \mathrm{C}(\mathbf{0}) - \mathrm{Cov}\big(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\big) \\ \gamma(\mathbf{s}_i - \mathbf{s}_j) &= \mathrm{C}(\mathbf{0}) - \mathrm{C}(\mathbf{s}_i - \mathbf{s}_j) \end{aligned} \tag{2.9}$$

that is, a second-order stationary process is intrinsically stationary, being the variogram given by

$$\gamma(\mathbf{h}) = \mathrm{C}(\mathbf{0}) - \mathrm{C}(\mathbf{h}) \tag{2.10}$$

2.3 Variogram and covariogram properties

The (semi-)variogram and/or the covariogram are functions generally used to model spatial dependency and, therefore, they must observe some properties.

2.3.1 Variogram properties

- The variogram is the expected value of the squared deviation between two observations, hence

$$\gamma(\mathbf{h}) \geq 0, \forall\, \mathbf{h} \in \mathbb{R}^d \tag{2.11}$$

and $\gamma(\mathbf{0}) = 0$.

- Also by the definition, the variogram is an even function,

$$\gamma(\mathbf{-h}) = \gamma(\mathbf{h}) \tag{2.12}$$

- When considering a continuous variable, one should expect the variogram to pass through the origin at a distance $\|\mathbf{h}\| = 0$. In practice, however, it is possible that the variogram approaches a positive value as $\|\mathbf{h}\|$ approaches zero, suggesting a discontinuous process,

$$\lim_{\|\mathbf{h}\|\to\mathbf{0}} \gamma(\mathbf{h}) = \tau^2 > 0 \tag{2.13}$$

  This discrepancy is known as the nugget variance. In this case, τ^2 is labeled as the *nugget effect*, as this discontinuity was identified in mining applications of the first studies in Geostatistics. The nugget effect occurs as a result from small scale variability between spatially correlated variables and/or measurement errors.

- Not all variogram functions are valid to perform inference or prediction. In order to a variogram to be a valid one, it must verify that

$$\sum_i \sum_j a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0 \tag{2.14}$$

  for any finite set of locations $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$ and any set of real constants $(a_1, \ldots, a_n)$ such that $\sum_{i=1}^n a_i = 0$. This condition is equivalent to say that the matrix $\Gamma = \big[\gamma(\mathbf{s}_i - \mathbf{s}_j)\big]_{i,j}$ is negative semidefinite.

2.3.2 Some further characteristics of the variogram

The **sill** is the maximum height, if existing, of the variogram curve. As $\|\mathbf{h}\|$, the distance between any two spatial points, becomes larger, the correlation (and hence the covariance) between the response at those points becomes negligible. That is, once $\lim_{\|\mathbf{h}\|\to\infty} 2\gamma(\mathbf{h}) = \lim_{\|\mathbf{h}\|\to\infty} \mathrm{Var}\big[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})\big] \simeq \mathrm{C}(\mathbf{0})$, the sill in the variogram curve corresponds to the variance of the process.

The **partial sill**, denoted by σ^2, is the difference between the sill and the nugget effect, $\sigma^2 = \mathrm{C}(\mathbf{0}) - \tau^2$.

It is common, in practice, that the correlation between $Z(\mathbf{s})$ and $Z(\mathbf{s}+\mathbf{h})$ vanishes when the distance $\|\mathbf{h}\|$ becomes too large. The **range** is the distance $\|\mathbf{h}\|$ such that pairs of spatial locations further than this distance apart are negligibly correlated. When this

condition holds, the variogram reaches the sill.

2.3.3 Covariogram properties

The covariogram share properties with the variogram, as it is valid, by (2.10), that

$$\mathrm{C}(\mathbf{h}) = \mathrm{C}(\mathbf{0}) - \gamma(\mathbf{h}).$$

Therefore,

- $\mathrm{C}(\mathbf{0}) = \mathrm{Var}\big[Z(\mathbf{s})\big] \geq 0$
- $\mathrm{C}(\mathbf{-h}) = \mathrm{C}(\mathbf{h})$ and, by Cauchy-Schwarz inequality, $|\mathrm{C}(\mathbf{h})| \leq \mathrm{C}(\mathbf{0})$
- $\lim\limits_{\|\mathbf{h}\|\to 0} \mathrm{C}(\mathbf{h}) = \sigma^2$
- $\lim\limits_{\|\mathbf{h}\|\to \infty} \mathrm{C}(\mathbf{h}) = 0$
- $\sum\limits_i \sum\limits_j a_i a_j \mathrm{Cov}\big(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\big) \geq 0$ for any finite set of locations $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$ and any set of real constants $(a_1, \ldots, a_n)$ such that $\sum_{i=1}^n a_i = 0$, that is, the matrix $\Sigma = \Big[\mathrm{Cov}\big(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\big)\Big]_{i,j}$ is positive semidefinite.

2.3.4 Isotropy and anisotropy

The assumption of isotropy has the advantage of greatly simplify the modelling of the spatial dependence. In many cases, there is no reason to expect that the spatial dependency has the same behavior in all directions. However, the assumption of isotropy is typically made out of convenience.

An intrinsically stationary spatial process $Z(\mathbf{s})$ is said to be isotropic if the variogram depends upon $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$ only through its length, not through its direction,

$$\mathrm{Var}\big[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})\big] = \gamma(\|\mathbf{h}\|),\ \forall\, \mathbf{h} \in \mathbb{R}^d,\ \mathbf{s}, \mathbf{s}+\mathbf{h} \in D. \tag{2.15}$$

In opposition, in a stationary anisotropic process, the spatial association depends upon the separation vector between locations both for its length and for its direction.

The condition of being an isotropic spatial process is not very demanding, since an anisotropic process can be reduced to isotropy by a linear transformation of the coordinates.

2.4 Parametric models for isotropic variograms

In the same way as when dealing with random variables, variograms can also be identified according to the family that they belong. There exist various families of models for variograms used in practice, some examples being listed next. In what follows, we are representing the range by ϕ, the nugget effect by τ^2 and the partial sill by σ^2.

Spherical

The spherical model is in general suitable for modelling a spatial correlation which decreases approximately linearly with the separation distance, being zero beyond a certain distance.

$$\gamma(\mathbf{h}) = \begin{cases} 0 & , \quad \mathbf{h} = 0 \\ \tau^2 + \sigma^2 \left(\frac{3}{2}\frac{\mathbf{h}}{\phi} - \frac{1}{2}\left(\frac{\mathbf{h}}{\phi}\right)^3 \right) & , \quad 0 < \mathbf{h} \leq \phi \\ \tau^2 + \sigma^2 & , \quad \mathbf{h} > \phi \end{cases} \tag{2.16}$$

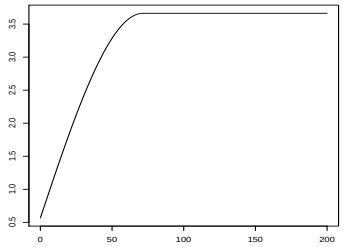

Figure 2.1: Example of spherical model

Exponential

The exponential model shows a shape similar to the spherical model and reaches the sill only asymptotically as $\|\mathbf{h}\| \to +\infty$.

$$\gamma(\mathbf{h}) = \begin{cases} 0 & , \quad \mathbf{h} = 0 \\ \tau^2 + \sigma^2 \left(1 - \exp\left(-\frac{\mathbf{h}}{\phi} \right) \right) & , \quad \mathbf{h} > 0 \end{cases} \tag{2.17}$$

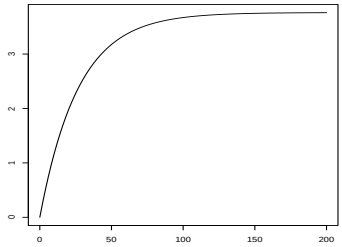

Figure 2.2: Example of exponential model

Linear

The linear model does not reach a sill, so the use of σ^2 is not appropriate, being replaced by $b \geq 0$. This model doesn't correspond to a stationary process.

$$\gamma(\mathbf{h}) = \begin{cases} 0 & , \quad \mathbf{h} = 0 \\ \\ \tau^2 + b\|\mathbf{h}\| & , \quad \mathbf{h} > 0 \end{cases} \tag{2.18}$$

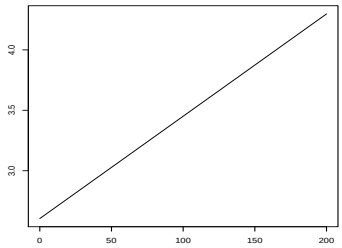

Figure 2.3: Example of linear model

Matérn

The Matérn family are highly flexible models, and so are suitable for modelling com-

plicated behaviors.

$$\gamma(\mathbf{h}) = \begin{cases} 0 & , \quad \mathbf{h} = 0 \\ \\ \tau^2 + \sigma^2 \left(1 - \rho(\mathbf{h})\right) & , \quad \mathbf{h} > 0 \end{cases} \tag{2.19}$$

where the correlation function $\rho(.)$ is given by

$$\rho(\mathbf{h}) = \left(2^{\nu-1}\Gamma(\nu)\right)^{-1} \left(\frac{\mathbf{h}}{\phi}\right)^{\nu} K_\nu \left(\frac{\mathbf{h}}{\phi}\right) \tag{2.20}$$

with $\nu > 0$ and $K(.)$ is the Bessel function of order ν.
In particular, if the order ν in (2.20) is 0.5, the Matérn model and the exponential model coincide. Also, as a limit case, if $\nu \to +\infty$, the Matérn model is also known as the **Gaussian** model, where the correlation function is

$$\rho(\mathbf{h}) = \exp\left(-\left(\frac{\mathbf{h}}{\phi}\right)^2\right)$$

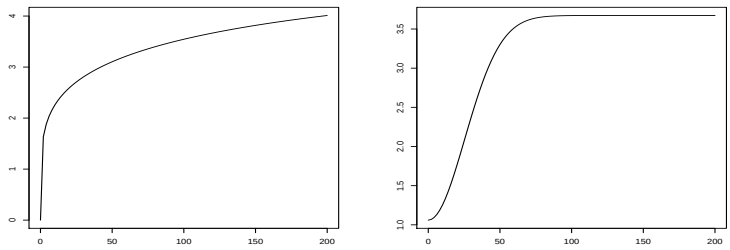

Figure 2.4: Example of Matérn model (left) and Gaussian model (right)

2.5 Parameter estimation and spatial predictions

In practical applications, after collecting a discrete set of observations $\{Z(\mathbf{s}_i), \mathbf{s}_i \in D, i = 1, \ldots, n\}$, the inference process aims at estimating the parameters of the variogram (or the covariogram) from the sample information.
Under second order stationarity assumption, the variogram function can be written as

$$\gamma(\mathbf{h}) = \frac{1}{2}\mathrm{E}\left[\left(Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})\right)^2\right] \tag{2.21}$$

Replacing in (2.21) the expected value by its empirical counterpart is a way to use the available data to estimate the variogram. The first proposal of an empirical variogram estimator is due to Matheron, usually known as the classical estimator

$$\widehat{\gamma}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{N(\mathbf{h})} \big(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\big)^2 \tag{2.22}$$

where $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \|\mathbf{s}_i - \mathbf{s}_j\| = \mathbf{h}\}$ and $|N(\mathbf{h})|$ is the cardinality of $N(\mathbf{h})$. The empirical variogram is, then, the primary tool used for inference on the model parameters. Having identified a covariance model of spatial dependence, one can proceed with predicting the spatially continuous process at an unsampled location $\mathbf{s}_0$. The process of spatial prediction, eventually to the whole study area, is generally mentioned as Kriging.

Kriging is a **L**inear interpolation method, since the estimated values are weighted linear combinations of the observed data, **U**nbiased once the mean of the errors is zero, and **B**est since it aims at minimizing the variance of the errors. That is, Kriging is a **BLUE** method.

Depending on the knowledge about the mean of the process under observation, one can have Simple, Ordinary or Universal Kriging. Simple Kriging assumes a known constant trend throughout the study area, Ordinary Kriging assumes an unknown constant trend and Universal Kriging assumes a varying, unknown trend.

According to Goovaerts (1997), given the n observations of the random process $Z(\mathbf{s})$, $\{z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n)\}$, the Kriging estimators are but variants of the linear regression estimator $\widehat{Z}(\mathbf{s})$, defined as

$$\widehat{Z}(\mathbf{s}) - \mu(\mathbf{s}) = \sum_{i=1}^{n(\mathbf{s})} \lambda_i(\mathbf{s})\big(Z(\mathbf{s}_i) - \mu(\mathbf{s}_i)\big) \tag{2.23}$$

where $\lambda_i(\mathbf{s})$ is the weight assigned to each datum $z(\mathbf{s}_i)$, $i = 1, \ldots, n(\mathbf{s})$. The number of data involved in the estimator (2.23) is specific to each location, in particular only the $n(\mathbf{s})$ data points closest to the location $\mathbf{s}$ being estimated are considered.

The main objective is to minimize the estimation of error variance under the constraint of unbiasedness,

$$\begin{aligned} &min\ \sigma^2(\mathbf{s}) = \mathrm{Var}\big[\widehat{Z}(\mathbf{s}) - Z(\mathbf{s})\big] \\ &\text{subject to} \\ &\mathrm{E}\big[\widehat{Z}(\mathbf{s}) - Z(\mathbf{s})\big] = 0 \end{aligned} \tag{2.24}$$

The derivation of the Kriging equations for each of the Simple, Ordinary or Universal cases, is well described in the literature. See, *e.g.* Goovaerts (1997), Chilès & Delfiner (2012) or Isaacs & Srivastava (1989). The predicted value of the spatial process at the unsampled location $\mathbf{s}_0$ is

$$\widehat{Z}(\mathbf{s}_0) = \mu(\mathbf{s}_0) + \mathbf{c}_0^T \mathbf{C}_Z^{-1}\big(\mathbf{Z} - \boldsymbol{\mu}\big) \tag{2.25}$$

being the variance of the prediction given by

$$\sigma^2(\mathbf{s}_0) = \mathbf{C}_0 - \mathbf{c}_0^T \mathbf{C}_Z^{-1} \mathbf{c}_0 \tag{2.26}$$

where $\mathbf{c}_0 = \mathrm{Cov}\big(Z(\mathbf{s}_0), Z(\mathbf{s})\big)$, $\mathbf{C}_0 = \mathrm{Var}\big[Z(\mathbf{s}_0)\big]$ and $\mathbf{C}_Z = \mathrm{Cov}\big(Z(\mathbf{s}_i), Z(\mathbf{s}_i)\big)$, $\mathbf{s}_i, \mathbf{s}_j \in D$.

2.6 Case study: water quality monitoring

This section illustrates an application (Margalho *et al.* (2011)) of the mentioned concepts to a real data set of measurements of NO_3 concentration in water samples. The data will be used just in this section with illustrative purposes, and no more elsewhere in the text.

The NO_3 concentration measurements were collected in four different moments independently of each other, at the Esposende-Vila do Conde aquifer, situated in the Cávado river basin, on the northwest region of Portugal (Figure 2.5, black dots represent sampling locations). The area in question is labeled as vulnerable, therefore it is important to know the behavior of water quality along time. The interest in a study like this derives from the fact that groundwater quality is regulated by the European law 2006/118/CE concerning groundwater protection against pollution and, particularly in Portugal for human purposes, by the Portuguese law 306/2007.

Exploratory Analysis of Nitrate Pollution Data

The following description is based on a set of 79 measurements of NO_3 concentration, performed over 25 different monitoring stations of groundwater quality during four different moments, Spring and Fall, of 2008 and 2009, although not every stations have

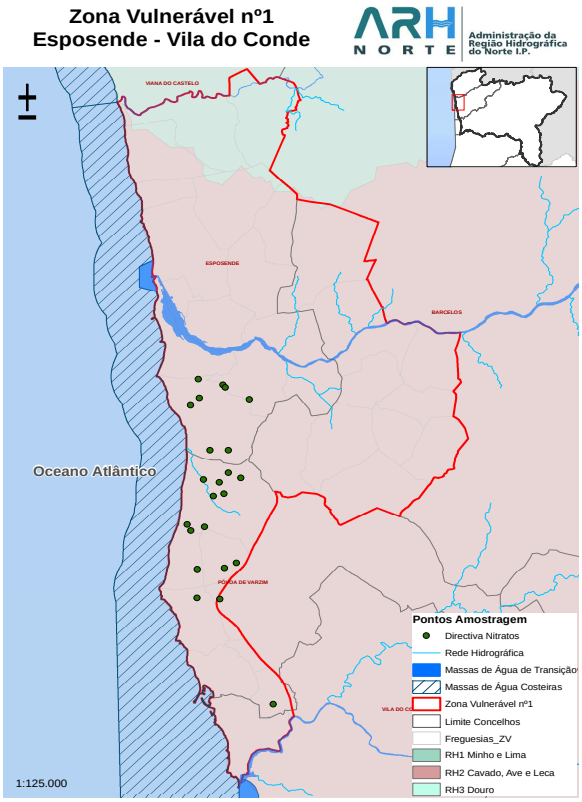

Figure 2.5: Map of vulnerable zone (delimited by the red line) with sampling points (black dots)

been monitored at all times. Table 2.1 shows some descriptive statistics related to these measurements.

Table 2.1: Descriptive statistics related to NO_3 measurements

NO_3(mg/L)	Spring'08	Fall'08	Spring'09	Fall'09
N	22	20	19	18
Min	5.50	2.00	2.10	2.20
Quart 1	48.43	35.15	64.25	50.75
Median	90.75	104.00	104.00	99.40
Quart 3	147.50	118.00	131.50	141.00
Max	229.00	382-00	331.00	277.00
Mean	101.70	102.60	112.60	109.00
St. Dev.	67.12	92.40	78.18	77.54

For the four periods of observation, the mean values of NO_3 concentrations are equivalent, although 2008 present mean values lower than 2009. The minimum observed values varies from 2 to 5.5 mg/L and the maximum values from 229 mg/L to 382 mg/L. The lowest range, 223.5 mg/L, occurred on Spring of 2008. Figure 2.6 depicts graphically the observed values. As stated before, Spring 2008 was the period with lowest range, but this was also the period with largest interquartile range.
Shapiro-Wilk tests revealed that Fall 2008 and Spring 2009 data fail normality. However, by using the square-root transform, new data could be assumed to behave like Gaussian. In the spatial analysis that follows, the square root transformed values will be considered for the four moments.

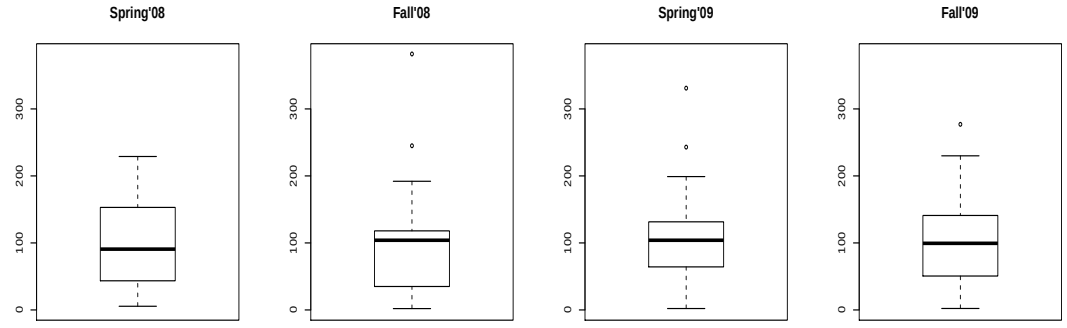

Figure 2.6: Boxplots of NO_3 concentration measurements in different moments

Spatial analysis of nitrate pollution data

Figure 2.5 shows that the sampling locations are not distributed over the whole area. As a consequence of that, for the spatial analysis that follows a restriction to the region containing sampled locations was considered.
After the computation of the empirical semivariogram for each period of observation, several models of isotropic semivariograms were adjusted. Table 2.2 represents the estimated parameters of spherical models (illustrated in Figure 2.7), obtained by the weighted least squares method.

Table 2.2: Estimated spherical semivariogram parameters obtained by Weighted Least Squares Method

	Spring'08	Fall'08	Spring'09	Fall'09
Nugget (τ^2)	0.58	11.12	2.65	2.38
Partial Sill (σ^2)	10.58	11.18	12.64	12.37
Range (ϕ)	700.00	700.00	700.00	700.00

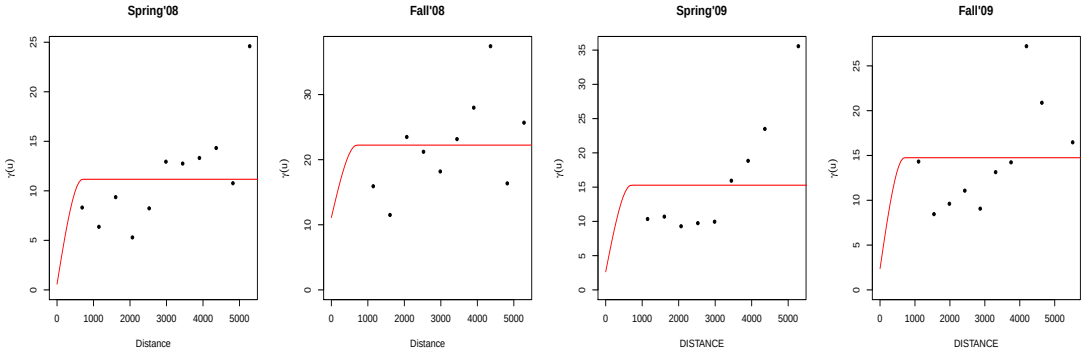

Figure 2.7: Empirical variograms for transformed NO_3 data with spherical model adjusted

The existence of a model of spatial correlation enables the interpolation of measurements to non observed locations, via Kriging methodology.

For each time period, predicted square root transformed values and the associated interpolation errors, defined as the square root of the interpolation variance (2.26) are summarized in Table 2.3. As for the sampled measurements, lower values for the mean

Table 2.3: Summary of square root predicted NO_3 concentration and associated interpolation error

	Spring'08		Fall'08		Spring'09		Fall'09	
NO_3	Pred.	Error	Pred.	Error	Pred.	Error	Pred.	Error
Min	3.11	1.20	5.18	3.99	3.14	2.15	3.43	2.09
Median	8.99	3.36	9.17	4.81	9.60	3.96	9.82	3.89
Max	14.87	3.50	13.69	4.89	16.14	4.09	15.25	4.02
Mean	9.03	3.21	9.17	4.74	9.62	3.83	9.83	3.77
St. dev.	1.18	0.38	0.76	0.17	1.04	0.33	1.13	0.32

and the median are registered in 2008. Also for this year, the range of predicted transformed concentrations is lower than for 2009, being the predictions related to Fall of 2008 the ones exhibiting less spreaded values. In terms of interpolation error, the mean and median values are of similar magnitude, except for Fall of 2008, where one can find larger values.

Maps are a valuable tool to better understand the spatial distribution of predicted concentrations across the study region. As can be observed in Figure 2.8 lower values (represented by the green shades) occur on the northern part, probably due to the proximity to a river. In the southern part, once there are no sampling locations nearby, the expected NO_3 concentrations exhibit no variation.

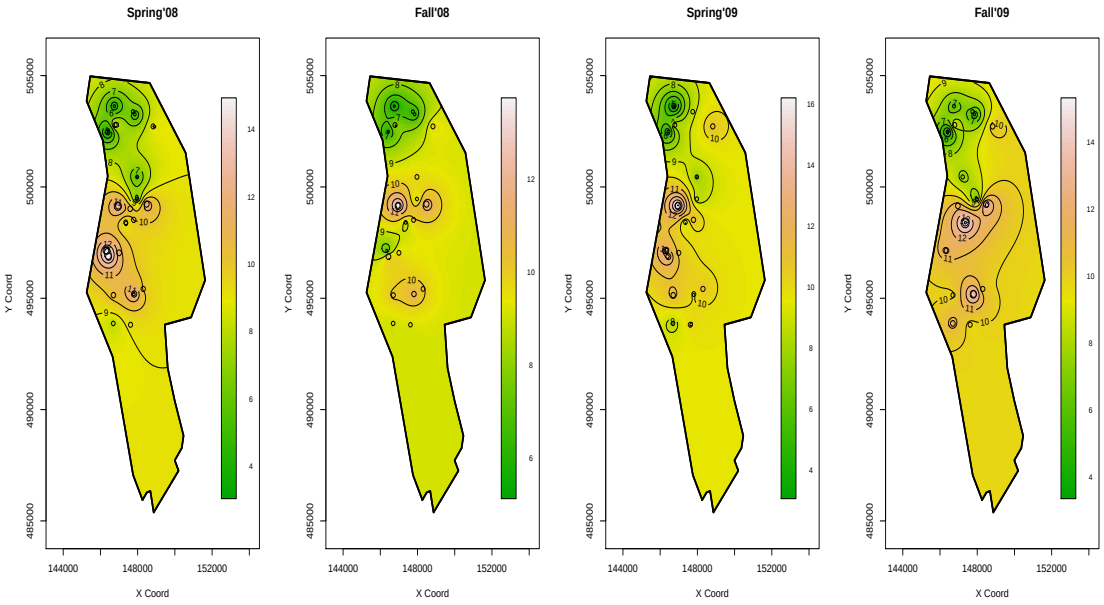

Figure 2.8: Prediction maps of square root transformed NO_3 concentration for each time period

As a measure of accuracy of the predicted values, maps of interpolation errors (Figure 2.9) show, as was to be expected, that more accurate values are those near the sampling locations. It is worthwhile to mention the fact that higher values of interpolation error occur in the southern part of the region under consideration, where less sampling locations are available.

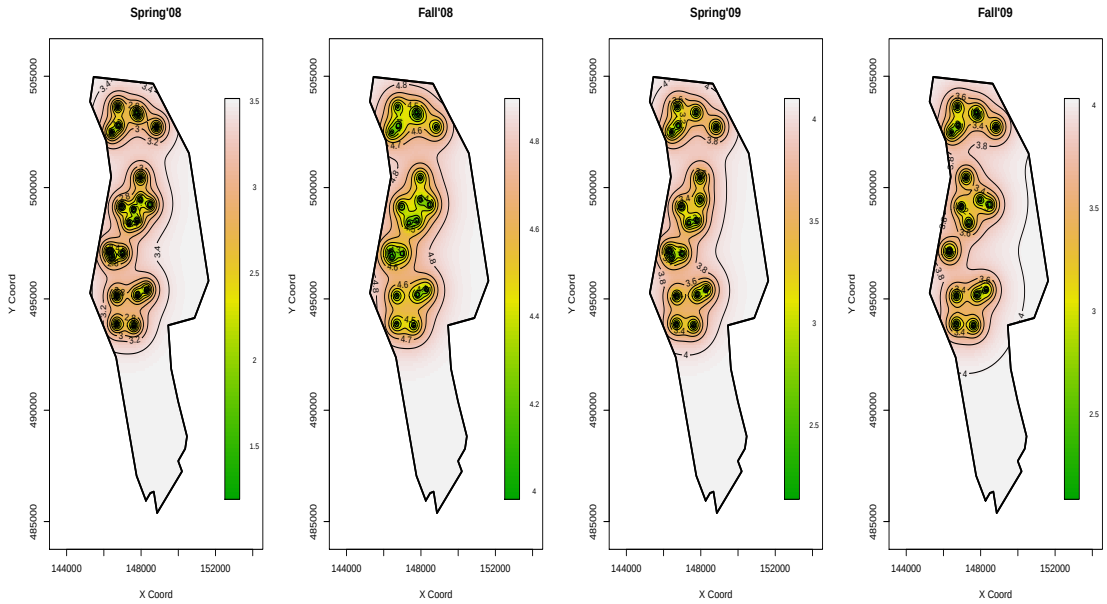

Figure 2.9: Interpolation error maps of square root transformed NO_3 concentration for each time period

Estimating the risk of exceeding regulatory thresholds

It may be the case that the primary objective of a given study is to predict the risk of exceeding particular values, such as regulatory thresholds of environmental contamination. For that, geostatistics is increasingly used to estimate and map that risk, as one possible way of identifying polluted areas is mapping pollutant concentrations. Interpretation of probability maps is based on a level of risk above which appropriate actions should be taken.

For NO_3 concentration, Portuguese law 306/2007 state that water with concentration above 50 mg/L should not be used for human purposes. In a way similar to the one used previously to construct prediction maps, risk maps are generated by Indicator Kriging. This procedure computes, using the samples in the neighborhood, the probability of data values in a given area being greater than the imposed threshold.

To do so, data values are transformed into indicator values: original values which exceed the chosen threshold value are coded 1, and those below the threshold value are coded 0. With these new indicator data, Indicator Kriging is conducted with the same

algorithm as Ordinary Kriging. The resulting risk maps are presented in Figure 2.10.

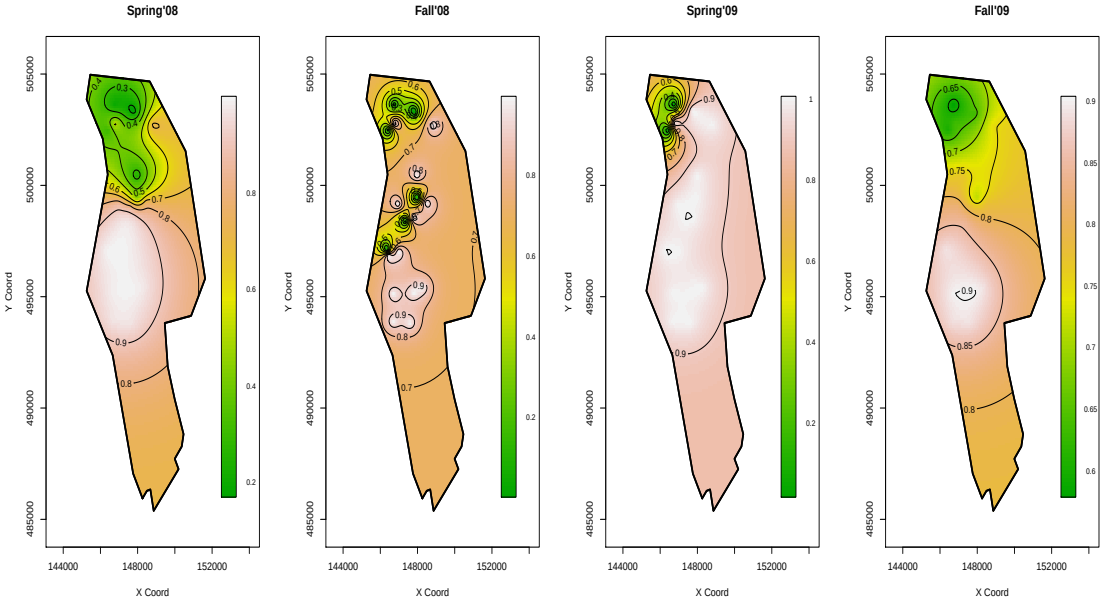

Figure 2.10: Contamination by NO_3 risk maps for each time period

By comparing these maps, we can see that for both years, Spring has a higher risk than Fall, probably because observations were collected after land chemical preparation for agricultural purposes. Also, for all observation periods, the risk of exceeding regulatory threshold is higher in the southern area of the observed region, which is in the same way as in the prediction maps. This is probably due not only to the existence of a river in the northern part of the study region, which can help to remove some of the groundwater pollution, but also to the reduced number of sampling locations in that area.

As a consequence of this study, the identification of a spatio-temporal model for this area would probably be a way to better understand the behavior over time of this vulnerable zone. That will not be done here, once this application was developed only for a illustrative purpose of a spatial analysis.

3

Environmental biomonitoring in mainland Portugal

3.1 Introduction

Among the pollutants affecting the environment, heavy metals belong to the most serious ones. The international mapping project *Atmospheric Heavy Metal Deposition in Europe* is surveying the atmospheric deposition of heavy metals using moss species as biomonitors, with the aim of investigate the existence of correlations between heavy metal concentrations in mosses.

Mosses are widely used as biomonitors of atmospheric heavy metal deposition. In Europe, they have been used since 1990, with the aim of map spatial and temporal patterns of accumulation in ecosystems (Holy *et al.* (2009)). Some examples of studies related to the use of moss samples are Diggle *et al.* (2010) using moss data from Galicia, northern Spain, on the context of analyzing the effect of preferential sampling on prediction, but not taking into account the temporal representativeness of data, Harmens *et al.* (2010) considering data from several countries across Europe, Steinnes *et al.* (2003) and Steinnes *et al.* (2011) concerning Norway data, Zechmeister *et al.* (2008) with data from Austria.

Uyar *et al.* (2007) mentions several advantages of using moss samples: the vast geographical distribution and the abundant grow in various natural habitats; the non-existence of epidermis or cuticle enabling their cell walls to be easily penetrable for metal ions; as mosses have no root systems, they obtain minerals mainly from air and

precipitation; the effect and contamination of soil by heavy metals is negligible for most mosses, so they show the concentrations of the most metals correlated to the amount of atmospheric deposition. The simple procedure of sampling and of cheap chemical analysis also makes mosses especially suitable organisms for the purpose of monitoring. Originally the biomonitoring network for the mapping project above mentioned, was established in 1980 after a Swedish initiative. Since then, the number of participant countries have increased, reaching twenty-eight European countries and over 6,000 sampling sites in the 2005 survey. The responsibility for the coordination of the survey, since 2001, belongs to the International Cooperative Programme, Vegetation Programme Coordination Centre at the Centre for Ecology and Hydrology (CEH), UK (Harmens *et al.* (2010)).

Portugal was one of the participating countries in the *Atmospheric Heavy Metal Deposition in Europe* project, performing surveys every 5 years since the beginning of the project in 1990. Moss samples of species *Hypnum cupressiforme* and *Scleropodium touretti* were collected in three nationwide surveys across mainland Portugal, referred to as the 1992, 1996 and 2002 surveys, and in an additional survey restricted to the central part of mainland Portugal performed in 2006. Due to the fact that this last survey was restricted to the center part of mainland Portugal and at different locations from the previous, it was not included in what follows.

Although the number of sampling locations was not the same throughout the surveys, 146 of those were common to the first three surveys (Figure 3.1). Sampling locations were selected in order to be representative of background areas, collected in a $30 \times 30km$ grid, although near large urban or industrial areas the sampling design was intensified, being the grid adjusted to $10 \times 10km$.

Chemical analysis yielded concentration measurements of Cadmium (Cd), Chromium (Cr), Copper (Cu), Iron (Fe), Manganese (Mn), Nickel (Ni), Lead (Pb) and Zinc (Zn). Further details on the sampling and analysis procedure can be found in Figueira *et al.* (2002) or Martins *et al.* (2012).

As stated before, the sampling design was not uniform throughout the whole region under study, opposite to the examples presented in Boquete *et al.* (2011) and Steinnes *et al.* (2011). There are cases where sampling is intensified in subregions where a large gradiente of the measured variable is expected (Diggle *et al.* (2010)). In fact, one possible aim of an air monitoring network might be to identify larger values of pollution, so

sampling locations could be selected based on high values rather than randomly (Guttorp & Loperfido (2008)). Specifically, in the Portuguese case, sampling locations are in

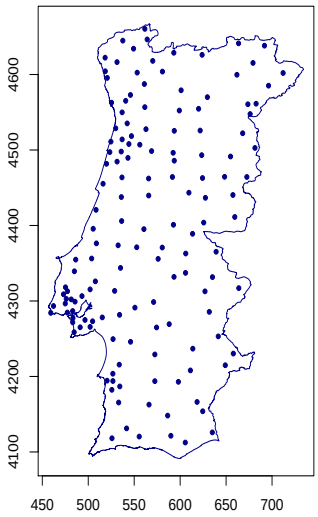

Figure 3.1: Map of Portugal showing the sampling locations shared among the three surveys

a larger number at regions with high industrial or urban density, such as in the region of Lisboa and Tejo valley, in the area between Porto and Aveiro, and near Sines oil refinery. This motivates the use of a function of the sampling intensity as explanatory variable when modelling air pollution data from Portugal.

The issue of using covariates in spatio-temporal models and testing for its significance is addressed in Díaz-Avalos *et al.* (2014) in the context of point processes. Particularly in the applications that follow, the inclusion of this variable will be obtained by a spatial kernel smoothing of the sampling locations density. As it should be expected, the larger values of this smooth function, ranging from 3.14 to 27.48 and with mean value equal to 11.62, occur in regions with higher sampling intensity (Fig. 3.2), where the industrial or urban areas are located.

Next, a description of the Mn and Pb data is presented.

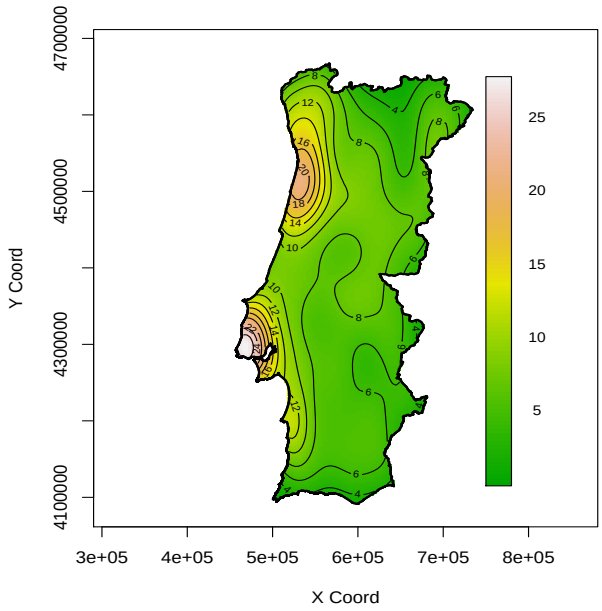

Figure 3.2: Map of sampling locations intensity

3.2 Exploratory data analysis

Among the various heavy metals found in nature, manganese is one of the most abundant and widely distributed, being found in waters, rocks and soils (Pinsino *et al.* (2012)). The presence of this metal has not only a natural cause as a result of mineral weathering and atmospheric deposition, but can also have anthropogenic origins, such as municipal wastewater discharges, mining and mineral processing, combustion of fossil fuels or emissions from the combustion of fuel additives (Howe *et al.* (2004)). Although essential for humans, at higher levels of contamination manganese can become toxic. Several studies relate chronic manganese excess with disturbances in the central nervous system, with symptoms resembling those of Parkinsons disease (Perl & Olanow (2007), Rocks & Levy (2008)).

Lead occurs naturally in the environment as well as in manufactured products. However, most lead concentrations that are found in the environment are a result of human activities. The major sources of lead emissions have historically been from fuels and industrial sources, such as mining and metal manufacturing, waste incinerators, battery

recycling, among others. Airborne lead can be deposited on soil and water, thus reaching humans via the food chain, causing several unwanted effects, such as anaemia, rise in blood pressure, disruption of nervous systems, or diminished intellectual capacity in children (Järup (2003)).

3.2.1 Distribution of the sample data

A preliminary descriptive analysis of both Mn and Pb concentration data for the three considered surveys, with values expressed in units of $mg(metal)/kg(moss)$, showed the presence of outlier values. A Box-Cox transformation of data, with parameter $\lambda = 0.15$ for Mn, and $\lambda = 0.06$ for Pb, was carried out, in order to reduce the effect that these values could cause in the estimation process.

Mn

The Mn concentrations in the original scale (Table 3.1) ranged from a minimum of $4.0mg/kg$, recorded in the second survey, and a maximum value of $970.0mg/kg$, which occurred in the first survey. One can observe that the mean value has increased from around $160mg/kg$, in the first survey to almost $180mg/kg$ in the second survey, and decreased to less than $150mg/kg$ in the third survey. A similar behavior occurred with the median, and one can also observe that the variability of Mn concentration values was always decreasing. It should be noticed the maximum recorded value for each survey, particularly in the first one when the maximum was about six times the mean value. This difference between the mean and the maximum value was not so marked in the second and third surveys, although being also strong. Histograms for the original data, for the Box-Cox transformed data and QQ-plots are in Figure 3.3, respectively on the first, second and third row. As a consequence of outlier values, the distribution of Mn data exhibits an asymmetry to the right. The first survey is strongly asymmetric, but over time the asymmetry becomes not so marked.
After performing the Box-Cox transformation of data, the asymmetry has faded away and the resulting distribution behaves like a Gaussian one (p-values in a Shapiro-Wilk test for normality of 43.46%, 19.94% and 7.98%, respectively for the first, second and third surveys.)

Table 3.1: Data summary of Mn concentration (observed and Box-Cox transformed values)

	Survey					
	1992		1996		2002	
	Observ.	Transf.	Observ.	Transf.	Observ.	Transf.
Min	16.00	3.39	4.03	1.54	23.14	3.96
Median	123.50	6.90	149.18	7.28	123.20	6.89
Max	970.00	11.63	685.55	10.73	503.11	9.97
Mean	161.62	6.88	178.62	7.13	147.12	6.85
St. dev.	147.97	1.69	136.01	1.65	99.28	1.41

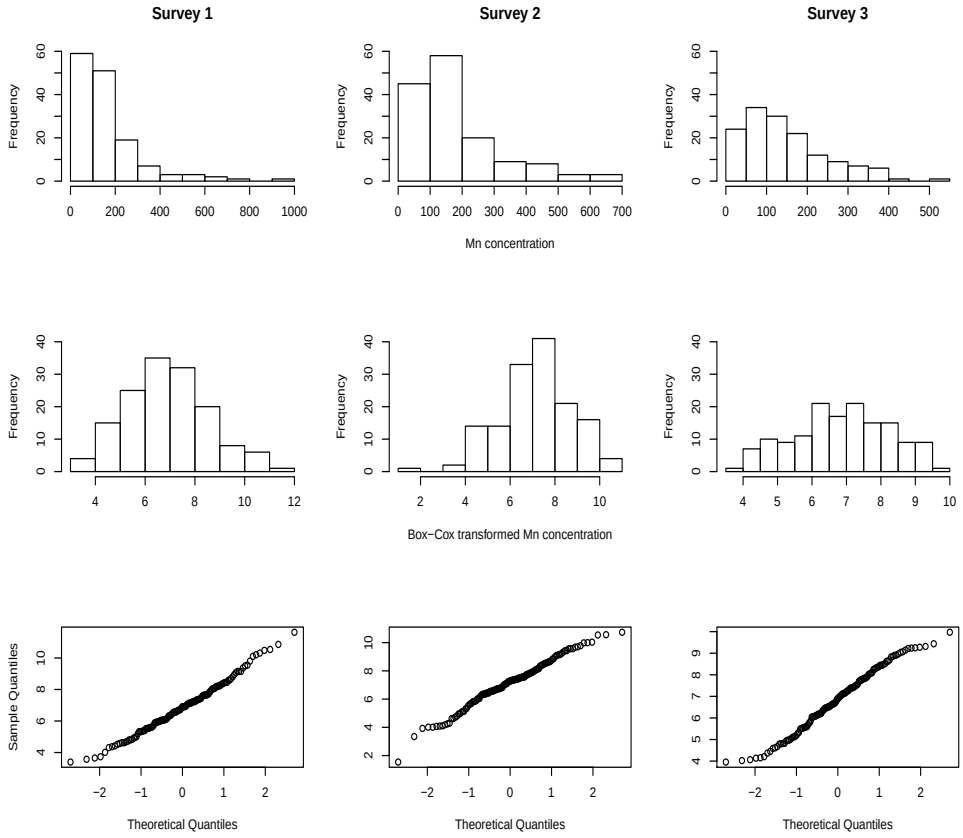

Figure 3.3: Histograms for the original (top row) and Box-Cox transformed Mn concentration (middle row), and QQ-plots for transformed data

Pb

In what respects the observed Pb concentrations (Table 3.2), the minimum value of $0.5mg/kg$ was recorded in the first survey and the maximum value, of over $191mg/kg$, was recorded in the second survey. For all the descriptive measures presented in Table 3.2, there was an increase from the first to the second survey and a broad decrease from the second to the third survey. This was an effect probably caused by the more frequent use of unleaded fuel, which was forced by Portuguese legislation at that time. The variability of Pb concentration values also shared the same pattern of decreasing values from the first to the third survey.

Table 3.2: Data summary of Pb concentration (observed and Box-Cox transformed values)

	Survey					
	1992		1996		2002	
	Observ.	Transf.	Observ.	Transf.	Observ.	Transf.
Min	0.50	-0.67	2.00	0.71	0.68	-0.38
Median	13.00	2.78	15.70	2.99	3.11	1.17
Max	172.00	6.04	191.17	6.18	109.93	5.44
Mean	16.40	2.75	21.72	3.03	5.40	1.29
St. dev.	16.91	0.86	24.05	0.92	10.39	0.90

Histograms for data on the original scale, for the Box-Cox transformed data and QQ-plots are in Figure 3.4, respectively on the first, second and third row. The presence of outlier observations is even more notorious than was observed for Mn data, resulting in right asymmetric distributions for the three surveys. The asymmetry, contrary to what has been registered for Mn, persisted for the three surveys. In fact, one half of the recorded measurements are under $3.11mg/kg$ for the third survey, while for the first and the second surveys are, respectively, under 13.00 and $15.70mg/kg$, indicating that the asymmetry is even more notorious in the last period of observations. The performed Box-Cox transformation of data mitigated the presence of outlier values, nevertheless only for the second survey the resulting distribution has the behavior of a Gaussian distribution (p-value of 5.23%). For this reason, in the application of the spatio-temporal model to be introduced next in Chapter 6, derived in order to perform

predictions of Gaussian data, the information about this heavy metal will not be considered.

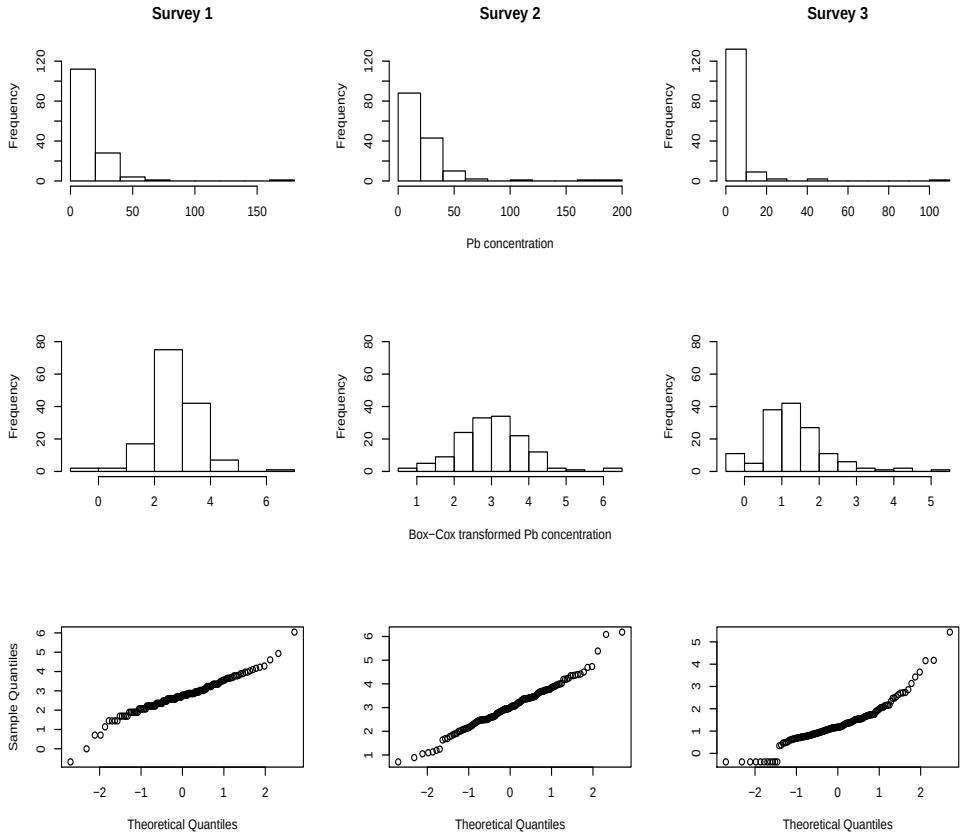

Figure 3.4: Histograms for the original (top row) and Box-Cox transformed Pb concentration (middle row), and QQ-plots for transformed data

3.2.2 Spatial behavior of data

To better understand the spatial behavior of data, transformed concentration values for both metals were predicted over a fine 300×100 grid covering mainland Portugal, allowing to have a grid point at a distance of around 2 km from each other, and interpolation errors were computed. The predicted concentration values were obtained by using Ordinary Kriging (OK), as defined in Section 2.5. By (2.24) and assuming

that the concentration $Z(\mathbf{s})$, at location $\mathbf{s}$, has constant but unknown expected value $E\big[Z(\mathbf{s})\big] = \mu$, the predicted concentration is given by

$$\widehat{Z}(\mathbf{s}_0) = \sum_{i=1}^{146} \lambda_i(\mathbf{s}_0) Z(\mathbf{s}_i) \tag{3.1}$$

where $\mathbf{s}_0$ is any location goal of prediction and λ_i, $i = 1, \ldots, 146$ are the Kriging weights chosen to satisfy the constraint $\sum_{i=1}^{146} \lambda_i = 1$.
With this aim, parametric exponencial models for the spatial dependence structure of $Z(\mathbf{s})$, whose parameters are detailed in Table 3.3, were adjusted to empirical variograms.

Table 3.3: Exponential model parameters for $Z(\mathbf{s})$, regarding Mn and Pb data and for each survey

		1992	1996	2002
Mn				
	$\widehat{\tau}^2$	0.84	1.45	0.90
	$\widehat{\sigma}^2$	2.35	1.53	1.17
	$\widehat{\phi}$	99106.80	75000.00	75000.00
Pb				
	$\widehat{\tau}^2$	0.06	0.56	0.39
	$\widehat{\sigma}^2$	0.73	0.33	0.48
	$\widehat{\phi}$	6471.80	18576.09	38615.67

Larger values for both the nugget effect, τ^2, and the partial sill, σ^2, are found for Mn, for all surveys, although the larger values of τ^2 are found for the second survey while the larger values for σ^2 are the ones related with the first survey. Also for the radius of influence ϕ, Mn data produce larger estimates than Pb data, being the larger value of almost $100km$ for Mn for the 1992 survey, and the minimum value of around $6.5km$ for Pb data also for the 1992 survey. Parameter estimates were obtained by maximum likelihood for Mn, and by ordinary least squares method for Pb, as for this metal only the second survey behaves according to a Gaussian distribution.

Mn

Table 3.4 show a summary of the predicted Mn transformed concentrations for each survey and the resulting prediction map (left) and interpolation error map (right). One

can find the larger range of predicted transformed values for the first survey, taking the value of over $5mg/kg$ and the lower range, of less than $4mg/kg$, for the third survey. The mean predicted value slightly increases from the first to the second survey and decreases in the third survey. The interpolation error is of similar magnitude for all surveys, although the second survey presents values slightly larger.

Table 3.4: Predicted Mn transformed concentration for each survey and associated interpolation error

	Survey					
	1992		1996		2002	
	Predicted	Error	Predicted	Error	Predicted	Error
Min	4.66	1.02	4.81	1.30	4.88	1.04
Median	7.31	1.36	7.44	1.54	6.99	1.26
Max	9.71	1.82	8.98	1.77	8.59	1.47
Mean	7.20	1.39	7.29	1.54	6.97	1.26
St. dev.	0.98	0.21	0.68	0.13	0.63	0.12

It is known that contamination by Mn is mainly associated with the soil typology (Figueira *et al.* (2002)). In accordance to this, for each survey, higher predicted

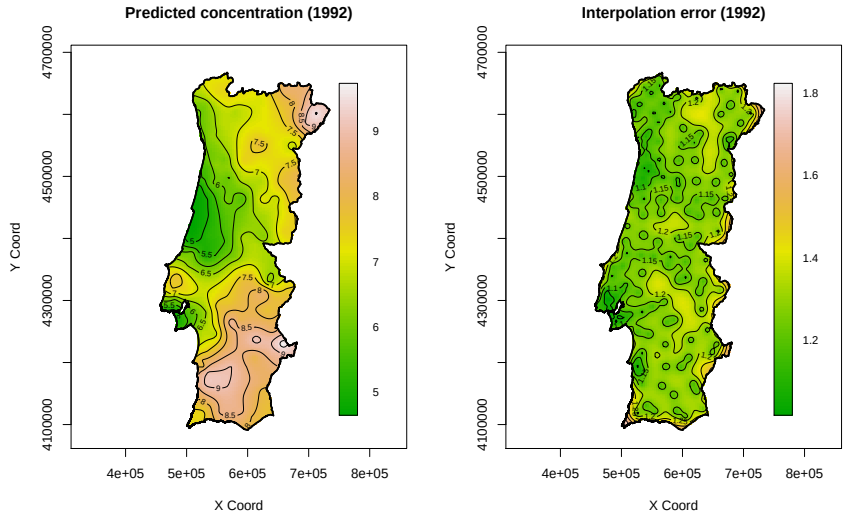

Figure 3.5: Predicted Mn transformed concentration map and the associated interpolation error map for the 1992 survey

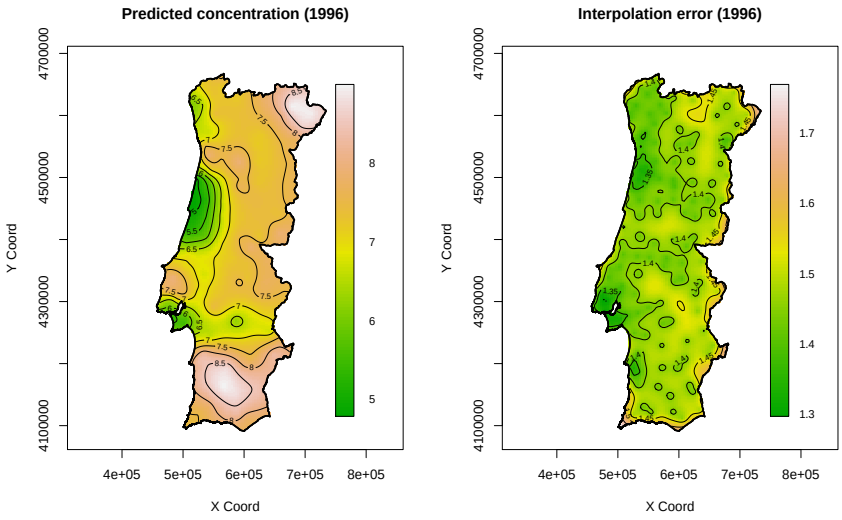

Figure 3.6: Predicted Mn transformed concentration map and the associated interpolation error map for the 1996 survey

values, marked with light pink color in Figures 3.5, 3.6 and 3.7, occur in eastern and

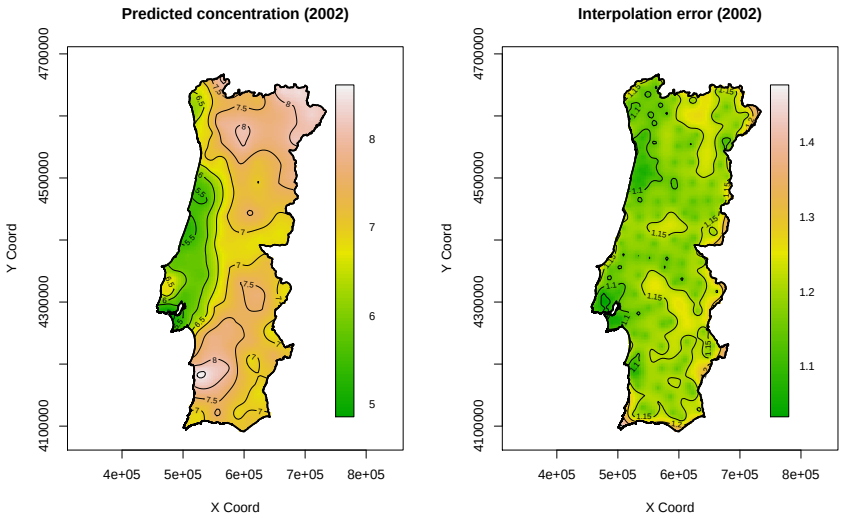

Figure 3.7: Predicted Mn transformed concentration map and the associated interpolation error map for the 2002 survey

south-western Portugal, regions with less forestry and hence with more soil erosion.

Pb

Regarding the predicted Pb concentration, the values obtained for each survey are summarized in Table 3.5. As was to be expected, the mean (and the median) value increases

Table 3.5: Predicted Pb transformed concentration for each survey and associated interpolation error

	Survey					
	1992		1996		2002	
	Predicted	Error	Predicted	Error	Predicted	Error
Min	0.01	0.37	2.12	0.84	0.19	0.70
Median	2.68	0.89	2.96	0.95	1.19	0.88
Max	5.52	0.89	4.23	0.95	3.01	0.94
Mean	2.68	0.87	2.99	0.93	1.20	0.86
St. dev.	0.17	0.04	0.18	0.02	0.28	0.07

from the first to the second survey, but for the third survey, the mean value is quite lower. Like for Mn, the largest range of predicted transformed values was registered in the first survey, being over $5.5 mg/kg$.

For each survey, the graphical representation of predicted values and associated interpolation errors are in Figures 3.8, 3.9 and 3.10.Contamination by this heavy metal is known to be mainly due to anthropogenic causes, particularly by emissions to ambient air resulting from fuel combustion (Järup (2003)), thus larger predicted values, marked with light pink color, are expected in areas where road traffic is more intense, near major cities (Lisboa and Porto) or more densely industrialized (near Sines oil refinery). This pattern is more notorious for the 2002 survey, where higher values are expected to occur as well in the north-eastern area, near an important borderland with Spain. It should be noticed the particular case of the interpolation errors map for the 1992 survey. From Table 3.3, it can be seen that the radius of influence for this survey is under $6.5 km$, about one third of the value for the second survey, and even less if comparing with the third survey. This is an indicatory sign for the pattern in the error map, which can be also checked in the second column of Table 3.5, where the median of errors equals the maximum error.

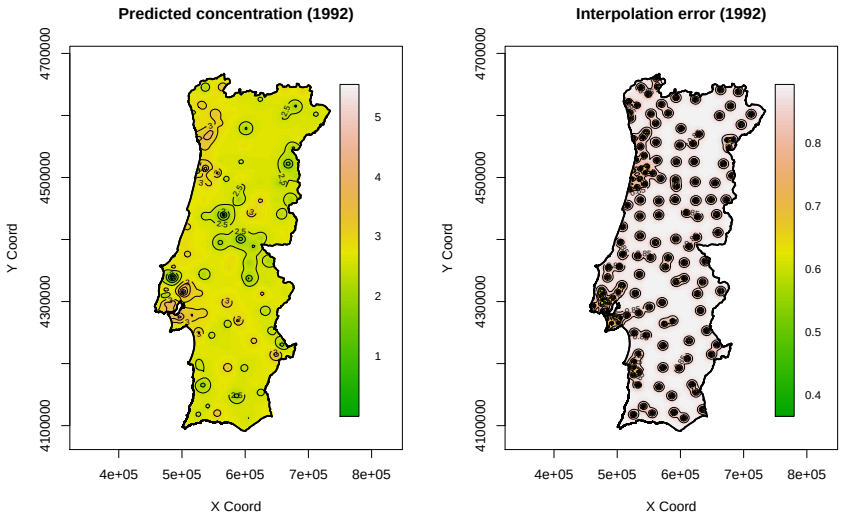

Figure 3.8: Predicted Pb transformed concentration map and the associated interpolation error map for the 1992 survey

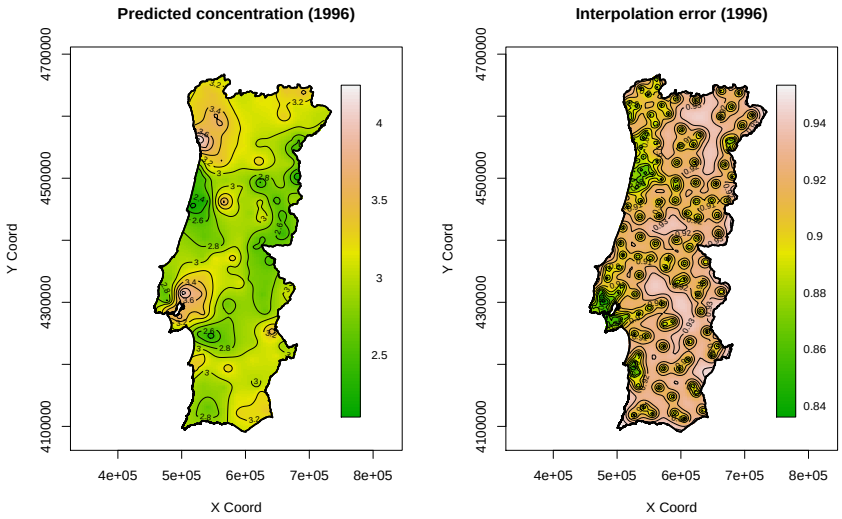

Figure 3.9: Predicted Pb transformed concentration map and the associated interpolation error map for the 1996 survey

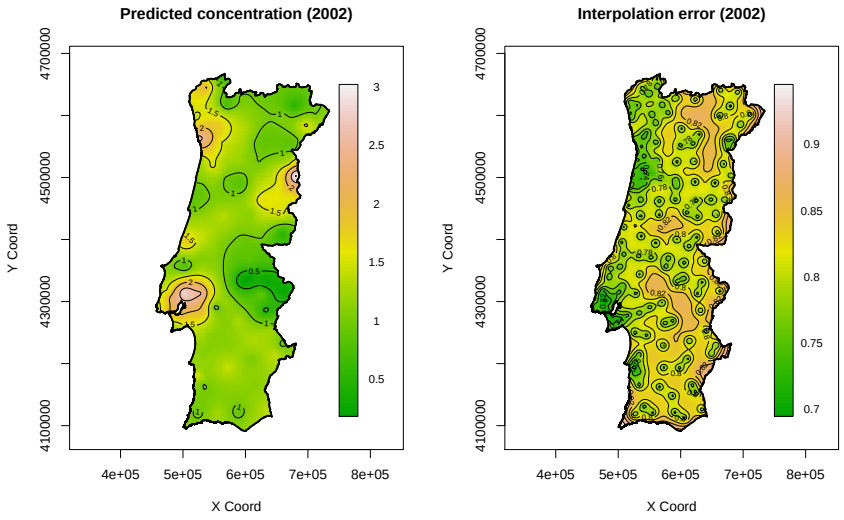

Figure 3.10: Predicted Pb transformed concentration map and the associated interpolation error map for the 2002 survey

3.3 Spatial prediction considering covariates

Once the aims of this work are centered in the spatio-temporal behavior of data, the predictions to be made in the following chapters will be restricted to the most recent survey. For that reason, in the present section predictions will also be made only for the 2002 survey.

The effect of considering country specific information on predicted concentration values and on interpolation error, is now analyzed. For the particular case of the Portuguese data here considered, several covariates were tested for significance, namely location coordinates themselves, as well as the intensity of sampling locations. Results showed that only the later covariate was significant.

The spatial analysis using Ordinary Kriging, presented in Section 3.2.2, may now be deepened. When considering the sampling location intensity as explanatory variable, the expected value $\mathrm{E}\big[Z(\mathbf{s})\big]$ is a function $\mu(\mathbf{s})$ of the observed location. This way, the interpolation procedure corresponds, according to Cressie (1993), to the Universal Kriging (UK), or Kriging with external trend as Goovaerts (1997) defines. Details on $\mu(\mathbf{s})$ modelling will be given latter in Chapter 5, when introducing an extension of an existing spatio-temporal model.

To proceed with the interpolation task, parametric models for the spatial dependence structure of $Z(\mathbf{s}) - \mu(\mathbf{s})$ were fitted to empirical variograms, both for Mn and Pb, whose parameters are detailed in Table 3.6. The nugget effect τ^2 and the partial sill σ^2 are

Table 3.6: Exponential model parameters for $Z(\mathbf{s}) - \mu(\mathbf{s})$, restricted for 2002 data

	$\widehat{\tau}^2$	$\widehat{\sigma}^2$	$\widehat{\phi}$
Mn	0.89	0.92	49999.90
Pb	0.33	0.44	26587.30

of similar magnitude for each metal, although the estimates for Pb are about half of the corresponding estimates for Mn. This same behavior is present for the radius of influence ϕ, about 50 km for Mn, but only around half of this value for Pb.
Figure 3.11 shows the empirical variogram with adjusted exponential covariance model.

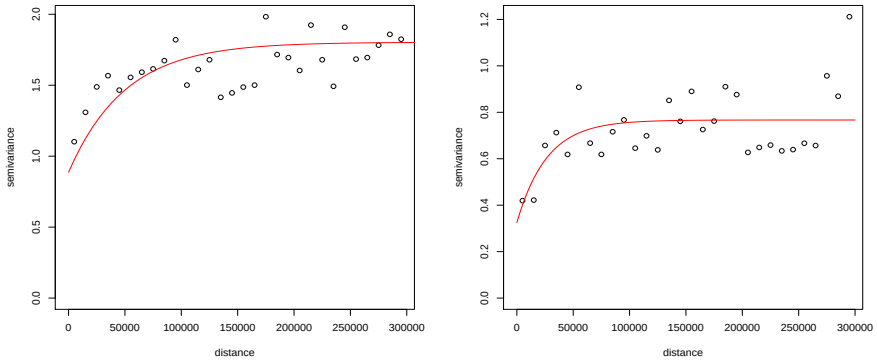

Figure 3.11: Empirical variogram with exponential fitted variogram, restricted for 2002 Mn (left) and Pb (right) transformed data, after removing the covariate information (given by the sampling location intensity)

The identification of these covariance models allowed to obtain the predicted transformed concentration for both metals, at unobserved locations placed over the same grid as in the previous application. The resulting predicted concentration values are summarized in Table 3.7.
A simple analysis of these values reveals a symmetric behavior of the predicted concentrations for the two metals under consideration. The mean and the median are of

Table 3.7: Predicted Mn and Pb transformed concentration considering covariates for the 2002 survey and associated interpolation error

	Mn		Pb	
	Predicted	Error	Predicted	Error
Min	4.39	1.03	0.13	0.66
Median	7.57	1.25	0.89	0.85
Max	8.58	1.37	3.25	0.89
Mean	7.40	1.24	0.99	0.83
St. dev.	0.63	0.09	0.33	0.05

similar magnitude in each metal. However, while for Mn they are closer to the maximum predicted value, suggesting a left asymmetric distribution, for Pb the results show the opposite.

Maps of predicted values, as well as the associated interpolation error, are in Figures 3.12 and 3.13. Like in the prediction maps obtained when not considering the covariate information, for Mn the higher predicted values are expected in the eastern and southern regions of mainland Portugal, while for Pb higher predicted values are identified to occur near major cities and near the borderline already mentioned.

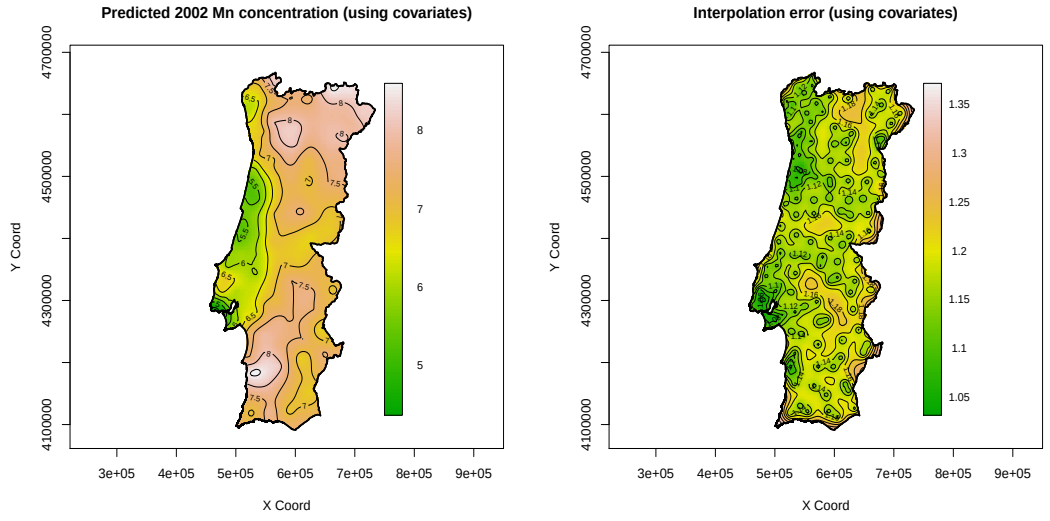

Figure 3.12: Predicted Mn transformed concentration map for the 2002 survey (left) and the associated interpolation error map (right)

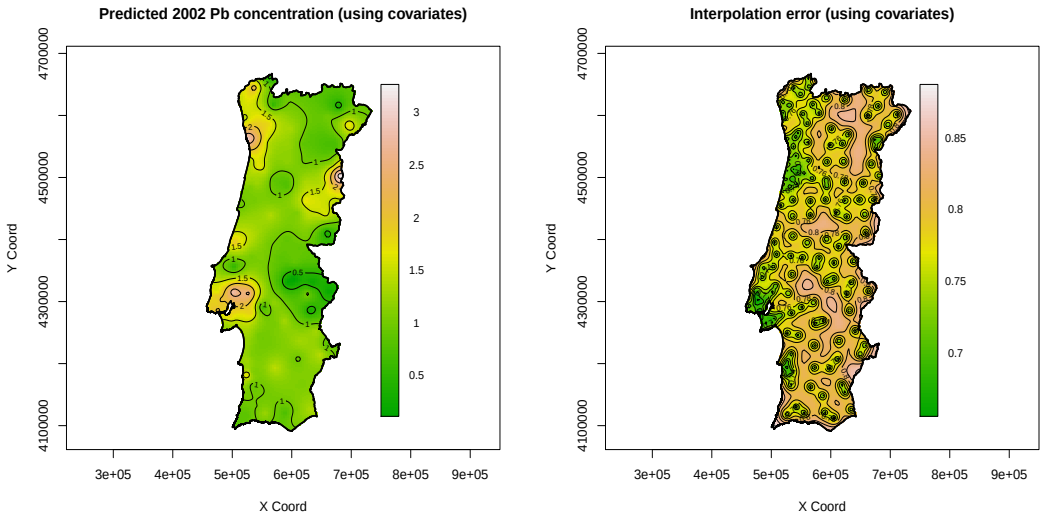

Figure 3.13: Predicted Pb transformed concentration map for the 2002 survey (left) and the associated interpolation error map (right)

3.4 Comparison of spatial prediction results

In the previous sections, predictions at unobserved locations were obtained for the 2002 survey, both for Mn and for Pb, first without considering the existence of explanatory covariates and, after, considering the possible existence of explanatory covariates for the process under observation. Particularly, the explanatory covariate considered was related with the sampling design used when collecting the data.

Comparing the prediction results obtained by the two processes for the most recent survey, summarized in the last columns of Tables 3.4 and 3.5, and in Table 3.7, one can register that the predicted Mn values are of larger magnitude when considering the covariate, while for Pb the larger values are encountered when not considering the covariate. This means that by incorporating this information in the prediction spatial model, the effect that would result in predicted values by ignoring a sampling design not evenly representative of the area under observation, was mitigated. In the application to the Portuguese moss data, where areas with more sampled locations are related to lower values of Mn, while for Pb the behavior is just opposite, the Mn concentration was being under-estimated and the Pb concentration was being over-estimated.

3. ENVIRONMENTAL BIOMONITORING IN MAINLAND PORTUGAL

4

Spatio-temporal geostatistics

4.1 Introduction

Initially, the main focus of geostatistics was in the modelling of variables which are distributed in space, named by Matheron as regionalized variables. However, more recently the focus turned on to the modelling of variables varying in both space and time. The idea behind this evolution is that when data are collected in locations within a spatial region and at varying times, the locations and the time themselves may help to explain the data variability.

The extension of spatial geostatistical techniques into the space-time domain is not straightforward. Difficulties do not arise from the fact that there is one more dimension to be incorporated in the model, but as a consequence that the spatial and the temporal dimensions are completely different. For example, temporal characteristics of the process, such as seasonality, are usually known. However, in space such a periodic behavior is not possible to consider.

A realistic geostatistical spatio-temporal model should be able to take into account the inherent differences between variation in space and in time, being this done through the covariance function.

4.2 Spatio-temporal geostatistics

Generalizing from the spatial framework introduced in section 2.2, let us consider a spatio-temporal random field

$$\big\{Z(\mathbf{s},t),\, \mathbf{s}\in D\,,\, t\in T\big\} \tag{4.1}$$

where $\mathbf{s}$ are locations within the observation region $D\subset \mathbb{R}^d$, observed at times $t\in T$. Typically, $d=1,2$ or 3, and often T is a subset of the positive integers, $T\subset \mathbb{Z}^+$. The spatio-temporal process (4.1) is generally characterized by its cumulative distribution function

$$F_{\mathbf{s}_1,\ldots,\mathbf{s}_n,t_1,\ldots,t_m}(z_1,\ldots,z_{nm}) = P\big(Z(\mathbf{s}_1,t_1)\le z_1,\ldots,Z(\mathbf{s}_n,t_m)\le z_{nm}\big) \tag{4.2}$$

Moments

The moment of order k of the random field $Z(\mathbf{s},t)$, defined at any spatio-temporal location $(\mathbf{s},t)\in D\times T$ is

$$\mathrm{E}\big[\big(Z(\mathbf{s},t)\big)^k\big] = \int x^k dF_{\mathbf{S},T}(x) \tag{4.3}$$

provided this integral exists.

Expected value

The expected value of a random field $Z(\mathbf{s},t)$ is defined to be the order-one moment,

$$\mu(\mathbf{s},t) = \mathrm{E}\big[Z(\mathbf{s},t)\big] \tag{4.4}$$

for any spatio-temporal location $(\mathbf{s},t)\in D\times T$.

Variance and Covariance

The variance of a random field $Z(\mathbf{s},t)$ is defined as the second-order moment about the expected value $\mu(\mathbf{s},t)$,

$$\mathrm{Var}\big[Z(\mathbf{s},t)\big] = \mathrm{E}\big[\big(Z(\mathbf{s},t)-\mu(\mathbf{s},t)\big)^2\big] \tag{4.5}$$

for any location $(\mathbf{s}, t) \in D \times T$. The covariance is defined by

$$\mathrm{Cov}_{ST}\big(Z(\mathbf{s}_i, t_k), Z(\mathbf{s}_j, t_l)\big) = \mathrm{E}\big[\big(Z(\mathbf{s}_i, t_k) - \mu(\mathbf{s}_i, t_k)\big)\big(Z(\mathbf{s}_j - t_l) - \mu(\mathbf{s}_j, t_l)\big)\big] \quad (4.6)$$

for any spatio-temporal locations $(\mathbf{s}_i, t_k)$ and $(\mathbf{s}_j, t_l)$ in $D \times T$.

Strict stationarity

One random field $Z(\mathbf{s}, t)$ is said to be **strictly stationary** if its spatio-temporal cumulative distribution function is invariant by any translation $(\mathbf{h}_S, h_T) \in \mathbb{R}^d \times \mathbb{R}$, that is,

$$F_{\mathbf{s}_1+\mathbf{h}_S,\cdots,\mathbf{s}_n+\mathbf{h}_S,t_1+h_T,\ldots t_m+h_T}(z_1, \cdots, z_{nm}) = F_{\mathbf{s}_1,\ldots,\mathbf{s}_n,t_1,\ldots,t_m}(z_1, \cdots, z_{nm}) \quad (4.7)$$

Second-order stationarity

One random field $Z(\mathbf{s}, t)$ is said to be **second-order stationary** if its moments up to order two exist, and are such that

- the mean function is modeled as a constant,

$$E\big[Z(\mathbf{s}, t)\big] = \mu, \forall (\mathbf{s}, t) \in D \times T$$

- the space-time covariance function depends only on the spatial and temporal lags $\mathbf{h}_S = \mathbf{s}_j - \mathbf{s}_i$ and $h_T = t_l - t_k$,

$$\mathrm{Cov}_{ST}\big(Z(\mathbf{s}_i, t_k), Z(\mathbf{s}_j, t_l)\big) = \mathrm{C}_{ST}(\mathbf{h}_S, h_T)$$

The function $\mathrm{C}_{ST}(.,.)$ is usually known as the spatio-temporal covariogram.

Intrinsic stationarity

As the second-order stationarity assumption may not be met in practical applications, a weaker version, of intrinsic stationarity, is also available for spatio-temporal random

fields.
$Z(\mathbf{s}, t)$ is said to be **intrinsically stationary** if the expected value and the variance of the increments $Z(\mathbf{s}, t) - Z(\mathbf{s} + \mathbf{h}_S, t + h_T)$ exist and are such that

- $\mathrm{E}\big[Z(\mathbf{s}, t) - Z(\mathbf{s} + \mathbf{h}_S, t + h_T)\big] = 0$
- $\mathrm{Var}\big[Z(\mathbf{s}, t) - Z(\mathbf{s} + \mathbf{h}_S, t + h_T)\big] = 2\gamma(\mathbf{h}_S, h_T)$

where the function $2\gamma(.,.)$ is the spatio-temporal variogram function.

The correlogram

The correlogram, $\rho_{ST}(.,.)$ is, like in the spatial framework, the standardized version of the covariogram,

$$\mathrm{Corr}_{ST}\big(Z(\mathbf{s}_i, t_k), Z(\mathbf{s}_j, t_l)\big) = \frac{\mathrm{Cov}_{ST}\big(Z(\mathbf{s}_i, t_k), Z(\mathbf{s}_j, t_l)\big)}{\sqrt{\mathrm{C}_{ST}(\mathbf{0}, 0)}} = \rho_{ST}(\mathbf{s}_i - \mathbf{s}_j, t_l - t_k) \quad (4.8)$$

Relationship between the covariogram, the variogram and the correlogram

In a similar way as in the purely spatial setting,

$$\begin{aligned} \gamma(\mathbf{h}_S, h_T) &= \mathrm{C}_{ST}(\mathbf{0}, 0) - \mathrm{C}_{ST}(\mathbf{h}_S, h_T) \\ \rho(\mathbf{h}_S, h_T) &= 1 - \frac{\gamma(\mathbf{h}_S, h_T)}{\mathrm{C}_{ST}(\mathbf{0}, 0)} \end{aligned} \quad (4.9)$$

Separability and full symmetry

According to Gneiting *et al.* (2007), although prediction in the space-time context only requires the appropriate specification of the covariance structure, simplifying conditions of stationarity, separability, and full symmetry are needed for estimation and modelling. The need for these simplifying conditions results from the fact that observations are made on the joint spatio-temporal process, not on one spatial and one temporal separate processes. Thus, a separable formulation of the covariance structure as the product of

a purely spatial component by a purely temporal one, allows for a computationally efficient way to proceed with the inference and estimation task. Namely, the covariance matrix can be expressed as the Kronecker product of two smaller dimension matrices which arise from the purely spatial and the purely temporal components, turning its determinant and inverse more easily computed. As a consequence of the definition of separability, separable covariance models are being used even in applications where they are not physically justifiable (Cressie & Huang (1999)).

The random field $Z(\mathbf{s}, t)$ is said to have separable covariance structure if there exist purely spatial and purely temporal covariance functions, Cov_S and Cov_T, such that

$$\forall(\mathbf{s}_i, t_k), (\mathbf{s}_j, t_l) \in D \times T, \mathrm{Cov}_{ST}\big(Z(\mathbf{s}_i, t_k), Z(\mathbf{s}_j, t_l)\big) = \mathrm{Cov}_S(\mathbf{s}_i, \mathbf{s}_j) \times \mathrm{Cov}_T(t_k, t_l) \quad (4.10)$$

Furthermore if, for all spatio-temporal locations $(\mathbf{s}_i, t_k), (\mathbf{s}_j, t_l) \in D \times T$, the covariance function is such that

$$\mathrm{Cov}\big(Z(\mathbf{s}_i, t_k), Z(\mathbf{s}_j, t_l)\big) = \mathrm{Cov}\big(Z(\mathbf{s}_i, t_l), Z(\mathbf{s}_j, t_k)\big) \quad (4.11)$$

the space-time process $Z(\mathbf{s}, t)$ is said to have fully symmetric covariance structure.

It can be shown that a separable covariance must be fully symmetric, but fully symmetric covariances may not be separable ones.

To assess the appropriateness of separability of a spatio-temporal covariance model, several procedures are available. Fuentes (2006), considering a spectral interpretation of separability and only one realization of the spatio-temporal process, reduces the test for separability to a simple two-way ANOVA procedure. Mitchell *et al.* (2006) propose a likelihood ratio test of separability in the context of multivariate repeated measures, decomposing the covariance matrix as a Kronecker product. Crujeiras *et al.* (2010) propose a nonparametric test based on the additivity of the log-spectrum considered as a regression function. Scaccia & Martin (2005) propose a spectral method to test, first, axial symmetry, and then, if this hypothesis is not rejected, to test separability for spatial lattice processes. Applications of these tests of separability, in each one of the mentioned works, require repeated measures or a sufficiently large number of time observations.

Although the use of separable models is computationally desirable, the assumption of separability is not always easily justifiable. This difficulty has led to the development of nonseparable models, allowing for a wider range of space-time correlation structures

to be considered. Some examples are Cressie & Huang (1999), de Iaco *et al.* (2002), Bruno *et al.* (2003), Kolovos *et al.* (2004).

4.3 Spatio-temporal covariance models

In a similar way to the purely spatial setting, in order to a covariance function to be a permissible one, it must be nonnegative definite, which is equivalent to say that the (semi)variogram function needs to be nonpositive definite.
Although being possible to propose a spatio-temporal model and test for its permissibility, as will be proposed latter on, in practical applications it is often the case to choose one model among a set of models that are known to be permissible. A large number of spatio-temporal covariance models, both separable and non-separable, have been proposed in the literature. See, *e.g.*, Kyriakidis & Journel (1999), Sherman (2011), Cressie & Huang (1999), de Iaco (2010), Stein (2005), or Gneiting (2002).
The next short list of spatio-temporal covariance models is, by no means, exhaustive. In fact, the choice of the models to mention at this stage was related to the particular application to perform later in Chapter 6.

Metric model

The metric spatio-temporal covariance model (Dimitrakopoulos & Luo (1994)) is given by

$$\mathrm{C}_{ST}(\mathbf{h}_S, h_T) = \mathrm{C}\big(a_1\|\mathbf{h}_S\|^2 + a_2|h_t|^2\big) \tag{4.12}$$

where the coefficients a_1, $a_2 \in \mathbb{R}$ enable the comparison between space and time. The metric model assumes the same type of covariance structure for the spatial and temporal covariances, with possible changes in the range, which makes this a restrictive model. As this model can be thought of as a spatial covariance model with an extra temporal dimension, the permissible spatial models already known are available for use in (4.12). In terms of the spatio-temporal variogram, the metric model is represented, by (4.9)

and (2.10), as

$$\gamma_{ST}(\mathbf{h}_S, h_T) = \gamma\big(a_1\|\mathbf{h}_S\|^2 + a_2|h_t|^2\big) \tag{4.13}$$

where $\gamma(.)$ is an isotropic spatial variogram.

Product model

The product (or separable) space-time covariance model (Rodriguez-Iturbe & Mejia (1974), de Cesare *et al.* (2001)) is given by

$$\mathrm{C}_{ST}(\mathbf{h}_S, h_T) = k \cdot \mathrm{C}_S(\mathbf{h}_S) \cdot \mathrm{C}_T(h_t) \tag{4.14}$$

where $k \in \mathbb{R}$, and C_S and C_T are admissible spatial and temporal covariance models, which can be combined in product form to give spatio-temporal covariance models. This model separates the spatial dependence from the temporal one. The parameter k is computed using (4.14) by setting both $\mathbf{h}_S$ and h_T equal to zero,

$$k = \frac{\mathrm{C}_{ST}(\mathbf{0}, 0)}{\mathrm{C}_S(\mathbf{0}) \cdot \mathrm{C}_T(0)}$$

Opposite to the metric model, the product model allows for different spatial and temporal covariance structures.
This model can be proposed in terms of the spatio-temporal variogram $\gamma_{ST}(\mathbf{h}_S, h_T)$ as

$$\gamma_{ST}(\mathbf{h}_S, h_T) = \mathrm{C}_T(0)\gamma_S(\mathbf{h}_S) + \mathrm{C}_S(\mathbf{0})\gamma_T(h_T) - \gamma_S(\mathbf{h}_S)\gamma_T(h_T) \tag{4.15}$$

being γ_S and γ_T, respectively, spatial and temporal variograms, and C_S and C_T, respectively, spatial and temporal covariances.
Casal (2003) considers, as a particular case of this model, the separable exponential semivariogram

$$\gamma_{ST}(\mathbf{h}_S, h_T) = \begin{cases} \tau^2 + \sigma^2\Big(1 - exp\Big\{ - \dfrac{h_T}{a} - \dfrac{\mathbf{h}_S}{b}\Big\}\Big) & , \mathbf{h}_S \neq 0 \vee h_T \neq 0 \\ 0 & , \mathbf{h}_S = \mathbf{0} \wedge h_T = 0 \end{cases} \tag{4.16}$$

with $\tau^2 \geq 0$ the nugget effect, $a \geq 0$ a temporal scale parameter, $b \geq 0$ a spatial scale parameter and $\sigma^2 > 0$ the partial sill.

Product-sum model

The product-sum space-time covariance model is an extension of the product model (4.14), including an additional component involving the product of the spatial and the temporal covariance models. The product-sum model (de Cesare *et al.* (2001)) is given by

$$\mathrm{C}_{ST}(\mathbf{h}_S, h_T) = k_1 \cdot \mathrm{C}_S(\mathbf{h}_S) \cdot \mathrm{C}_T(h_t) + k_2 \cdot \mathrm{C}_S(\mathbf{h}_S) + k_3 \cdot \mathrm{C}_T(h_t) \tag{4.17}$$

where $k_1 > 0$, $k_2 \geq 0$ and $k_3 \geq 0$, for the model to be permissible. The product-sum model allows for the specification of different types of covariance models for the spatial and temporal components and, also, provides a mechanism for the interaction of the space and time components, thereby offering more flexibility than the metric or the product models (Denham (2012)).
As in the product model, also the product-sum model can be rewritten in terms of the spatio-temporal variogram,

$$\gamma_{ST}(\mathbf{h}_S, h_T) = \big[k_2 + k_1\mathrm{C}_T(0)\big]\gamma_S(\mathbf{h}_S) + \big[k_3 + k_1\mathrm{C}_S(\mathbf{0})\big]\gamma_T(h_T) - k_1\gamma_S(\mathbf{h}_S)\gamma_T(h_T) \tag{4.18}$$

4.4 Spatio-temporal parameter estimation and prediction

To estimate the parameters of the covariogram, similar methods to spatial correlation estimation may be applied here. The semi-variogram for a stationary spatio-temporal

process is

$$\gamma_{ST}(\mathbf{h}_S, h_T) = \frac{1}{2}\mathrm{E}\big[\big(Z(\mathbf{s},t) - Z(\mathbf{s}+\mathbf{h}_S, t+h_T)\big)^2\big] \tag{4.19}$$

The corresponding empirical spatio-temporal semi-variogram estimator is the primary tool for inference and is defined by

$$\widehat{\gamma}_{ST}(\mathbf{h}_S, h_T) = \frac{1}{2|N(\mathbf{h}_S, h_T)|}\sum_{N(\mathbf{h}_S,h_T)}\big(Z(\mathbf{s}_i,t_j) - Z(\mathbf{s}_i+\mathbf{h}_S, t_j+h_T)\big)^2 \tag{4.20}$$

where the set $N(\mathbf{h}_S, h_T)$ consists of the points that are within spatial distance $\mathbf{h}_S$ and time lag h_T of each other.

Once having identified a spatio-temporal covariance model, one can estimate the continuous process under observation at an unsampled space-time location $(\mathbf{s}, t)$. Generalizing the spatial setting (2.23) introduced in Section 2.5, the spatio-temporal kriging technique produces estimates over a weighted linear combination of a subset of the available data $\{z(\mathbf{s}_i, t_j),\ i = 1,\ldots,N,\ j = 1,\ldots,T\}$, or more specifically, a subset of the residuals $\{z(\mathbf{s}_i, t_j) - \mu(\mathbf{s}, t),\ i = 1,\ldots,N,\ j = 1,\ldots,T\}$, which are dependent on the specification of the mean function. he data to be considered in such a weighted linear combination are selected according to the spatial and temporal distance from the estimation datum, and the weights are computed by taking into account the proximity of each observation to the prediction location.

Spatio-temporal kriging has the same principle of interpolation as the spatial kriging has, that is, its a BLUE method.

As in the spatial setting, the objective of the estimator is to minimize the error variance under the constraint of unbiasedness,

$$\begin{aligned} &min\ \sigma^2(\mathbf{s},t) = \mathrm{Var}\big[\widehat{Z}(\mathbf{s},t) - Z(\mathbf{s},t)\big] \\ &\text{subject to} \\ &\mathrm{E}\big[\widehat{Z}(\mathbf{s},t) - Z(\mathbf{s},t)\big] = 0 \end{aligned} \tag{4.21}$$

where the estimator $\widehat{Z}(\mathbf{s}, t)$ is defined by

$$\widehat{Z}(\mathbf{s},t) - \mu(\mathbf{s},t) = \sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t)\big(Z(\mathbf{s}_i,t_i) - \mu(\mathbf{s}_i,t_i)\big) \tag{4.22}$$

which varies depending on the chosen covariance model adjusted to the empirical semi-variogram.

Typically the spatio-temporal random field $Z(\mathbf{s},t)$ is decomposed as the sum of a trend component $\mu(\mathbf{s},t)$ and a stationary, zero mean random field residual component $R(\mathbf{s},t)$, with covariance function $\mathrm{C}_R(\mathbf{h}_S, h_T)$,

$$Z(\mathbf{s},t) = \mu(\mathbf{s},t) + R(\mathbf{s},t) \tag{4.23}$$

and such that $\mathrm{E}\big[Z(\mathbf{s},t)\big] = \mu(\mathbf{s},t)$.

Simple Kriging

The Simple Kriging technique assumes the mean component known and constant, $\mu(\mathbf{s},t) = \mu,\ \forall(\mathbf{s},t) \in D \times T$, meaning that this technique does not adapt to local trends. This allows to rewrite the estimator (4.22) as

$$\begin{aligned} \widehat{Z}(\mathbf{s},t) &= \sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t)\big(Z(\mathbf{s}_i,t_i) - \mu\big) + \mu \\ &= \sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t) Z(\mathbf{s}_i,t_i) + \mu \cdot \Big[1 - \sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t) Z(\mathbf{s}_i,t_i)\Big] \end{aligned} \tag{4.24}$$

The weights are computed in order to satisfy the minimization problem (4.21), using the system of equations given by

$$\sum_{j=1}^{n(\mathbf{s},t)} \lambda_j(\mathbf{s},t)\, \mathrm{C}_R(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j) = \mathrm{C}_R(\mathbf{s}_i - \mathbf{s}, t_i - t),\ i = 1,\ldots,n(\mathbf{s},t) \tag{4.25}$$

considering the residual covariance between observations, and the residual covariance between observations and the location goal of prediction. The resulting minimum error

variance is

$$\sigma^2 = \mathrm{C}_R(\mathbf{0}) - \sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t)\,\mathrm{C}_R(\mathbf{s}_i - \mathbf{s}, t_i - t) \tag{4.26}$$

Ordinary Kriging

Ordinary Kriging accounts for cases where the mean component is unknown but constant, being this component estimated simultaneously with the residual component. Similarly to the Simple Kriging, the estimator (4.22) can be written as

$$\begin{aligned} \widehat{Z}(\mathbf{s},t) &= \sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t)\big(Z(\mathbf{s}_i,t_i) - \mu(\mathbf{s},t)\big) + \mu(\mathbf{s},t) \\ &= \sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t) Z(\mathbf{s}_i,t_i) + \mu(\mathbf{s},t)\cdot\Big[1 - \sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t) Z(\mathbf{s}_i,t_i)\Big] \end{aligned} \tag{4.27}$$

where the weights are forced to sum to 1. Thus, (4.27) is equivalent to

$$\widehat{Z}(\mathbf{s},t) = \sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t) Z(\mathbf{s}_i,t_i) \;\text{ with }\; \sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t) = 1 \tag{4.28}$$

The system of equations to solve the minimization problem (4.22) is

$$\begin{cases} \displaystyle\sum_{j=1}^{n(\mathbf{s},t)} \lambda_j(\mathbf{s},t)\,\mathrm{C}_R(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j) + \mu_{OK}(\mathbf{s},t) = \mathrm{C}_R(\mathbf{s}_i - \mathbf{s}, t_i - t),\ i = 1,\ldots,n(\mathbf{s},t) \\ \displaystyle\sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t) = 1 \end{cases} \tag{4.29}$$

where μ_{OK} is a Lagrange parameter that accounts for the constraint on the weights.

The resulting minimum error variance is

$$\sigma^2 = \mathrm{C}_R(\mathbf{0}) - \sum_{i=1}^{n(\mathbf{s},t)} \lambda_i(\mathbf{s},t)\, \mathrm{C}_R(\mathbf{s}_i - \mathbf{s}, t_i - t) - \mu_{OK}(\mathbf{s},t) \tag{4.30}$$

4.5 One particular spatio-temporal model

The previous sections presented particularities in the modelling of spatio-temporal processes. Many models have been proposed in the literature, however the interest in choosing a particular one focuses on its behavior in the prediction task.
This section describes one of those spatio-temporal models, introduced in Høst *et al.* (1995). Authors proposed a framework to accurately represent interpolation errors when data are available from repeated observations of monitoring networks. This framework takes into account that the use of different interpolation methods may provide similar predicted values, but the inherent interpolation errors may be not comparable. With this model description, the main goal is to assess the gain achieved in the predictive accuracy at unsampled locations, when comparing with predictions based only on the spatial information.

4.5.1 The model

The mentioned model states that the response variable value $Z(\mathbf{s},t)$ at location $\mathbf{s}$ and time t can be written as

$$Z(\mathbf{s},t) = \mu(\mathbf{s},t) + \omega(\mathbf{s},t)\varepsilon(\mathbf{s},t) \tag{4.31}$$

where $\mu(\mathbf{s},t)$, $\omega(\mathbf{s},t)$ and $\varepsilon(\mathbf{s},t)$ are space-time mutually independent random fields representing mean, scale and residuals. This general formulation, for practical purposes, is usually decomposed between spatial and temporal effects in the following way:

$$\mu(\mathbf{s},t) = M_1(\mathbf{s}) + m_2(t), \tag{4.32}$$

the sum of a purely spatial component $M_1(\mathbf{s})$, modelling the spatial mean variation, and $m_2(t)$, a temporal modulation at discrete times, which corresponds to consider an additive separability in the mean field $\mu(\mathbf{s}, t)$. Moreover, one may assume multiplicative separability in the scale field,

$$\omega(\mathbf{s}, t) = S_1(\mathbf{s}) s_2(t), \tag{4.33}$$

that is, rewrite $\omega(\mathbf{s}, t)$ as the product of a purely spatial $S_1(\mathbf{s})$ component and a purely temporal $s_2(t)$ component. In practical terms, $M_1(\mathbf{s})$ represents a mean random effect in space and $m_2(t)$ the associated time correction, being approximately equal to zero. In a similar way $S_1(\mathbf{s})$ represents a scale random effect in space and $s_2(t)$ the associated time correction being approximately one. The random component $\varepsilon(\mathbf{s}, t)$ identifies the remaining space-time interactions not captured by the foregoing components.
This model aims to be simple, appealing and, mainly, it may cover a wide range of practical situations, where it is reasonable to assume time and space separability.
$M_1(\mathbf{s})$ and $S_1(\mathbf{s})$ are considered as realizations of second-order stationary random fields, with the first-order moments of fields M_1 and S_1 such that $\mathrm{E}\big[M_1(\mathbf{s})\big] = \mu$ and $\mathrm{E}\big[S_1(\mathbf{s})\big] = \nu$, and furthermore $\mathrm{E}\big[\varepsilon(\mathbf{s}, t)\big] = 0$ and $\mathrm{Var}\big[\varepsilon(\mathbf{s}, t)\big] = 1$.
It is of interest to note that, once the space-time mean random field is mainly decomposed by a spatial mean component, added with a time correction, this model becomes suitable for cases where observations are in a larger number in space than in time.

4.5.2 Variance of predictions

In Høst *et al.* (1995), the authors also propose how to compute the interpolation error, defined as the interpolation standard deviation, at an unmonitored location s_0 and a monitored time t_i,

$$\begin{aligned}\mathrm{Var}\big[Z(\mathbf{s}_0, t_i) - \widehat{Z}(\mathbf{s}_0, t_i)\big] = \;& \mathrm{Var}\big[M_1(\mathbf{s}_0) - \widehat{M_1}(\mathbf{s}_0)\big] + \\ & s_2^2(t_i)\, \nu^2\, \mathrm{Var}\big[\varepsilon(\mathbf{s}_0, t_i) - \widehat{\varepsilon}(\mathbf{s}_0, t_i)\big] + \\ & s_2^2(t_i)\sigma_{S_1}^2 \Big[1 - \mathrm{Corr}\big(S_1(\mathbf{s}_0), \widehat{S}_1(\mathbf{s}_0)\big)\mathrm{Corr}\big(\varepsilon(\mathbf{s}_0, t_i), \widehat{\varepsilon}(\mathbf{s}_0, t_i)\big)\Big]\end{aligned} \tag{4.34}$$

denoting $\sigma_{S_1}^2$ the variance of the scale field, and, from here, some special cases depending on the particular mean, scale and residual fields obtained from the data. Among these,

authors refer the cases where

(a) there is a known spatial mean field and a constant space scale field,

$$\mathrm{Var}\big[Z(\mathbf{s}_0,t_i)-\widehat{Z}(\mathbf{s}_0,t_i)\big]=s_2^2(t_i)\,\nu^2\,\mathrm{Var}\big[\varepsilon(\mathbf{s}_0,t_i)-\widehat{\varepsilon}(\mathbf{s}_0,t_i)\big]$$

or

(b) there is no structure in the spatial scale field,

$$\begin{aligned}\mathrm{Var}\big[Z(\mathbf{s}_0,t_i)-\widehat{Z}(\mathbf{s}_0,t_i)\big] &= \mathrm{Var}\big[M_1(\mathbf{s}_0)-\widehat{M}_1(\mathbf{s}_0)\big]\\ &+ s_2^2(t_i)\big(\nu^2+\sigma_{S_1}^2\big)\mathrm{Var}\big[\varepsilon(\mathbf{s}_0,t_i)-\widehat{\varepsilon}(\mathbf{s}_0,t_i)\big]\end{aligned}$$

or

(c) the scale field variance is negligible,

$$\begin{aligned}\mathrm{Var}\big[Z(\mathbf{s}_0,t_i)-\widehat{Z}(\mathbf{s}_0,t_i)\big] &= \mathrm{Var}\big[M_1(\mathbf{s}_0)-\widehat{M}_1(\mathbf{s}_0)\big]\\ &+ s_2^2(t_i)\nu^2\mathrm{Var}\big[\varepsilon(\mathbf{s}_0,t_i)-\widehat{\varepsilon}(\mathbf{s}_0,t_i)\big]\end{aligned}$$

or

(d) the scale field variance and residual field variance are negligible,

$$\mathrm{Var}\big[Z(\mathbf{s}_0,t_i)-\widehat{Z}(\mathbf{s}_0,t_i)\big]=\mathrm{Var}\big[M_1(\mathbf{s}_0)-\widehat{M}_1(\mathbf{s}_0)\big] \tag{4.35}$$

The predicted value $\widehat{Z}(\mathbf{s}_0,t_i)$ at an unsampled location $\mathbf{s}_0$ and an observed time $t_i,\ i=1,\ldots,T$ according to (4.31), is

$$\widehat{Z}(\mathbf{s}_0,t_i)=\widehat{M}_1(\mathbf{s}_0)+\widehat{m}_2(t_i)+\widehat{S}_1(\mathbf{s}_0)\widehat{s}_2(t_i)\widehat{\varepsilon}(\mathbf{s}_0,t_i) \tag{4.36}$$

where $\widehat{M}_1(\mathbf{s}_0)=\sum_{j=1}^{n}\lambda_j\overline{M}_1(\mathbf{s}_j)$, being $\overline{M}_1(\mathbf{s}_j)$ the mean value of the temporal observations collected at location $\mathbf{s}_j$ and λ_j the Kriging weights associated to the mean field; $\widehat{m}_2(t_i)=\sum_{j=1}^{n}\lambda_j\left(Z(\mathbf{s}_j,t_i)-\overline{M}_1(\mathbf{s}_j)\right)$; $\widehat{S}_1(\mathbf{s}_0)=\sum_{j=1}^{n}\theta_j\sqrt{S^2(\mathbf{s}_j)}$, being $S^2(\mathbf{s}_j)$ the estimated variance at each sampled location $\mathbf{s}_j$, and θ_j the Kriging weights associated to the scale field; and, at last, $\widehat{\varepsilon}(\mathbf{s}_0,t_i)=\sum_{j=1}^{n}\alpha_{ji}\,\varepsilon(\mathbf{s}_j,t_i)$, being α_{ji} the Kriging weights associated to the residuals field. Stated another way: $\widehat{M}_1(\mathbf{s}_0)$, $\widehat{S}_1(\mathbf{s}_0)$ and $\widehat{\varepsilon}(\mathbf{s}_0,t_i)$ represent the Ordinary Kriging estimates at $\mathbf{s}_0$ for fields M_1, S_1 and ε; $m_2(t_i)$ and $s_2(t_i)$ represent the corresponding time corrections at time t_i.

4.6 Spatio-temporal prediction of manganese and lead data

Section 4.5 introduces one spatio-temporal model, allowing to make predictions at unobserved locations and taking into account the temporal information of the random field under study.
This model will now be considered to obtain the prediction map, as well as the interpolation error map, of manganese and lead concentration for the most recent survey.

4.6.1 Inference on model components

To obtain predicted concentration values, each component of $Z(\mathbf{s}, t)$ described in (4.31), was estimated according to the details presented in Høst *et al.* (1995). For that, as required by the characterization of this particular model, the independence of model components was checked by inspection of pairwise scatterplots of estimated M_1, S_1 and ε fields. Additionally, it is reasonable to assume separability in space and time in many practical cases. As mentioned in Cressie & Huang (1999), *separable models are often chosen for convenience rather than for their ability to fit the data well.*
For each of the mean, scale and residuals fields, parametric covariance models were fitted to empirical variograms (Figure 4.1). In Table 4.1, we can find the resulting parameters estimates. The estimated values for both the nugget effect and the partial sill in the scale field are the lowest. As expected, the smallest radius of influence ϕ, is found in the residuals field, pointing to a smaller spatial correlation. The estimated values for the nugget effect in the mean field and the scale field for Mn are about two times the correspondent ones for Pb, but the residuals field does not share this

Table 4.1: Exponential model parameters estimates for fields M_1 (mean), S_1 (scale) and ε (residuals)

	Mn			Pb		
	$\widehat{\tau}^2$	$\widehat{\sigma}^2$	$\widehat{\phi}$	$\widehat{\tau}^2$	$\widehat{\sigma}^2$	$\widehat{\phi}$
Mean	0.69	1.32	87469.40	0.30	0.18	50000.00
Scale	0.15	0.04	50000.00	0.08	0.05	50000.00
Residuals	0.74	0.30	15000.00	0.85	< 0.01	822.80

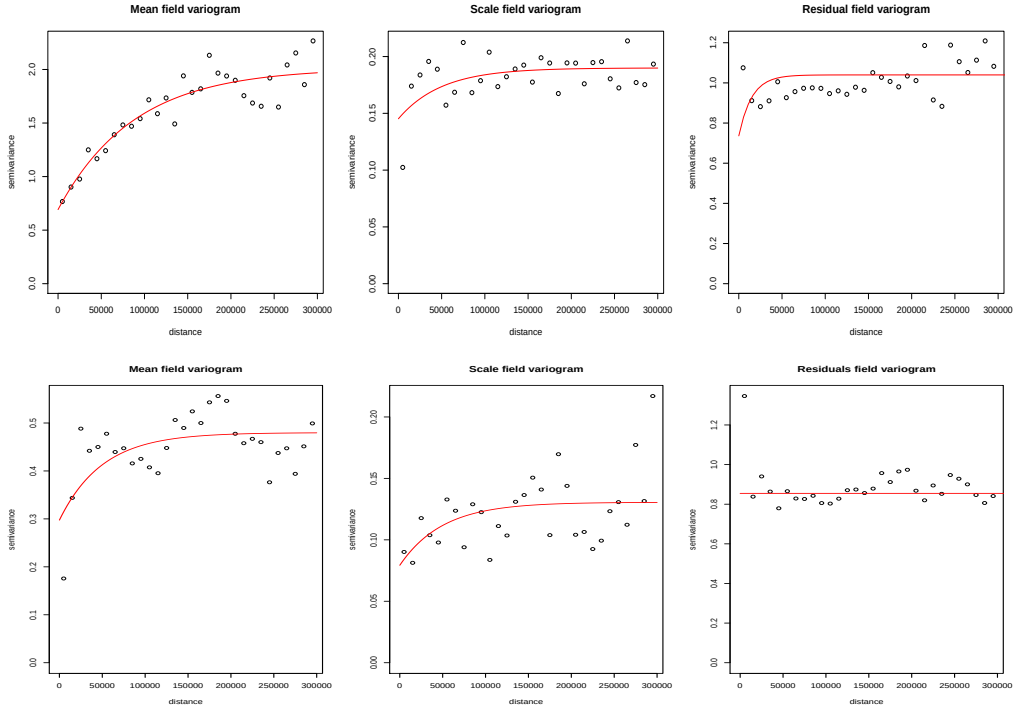

Figure 4.1: Empirical mean (left), scale (center) and residuals (right) field variogram with exponential parametric model, for Mn (top panel) and Pb (bottom panel)

same pattern. For σ^2, the scale field presents similar values for both metals, but for the mean field and the residuals field the estimated values for Mn are higher than for Pb. It should be mentioned an approximately null value for the partial sill σ^2 for the residuals field of Pb, leading to an almost flat semivariogram (Figure 4.1, bottom right panel).

The nugget effect τ^2 can be attributed to measurement errors, or to small scale spatial variability. Comparing estimates of this parameter for Mn, either the mean field and the residuals field show values more than four times larger than the scale field, suggesting more variability for the former fields. In case of Pb, the difference in values of small scale variability is even more notorious, where the residuals field presents a value of τ^2 ten times larger than the scale field. Regarding the radius of influence, in the mean field of Mn data are correlated up to a distance of almost $90km$, whereas at the residuals field, autocorrelation is detected up to $15km$. For Pb, it is also in the case of the residuals field that the shortest distance, of less than $1km$, for autocorrelated data is detected.

From here, estimated mean, scale and residual field maps (Figures 4.2 and 4.3) were obtained over the prediction grid, based on the Ordinary Kriging approach as explained before. Regarding the mean field prediction map for Mn and Pb, we can find a behavior as in the prediction maps in Figures 3.12 and 3.13, revealing higher predicted Mn values in regions with more soil erosion, and higher predicted Pb values in regions with more urban or industrial intensity. Once the variogram for the Pb residuals field suggested a pure nugget effect model, consequence of the nearly null estimative for σ^2, it was expected to obtain approximately constant values for this field. This pattern was not observed for Mn, once the partial sill σ^2 in the residuals field presents a much higher value than for Pb, leading to more variability in estimated values.

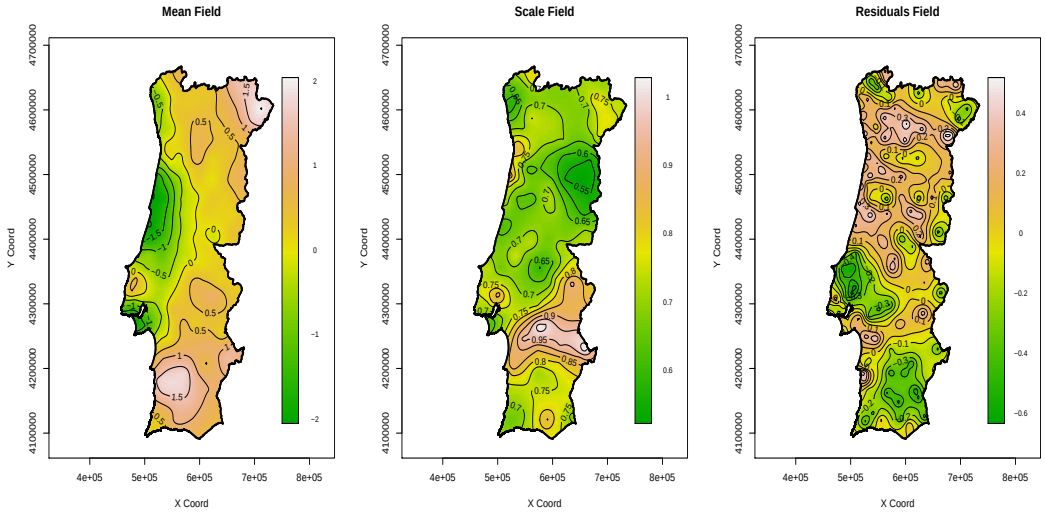

Figure 4.2: Estimated mean (left), scale (center) and residuals (right) maps for Mn

4.6.2 Spatio-temporal prediction

The identification of the components of model (4.31) allowed to compute the predicted transformed concentration values for the 2002 survey, which are summarized in Table 4.2. Comparing results with the ones obtained previously (right most columns in Tables (3.4) and (3.5)), one can find similarity between predicted values by means of a spatial model or a spatio-temporal model. It is worthwhile noticing that the spatio-temporal

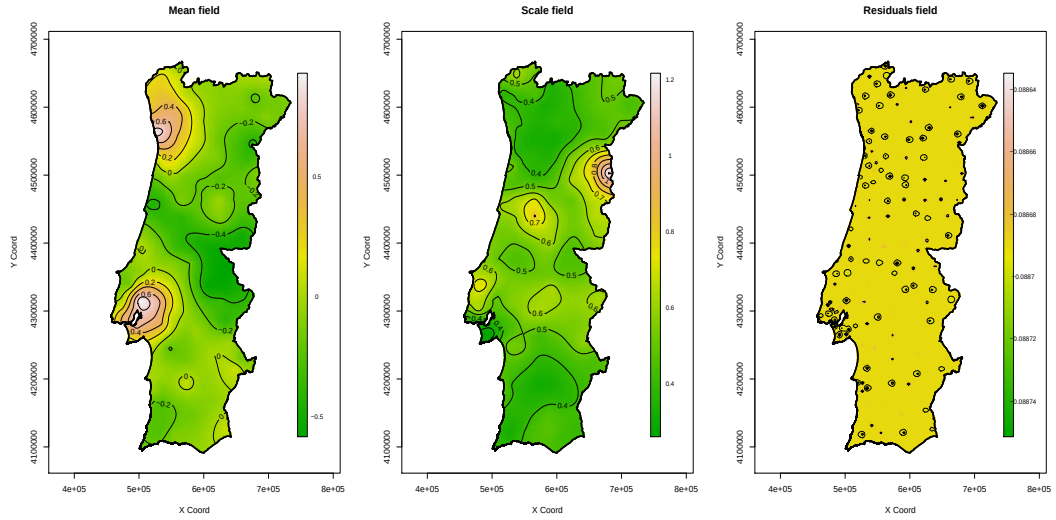

Figure 4.3: Estimated mean (left), scale (center) and residuals (right) maps for Pb

model provides predicted transformed concentrations with lower amount of interpolation error than the spatial one. A detailed interpretation of these error values will be given latter in Chapter 5, when comparing results from a full set of prediction models.

Table 4.2: Predicted transformed concentration values obtained according to model 4.31 for the 2002 survey and associated interpolation error

	Mn		Pb	
	Predicted	Error	Predicted	Error
Min	4.87	0.92	0.67	0.59
Median	7.21	1.16	1.15	0.66
Max	8.72	1.45	2.16	0.70
Mean	7.03	1.17	1.16	0.66
St. dev.	0.73	0.14	0.19	0.03

The spatial pattern identified previously in Figures 3.7 and 3.10 is also captured by this spatio-temporal model: higher predicted values in eastern territory for Mn, and near major cities or industrialized regions for Pb.

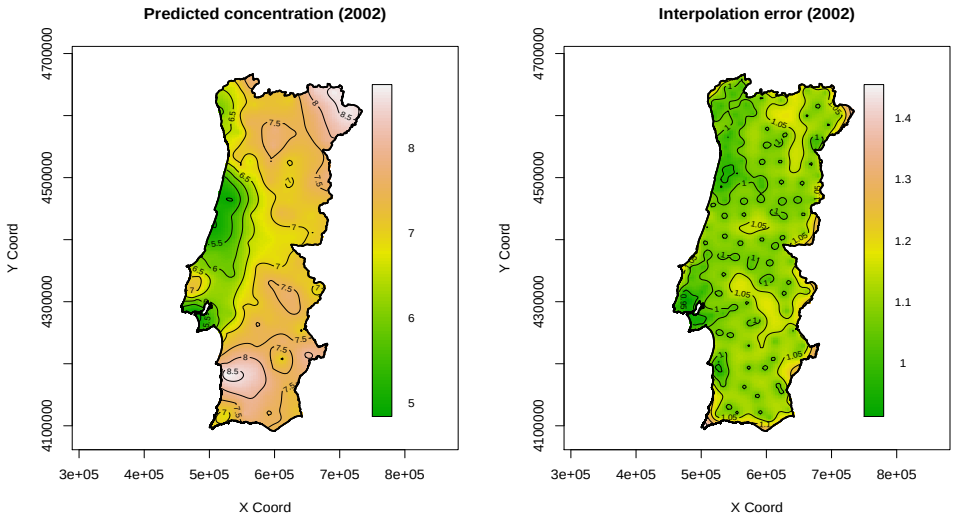

Figure 4.4: Predicted Mn concentration map for the 2002 survey (left) and the associated interpolation error map (right)

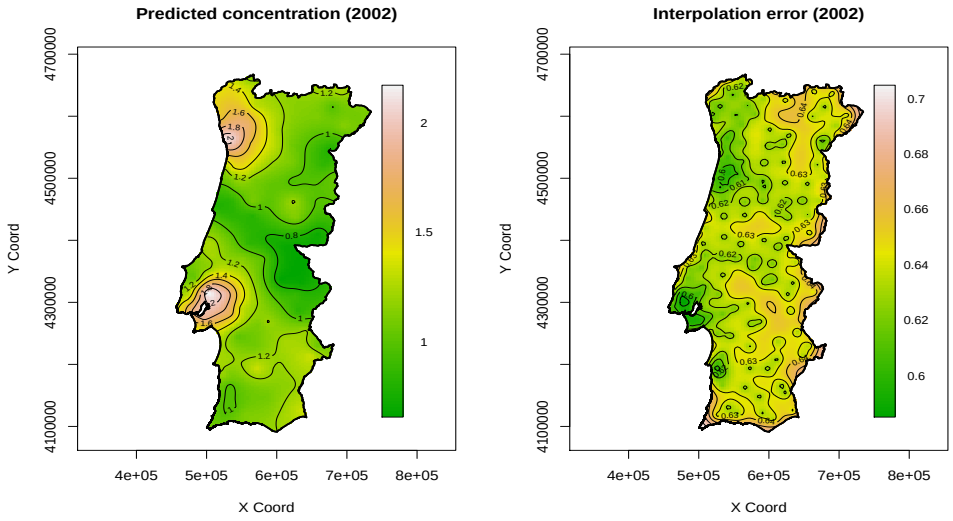

Figure 4.5: Predicted Pb concentration map for the 2002 survey (left) and the associated interpolation error map (right)

5

Extension of Høst model

5.1 Introduction

Preferential sampling refers to situations for which the sampling design process is stochastically dependent of the spatial process. In such a situation, the geostatistical modelling of the process under observation should take into account the information about the sampling design (Diggle *et al.* (2010)).

Previously, in Chapter 3, the monitoring network concerning the Portuguese contribution to the project *Atmospheric Heavy Metal Deposition in Europe* was introduced. The network, approximately kept the same along the three surveys, was not equally distributed over the observation region, as more data were collected close to industrialized areas. This is revealing of the presence of preferentially sampled locations or, stated another way, means that the sampling design doesn't conform to a *complete spatial randomness*, according to Diggle (2003). This is an example of a feature specific of this country's surveys which should be considered when modelling data pollution.

The non randomness of sampling site selection is an issue still raising debate, originating several publications on the subject. Menezes *et al.* (2008) suggest a kernel correction dependent on the neighborhood density for the specific case of the variogram estimation. Brus & de Gruijter (1997) present a comparison on some basic notions and terminology of the design-based and the model-based approaches for spatial prediction. Bruno *et al.* (2013) develop a conceptual design-based framework which emphasizes the use of geographical information on population sites of the location where a value has to be predicted from all other locations in the population. Shaddick & Zidek (2012)

continue the work of Diggle *et al.* (2010) on preferential sampling, extending it to the spatio-temporal domain. Gelfand *et al.* (2012) state that major differences can be found in spatial prediction if using preferentially chosen or random sampled locations. The issue of the choice of the sampling design is also addressed in Mateu & Müller (2012), where a comprehensive state-of-the-art is presented for network designing and planning for spatial and spatio-temporal data acquisition, giving some detail on the choice of a particular design criterion reflecting the purpose of the study, and mentioning environmental application examples of spatio-temporal monitoring network design.

5.2 Extension of Høst model

Margalho *et al.* (2014) propose an extension of model (4.31)

$$\begin{aligned} Z(\mathbf{s},t) &= \mu(\mathbf{s},t) + \omega(\mathbf{s},t)\cdot\varepsilon(\mathbf{s},t) \\ &= M_1(\mathbf{s}) + m_2(t) + S_1(\mathbf{s})\cdot s_2(t)\cdot\varepsilon(\mathbf{s},t) \end{aligned} \tag{5.1}$$

which consists in the expansion of the spatial component $M_1(\mathbf{s})$ of the mean field $\mu(\mathbf{s},t)$, as a linear combination of spatial effects defined by a suitable number p of functions f,

$$M_1(\mathbf{s}) = \sum_{i=1}^{p} \beta_i f_i(\mathbf{s}) \tag{5.2}$$

This extension is intended as a generalized regression model in the spatio-temporal context, with regression coefficients β_i, $i = 1,\ldots,p$, allowing to take into account the existence of country specific relevant covariates explaining the survey process, for instance the longitude or the latitude of sampling sites, or some index of industrialization. The effect of the inclusion of these environmental covariates, occurring through the spatial trend component, can be estimated in terms of the corresponding regression coefficients β_i. Note that by (5.2), the use of smoothing functions in the linear predictor is enabled. When proceeding with prediction at an unobserved location and monitored time, the estimation of each component of the proposed extension should occur as described for equation (4.36). This approach relies on the Kriging interpolation technique, which is known to provide linear and unbiased estimators.

5.3 Simulation study

With the purpose of validating the results of prediction obtained by means of the proposed extension, we proceed with a simulation study where different prediction models are applied to simulated data.

Two scenarios are considered, next identified as *reduced number of times* and *large number of times*. First, in a context similar to that of the real data set described in Chapter 3, where observations are available from three surveys, we simulate three data sets, one per each survey, represented by $n = 100$ locations in the square $[0, 10] \times [0, 10] \subset \mathbb{R}^2$. Second, using $n = 70$ locations, the number of data sets was augmented to thirty six, allowing for the presence of a seasonal behavior. The latter scenario illustrates the situation where one has monthly surveys happening along a total of three years.

For each scenario, a total of four models are under comparison. Two of such models correspond to perform Ordinary Kriging for the latest survey data, not considering explanatory covariates information, and to Universal Kriging, considering explanatory covariates. The other two models are the one given in (4.31) and its proposed extension (5.1)including covariates. From now on, these four models are referred to as: *(i)* spatial model without covariates, *(ii)* spatial model with covariates, *(iii)* spatio-temporal model without covariates and *(iv)* spatio-temporal model with covariates. The inclusion of such covariates is performed according to the generalized additive model (5.2) for the mean of the process.

Initially, we started by generating one stationary gaussian random field $Z(\mathbf{s})$, with expected value equal to 5. The covariance function was assumed to be given by a spherical or an exponential model with partial sill equal to 2.25, a range equal to 4, and assuming the nugget effect τ^2 null or equal to 0.1.

The sampling locations were elected aiming to reproduce the situation in which sample data are collected where higher values of pollution are expected to be found. At this stage, for each combination of the theoretical covariance model and nugget value, one has a sample data set identifying the latest survey.

The location goal of prediction was defined as the central point of the observation region, that is, the point with coordinates (5,5), for the most recent survey.

Reduced number of times

Having one data set, two other data sets were generated according to an auto-regressive model of order 1, AR(1), which corresponds to a total of three surveys happening at the same locations. Hence, the application of spatio-temporal models was enabled, as we have observed values at 100 spatial points along three time points.
For each model given above from *(i)* to *(iv)*, a total of 100 independent replicates were generated, each of which returning an estimate of $Z(\mathbf{s})$, if the model is restricted to space, or $Z(\mathbf{s}, t)$, if the model involves space-time data, for the last survey, at the spatial point goal of prediction. The covariate included in this simulation study was the intensity of the sampling design. For the spatio-temporal models, the estimated value $Z(\mathbf{s}_0, t)$, where $\mathbf{s}_0 = (5, 5)$ and $t = 3$, was determined according to the description given in (4.36).
As a measure of accuracy, the absolute prediction error (APE)

$$\text{APE} = |Z(\mathbf{s}_0) - \widehat{Z}(\mathbf{s}_0)| \quad \text{or} \quad \text{APE} = |Z(\mathbf{s}_0, t) - \widehat{Z}(\mathbf{s}_0, t)| \tag{5.3}$$

the absolute difference between the observed and the predicted value, was computed. As a result, a set of sixteen values of the mean absolute prediction errors (MAPE) became available, once four prediction models, two covariance functions and two nugget values were considered. This measure of accuracy is one of the scoring functions mentioned in Gneiting (2011), which are suggested to be used to evaluate point forecasts.
Table 5.1 presents the results from this simulation study. The prediction models under comparison including the considered covariate exhibit lower values for MAPE, when compared with the models not including the covariate.
For each covariance function and for each value of τ^2, **bold** identifies the minimum MAPE value, which is attained by the spatio-temporal model with covariates. Moreover, the lowest values of the standard deviation of APE are also attained by the same model. This is an indicator outcome of the gain, in terms of accuracy of predictions, that results by including in the prediction model not only the historical information of the process, but also the information provided by the given covariate in use.

Table 5.1: Mean (and standard deviation) APE in the simulation study with a reduced number of times, based on 100 replicates

		Spatial Model		Spatio-temporal Model	
	τ^2	without covariate	with covariate	without covariate	with covariate
Sph	0	0.84 (0.67)	0.71 (0.52)	0.85 (0.67)	**0.70** (0.51)
	0.1	0.92 (0.71)	0.82 (0.62)	0.94 (0.71)	**0.80** (0.60)
Exp	0	0.61 (0.46)	0.52 (0.37)	0.62 (0.47)	**0.51** (0.36)
	0.1	0.63 (0.40)	0.55 (0.34)	0.64 (0.39)	**0.54** (0.33)

Large number of times

Under this second scenario, the goal was again to obtain predicted values for the last survey at the spatial location $\mathbf{s}_0 = (5, 5)$, considering the same four prediction models as previously.

For each location $\mathbf{s}$, selected after generating a stationary random field in the same manner as previously, a time series of length 36 was generated according to the process

$$Z(\mathbf{s}, t) = 0.2 \cos\left(\left(\frac{\pi}{6}\right) t\right) + 0.95 Z(\mathbf{s}, t-1) + \varepsilon_t$$

with $\varepsilon_t \sim N(0, 0.7)$.

The process to obtain the predicted value by each model under comparison was, for the possible combinations of covariance function and nugget value, as described before.

When analyzing the values of mean and standard deviation APE from the simulation study (Table 5.2), one can register a better performance if an exponential model for the covariance structure is considered, except in case of the spatio-temporal model including covariates with a non-null nugget effect, for which the model with a spherical covariance structure provided the lowest value of MAPE. These results follow the same direction as in the *reduced number of times* scenario, that is, the prediction model including not only the temporal information but also an explanatory variable leads to the lowest values of MAPE and corresponding standard deviation.

Table 5.2: Mean (and standard deviation) APE in the simulation study with a large number of times, based on 100 replicates

		Spatial Model		Spatio-temporal Model	
		without	with	without	with
	τ^2	covariate	covariate	covariate	covariate
Sph	0	1.31 (1.10)	1.15 (1.01)	1.94 (1.38)	**0.99** (1.00)
	0.1	1.13 (0.74)	0.98 (0.67)	1.44 (0.96)	**0.60** (0.46)
Exp	0	0.90 (0.69)	0.81 (0.63)	1.33 (0.99)	**0.73** (0.49)
	0.1	0.94 (0.64)	0.84 (0.58)	1.38 (0.88)	**0.78** (0.52)

5.4 Spatio-temporal prediction of manganese and lead data considering covariates

5.4.1 Inference on model components

With the aim of obtain interpolated concentrations at non-observed locations for the most recent survey, spatio-temporal Mn and Pb data was complemented with country specific information. The procedure adopted here was similar to the procedure adopted in the application developed in Chapter 4, but including the selected covariate through the spatial component $M_1(\mathbf{s})$ of the mean spatio-temporal random field $\mu(\mathbf{s}, t)$.
The inclusion of this covariate was done by a generalized additive model as in (5.2),

$$M_1(\mathbf{s}) = \beta_0 + \beta_1 f_1(\mathbf{s})$$

being $f_1(\mathbf{s})$ a spline of the covariate sampling intensity. Table 5.3 presents the related regression coefficients which, for each metal, were significant. The coefficients for Mn are larger than for Pb. Therefore, the baseline coefficient β_0 is capturing this different scale of values. The effect of the covariate in the response variable is negative for Mn, which suggests an under-estimation when not considering this covariate and for Pb, the results show the opposite behavior.

Table 5.3: Regression coefficients associated with the covariate sampling intensity

	Mn			Pb		
	Estimate	St. error	p-value	Estimate	St. error	p-value
$\widehat{\beta}_0$	8.12	0.26	4×10^{-66}	2.17	0.14	5×10^{-31}
$\widehat{\beta}_1$	-0.11	0.02	2×10^{-7}	0.05	0.01	1×10^{-5}

In Table 5.4 one finds estimates for the parameters of covariance models fitted to the mean, the scale and the residuals fields. The graphical representation of these empirical variograms with the respective fitted parametric covariance models is omitted, once it is equivalent to the one in Figure 4.1. We expect to obtain more precise estimates,

Table 5.4: Exponential model parameters estimates for fields M_1, S_1 and ε, when considering covariates (spherical model parameters estimates for ε field in case of Pb)

	Mn			Pb		
	Mean	Scale	Residuals	Mean	Scale	Residuals
$\widehat{\tau}^2$	0.50	0.15	0.81	0.32	0.04	0.91
$\widehat{\sigma}^2$	0.97	0.03	0.24	0.10	0.06	0.01
$\widehat{\phi}$	50000.00	50000.00	15000.00	50000.00	25000.00	5000.00

as the influence of a non-democratic sampling design was removed with the inclusion of the covariate. Here we can verify, as was noticed previously in Table 4.1, that the smallest radius of influence ϕ, occurs for the residuals field in both metals. For Pb, the estimated partial sill σ^2 has a value almost null in the residuals field, suggesting a pure nugget effect model. Regarding the nugget value τ^2, the behavior follows the same pattern as before, the mean field and the scale field with higher values for Mn, residuals field with lower value for Mn.

The selected covariance models allowed for the estimation of the mean, scale and residuals model components given in (4.31), and for the construction of the corresponding maps, which are presented in Figures 5.1 and 5.2.

When comparing these maps with the ones in Figures 4.2 and 4.3, the effect of in-

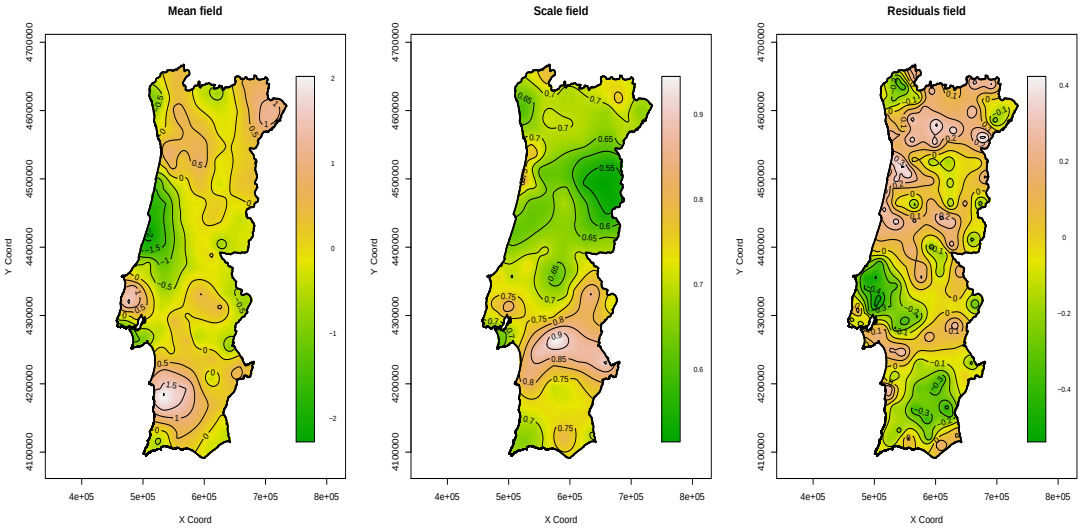

Figure 5.1: Estimated mean (left), scale (center) and residuals (right) maps for Mn, considering covariates

cluding country specific information seems to be almost negligible in the mean field for the Mn element, while for Pb the mean field map presents values that are more discretized across the prediction region. Also for Pb, the radius of influence ϕ in the residuals field is higher when considering the sampling intensity as covariate, hence the predicted values are not as constant as before when not considering this covariate.

5.4.2 Spatio-temporal prediction

Values obtained according to the prediction model defined in (4.36) are summarized in Table 5.5 and the corresponding maps are in Figure 5.3. When comparing the values in Table 5.5 with the ones in Table 4.2, one may observe that for Mn the range of predicted concentration is larger when not considering covariates. However, both the median and the mean of the predicted concentration are larger if the covariate is considered. Regarding the predicted Pb concentration values, the range of predicted values is also larger when not considering the covariate, being this same pattern shared for both the mean and the median.

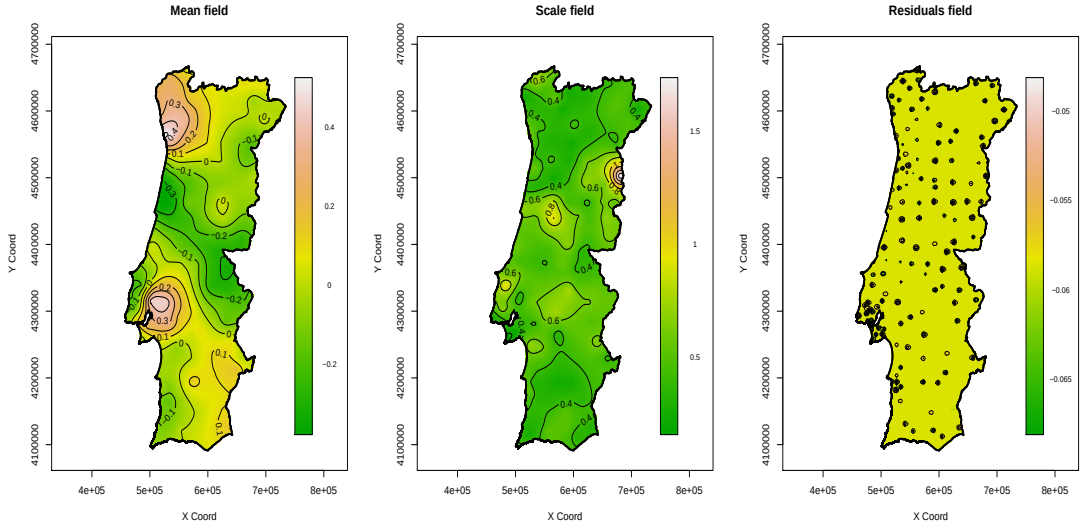

Figure 5.2: Estimated mean (left), scale (center) and residuals (right) maps for Pb, considering covariates

Table 5.5: Predicted transformed concentration obtained according to model (4.31) for the 2002 survey and associated interpolation error, when considering covariates

	Mn		Pb	
	Predicted	Error	Predicted	Error
Min	4.30	0.80	0.61	0.59
Median	7.53	1.07	0.87	0.63
Max	8.93	1.18	2.00	0.65
Mean	7.37	1.06	0.94	0.63
St. dev.	0.66	0.12	0.25	0.02

5.5 Comparison of results obtained so far

So far, predicted transformed concentration values for the 2002 survey were already obtained via four different prediction models,

- two purely spatial models, corresponding to Ordinary Kriging and Universal Kriging
- two spatio-temporal models, the one defined by (4.31) and the proposed extension detailed in (5.1) and (5.2) considering covariates

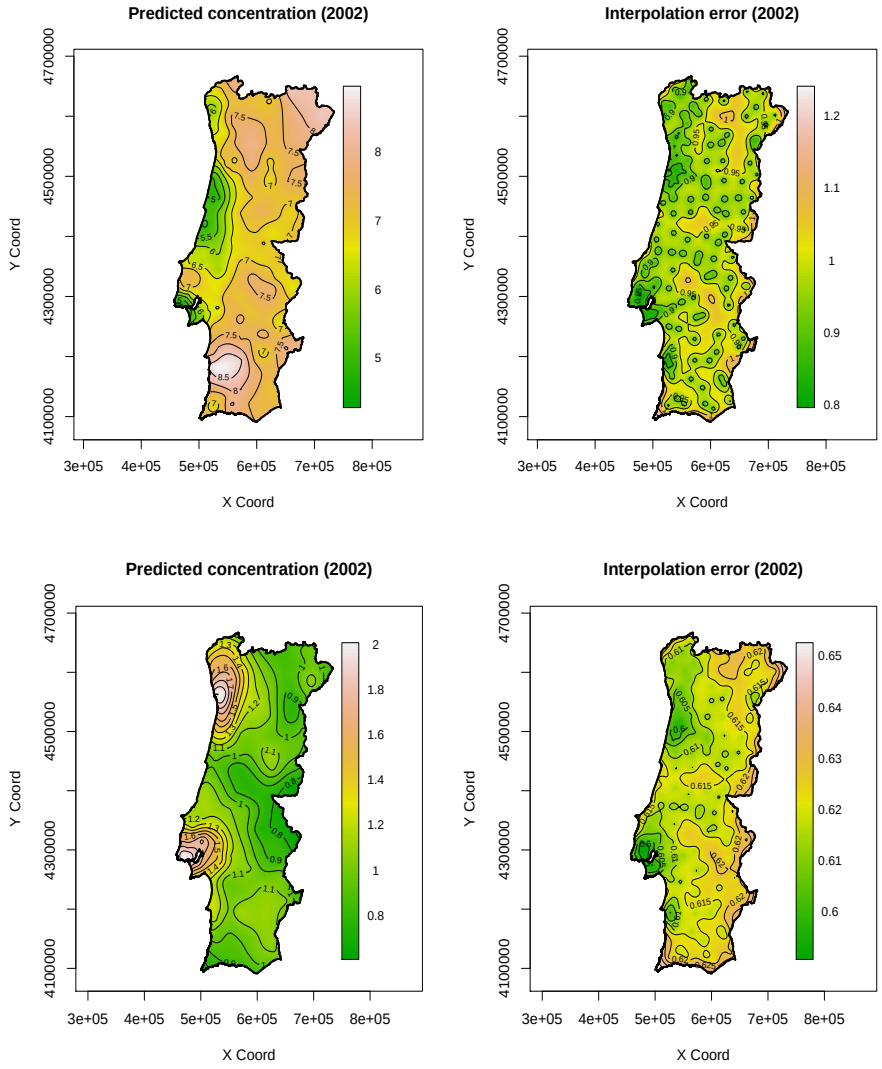

Figure 5.3: Predicted transformed concentration map for the 2002 survey (left) and the associated interpolation error map (right), for Mn (top panel) and Pb (bottom panel), when considering covariates

Table 5.6 synthesizes the values presented in Tables 3.4 and 3.5 (right-most columns), 3.7, 4.2 and 5.5. One may observe that the median and the mean predicted concentrations are larger for Mn when considering covariates, both in the spatial and the spatio-temporal models. In what respects the Pb predicted concentration, although the standard deviation of predicted values is of similar magnitude for the four models,

Table 5.6: Summary of predicted transformed concentration for the 2002 survey, via four different prediction models

		Spatial model		Spatio-temporal model	
		without covariate	with covariate	without covariate	with covariate
Mn					
	Min	4.88	4.39	4.87	4.30
	Median	6.99	7.57	7.21	7.53
	Max	8.59	8.58	8.72	8.93
	Mean	6.97	7.40	7.03	7.37
	St. dev.	0.63	0.63	0.73	0.66
Pb					
	Min	0.19	0.13	0.67	0.61
	Median	1.19	0.89	1.15	0.87
	Max	3.01	3.25	2.16	2.00
	Mean	1.20	0.99	1.16	0.84
	St. dev.	0.28	0.33	0.19	0.25

the median and the mean take lower values when predictions are made considering covariates.

These results underline the importance of considering the information about the sampling design as a covariate. Otherwise, predicted values would be underestimated in case of Mn, and overestimated in case of Pb.

5.5.1 Cross validation of sample data

As one may observe in the summary Table 5.6, different prediction methodologies lead to different predicted concentration values. Cross validation, with the aim of comparing different prediction methodologies, has been used in spatial statistics for the analysis of prediction errors (Olea (2012)). In the application under consideration, to understand the accuracy of the interpolated values obtained for the most recent survey, when using each one of the methods described so far, an exercise of cross-validation was performed. This means that one omits at each time one sampled location and interpolates the value at that location as a function of the observed values at all other locations, resulting in a set of 146 cross-validated values. Then, at each monitoring location, the APE,

as defined in (5.3), was used as a discrepancy measure between the observed and the predicted value.

The results obtained when comparing the mentioned methods are given in Table 5.7. As it was to be expected, the lowest values of mean APE (MAPE) are produced by the

Table 5.7: Mean and standard deviation of the APE

		Spatial Model		Spatio-temporal Model	
		without covariates	with covariates	without covariates	with covariates
Mn					
	MAPE	0.95	0.93	0.94	0.92
	Sd(APE)	0.65	0.65	0.64	0.65
Pb					
	MAPE	0.61	0.59	0.57	0.56
	Sd(APE)	0.63	0.62	0.63	0.63

most informative model, *i.e.*, the spatio-temporal model with covariates. We can also observe that when comparing the spatial models with the spatio-temporal models, lower values of MAPE are provided by the cases where covariates are considered. Values of the standard deviation of the APE (Sd(APE)) are similar for the four models compared.

5.5.2 Assessing interpolation errors

Considering now the assessment of interpolation error for each one of the four models previously mentioned, the resulting interpolation errors obtained when predicting Mn and Pb concentrations for the most recent survey were compared.

At each unobserved location goal of prediction, the estimated interpolation error is defined as the kriging standard deviation, if the prediction is made by means of a model considering only spatial data, or as the positive square root of (4.34) or any of its particular cases, when considering spatio-temporal data. For the data concerning both Mn and Pb concentrations, the interpolation error was estimated according to the specific case of (4.35), which is suitable when the variances for the scale and the residuals fields are negligible. Specifically, the values for the scale field variance and residual field

variance were, when not considering covariates, 1.5×10^{-3} and 1.8×10^{-2} and, when considering covariates, 1.64×10^{-1} and 1.17×10^{-2}. For Pb the same procedure was adopted to estimate the interpolation error, as the scale field variance and residual field variance were 8.72×10^{-3} and 2.48×10^{-12} or 1.18×10^{-1} and 3.02×10^{-7}, respectively when considering or nor considering covariates.

Table 5.8 summarizes, from the information in Tables 3.4 and 3.5 (right-most columns), 3.7, 4.2 and 5.5, the obtained interpolation error values, for the same grid 300×100 under consideration covering mainland Portugal, which is composed by 30.000 grid points at a distance of around 2 km from each other. It shows a decrease both in the central

Table 5.8: Interpolation error values summary when predicting 2002 Mn and Pb concentrations

		Spatial Model		Spatio-temporal Model	
	Interpolation Error	without covariates	with covariates	without covariates	with covariates
Mn					
	Min	1.04	1.03	0.92	0.80
	Median	1.26	1.25	1.16	1.07
	Max	1.47	1.37	1.45	1.18
	Mean	1.26	1.24	1.17	1.06
	St. dev.	0.12	0.09	0.14	0.12
Pb					
	Min	0.70	0.66	0.59	0.59
	Median	0.88	0.85	0.66	0.63
	Max	0.94	0.89	0.70	0.65
	Mean	0.86	0.83	0.66	0.63
	St. dev.	0.07	0.05	0.03	0.02

tendency and in the dispersion measures for the interpolation error, with exception for the standard deviation, from the "worst" scenario (spatial model without covariates) to the "best" scenario (spatio-temporal model with covariates). Comparing the spatial with the spatio-temporal model, less dispersion is found if covariates are taken into account.

In summary, if the emphasis is put on the prediction accuracy rather than on the prediction itself, these results emphasize that the most informative model, *i.e.* the one

using explanatory covariates together with temporal information, leads to the lower amount of prediction error.

6

A multivariate spatio-temporal model

6.1 Introduction

Nowadays, due to technology developments and worldwide policies, environmental monitoring networks are providing large amounts of data exhibiting a spatial and a temporal correlated nature, and as a consequence a large number of models and techniques to analyze this sort of data has emerged. Some references on the subject of spatio-temporal modelling are Kyriakidis & Journel (1999) reviewing stochastic models involving the extension of spatial analysis tools to include the time dimension, de Cesare *et al.* (2001) with a discussion on some classes of models and the introduction of the so-called product-sum model, Sahu & Mardia (2005) with a review on methods for modelling spatio-temporal point referenced data, Sherman (2011) with a brief survey of several types of spatio-temporal covariance models, Huang *et al.* (2007) comparing four spatio-temporal models covering the situations of separability and non-separability, Cameletti *et al.* (2011) with a comparison of six alternative spatio-temporal models belonging to the class of Bayesian hierarchical models and providing criteria to choose among them.

In environmental sciences, typically data are collected through monitoring stations. Shaddick & Wakefield (2002) uses daily pollution data collected at eight monitoring sites within London, measuring the pollutants particulate matter PM_{10}, carbon monoxide, nitrogen oxide and sulphur dioxide over the period from 1994 to 1997, Fanshawe

et al. (2008) models PM_4 levels in the United Kingdom, using data routinely collected from 20 monitoring stations between 1961 and 1992, Lindstrom *et al.* (2011) considers data from 20 sites in and around Los Angeles from an Air Quality System network measuring ambient concentrations of air pollutants collected at a 2-week timescale, from the beginning of 1999 until September 2009, Cameletti *et al.* (2011) models PM_{10} concentration in Piemonte region, analyzing daily data collected during 182 day from 24 sites.

However, data may also be collected through biomonitoring surveys covering extensive areas. Some examples of studies involving moss samples as biomonitors of atmospheric heavy metal deposition are Aboal *et al.* (2006) and Diggle *et al.* (2010) with data from Galicia, northern Spain, Harmens *et al.* (2010) considering data from several countries across Europe, Steinnes *et al.* (2003) and Steinnes *et al.* (2011) concerning Norway data, Zechmeister *et al.* (2008) with data from Austria.

The aim of these different studies, however, is basically the same: to predict one attribute of interest at an unmonitored location or time. The task of modelling such data, with the concern for the accuracy of predictions, has to account for both spatial and temporal interactions. To deal with this dependencies, different approaches can be considered:

(i) a spatial-temporal data analysis with methods for random fields in $\mathbb{R}^{d+1}$,

(ii) a multivariate spatial analysis for each time point, if the spatial dimension is larger than the temporal, or

(iii) a multivariate temporal analysis for each location, if the temporal dimension is larger than the spatial.

Although the first approach is less appropriate since space, which has not past, present, or future, and time are not directly comparable, one can identify drawbacks also for the second and the third approaches. Namely the fact that with the later, a possibly existing temporal correlation can not be incorporated in predictions, and the fact that with the former, predictions made for the last time point don't take into account historic data. Moreover, once the interpolation of observations in a continuous space-time process should take into account the interaction between the spatial and the temporal

components and allow for predictions both in unmonitored time and/or space, to perform separate analysis in time allows for predictions in space only, and reciprocally.
It is common to have studies, such as the ones mentioned before related to monitoring stations, involving environmental spatio-temporal data containing a dense time dimension but only a sparse spatial one, as a result of the easiness of gathering data enabled by modern technologies. That is not the case of the biomonitoring data being used in this work, where measurements of heavy metal concentrations were made at 146 spatial locations in only 3 surveys.
Our aim is to propose a naive spatio-temporal framework which incorporates into the model both time and space correlations, capable to fit spatio-temporal data containing a reduced number of time observations. Due to this particular characteristic of having few temporal records, and under the hypothesis of separability of the correlation structure, it may be the case that the number of parameters to estimate in the temporal correlation function equals the number of temporal observations, which corresponds to have a saturated correlation model in the time dimension, *i.e.*, a model perfectly reproducing the data.

6.2 The model

We propose a spatio-temporal model for Gaussian data, collected at location $\mathbf{s} \in \mathbb{R}^2$ and time $t \in \mathbb{N}$, defined as

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + Z(\mathbf{s}, t) + \varepsilon(\mathbf{s}, t) \tag{6.1}$$

Considering N locations observed at T surveys, the mean component $\mu(\mathbf{s}, t)$, depending on possibly observed covariates $f(\mathbf{s}, t)$, indexed in space or in time, will be considered as

$$\mu(\mathbf{s}_i, t_k) = \sum_{j=1}^{p} \beta_j f_j(\mathbf{s}_i, t_k) \tag{6.2}$$

where $\mathrm{E}\left[Y(\mathbf{s}_i, t_k)\right] = \mu(\mathbf{s}_i, t_k)$, $i = 1, \ldots, N, k = 1, \ldots, T$. Under matrix notation, one

has

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{Z} + \boldsymbol{\varepsilon} \tag{6.3}$$

with

$$\boldsymbol{\mu} = M \cdot \boldsymbol{\beta} \tag{6.4}$$

the product of the $NT \times (p+1)$ design matrix M (being p the number of considered covariates), and $\boldsymbol{\beta}$, the vector of regression coefficients in the mean component (6.2). The non-observed spatio-temporal process $Z(\mathbf{s}, t)$ is such that

$$\boldsymbol{Z} \sim MVN\big(0, \Sigma\big) \tag{6.5}$$

and $\varepsilon(\mathbf{s}, t)$ represents Gaussian space-time measurements errors,

$$\boldsymbol{\varepsilon} \sim N\big(\mathbf{0}, \tau^2 I_{NT}\big) \tag{6.6}$$

with I_{NT} the identity matrix of order NT.
In the spatio-temporal process (6.5), Σ is a $NT \times NT$ symmetric matrix, which can be interpreted as a $T \times T$ block matrix such that for k, $l = 1, \ldots, T$, the element on line i and column j is

$$\Sigma^{i,j,k,l} = \mathrm{Cov}_{ST}\big[Z(\mathbf{s}_i, t_k), Z(\mathbf{s}_j, t_l)\big] \ , \quad i, j = 1, \ldots, N. \tag{6.7}$$

The proposed model assumes an isotropic and separable covariance structure, so we define purely spatial and purely temporal covariance functions, Cov_S and Cov_T, resulting in

$$\begin{aligned} \Sigma^{i,j,k,l} &= \mathrm{Cov}_S\left(\|\mathbf{s}_i - \mathbf{s}_j\|\right) \times \mathrm{Cov}_T\left(|t_k - t_l|\right) \\ &= \mathrm{Cov}_S\left(\mathbf{h}_S\right) \times \mathrm{Cov}_T\left(h_T\right). \end{aligned} \tag{6.8}$$

Section 4.2 introduces a brief review on available tests for space-time separability. The definition of separability, according to (Bruno *et al.* (2003)), satisfies the two following statements:

(i) the spatial covariance function is constant in time, so that

$$\mathrm{Cov}\big[Z(\mathbf{s}_i, t_k), Z(\mathbf{s}_j, t_k)\big] = \mathrm{Cov}\big[Z(\mathbf{s}_i, t_l), Z(\mathbf{s}_j, t_l)\big]$$

for all $(t_k, t_l), \mathbf{s}_i, \mathbf{s}_j \in \mathbb{R}^2$, and

(ii) the temporal covariance function is the same at all monitoring locations, regardless of the displacement between locations,

$$\mathrm{Cov}\big[Z(\mathbf{s}_i, t_k), Z(\mathbf{s}_i, t_l)\big] = \mathrm{Cov}\big[Z(\mathbf{s}_j, t_k), Z(\mathbf{s}_j, t_l)\big]$$

for all $(\mathbf{s}_i, \mathbf{s}_j), t_k, t_l = 1, \ldots, T$.

6.2.1 Inference on model parameters

Under the assumption of second order stationarity, we allow two different interpretations for the covariance function. The most common one (*e.g.*, Rodriguez-Iturbe & Mejia (1974), Sherman (2011)) considers a scale parameter σ^2_{total} representing the overall variance, being the covariance matrix given by

$$\Sigma = \sigma^2_{total}\, R_S \otimes R_T \tag{6.9}$$

$\otimes$ is the Kronecker product of matrices, and R_S and R_T are, respectively, the $N \times N$ spatial correlation matrix and the $T \times T$ temporal correlation matrix. As an alternative, we propose to take into account different scale parameters for the spatial and the temporal components,

$$\Sigma = \sigma^2_S R_S \otimes \sigma^2_T R_T \tag{6.10}$$

being σ^2_S the spatial variance and σ^2_T the temporal variance. Basically, this corresponds to perform a re-parametrization of the overall variance, decomposing it as the product of a spatial variance by a temporal one.

Regarding the spatial correlation matrix, and if a member of the exponential spatial correlation function is considered, the element on line $i = 1, \cdots, N$ and column $j = 1, \cdots, N$ is

$$\big[R_S(\phi_S)\big]_{ij} = \left[\exp\left\{-\frac{1}{\phi_S}\|\mathbf{s}_i - \mathbf{s}_j\|\right\}\right] \tag{6.11}$$

Having in mind that only a reduced number of time observations is available, the temporal correlation model is saturated, meaning that R_T is a $T \times T$ matrix whose elements, denoted by $\rho_T|t_k - t_l| = \rho_{k,l}$, $k, l = 1, \ldots, T$, measure the association between observations at time t_k and t_l.

Denoting by $\mathbf{C}_Y$ the covariance matrix of $Y(\mathbf{s}, t)$, which depends on the parameters τ^2

and $\boldsymbol{\theta} = \left(\sigma^2_{total}, \phi_S, \rho_{k,l}\right)$ or $\boldsymbol{\theta} = \left(\sigma^2_S, \sigma^2_T, \phi_S, \rho_{k,l}\right)$, ($\rho_{k,l}$ are $\frac{T\times(T-1)}{2}$ parameters, with $l > k$), we have

$$\mathbf{C}_Y(\tau^2, \boldsymbol{\theta}) = \Sigma(\boldsymbol{\theta}) + \tau^2 I_{NT}$$

Following standard results from the Gaussian distribution theory, the log-likelihood function is

$$\begin{aligned} \log\, L(\tau^2, \boldsymbol{\theta}) = &- \frac{NT}{2}\, \log(2\pi) - \frac{1}{2}\, \log\Big(det\big(\mathbf{C}_Y(\tau^2, \boldsymbol{\theta})\big)\Big) - \\ &- \frac{1}{2}\Big(\boldsymbol{Y} - M\,\boldsymbol{\beta}\Big)^t \mathbf{C}_Y^{-1}(\tau^2, \boldsymbol{\theta})\Big(\boldsymbol{Y} - M\,\boldsymbol{\beta}\Big) \end{aligned} \tag{6.12}$$

being the Maximum Likelihood Estimator for $\boldsymbol{\beta}$ given by

$$\widehat{\boldsymbol{\beta}} = \Big(M^t \mathbf{C}_Y^{-1}(\tau^2, \boldsymbol{\theta}) M\Big)^{-1} M^t \mathbf{C}_Y^{-1}(\tau^2, \boldsymbol{\theta})\, \boldsymbol{Y} \tag{6.13}$$

The log-likelihood function (6.12) depends on $6 + (p + 1)$ parameters, if one considers (6.9), or $7 + (p + 1)$ in case of (6.10). The advantage resulting from the knowledge of an analytic estimator for the parameter vector $\boldsymbol{\beta}$ is a reduction on the computational effort required for the parameter estimation.
The difference on the number of parameters in the log-likelihood function results from the re-parametrization of the overall variance σ^2_{total} in (6.9) by $(\sigma_S \cdot \sigma_T)^2$. Hence, it should be expected to find an estimate $\widehat{\sigma}^2_{total}$ close to the product $\widehat{\sigma}^2_S \cdot \widehat{\sigma}^2_T$.

6.2.2 The theoretical semi-variogram

The space-time semi-variogram $\gamma(\mathbf{h}_S, h_T)$ of the non-observed spatio-temporal process $Z(\mathbf{s}, t)$ is

$$\gamma(\mathbf{h}_S, h_T) = \frac{1}{2}\mathrm{Var}\big[Z(\mathbf{s}, t) - Z(\mathbf{s} + \mathbf{h}_S, t + h_T)\big] \tag{6.14}$$

The relationship

$$\gamma(\mathbf{h}_S, h_T) = \mathrm{Cov}_{ST}(\mathbf{0}, 0) - \mathrm{Cov}_{ST}(\mathbf{h}_S, h_t)$$

under the second order stationarity conditions, between $\gamma(\mathbf{h}_S, h_T)$ and the space-time covariance function is well known (see, *e.g.*, Cressie & Wikle (2011), Sherman (2011)), and was previously indicated in (4.9).
If the covariance structure (6.9) is considered, then

$$\begin{aligned} \text{Cov}_{ST}(\mathbf{0}, 0) = &\quad \text{Var}\left[Z(\mathbf{s}, t)\right] \\ = &\quad \sigma^2_{total} \end{aligned}$$

and

$$\text{Cov}_{ST}(\mathbf{h}_S, h_t) = \sigma^2_{total} \cdot \rho_S(\mathbf{h}_S) \cdot \rho_T(h_T)$$

where $\rho_S(\mathbf{h}_S) = \rho_S(\|\mathbf{s}_i - \mathbf{s}_j\|)$ and $\rho_T(h_T) = \rho_T(|t_k - t_l|)$ represent, respectively, the elements of the spatial and temporal correlation matrices.
Hence, assuming an exponential spatial model,

$$\begin{aligned} \gamma(\mathbf{h}_S, h_t) &= \sigma^2_{total} - \left(\sigma^2_{total} \cdot \rho_S(\mathbf{h}_S) \cdot \rho_T(h_T)\right) \\ &= \begin{cases} 0 & , \ \mathbf{h}_S = 0 \wedge h_T = 0 \\ \sigma^2_{total}\left(1 - \exp\left\{-\dfrac{1}{\phi_S}\mathbf{h}_S\right\} \cdot \rho_T(h_T)\right) & , \ \mathbf{h}_S \neq 0 \vee h_T \neq 0 \end{cases} \end{aligned} \tag{6.15}$$

Equivalently, if (6.10) is assumed, one has

$$\gamma(\mathbf{h}_S, h_T) = \begin{cases} 0 & , \ \mathbf{h}_S = 0 \wedge h_T = 0 \\ \sigma^2_S\sigma^2_T\left(1 - \exp\left\{-\dfrac{1}{\phi_S}\mathbf{h}_S\right\} \cdot \rho_T(h_T)\right) & , \ \mathbf{h}_S \neq 0 \vee h_T \neq 0 \end{cases} \tag{6.16}$$

Again, these two versions for the semi-variogram are analogous, only differing on the re-parametrization of the variance. The resulting advantage of considering this decomposition comes from the capacity to assign different magnitudes to the spatial and the temporal variability, which is not allowed in (6.15).

6.2.3 Prediction at unsampled locations

The model described in (6.1) assumes that the hidden process $Z(\mathbf{s},t)$ and the measurement error $\varepsilon(\mathbf{s},t)$ are Gaussian ((6.5) and (6.6)). It is well known (*e.g.* Cressie & Wikle (2011)), that for a non-observed location $\mathbf{s}_0$ and a time t_0, the joint distribution of $Y(\mathbf{s}_0,t_0)$ and $\boldsymbol{Y}$ is

$$\begin{bmatrix} Y(\mathbf{s}_0,t_0) \\ \boldsymbol{Y} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu(\mathbf{s}_0,t_0) \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} C_{0,0} & \mathbf{c}_0^T \\ \mathbf{c}_0 & \mathbf{C}_Y(\tau^2,\boldsymbol{\theta}) \end{bmatrix}\right) \tag{6.17}$$

where $\mu(\mathbf{s}_0,t_0) = \sum_{i=1}^{p} \widehat{\beta}_i f_i(\mathbf{s}_0,t_0)$ and $\boldsymbol{\mu}$ is defined by (6.2), $C_{0,0} = \text{Var}\left[Y(\mathbf{s}_0,t_0)\right]$, $\mathbf{c}_0 = \text{Cov}\left[Y(\mathbf{s}_0,t_0),\boldsymbol{Y}\right]$, and $\mathbf{C}_Y(\tau^2,\boldsymbol{\theta}) = \Sigma(\boldsymbol{\theta}) + \tau^2 I_{NT}$ with $\Sigma(\boldsymbol{\theta})$ as in (6.7).
The predicted value $\widehat{Y}(\mathbf{s}_0,t_0)$ at an unsampled location can be obtained from (6.17), being given by

$$\widehat{Y}(\mathbf{s}_0,t_0) = \text{E}\big[Y(\mathbf{s}_0,t_0)|\boldsymbol{Y}\big] = \mu(\mathbf{s}_0,t_0) + \mathbf{c}_0^T\,\mathbf{C}_Y^{-1}(\tau^2,\boldsymbol{\theta})\,\big(\boldsymbol{Y}-\boldsymbol{\mu}\big) \tag{6.18}$$

The variance of the prediction, also resulting from (6.17), is

$$\sigma^2(\mathbf{s}_0,t_0) = \text{E}\left[Y(\mathbf{s}_0,t_0) - \widehat{Y}(\mathbf{s}_0,t_0)\right]^2 = C_{0,0} - \mathbf{c}_0^T\,\mathbf{C}_Y^{-1}(\tau^2,\boldsymbol{\theta})\,\mathbf{c}_0 \tag{6.19}$$

In particular, if an exponential spatial model is assumed, $\mathbf{c}_0$ is the concatenation of T vectors c_1, c_2, …, c_T, each with dimension $N \times 1$, where

$$c_i = \widehat{\sigma}^2_{total}\left(\exp\big(-\frac{1}{\widehat{\phi}_S}\|\mathbf{s}_0-\mathbf{s}_1\|\big)\widehat{\rho}_T(|t_i-t_0|),\ \ldots\ ,\exp\big(-\frac{1}{\widehat{\phi}_S}\|\mathbf{s}_0-\mathbf{s}_N\|\big)\widehat{\rho}_T(|t_i-t_0|)\right)^t$$

or

$$c_i = \widehat{\sigma}^2_S\,\widehat{\sigma}^2_T\left(\exp\big(-\frac{1}{\widehat{\phi}_S}\|\mathbf{s}_0-\mathbf{s}_1\|\big)\widehat{\rho}_T(|t_i-t_0|),\ \cdots\ ,\exp\big(-\frac{1}{\widehat{\phi}_S}\|\mathbf{s}_0-\mathbf{s}_N\|\big)\widehat{\rho}_T(|t_i-t_0|)\right)^t$$

with $i = 1,2,\ldots,T$, depending on assuming the covariance structure in (6.9) or in (6.10).

6.3 Simulation study

For model validation purposes, a simulation study was conducted. Gaussian data spatially correlated, with zero mean, was generated on a set of $N = 50$ randomly chosen locations considered in the square $[0, 1]^2$, and at $T = 3$ time points according to an AR(2) model, in order to be representative of a strong temporal correlation among the three time points. To replicate the behavior of the real data set described in Chapter 3, having a region with more intensified sampling density, 15 of those locations belong to the square $[0.45, 0.55]^2$.

The mean component (6.2) includes the covariates *intensity of sampling locations*, $int(\mathbf{s})$, and the specific contribution of a given survey, $v_i(t)$, resulting in

$$\mu(\mathbf{s}, t) = \beta_0 + \beta_1 int(\mathbf{s}) + \beta_2 v_2(t) + \beta_3 v_3(t) \tag{6.20}$$

where

$$v_i(t) = \begin{cases} 1 & \text{if } t = i \\ & \qquad\qquad i = 2, 3 \\ 0 & \text{otherwise} \end{cases} \tag{6.21}$$

The reason to consider the covariate sampling intensity was previously detailed in Chapter 4. The covariates v_1 and v_2 intend to accommodate situations in which, for instance, a change in policies is expected to modify some outcome. That is the case, as mentioned in Chapter 3, of the observed reduction on the scale of Pb measurements at the 2002 survey, which could probably be attributed to new legislation forcing the use of unleaded fuel.

The two different space-time covariance functions described in (6.9) and (6.10) (from now on, called Scenario 1 and Scenario 2) were compared. The particular choice of the parameter values for the mean component (6.20) were: $\beta_0 = 0$, which indicates that the expected value for the first survey is zero; $\beta_1 = 1$, once the inclusion of the covariate *sampling intensity* in the design matrix M in (6.2) was made by subtracting the mean intensity of all locations to the intensity of each location, thus an unitary coefficient turn the contribution of the covariate to the mean dependent on the sign of that difference; $\beta_2 = \beta_3 = 0$, which means that the diference between the expected values of, respectively, the second and the first survey, and the third and the first survey, is zero. The parameters for the covariance function were $\rho_T(|t_1 - t_2|) = \rho_T(|t_2 - t_3|) = 0.857$ and $\rho_T(|t_1 - t_3|) = 0.814$, obtained from the autocorrelation function of an AR(2)

model with $\phi_1 = 0.6$ and $\phi_2 = 0.3$; $\sigma_S^2 = 5$, representing the mean of spatial variances along time and $\sigma_T^2 = 0.75$ representing the mean of temporal variances at each sample location; $\sigma^2 = 4$, representing the variance of all observations along time and space; $\tau^2 = 0.25$; $\phi = 0.3$, being approximately 30% of the maximum distance between two sampled locations.

For each scenario 100 replicates were computed. For each replicate, the estimates of β's were computed according to (6.13) and the optimization method L-BFGS-B of the function *optim* in R was used to obtain estimates for the other parameters. To assess the predictive performance of each scenario, the mean and the standard error of the estimates were computed. Results on the estimation of the model parameters may be found in Tables 6.1 and 6.2.

Table 6.1: Estimates for the model parameters when considering the space-time covariance function in (6.9)

True Param.	Mean	Std. Error
$\beta_0 = 0$	0.042	0.102
$\beta_1 = 1$	0.998	0.007
$\beta_2 = 0$	-0.015	0.051
$\beta_3 = 0$	-0.012	0.064
$\rho_T(\lvert t_1 - t_2\rvert) = 0.857$	0.838	0.006
$\rho_T(\lvert t_1 - t_3\rvert) = 0.814$	0.810	0.006
$\rho_T(\lvert t_2 - t_3\rvert) = 0.857$	0.844	0.005
$\log(\sigma^2) = 1.386$	1.078	0.029
$\log(\tau^2) = -1.386$	-1.530	0.031
$\log(\phi) = -1.203$	-1.564	0.035

Although with no major differences, these estimates show lower values of standard errors for the majority of the parameters if one considers separately the spatial and the temporal variance contribution.

6.4 Application to environmental data

The model previously described in Section 6.2 will now be used to obtain predicted values of heavy metal concentration at each point of the grid covering mainland Por-

Table 6.2: Estimates for the model parameters when considering the space-time covariance function in (6.10)

Param	Mean	Std. Error
$\beta_0 = 0$	-0.043	0.096
$\beta_1 = 1$	0.992	0.009
$\beta_2 = 0$	-0.054	0.049
$\beta_3 = 0$	-0.032	0.053
$\rho_T(\lvert t_1 - t_2\rvert) = 0.857$	0.847	0.005
$\rho_T(\lvert t_1 - t_3\rvert) = 0.814$	0.835	0.005
$\rho_T(\lvert t_2 - t_3\rvert) = 0.857$	0.853	0.006
$\log(\sigma_S^2) = 1.609$	1.465	0.015
$\log(\sigma_T^2) = -0.287$	-0.432	0.015
$\log(\tau^2) = -1.386$	-1.498	0.027
$\log(\phi) = -1.203$	-1.645	0.035

tugal, as in the previous application in Section 4.6 of Chapter 4, which corresponds to have a prediction point every $2km$. The development of this model was done assuming data from a Gaussian distribution, hence predictions for the most recent survey will be made only for Mn, as for Pb the Box-Cox transformed concentrations don't fit this distribution.

For this application, the assumption of separability considered in the definition of the model will now to be assessed.

6.4.1 Assessing the separability assumption

An importante assumption in the model definition is the separability of the covariance structure. This is a simplifying assumption frequently considered in environmental modelling, as considerable computational benefits arise from this convenient property. For the application under consideration, due to the reduced number of time observations, a formal test to ascertain the separability of the covariance structure is not possible to perform.

Alternatively, we use the aid of the empirical spatio-temporal semivariogram to adjust a spatio-temporal separable model (4.14), a metric model (4.12) and a product-sum model (4.17). For each model, the resulting parameter estimates are in Table 6.3. By

default, exponential models were assumed in the covariance functions to obtain these estimates.

Table 6.3: Parameter estimates for the adjusted variograms in Figure 6.1 ($\widehat{\phi}_S$ in meters and $\widehat{\phi}_T$ in years)

Model	Component	$\widehat{\tau}^2$	$\widehat{\sigma}^2$	$\widehat{\phi}$
Separable	Spatial	0.41	0.59	53108.81
	Temporal	0.24	0.76	2.69
	Joint	—	2.45	—
Metric	Joint	1.13	0.91	—
ProductSum	Spatial	—	1.58	69065.92
	Temporal	—	1.94	2.04
	Joint	0.99	3.49	—

If a comparison is made with, for instance, the parameter estimates obtained for the mean field in model (5.1)[1], represented in the left-most column of Table 5.4, one can find more similarity in the estimates for the nugget and the range of the spatial component of the separable model.

There are some graphical differences between these several adjusted theoretical models, which are being represented in Figure 6.1. The separable model captures more precisely the behavior shown by the empirical semivariogram, detecting an increase of variances from the first to the second survey, and a decrease for the third. Cressie & Huang (1999) state that *separable models are often chosen for convenience rather than for their ability to fit the data well*, so the assumption of separability will be assumed for the Mn data.

As stated before, the definition of separability satisfies the statements *(i)* and *(ii)* in page 80. In our application case, *(ii)* reinforces the fact of having a saturated temporal correlation model. On the other hand, *(i)* means that the spatial structure is the same for all surveys. Figure 6.2 shows in the left the empirical semivariograms related with each survey and, in the right, cross semivariograms for time lags of 0, 1 and 2 surveys.

[1]The reason to choose this model is that it also considers the inclusion of covariates and the separability on the correlation structure.

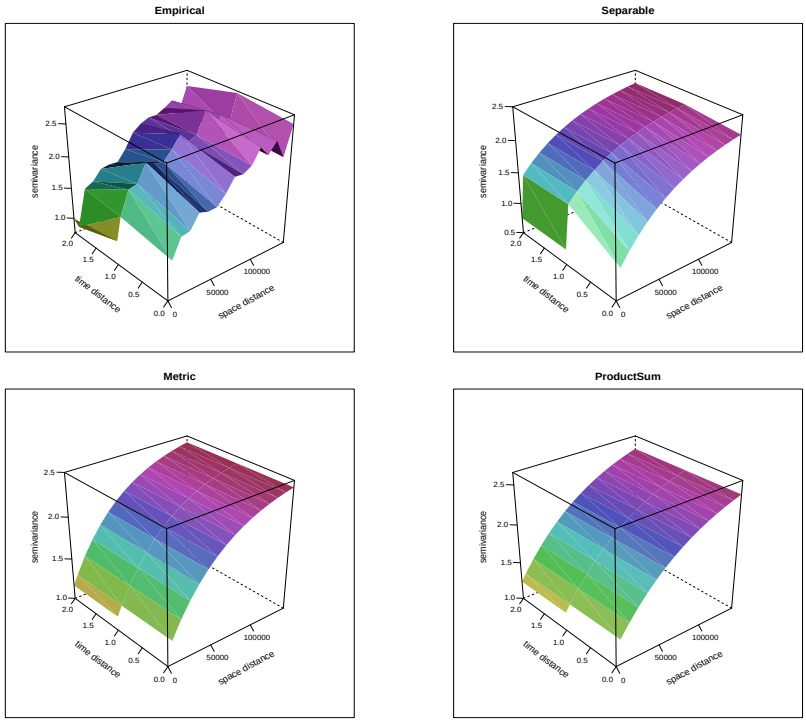

Figure 6.1: Empirical (top left), separable (top right), Metric (bottom left) and ProductSum (bottom right) spatio-temporal semivariograms for Mn transformed data

The structure revealed by the cross semivariograms is similar independently of the lag being considered, corroborating the assumption of equality of spatial structure for the performed surveys.

6.4.2 Maximum likelihood estimates of the parameters

The parameter estimates of model (6.1), considering the covariance functions defined by (6.9) and by (6.10), are given in Table 6.4. In particular, one can observe that, in either case, the baseline effect estimate β_0 is above seven units, being of similar magnitude to the mean of the transformed Mn concentration. The inclusion of the covariate *sampling intensity*, as detailed in the simulation study, was done considering the difference between the intensity of each location and the mean of the intensity for all locations. This way, the estimate of β_1 resultant from (6.10) reflects more precisely the fact that larger values of Mn concentration occur in regions with low urban or industrial density,

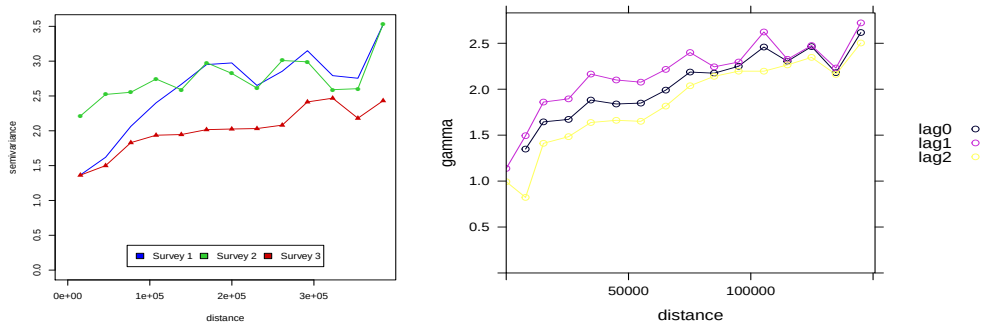

Figure 6.2: Empirical semivariograms (left) and cross semivariograms for lags 0, 1 and 2 (right) for Mn transformed data

which corresponds also to regions with lower sampling locations intensity. The contribution of the second and of the third surveys are estimated to be of different signs, namely the second survey with a positive contribution and the third with a negative one. However, the estimate of β_3 for the covariance function described in (6.10) reinforces the contribution of the third survey to the decrease of predicted concentrations. As was to be expected, the parameters related to the temporal correlation are equal in either approach, denoting data strongly time correlated. Similarly, the nugget effect is estimated equally by either approach. The covariance function given by (6.10) assumes different spatial and temporal variances, contributing as a product to the covariance of the process. It is worthwhile to notice that the product of the estimates of these parameters equals the overall variance estimate for (6.9). The estimates for the radius of influence ϕ are also of similar magnitude for either approach, being approximately 55.5 km. The computation of the standard errors was made via Monte-Carlo simulation.

6.4.3 Prediction of Mn concentration for the most recent survey

These parameter estimates, together with (6.18) and (6.19), allowed to predict, for the most recent survey, the Box-Cox transformed Mn concentration values and the associated interpolation errors. This corresponds to adopt plug-in estimates of the parameters that define the mean and the covariance structure of the Mn variable to proceed with prediction. A brief summary of those predicted values is in Table 6.5. One

Table 6.4: Model (6.1) parameter estimates with standard errors

	Covariance function in (6.9)		Covariance function in (6.10)	
Parameter	Estimate	St. Error	Estimate	St. Error
β_0	7.39	0.04	7.45	0.04
β_1	-0.01	<0.01	0.01	<0.01
β_2	0.21	0.01	0.15	0.01
β_3	-0.14	0.02	-0.24	0.02
$\rho_T(\lvert t_1 - t_2 \rvert)$	0.97	0.01	0.97	0.01
$\rho_T(\lvert t_1 - t_3 \rvert)$	0.91	0.01	0.91	0.01
$\rho_T(\lvert t_2 - t_3 \rvert)$	0.96	0.01	0.96	0.01
σ^2_{total}	1.45	0.03	—	—
σ^2_S	—	—	0.98	0.01
σ^2_T	—	—	1.48	0.02
τ^2	1.02	0.01	1.02	0.01
$\phi\,(m)$	58596.89	1421.10	58624.08	1470.20

can observe that the range of predicted values, in either situation, is similar, although with larger predicted values when considering spatial and temporal scale parameters. In fact, all the measures, with exception of the standard deviation, of the predicted values according to (6.10) are larger than the ones obtained considering (6.9). Comparing these results with the ones resulting from the extension of Høst model (Table 5.5), one can register that model 6.1 produces lower median and mean values, even though either the minimum and the maximum predicted Mn transformed concentration considering separate spatial and temporal scale parameters, are larger. If predictions are computed considering an overall variance, one can register that the minimum value, the median and the mean are the lowest of all.

Figure 6.3 represents, on the left panel, the Mn predicted concentration map, while the associated prediction error is in the right panel. As was to be expected, once contamination by Mn is more associated to parameters describing soil typology and sampling site conditions, and less related to anthropogenic contamination sources, the higher predicted contamination values occur in the eastern part of mainland Portugal. This spatial structure is captured by either approach for the covariance function.

Table 6.5: Predicted Mn concentration for the 2002 survey and interpolation error values

	Covariance function in (6.9)		Covariance function in (6.10)	
Parameter	Predicted	Error	Predicted	Error
Min	4.05	0.23	4.53	0.41
Median	7.10	0.72	7.21	0.79
Max	9.05	1.19	9.16	1.19
Mean	6.92	0.78	7.09	0.85
St. dev.	0.92	0.23	0.74	0.19

6.4.4 Cross validation study

Similarly to what was performed before in Section (5.5.1), an exercise of cross-validation of results was developed to assess the accuracy of the predicted concentrations obtained considering the two approaches for the covariance structure in (6.9) and in (6.10). The technique in use for this cross-validation study was the same as before, that is, to leave out at each time one sampled location and interpolate the value at that location as a function of all other locations .

The APE was again considered to measure the discrepancy between the observed and the interpolated value at each location. Results revealed the value 1.02 for the mean APE, with a standard deviation of 0.81, no matter which approach for the covariance function is considered.

This is not a surprising result, once the two sets of predicted values were obtained from the same model, only with a different parametrization of the separable spatio-temporal covariance structure. In fact, this finding reinforces the advantage of being able to identify two distinct parameters for variability, one for the spatial dimension and the other for the temporal dimension, without compromising the accuracy of predicted values.

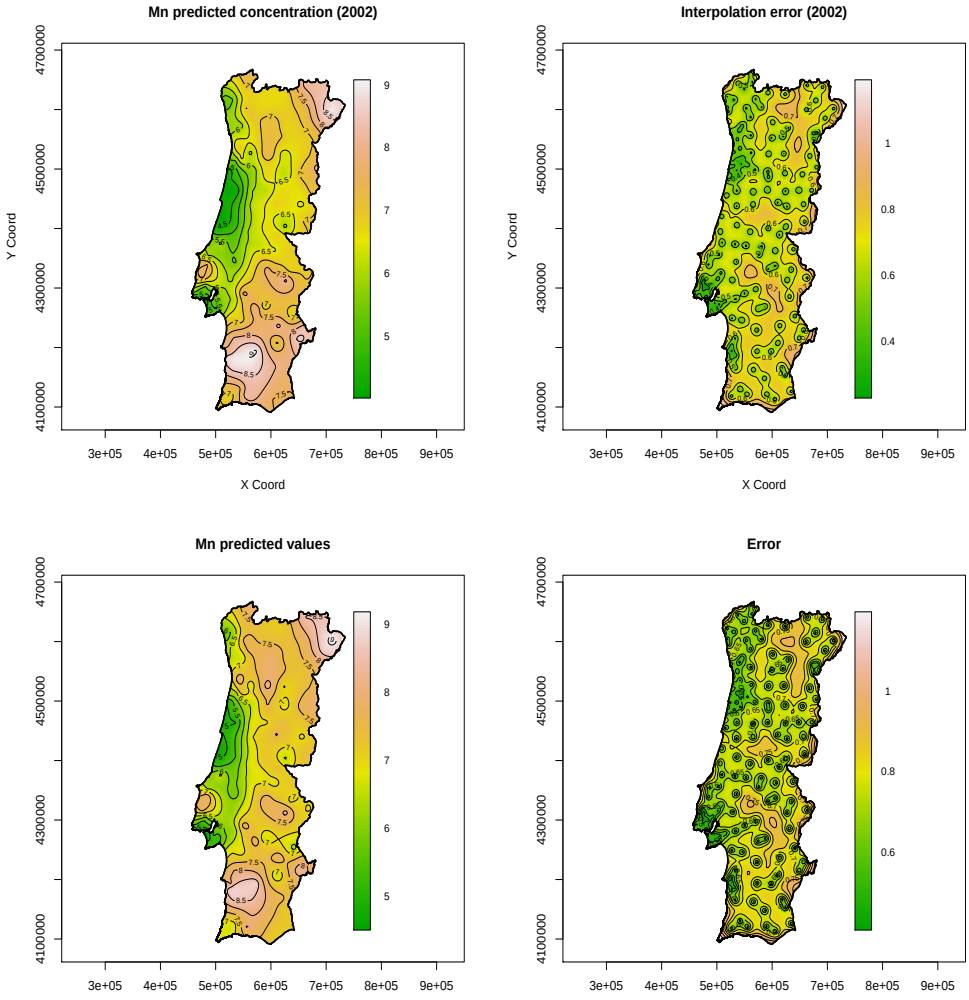

Figure 6.3: Prediction map for the 2002 survey (left) and the associated interpolation error map (right) for Mn transformed data, considering the covariance function given in (6.9) (top row) or (6.10) (bottom row)

7

Conclusions and future work

7.1 Conclusions

The main objectives of this work were centered in assessing the gain achieved in terms of spatial prediction accuracy, when incorporating into the prediction model not only the temporal dimension of data but also the information provided by explanatory variables of the process under observation. The motivation for this assessment resulted from the need to apply spatio-temporal prediction techniques to data sparse in the time dimension.

Typically, to deal with a reduced number of time observations, the temporal dimension is treated as a covariate, or by interpreting the response random variable as resulting from a multivariate spatial process. This work proposed to search for alternative ways of incorporating in the modelling the temporal correlation itself. As a result of this search, two different approaches were identified.

After introducing the fundamental concepts of spatial geostatistics, a practical illustration of them was performed by considering data related to water quality monitoring, gathered along four different time periods. Results suggested the need to incorporate in the prediction model the time dimension, to better understand the behavior of water quality in the whole sampling region.

The fundamental concepts of spatio-temporal geostatistics, introduced in Chapter 4, were also illustrated by means of an existing spatio-temporal prediction model. This model focuses on the spatial dimension by defining random fields for the mean, the scale and the residuals components, and incorporating the time dimension by means of

strictly temporal random fields, which work as corrections for the temporal evolution of the process. When geo-referenced information about the process under observation is available, it could be incorporated in the mean field component, as pointed in Chapter 5. The application of the mentioned model allowed to derive the predicted spatial pattern of pollution by heavy metals over mainland Portuguese territory. Specifically, data concerning concentrations of manganese and lead in moss samples, collected at 146 locations in three nationwide surveys, were used in the prediction procedure. Latter, the sampling locations intensity was considered as covariate, as the sampling was intensified near urban or industrialized areas.

This application also allowed to compare, first, the effect induced by the inclusion of the proposed covariate in the predicted concentrations and, second, the accuracy of predicted values. When comparing predictions obtained considering the covariate with the ones obtained when not considering, we have concluded that the inclusion of this covariate corrected the predicted values in opposite directions, as the pollution by Mn was being underestimated and by Pb was being overestimated, when the information of the sampling intensity was not taken into account. In what respects the accuracy of predictions, Table 5.8 is illustrative of the gain achieved not only by considering the temporal information, but also the information related to the covariate in use.

Chapter 6 proposed a model-based approach, assuming the separability of the Gaussian process under observation, in a way that the reduced number of time observations led to a saturated temporal correlation model. The proposed model was derived in order to accommodate not exclusively geo-referenced covariates, but also covariates associated to the temporal behaviour of the process.

The data set considered to illustrate the application of the model was the one concerning heavy metal concentration biomonitoring. Once the model requires the Gaussianity of data, only Mn data was used. The predicted concentration values for the most recent survey obtained by both approaches were similar, being the same conclusion valid for the spatial pattern. However, in what respects the accuracy of predictions, the model-based approach revealed to perform better.

In summary, the main contributions of this work reside in the extension of the model proposed in Høst *et al.* (1995), allowing to accommodate relevante information for the process under observation, and in the model-based approach which allows to incorporate, under the assumption of separability of the correlation structure, different

parameters for spatial and for temporal variability.

7.2 Future work

The spatio-temporal model introduced in Chapter 6 was developed considering a reduced number of time observations. In the application of this model, the separable spatio-temporal correlation function was factored by a purely exponential spatial correlation function, and a purely temporal correlation function corresponding to a saturated correlation function.
In the future, it would be of interest to apply this same model to predict concentrations of other heavy metals, as well to investigate different spatial models rather than the exponential for the spatial dependence structure.
It is also our concern to analyze the possibility of use different temporal correlation models rather than a saturated one.
The data giving rise to the applications in Chapters 4 and 6 were originated by biomonitoring studies. It is also of our interest to proceed with spatio-temporal modelling of Portuguese environmental data from monitoring stations, with several observations in time and incorporating information about the sampling process over time.

References

Aboal, J., Real, C., Fernández, J. & Carballeira, A. (2006). Mapping the results of extensive surveys: The case of atmospheric biomonitoring and terrestrial mosses. *Science of The Total Environment*, **356**, 256–274. 78

Boquete, M., Fernández, J., Aboal, J. & Carballeira, A. (2011). Analysis of temporal variability in the concentrations of some elements in the terrestrial moss *pseudoscleropodium purum*. *Environmental and Experimental Botany*, **72**, 210–217. 26

Bowman, A. & Crujeiras, R. (2013). Inference for variograms. *Computational Statistics and Data Analysis*, **66**, 19–31.

Bruno, F., Guttorp, P., Sampson, P. & Cocchi, D. (2003). Non-separability of space-time covariance models in environmental studies. In J. Mateu, D. Holland & W. Manteiga, eds., *The ISI International Conference on Environmental Statistics and Health*, Universidade de Santiago de Compostela. 3, 48, 80

Bruno, F., Cocchi, D. & Vagheggini, A. (2013). Finite population properties of individual predictors based on spatial patterns. *Environmental and Ecological Statistics*, **20**, 467–494. 63

Brus, D. & de Gruijter, J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, **80**, 1–59. 63

Cameletti, M., Ignaccolo, R. & Bande, S. (2011). Comparing spatio-temporal models for particulate matter in piemonte. *Environmetrics*, **22**, 985–996. 77, 78

Casal, R. (2003). *Geoestadística espacio-temporal. Modelos flexibles de variogramas anisotrópicos no separables*. Ph.D. thesis, Universidade de Santigo de Compostela. 49

Chilès, J. & Delfiner, P. (2012). *Geostatistics: modeling spatial uncertainty*. Wiley Series in Probability and Statistics, Johnn Wiley and Sons, Inc., New York, 2nd edn. 18

Christensen, O. & Ribeiro, P. (2002). geoRglm - a package for generalised linear spatial models. *R-NEWS*, **2**, 26–28, iSSN 1609-3631. 5

Cocchi, D., Greco, F. & Trivisano, C. (2007). Hierarchical space-time modelling of PM_{10} pollution. *Atmospheric Environment*, **41**, 532–542. 3, 7

Cressie, N. (1993). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Ltd, New York. 8, 10, 38

Cressie, N. & Huang, H. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, **94**, 1330–1340. 7, 47, 48, 57, 88

REFERENCES

Cressie, N. & Wikle, C. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons Inc, Hoboken, New Jersey, revised edn. 8, 83, 84

Crujeiras, R., Fernández-Casal, R. & Manteiga, W. (2010). Nonparametric test for separability of spatio-temporal processes. *Environmetrics*, **21**, 382–399. 47

de Cesare, L., Myers, D. & Posa, D. (2001). Estimating and modeling space-time correlation structures. *Statistics and Probability Letters*, **51**, 9–14. 49, 50, 77

de Iaco, S. (2010). Space-time correlation analysis: a comparative study. *Journal of Applied Statistics*, **37**, 1027–1041. 48

de Iaco, S., Myers, D. & Posa, D. (2002). Nonseparable space-time covariance models: some parametric families. *Mathematical Geology*, **34**. 48

Denham, A. (2012). *Geostatistical Spatiotemporal Modelling with application to the western king prawn of the Shark Bay managed prawn fishery*. Ph.D. thesis, EDITH COWAN UNIVERSITY. 50

Díaz-Avalos, C., Juan, P. & Mateu, J. (2014). Significance tests for covariate-dependent trends in inhomogeneous spatio-temporal point processes. *Journal of Stochastic Environmental Research and Risk Assessment*, **28**, 593–609. 27

Diggle, P. (2003). *Statistical analysis of spatial point patterns*. Arnold, London, 2nd edn. 63

Diggle, P. & Ribeiro, P. (2007). *Model based geostatistics*. Springer series in statistics, Springer Verlag, New York. 8

Diggle, P., Menezes, R. & Su, T. (2010). Geostatistical inference under preferential sampling. *Applied Statistics*, **59**, 191–232. 25, 26, 63, 64, 78

Dimitrakopoulos, R. & Luo, X. (1994). Spatiotemporal modelling: covariances and ordinary kriging systems. In *Geostatistics for the next century*, 88–93, Springer. 48

Fanshawe, T., Diggle, P., Rushton, S., Sanderson, R., Lurz, P., Glinianaia, S., Pearce, M., Parker, L., Charlton, M. & Pless-Mulloli, T. (2008). Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. *Environmetrics*, **19**, 549–566. 77

Figueira, R., Sérgio, C. & Sousa, A. (2002). Distribution of trace metals in moss biomonitors and assessment of contamination sources in Portugal. *Environmental Pollution*, **118**, 153–163. 26, 34

Fischer, M. & Getis, A., eds. (2009). *Handbook of Applied Spatial Analysis. Software Tools, Methods and Applications*. Springer-Verlag Berlin Heidelberg. 4

Fuentes, M. (2006). Testing for separability of spatio-temporal covariance functions. *Journal of Statistical Planning and Inference*, **136**, 447–466. 47

Gadzała-Kopciuch, R., Berecka, B., Bartoszewicz, J. & Buszewski, B. (2004). Some considerations about bioindicators in environmental monitoring. *Polish Journal of Environmental Studies*, **13**, 453–462. 2

Gelfand, A., Sahu, S. & Holland, D. (2012). On the effect of preferential sam-

pling in spatial prediction. *Environmetrics*, **23**. 64

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, **97**, 590–600. 48

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106**, 746–762. 66

Gneiting, T., Genton, M. & Guttorp, P. (2007). *Geostatistical space-time models, stationarity, separability and full symmetry*, chap. 4, 151–175. Chapman & Hall/CRC, Boca Raton. 46

Gómez-Rubio, V., Ferrándiz-Ferragud, J. & Lopez-Quílez, A. (2005). Detecting clusters of disease with R. *Journal of Geographical Systems*, **7**, 189–206. 5

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Applied geostatistics series, Oxford University Press. 8, 17, 18, 38

Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of hydrology*, **228**, 113–129. 7

Guttorp, P. & Loperfido, N. (2008). Network bias in air quality monitoring design. *Environmetrics*, **19**, 661–671. 7, 27

Harmens, H., Norris, D., Steinnes, E., Kubin, E., Piispanen, J., Alber, R., Aleksiayenak, Y., Blum, O., Coskun, M., Dam, M., Temmerman, L.D., Fernández, J., Frolova, M., Frontasyeva, M., González-Miqueo, L., Grodzińska, K., Jeran, Z., Korzekwa, S., Krmar, M., Kubin, E., Kvietkus, K., Leblond, S., Liiv, S., Magnússon, S., Maňkovská, B., Pesch, R., Rüling, A., Santamaria, J., Schöder, W., Spiric, Z., Suchara, I., Thöni, L., Urumov, V., Yurukova, L. & Zechmeister, H. (2010). Mosses as biomonitors of atmospheric heavy metal deposition: Spatial patterns and temporal trends in Europe. *Environmental Pollution*, **158**, 3144–3156. 25, 26, 78

Hengl, T. (2007). *A practical guide to geostatistical mapping of environmental variables*. European Commission, Joint Research Centre, Institute for Environment and Sustainability, Luxembourg. 4

Holy, M., Pesch, R., Schöder, W., Harmens, H., Ilyin, I., Alber, R., Aleksiayenak, Y., Blum, O., Coskun, M., Dam, M., Temmerman, L.D., Fedorets, N., Figueira, R., Frolova, M., Frontasyeva, M., Goltsova, N., Miqueo, L., Grodzińska, K., Jeran, Z., Korzekwa, S., Krmar, M., Kubin, E., Kvietkus, K., Larsen, M., Leblond, S., Liiv, S., Magnússon, S., Maňkovská, B., Mocanu, R., Piispanen, J., Rühling, A., Santamaria, J., Steinnes, E., Suchara, I., Thöni, L., Turcsányi, G., Urumov, V., Wolterbeek, B., Yurukova, L. & Zechmeister, H. (2009). First thorough identification of factors associated with Cd, Hg and Pb concentrations in mosses sampled in the European surveys 1990, 1995, 2000 and 2005. *Journal of Atmospheric Chemistry*, **63**, 109–124. 25

REFERENCES

Høst, G., Omre, H. & Switzer, P. (1995). Spatial interpolation errors for monitoring data. *Journal of the American Statistical Association*, **90**, 853–861. 7, 54, 55, 57, 96

Howe, P., Malcolm, H. & Dobson, S. (2004). Manganese and its compounds: environmental aspects. Concise International Chemical Assessment Document 63, World Health Organization, Geneva. 28

Huang, H., Martinez, F., Mateu, J. & Montes, F. (2007). Model comparison and selection for stationary space-time models. *Computational Statistics and Data Analysis*, **51**, 4577–4596. 77

Isaacs, E. & Srivastava, R. (1989). *Applied Geostatistics*. Oxford University Press, New York. 7, 18

Järup, L. (2003). Hazards of heavy metal contamination. *British Medical Bulletin*, **68**, 167–182. 29, 36

Jing, L. & de Oliveira, V. (2015). geoCount: An R Package for the Analysis of Geostatistical Count Data. *Journal of Statistical Software*, **63**, 1–33. 5

Journel, A. & Huijbregts, C.J. (1997). *Mining geostatistics*. Academic Press Limited, Suffolk, 7th edn. 8

Kolovos, A., Christakos, G., Hristopoulos, D. & Serre, M. (2004). Methods for generating non-separable spatiotemporal covariance models with potential environmental applications. *Advances in Water Research*, **27**, 815–830. 48

Krige, D. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, **52**, 119–139. 7

Kyriakidis, P. & Journel, A. (1999). Geostatistical Space-Time Models: A Review. *Mathematical Geology*, **31**, 651–684. 48, 77

Kyriakidis, P., Kim, J. & Miller, N. (2001). Geostatistical mapping of precipitation from rain gauge data using atmospheric and terrain characteristics. *Journal of Applied Meteorology*, **40**, 1855–1877. 7

Lindstrom, J., Szpiro, A., Sampson, P., Sheppard, L., Oron, A., Ricahrds, M. & larson, T. (2011). A flexible spatio-temporal model for air pollution: Allowing for spatio-temporal covariates. UW Biostatistics Working Paper Series, working paper 370. 1, 78

Margalho, L., Menezes, R., Sousa, I. & Silva, J. (2011). Monitoring NO_3 contamination of aquifer system of Bacia do Cávado/Ribeiras Costeiras. In A. Stein, E. Pebesma & G. Heuvelink, eds., *Procedia Environmental Sciences*, vol. 7, 377–382, Elsevier Ltd. 18

Margalho, L., Menezes, R. & Sousa, I. (2014). Assessing interpolation error for spacetime monitoring data. *Journal of Stochastic Environmental Research and Risk Assessment*, **28**, 1307–1321. 64

Markert, B., Breure, A. & Zechmeister, H., eds. (2003). *Bioindicators & Biomonitors: Principles, Concepts, and Applications*, vol. 6 of *Trace metals and other contaminants in the environment*. Elsevier Ltd., UK. 2

Martins, A., Figueira, R., Sousa, A. & Sérgio, C. (2012). Spatio-temporal patterns of Cu contamination in mosses using geostatistical estimation. *Environmental Pollution*, **170**, 276–284. 26

Mateu, J. & Müller, W., eds. (2012). *Spatio-temporal design: advances in efficient data acquisition*. Chichester, John Wiley & Sons, Ltd. 64

Matheron, G. (1971). *The theory of regionalized variables and its applications*. Centre de Morphologie Mathématique Fontainebleau: Les cahiers du Centre de Morphologie Mathématique de Fontainebleau, École Nationale Supérieure des Mines de Paris. 7

Menezes, R., Garcia-Soidán, P. & Febrero-Bande, M. (2008). A kernel variogram estimator for clustered data. *Scandinavian Journal of Statistics*, **35**, 18–37. 63

Mililo, T. (2009). *Applying geostatistical interpolation methods to studies of environmental contamination of urban soils and image analysis of time-of-flight secondary ion mass spectrometry*. Phd thesis, State University of New York at Buffalo, New York. 3

Mitchell, M., Genton, M. & Gumpertz, M. (2005). Testing for separability of space-time covariances. *Environmetrics*, **16**, 819–831. 3

Mitchell, M., Genton, M. & Gumpertz, M. (2006). A likelihood ratio test for separability of covariances. *Journal of Multivariate Analisys*, **97**, 1025–1043. 47

Olea, R. (2012). Building on crossvalidation for increasing the quality of geostatistical modeling. *Journal of Stochastic Environmental Research and Risk Assessment*, **26**, 73–82. 7, 73

Pebesma, E. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, **30**, 683–691. 5

Pebesma, E., Bivand, R. & Ribeiro, P.J. (2015). Software for Spatial Statistics. *Journal of Statistical Software*, **63**. 5

Perl, D. & Olanow, C. (2007). The Neuropathology of Manganese-Induced Parkinsonism. *Journal of Neuropathology and Experimental Neurology*, **66**, 675. 28

Pinsino, A., Roccheri, M. & Matranga, V. (2012). *Manganese: a new emerging contaminant in the environment*. INTECH Open Access Publisher. 28

Ribeiro, P. & Diggle, P. (2001). geoR: a package for geostatistical analysis. *R-NEWS*, **1**, 14–18. 4

Rocks, S. & Levy, L. (2008). Manganese Health Research Program: Overview of Research into the Health Effects of Manganese (2007-2008). Report, Institute of Environment and Health, Cranfield University, Silsoe, Bedfordshire. 28

Rodriguez-Iturbe, I. & Mejia, J. (1974). The design of networks in time and space. *Water Resources Research*, **10**, 713–728. 4, 49, 81

Rouhani, S. & Wackernagel, H. (1990). Multivariate geostatistical approach to space-time data analysis. *Water Resources Research*, **26**, 585–591. 7

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 319–392. 5

Rühling, Å. & Steinnes, E. (1998). Atmospheric heavy metal deposition in europe 1995-1996. Technical Report 15, Nordic Council of Ministers, Copenhagen. 1

Sahu, S. & Mardia, K. (2005). Recent trends in modelling spatio-temporal data. In *Proceedings of the special meeting on Statistics and Environment*, 69–83. 77

Samet, J., Zeger, S., Dominici, F., Curriero, F., Coursac, I., Dockery, D., Schwartz, J. & Zanobetti, A. (2000). Morbidity, Mortality and Air Pollution Study, Part II: Morbidity and Mortality from Air Pollution in the United States. Research Report 84, Health Effects Institute, Cambrige MA. 2

Sarmento, S. (2012). *Application of Atmospheric Biomonitoring to Epidemiology: Issues in Data Quality, Sampling, Aggregation & Confounding*. IOS Press Inc. 2

Scaccia, L. & Martin, R. (2005). Testing axial symmetry and separability of lattice processes. *Journal of Statistical Planning and Inference*, **131**, 1939. 47

Schlather, M., Menck, P., Singleton, R., Pfaff, B. & and the R Team (2013). *RandomFields: Simulation and Analysis of Random Fields*. R package version 2.0.71. 5

Shaddick, G. & Wakefield, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of The Royal Statistical Society Series C-applied Statistics*, **51**, 351–372. 77

Shaddick, G. & Zidek, J. (2012). Unbiasing estimates from preferentially sampled spatial data. Technical Report 268, The University of British Columbia. 63

Sherman, M. (2011). *Spatial Statistics and Spatio-Temporal Data: covariance functions and directional properties*. John Wiley & Sons Inc, Chichester. 48, 77, 81, 83

Stein, M. (2005). Space-time covariance functions. *Journal of the American Statistical Association*, **100**, 310–321. 48

Steinnes, E., Berg, T. & Sjøbakk, T. (2003). Temporal and spatial trends in Hg deposition monitored by moss analysis. *The Science of the Total Environment*, **304**, 215–219. 25, 78

Steinnes, E., Berg, T. & Uggerud, H. (2011). Three decades of atmospheric metal deposition in Norway as evident from analysis of moss samples. *The Science of the Total Environment*, **412-413**, 351–358. 25, 26, 78

Uyar, G., Ören, M., Yildirim, Y. & Ince, M. (2007). Mosses as indicators of atmospheric heavy metal deposition around a coal-fired power plant in Turkey. *Fresenius Environmental Bulletin*, **16**, 182–192. 25

Venables, W. & Ripley, B. (2002). *Modern Applied Statistics with S*. Springer, New York, 4th edn. 5

Zechmeister, H., Hohenwallner, D., Hanus-Illnar, A., Hagendorfer, H., Roder, I. & Riss, A. (2008). Temporal patterns of metal deposition at various scales during the last two decades. *Atmospheric Environment*, **42**, 1301–1309. 25, 78